THE MODIFIED DISCRETE COSINE TRANSFORM: ITS IMPLICATIONS FOR AUDIO CODING AND ERROR CONCEALMENT

YE WANG AND MIIKKA VILERMO

Nokia Research Center, P.O. Box 100, FIN-33721 Tampere, Finland
ye.wang@nokia.com
miikka.vilermo@nokia.com

This paper presents a study of the Modified Discrete Cosine Transform (MDCT) and its implications for audio coding and error concealment from the perspective of Fourier frequency analysis. A relationship between the MDCT and DFT via Shifted Discrete Fourier Transform (SDFT) is established, which provides a possible fast implementation of MDCT employing an FFT routine. The concept of Time Domain Alias Cancellation (TDAC), the symmetric and non-orthogonal properties of MDCT are analyzed and illustrated with intuitive examples. This study gives some new insights for innovative solutions in audio codec design and MDCT domain audio processing such as error concealment.

INTRODUCTION

With rapid deployment of audio compression technologies, more and more audio content is stored and transmitted in compressed formats. The Internet transmission of compressed digital audio, such as MP3, has already shown a profound effect on the traditional process of music distribution. Recent developments in this field have made possible the reception of streaming digital audio with handheld network communication devices, for example.

Signal representation in the Modified Discrete Cosine Transform (MDCT) domain has emerged as a dominant tool in high quality audio coding because of its special properties: in addition to the energy compaction capability similar to DCT, MDCT simultaneously achieves critical sampling, reduction of block effect and flexible window switching.

In applications such as streaming audio to handhold devices, it is often necessary to have fast implementations and optimised codec structures. In certain situations it is also desirable to perform MDCT domain audio processing such as error concealment which mitigates the degradation of subjective audio quality. These are motivations for us to conduct this study.

The MDCT uses the concept of time domain alias cancellation [1][2], while the Quadrature Mirror Filterbank (QMF) uses the concept of frequency domain alias cancellation [3]. This can be viewed as the duality of MDCT and QMF. However, it should be noted that MDCT also cancels frequency domain aliasing, while the QMF does not cancel time domain aliasing. In other words, MDCT is designed to achieve perfect reconstruction, while QMF is not.

Before the introduction of MDCT, transform based audio coding techniques used DFT and DCT with window functions such as rectangular and sine-taper functions. However, these early coding techniques have failed to fulfil the contradictory requirements imposed by high-quality audio coding. For example, with a rectangular window the analysis/synthesis system is critically sampled, *i.e.*, the overall number of the transformed domain samples is equal to the number of time domain samples, but the system suffers from poor frequency resolution and block effects, which are introduced after quantization or other manipulation in the frequency domain. Overlapped windows allow for better frequency response functions but carry the penalty of additional values in the frequency domain, thus not critically sampled. MDCT is currently the best solution, which has satisfactorily solved the paradox. The concept of the window switching was introduced to tackle possible pre-echo problems in the case of insufficient time resolutions [4]. Nevertheless, it is worth mentioning that the mismatch between the MDCT and DFT based perceptual model of human auditory systems could still be behind certain coding artefacts at low bitrates [5].

The complex version of the MDCT has been investigated in [6][7][8] in terms of filterbank theory. Our research has approached the problem from a different perspective: Fourier spectrum analysis. We hope that the study presented in this paper can not only provide an intuitive tutorial in the concept of MDCT and TDAC, but also some stimulation for innovative solutions in applications such as MDCT domain audio processing and error concealment.

This paper is organised as follows. A relationship between the MDCT and DFT via SDFT is established in section 1. The symmetric property of MDCT and the TDAC concept are illustrated in section 2. The non-orthogonal property of MDCT is then discussed in section 3. The implications for audio coding and error concealment are outlined in section 4. Finally, section 5 concludes the paper with some discussions.

1 THE INTERCONNECTION BETWEEN MDCT, SDFT AND DFT

The direct and inverse MDCT are defined as [1][2]:

$$\alpha_r = \sum_{k=0}^{2N-1} \tilde{a}_k \cos\left[\pi \frac{(k+(N+1)/2)(r+1/2)}{N}\right] \quad (1)$$
$$r = 0,\ldots,N-1$$

$$\hat{a}_k = \frac{2}{N}\sum_{r=0}^{N-1} \alpha_r \cos\left[\pi \frac{(k+(N+1)/2)(r+1/2)}{N}\right] \quad (2)$$
$$k = 0,\ldots,2N-1$$

where $\tilde{a}_k = h_k a_k$ is the windowed input signal, a_k is the input signal of $2N$ samples. h_k is a window function. We assume an identical analysis-synthesis time window. The constraints of perfect reconstruction are [6][8]:

$$h_k = h_{2N-1-k} \quad (3)$$

$$h_k^2 + h_{k+N}^2 = 1 \quad (4)$$

A sine window is widely used in audio coding because it offers good stop-band attenuation, provides good attenuation of the block edge effect and allows perfect reconstruction. Other optimised windows can be applied as well [6]. The sine window is defined as:

$$h_k = \sin[\pi\,(k+1/2)/2N] \quad (5)$$
$$k = 0,\ldots,2N-1$$

$\hat{a}_k$ in (2) are the IMDCT coefficients of α_r, which contains time domain aliasing:

$$\hat{a}_k = \begin{cases} \tilde{a}_k - \tilde{a}_{N-1-k}, & k = 0,\ldots,N-1 \\ \tilde{a}_k + \tilde{a}_{3N-1-k}, & k = N,\ldots,2N-1 \end{cases} \quad (6)$$

The relationship between MDCT and DFT can be established via Shifted Discrete Fourier Transforms (SDFT). The direct and inverse SDFTs are defined as [13]:

$$SDFT_{u,v} = \alpha_r^{u,v} = \sum_{k=0}^{2N-1} a_k \exp[i2\pi(k+u)(r+v)/2N], \quad (7)$$

$$ISDFT_{u,v} = a_k^{u,v} = \frac{1}{2N}\sum_{r=0}^{2N-1} \alpha_r^{u,v} \exp[-i2\pi(k+u)(r+v)/2N], \quad (8)$$

where u and v represent arbitrary time and frequency domain shifts respectively. SDFT is a generalization of DFT that allows a possible arbitrary shift in position of the samples in the time and frequency domain with respect to the signal and its spectrum coordinate system.

We have proven that the MDCT is equivalent to the SDFT of a modified input signal [10][11].

$$\alpha_r = \frac{1}{2}\sum_{k=0}^{2N-1} \hat{a}_k \exp\left[i\pi \frac{(k+(N+1)/2)(r+1/2)}{N}\right] \quad (9)$$

The right side of (9) is $SDFT_{(N+1)/2,1/2}$ of the signal $\hat{a}_k$ formed from the initial windowed signal $\tilde{a}_k$ according to (6).

Physical interpretation of (6) is straightforward. MDCT coefficients can be obtained by adding the $SDFT_{(N+1)/2,1/2}$ coefficients of the initial windowed signal and the alias.

For real-valued signals, it is quite straightforward to prove that MDCT coefficients are equivalent to the real part of $SDFT_{(N+1)/2,1/2}$ of the input signal. That is

$$\alpha_r = real\{SDFT_{(N+1)/2,1/2}(\tilde{a}_k)\} \quad (10)$$

With reference to (6)(9) and Figure 1(f), the alias is added to the original signal in such a way that the first half of the window (the signal portion between points A and B) is mirrored in the time domain and then inverted, before being subsequently added to the original signal. The second half of the window (signal portion between points B and C) is also mirrored in the time domain and added to the original signal.

From (1)(2)(6)(9) and Figure 1(f) we can see that, in comparison with conventional orthogonal transforms, MDCT has a special property: the input signal cannot be perfectly reconstructed from a single block of MDCT coefficients. MDCT itself is a lossy process. That is, the imaginary coefficients of the $SDFT_{(N+1)/2,1/2}$ are lost in the MDCT transform, which is equivalent to a decimation operation. Applying a MDCT and then an

IMDCT converts the input signal into one that contains a certain twofold symmetric alias (see (6) and Figure 1(f)). The introduced alias is cancelled in the overlap-add process to achieve perfect reconstruction (see Figure 2).

The formulation in (9) is different in comparison with the Odd-DFT concept discussed in [6]. The Odd-DFT is $SDFT_{0,1/2}$ of the initial windowed signal $\tilde{a}_k$.

The $SDFT_{(N+1)/2,1/2}$ can be expressed by means of the conventional DFT as:

$$\sum_{k=0}^{2N-1} \hat{a}_k \exp\left[i2\pi \frac{(k+(N+1)/2)(r+1/2)}{2N}\right]$$
$$= \left\{\sum_{k=0}^{2N-1}\left[\hat{a}_k \exp\left(i2\pi\frac{k}{4N}\right)\right]\exp(i2\pi\frac{kr}{2N})\right\}\exp\left(i2\pi\frac{(N+1)r}{4N}\right)\exp\left(i\pi\frac{N+1}{4N}\right) \qquad (11)$$

To the right side of (11), the first exponential function corresponds to a modulation of $\hat{a}_k$ that results in a signal spectrum shift in the frequency domain by ½ of the frequency-sampling interval. The second exponential function corresponds to the conventional DFT. The third exponential function modulates the signal spectrum that is equivalent to a signal shift by (N+1)/2 of the sampling interval in the time domain. The fourth term is a constant phase shift. Therefore, $SDFT_{(N+1)/2,1/2}$ is the conventional DFT of this signal shifted in the time domain by (N+1)/2 of the sampling interval and evaluated with the shift of ½ the frequency-sampling interval. This formulation provides one possible fast implementation employing an FFT routine.

2 SYMMETRIC PROPERTY OF THE MDCT AND THE CONCEPT OF TIME DOMAIN ALIAS CANCELLATION

2.1 Symmetric Property of the MDCT

$SDFT_{(N+1)/2,1/2}$ coefficients exhibit symmetric properties:

$$\alpha_{2N-r-1}^{(N+1)/2,1/2} = (-1)^{N+1}\left(\alpha_r^{(N+1)/2,1/2}\right)^* \qquad (12)$$

where $*$ is the complex conjugate of the coefficients. Similarly, MDCT coefficients exhibit symmetric properties:

$$\alpha_{2N-r-1} = (-1)^{N+1}\alpha_r \qquad (13)$$

whereby, the MDCT coefficients are odd symmetric, if N is even, which is normally the case in audio coding applications.

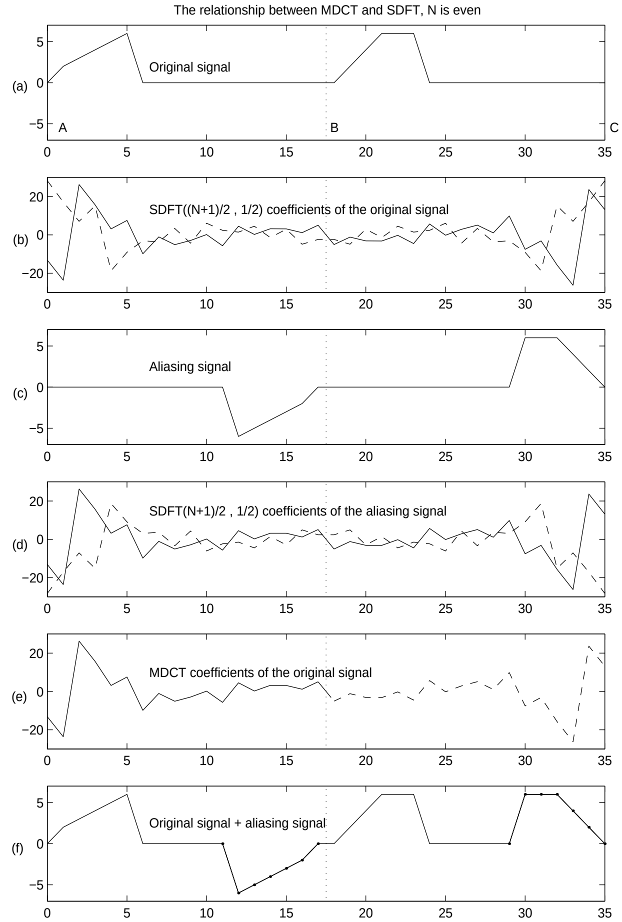


Figure 1. Illustration of the interconnection between MDCT and $SDFT_{(N+1)/2,1/2}$. (a) an artificial time domain signal of 36 samples; (b) $SDFT_{(N+1)/2,1/2}$ coefficients of the signal in (a); (c) The time domain alias; (d) $SDFT_{(N+1)/2,1/2}$ coefficients of the alias, where the solid lines are the real parts, the dotted lines are the imaginary parts in both (b) and (d); (e) MDCT coefficients of the time signal in (a), where the dotted line is odd symmetric to the solid line, and therefore it is redundant; (f) the alias embedded time signal.

We have proved that:

$$IMDCT(\alpha_r) = ISDFT_{(N+1)/2,1/2}(\alpha_r) \qquad (14)$$
$$r = 0,\ldots,2N-1$$

Due to the decimation of MDCT, we have N independent frequency components. That is, if we want to implement IMDCT using ISDFT, it is necessary to apply the symmetric property of MDCT in order to have 2N dependent frequency components to the ISDFT routine as shown in Figure 1(e).

In order to illustrate the symmetric properties of MDCT and the interconnection between MDCT and $SDFT_{(N+1)/2,1/2}$ in an intuitive way, we have employed an artificial time domain signals (N=18) as shown in Figure 1(a). The $SDFT_{(N+1)/2,1/2}$ coefficients of the original signal are shown in Figure 1(b). The time domain alias is illustrated in Figure 1(c). Its $SDFT_{(N+1)/2,1/2}$ coefficients are presented in Figure 1(d). In Figure 1(b, d) the solid lines are the real parts, the dashed lines are the imaginary parts. The MDCT coefficients are shown in Figure 1(e). They are equivalent to the real parts of the $SDFT_{(N+1)/2,1/2}$ coefficients of the original signals in Figure 1(a). The dashed line in Figure 1(e) is odd symmetric to the solid line and the dashed line represents the redundant coefficients, which are left out in the MDCT definition. The alias-embedded time signal is presented in Figure 1(f). It equals the IMDCT of the MDCT coefficients scaled by factor two. A rectangular window is used here for clarity.

2.2 Intuitive illustration of the Concept of Time Domain Alias Cancellation (TDAC)

Based on (1) (2) (6) (9), we have used a similar artificial time domain signal as in Figure 1(a) to illustrate the Time Domain Aliasing Cancellation (TDAC) concept in an intuitive way. The artificial signal of 54 samples is shown in Figure 2(a). The MDCT coefficients of the signal in Window 1 are shown in Figure 2(b). For illustration of the concept, a rectangular window is employed. Due to the 50% decimation in MDCT (from 2N time domain samples in Figure 2(a) to N independent frequency domain coefficients in Figure 2 (b)), the alias is introduced. This is illustrated in Figure 2(c). The IMDCT introduces redundancy (from N frequency domain coefficients in Figure 2(b) to 2N time domain samples in Figure 2(c)). The MDCT coefficients of the signal in Window 2 are presented in Figure 2(d). The corresponding IMDCT time domain signal is shown in Figure 2(e). If the overlap-add procedure is performed with Figure 2(c) and (e), perfect reconstruction (PR) of the original signal in the overlapped part (between points B and C) can be achieved.

It is clear that one cannot achieve perfect reconstruction (PR) for the first half of the first window and the second half of the last window as indicated in Figure 2.

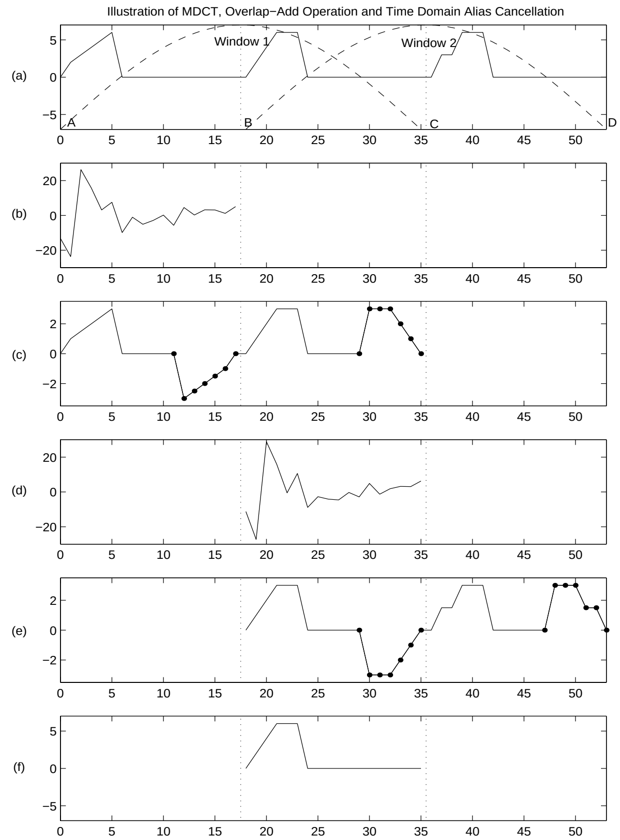


Figure 2. Illustration of the MDCT, overlap-add (OA) procedure and the concept of the Time Domain alias cancellation (TDAC). (a) An artificial time signal, dashed lines indicating the 50% overlapped windows; (b) MDCT coefficients of the signal in Window 1; (c) IMDCT coefficients of the signal in (b), the alias is shown by markers on the line; (d) The MDCT coefficients of the signal in Window 2; (e) IMDCT coefficients of the signal in (d), the alias shown by markers on the line; (f) The reconstructed time domain signal after the overlap-add (OA) procedure. The original signal in the overlapped part (between points B and C) is perfectly reconstructed.

In order to illustrate the TDAC concept during the window switching specified in the MPEG AAC ISO/IEC standard [9], we define two overlapping windows with window functions h_k and g_k. The conditions for perfect reconstruction are [4]:

$$h_{N+k} \cdot h_{2N-1-k} = g_k \cdot g_{N-1-k} \tag{15}$$

$$h_{N+k}^2 + g_k^2 = 1 \tag{16}$$

Using (6) one can easily see one of the important properties of MDCT: the time domain alias in each half of the window is independent, which allows adaptive window switching [4]. The TDAC concept during window switching in MPEG AAC is illustrated in Figure 3.

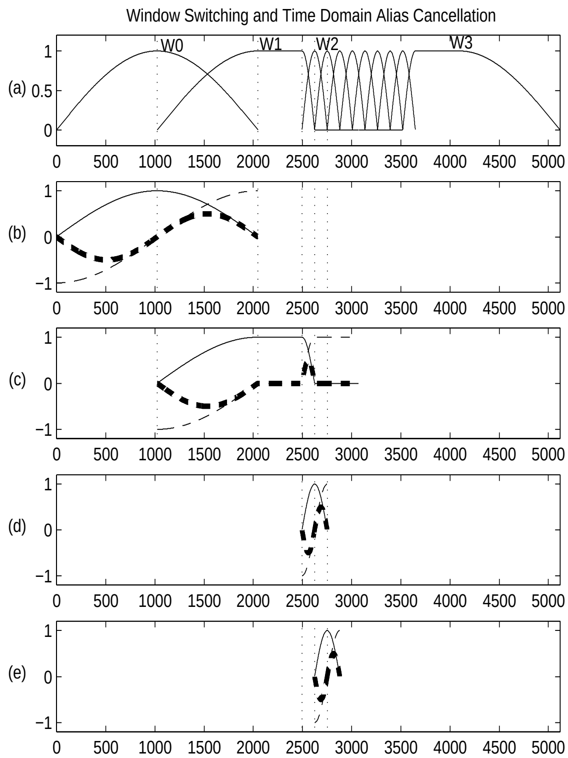


Figure 3. TDAC in the case of window switching. (a) four types of window shape in MPEG-2 AAC indicated with W0,..W3; (b) window function in the long window (solid line), time domain alias (thin dashed line), time domain alias after weighting with the window function (thick dashed line); (c) window function in the transition window (solid line), time domain alias (thin dashed line), time domain alias after weighting with the window function (thick dashed line); (d)(e) window function in the short window (solid line), time domain alias (thin dashed line), time domain alias after weighting with the window function (thick dashed line).

3 NON-ORTHOGONAL PROPERTY OF THE MDCT

3.1 Observation from a single transform block

If a signal exhibits local symmetry such that

$$\begin{cases} \tilde{a}_k = \tilde{a}_{N-k-1}, & k = 0,\ldots,N-1 \\ \tilde{a}_k = -\tilde{a}_{3N-k-1}, & k = N,\ldots,2N-1 \end{cases} \tag{17}$$

its MDCT degenerates to zero: $\alpha_r = 0$ for $r = 0,\ldots,N-1$. This property follows from (6). This is an example to show that MDCT does not fulfil Parseval's theorem, i.e. the time domain energy is not equal to the frequency domain energy (see Figure 4).

If a signal exhibits local symmetry such that

$$\begin{cases} \tilde{a}_k = -\tilde{a}_{N-k-1}, & k = 0,\ldots,N-1 \\ \tilde{a}_k = \tilde{a}_{3N-k-1}, & k = N,\ldots,2N-1 \end{cases}, \tag{18}$$

MDCT and IMDCT of a single transform block will perfectly reconstruct the original time domain samples. This property also follows from (6).

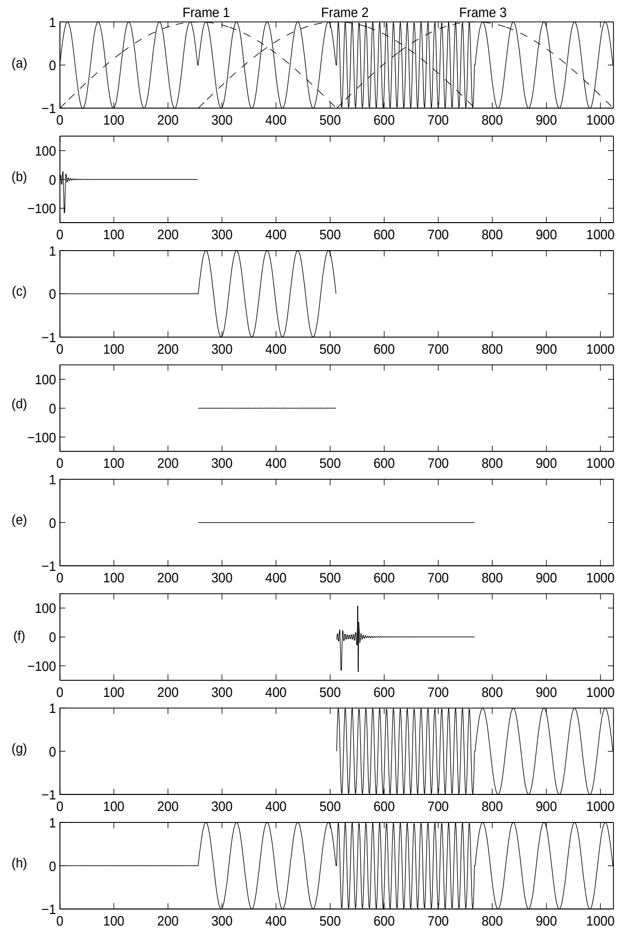


Figure 4. Illustration of signal analysis/synthesis with MDCT, overlap-add procedure and perfect reconstruction of time domain samples. (a) a phase/frequency-modulated time signal; (b)(d)(f) MDCT spectra in different time slots, indicated as frames 1, 2, 3 in (a); (c)(e)(g) reconstructed time domain samples (with IMDCT) of frames 1, 2, 3 respectively; (h) the reconstructed time samples after the overlap-add procedure.

In order to illustrate that MDCT does not fulfil Parseval's theorem in an intuitive way, we have designed a phase/frequency-modulated time signal in

Figure 4(a), which has two different frequency elements with the duration of half of the frame size (frame size = 512 samples). Dashed lines in Figure 4 (a) illustrate the 50% window overlap. However, MDCT spectra of different time slots in Figures 4(b)(d)(f) are calculated with rectangular windows for the purpose of illustration. The IMDCT time domain samples of frame 1, 2, 3 are shown in Figures 4(c)(e)(g) respectively. The reconstructed time domain samples after overlap-add (OA) procedure is shown in Figure 4(h). With frame 2 the condition in (17) holds, and the MDCT coefficients are all zero. Nevertheless, the time domain samples in frame 2 can still be perfectly reconstructed after the overlap-add procedure. With frame 3 the condition (18) holds, and the original time samples are perfectly reconstructed even without overlap-add procedure. These are, of course, very special occurrences, which are rare in real life audio signals, especially after proper windowing such as a sine window. If the signal is close to the condition in (17) however, MDCT spectrum will be very unstable in comparison with DFT spectrum. In this case, using the output of the DFT based psychoacoustic model to quantise MDCT coefficients could cause certain coding artefacts. This is a limitation of MDCT.

3.2 Observation from multiple transform blocks

As in (19), the matrix of the MDCT for transforming 2N input samples to N spectral components is of size $N \times 2N$ and therefore cannot be orthogonal. However, the underlying basis functions of MDCT (corresponding to the rows of the matrix) are orthogonal.

$$P = \begin{bmatrix} p_{1,1} & p_{1,2} & p_{1,3} & \cdot & \cdot & \cdot & \cdot & p_{1,2N} \\ p_{2,1} & p_{2,2} & p_{2,3} & \cdot & \cdot & \cdot & \cdot & p_{2,2N} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ p_{N,1} & p_{N,2} & p_{N,3} & \cdot & \cdot & \cdot & \cdot & p_{N,2N} \end{bmatrix} \qquad (19)$$

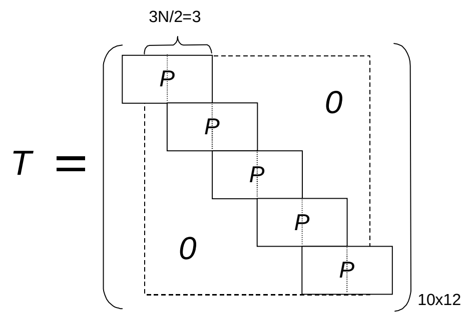
In the case of a continuous input stream x, a block-diagonal matrix T can be made of MDCT matrices P on the diagonal and zeros elsewhere (see (20)). This block-diagonal matrix T for transforming $(n+1)\cdot N$ input samples to $n \cdot N$ spectral components is of size $(n \cdot N)\times[(n+1)\cdot N]$. T becomes an orthogonal and square matrix if $n \to \infty$.

$$X_{n\cdot N} = \begin{bmatrix} P & & & & 0 \\ & P & & & \\ & & \ddots & & \\ & & & P & \\ 0 & & & & P \end{bmatrix}_{(n\cdot N)\times[(n+1)\cdot N]} \cdot x_{(n+1)\cdot N} = T \cdot x_{(n+1)\cdot N} \qquad (20)$$

where x is the input vector of the signal, X is the output vector of the MDCT coefficients. The orthogonality of T implies

$$T^T \cdot T = T \cdot T^T = I \qquad (21)$$

However, in the case of finite-length input signals, T is not anymore orthogonal. In order to illustrate this scenario in an intuitive way, let us observe a simple example with N=2 and n=5. In this case, the block-diagonal matrix appears as follows:



(22)

That is,

$$T \cdot T^T = \begin{bmatrix} \begin{bmatrix}1&0\\0&1\end{bmatrix} & & & & 0 \\ & \begin{bmatrix}1&0\\0&1\end{bmatrix} & & & \\ & & \begin{bmatrix}1&0\\0&1\end{bmatrix} & & \\ & & & \begin{bmatrix}1&0\\0&1\end{bmatrix} & \\ 0 & & & & \begin{bmatrix}1&0\\0&1\end{bmatrix} \end{bmatrix} = I \qquad (23)$$

and

$$T^T \cdot T = \begin{bmatrix} \begin{bmatrix}0.5&0.5\\0.5&0.5\end{bmatrix} & & & & & 0 \\ & \begin{bmatrix}1&0\\0&1\end{bmatrix} & & & & \\ & & \begin{bmatrix}1&0\\0&1\end{bmatrix} & & & \\ & & & \begin{bmatrix}1&0\\0&1\end{bmatrix} & & \\ & & & & \begin{bmatrix}1&0\\0&1\end{bmatrix} & \\ 0 & & & & & \begin{bmatrix}0.5&0.5\\0.5&0.5\end{bmatrix} \end{bmatrix} \qquad (24)$$

It is clear from (24) that the matrices of the first and last blocks are not unit matrices, though this usually does not pose a serious problem in audio coding applications. However, one should keep this effect in mind when manipulating audio signals in the compressed domain, such as editing, error concealment, etc.

With a regular block transform, we simply require that the number of samples in the signal is a multiple of the block size N. With MDCT, however, we need a slight modification in the computation of the transforms of the first and last blocks. Otherwise, the corresponding MDCT basis function would extend outside the region of support of the signal [8].

For a finite-length input vector, the infinite matrix T in (20) would be replaced by the finite matrix

$$T = \begin{bmatrix} P_a & & & & 0 \\ & P & & & \\ & & \ddots & & \\ & & & P & \\ 0 & & & & P_b \end{bmatrix}_{(n \cdot N) \times (n \cdot N)} \qquad (25)$$

Matrices of the first and last blocks are denoted as P_a and P_b respectively. The block-diagonal matrix T becomes square only after some special handling of P_a and P_b, which would generate basis functions of length 3N/2, because there can be no overlapping outside the signal region of support [8]. N is an even number. After this handling, the dimension of both P_a and P_b becomes $N \times \left(\frac{3N}{2}\right)$ as illustrated in (22), thus T becomes a square matrix as shown in (25). It corresponds to the square matrix shown in dashed line in (22).

This has some implications for practical applications such as MDCT domain error concealment, which is discussed in the following section.

4 IMPLICATIONS FOR AUDIO CODING AND ERROR CONCEALMENT

4.1 MDCT based Audio Coding

An important implication of (10) is that MDCT can be implemented via a SDFT. That is, the complex frequency domain values required in the psychoacoustic model and the MDCT values could be calculated from the same set of computation, thus reducing the computational complexity of the encoder.

The quantizer connects the MDCT and the DFT based psychoacoustic model, which could present a mismatch problem. This MDCT-DFT mismatch problem can be illustrated with a practical example of an MP3 audio coder. The output of a psychoacoustic model is the signal-to-masking ratio (*SMR*) calculated in the DFT domain. The maximal inaudible quantization error (*EN*) is calculated according to

$$EN = \frac{ES}{SMR} \qquad (26)$$

where *ES* is the MDCT domain signal energy. Using a sinusoid as a test signal, the *SMR* is stable over time because DFT is an orthogonal transform. However, *ES* can fluctuate over time because MDCT does not obey Parseval's theorem, thus causing an undesirable fluctuation of the *EN* over time. This phenomenon is referred to as the MDCT-DFT mismatch phenomenon. This could lead to coding artifacts.

4.2 MDCT Domain Error Concealment

In the transmission of compressed audio, one of the most significant challenges today is the need to handle errors in lossy channels.

Error concealment is usually referred as the last resort to mitigate the degradation of audio quality in real-time streaming applications.

For speech communications in a packet network, the use of repetition is recommended as offering a good compromise between achieved quality and excessive complexity [15]. However, simple repetition can pose problems in streaming music, which often contains percussive sounds such as drumbeats.

If a drumbeat is replaced with other signals such as singing from the neighbouring packet, the drumbeat is simply eliminated. On the other hand, if the drumbeat is copied to its following packet, it may result in a subjectively very annoying distortion defined as a *double-drumbeat effect*. The degree of annoyance of the double-drumbeat effect depends on the time-frequency structure of the drumbeat. And it also depends on the distance between the original drumbeat and that generated due to packet repetition [16].

Due to the non-orthogonal property of MDCT, the repetition violates the TDAC conditions. Consequently, the alias distortions in the overlapped parts cannot cancel each other out (see Figure 5). However, the MDCT window functions enable a natural fade-in and fade-out in the overlap-add operation in the time domain. The un-cancelled alias is normally not perceptible if the signal is stationary and the lost data unit is short enough.

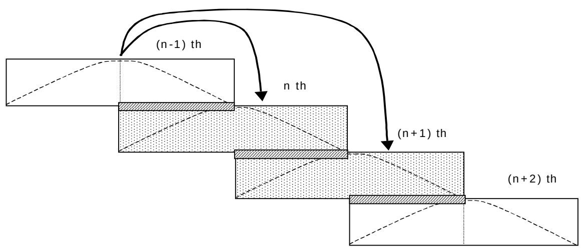


Figure 5. Illustration of a special problem with repetition scheme in the MDCT domain. Shaded rectangles represent corrupted data units. Blank rectangles represent error-free ones. Dashed lines indicate the window shape. The arrows indicate packet repetition operations. Heavily shaded rectangles represent the un-cancelled alias. n is an integer number that represents the data unit index.

Another potential problem is that simple repetition does not consider the window switching commonly used in state-of-the-art audio codecs. Therefore, it leads to a possible *window type mismatch phenomenon*, which is illustrated with the help of Figure 6.

Window switching is an important concept to reduce pre-echo in an MDCT based audio codec such as MP3 and AAC. Both MP3 and AAC use four different window types: long, long-to-short, short and short-to-long which are indexed with 0, 1, 2, 3 respectively. The short window is introduced to tackle transient signals better. 50% window overlap is used with the MDCT.

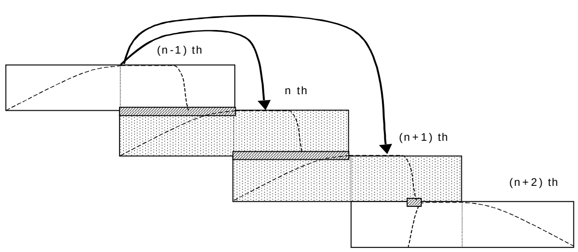


Figure 6. An example of the window type mismatch problem

If the two consecutive short window frames indexed as 22 in a window-switching sequence of 1223 are lost in a transmission channel, it is easy to deduce their window types from their neighbouring frames. This information could be used in error concealment [16]. However, if we disregard the window-switching information available from the audio bitstream and perform simple repetition, it could result in window-switching patterns of 1113 (see Figure 6). In this case, not only are the TDAC conditions violated in the window overlapped areas, but also we will have some undesired energy fluctuation, since the squares of the two overlapping window functions do not add up to a constant [4]. This may create annoying artifacts. This phenomenon is defined as *window type mismatch phenomenon*.

In order to enhance coding efficiency, the state-of-the-art audio coding techniques tend to use longer transform block length, e.g., 1024 MDCT coefficients in AAC versus 576 MDCT coefficients in MP3. For the same reason, AAC tends to use less window-switching than MP3. As a result, a significant amount of transient signals such as beats are still coded with long window (2048 PCM samples) in an AAC encoder according to our examinations of AAC bitstreams. The reduced time resolution increases the effect of double-drumbeat problems, if using simple repetition or drumbeat replacement [16]. Figure 7 illustrates potential problems with our previous method described in [16], if the locations of the original beat and the replacement beat are not consistent.

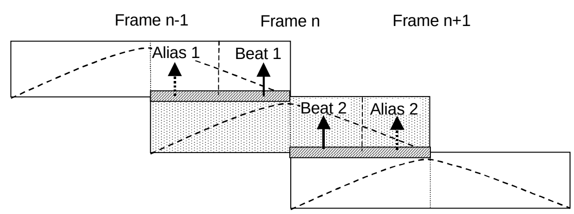


Figure 7. Illustration of the possible quadruple-drumbeat problem in the case of beat replacement when using a long MDCT transform block.

It is impossible to solve the problem with the time resolution of the AAC frame length (2048 samples). However, if the beat detection is performed with an increased time resolution as illustrated in Figure 8, we will have a better chance to tackle the double/quadruple drumbeat problem.

In order to increase the time resolution of the beat detector, we perform a parallel signal analysis with the short windows, whose time resolution is improved by factor of 8 as shown in Figure 8. In this case, we will know the more precise position of a beat within each frame. If the sampling frequency is 44.1 kHz, the original time resolution is about 23 ms, the improved time resolution is about 3 ms, which is close to the time resolution of the human ear [17]. With the improved time resolution, we not only know the more precise location of the beat, but also the location of its alias according to the symmetric property of MDCT. With this information we can boost the desired beat and attenuate the undesired ones accordingly in order to improve the performance of the error concealment method in [16]. A detailed description of the new method will be published elsewhere.

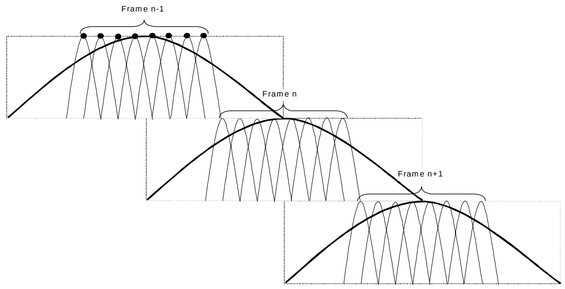


Figure 8. Illustration of the improved time resolution of the beat detector. The 8 dots represent the central position of each short window, which indicate the finer time grids.

5 DISCUSSION

This paper has presented a study of the Modified Discrete Cosine Transform (MDCT) and its implications for audio coding and error concealment from the perspective of Fourier frequency analysis.

Some remarks on MDCT based on our study:

- MDCT becomes an orthogonal transform, if the signal length is infinite. This is different from the traditional definition of orthogonality, which requires a square transform matrix.

- The MDCT spectrum of a signal is the Fourier spectrum of the signal mixed with its alias. This compromises the performance of MDCT as a Fourier spectrum analyser and leads to possible mismatch problems between MDCT and DFT based perceptual models. Nevertheless, MDCT has been successfully applied to perceptual audio compression without major problems if a proper window such as a sine window is employed.

- The TDAC of an MDCT filterbank can only be achieved with an overlap-add (OA) process in the time domain. Although MDCT coefficients are quantized in an individual data block, it is usually analyzed in the context of a continuous stream. In the case of discontinuity such as editing or error concealment, the aliases of the two neighboring blocks in the overlapped area are not able to cancel each other.

- MDCT can achieve perfect reconstruction only without quantization, which is never the case in coding applications. If we model the quantization as a superposition of quantization noise to the MDCT coefficients, then the time domain alias of the input signal is still cancelled, but the noise components will be extended as additional "noise alias". In order to have 50% window overlap and critical sampling simultaneously, the MDCT time domain window is twice as long as that of ordinary orthogonal transforms such as DCT. Because of the increased time domain window length, the quantization noise is spread to the whole window, thus making pre-echo more likely to be audible. Well-known solutions to this problem are window switching [4] and temporal noise shaping (TNS) [14].

- In very low bitrate coding, the high frequency components are often removed. This corresponds to a very steep lowpass filter. Due to the increased window size, the ringing effect caused by high frequency cutting is longer.

Two cases of applications are studied: MDCT domain audio coding and error concealment. Some challenges are presented with possible solutions.

REFERENCES

[1] Princen, J. P., Bradley, A. B., "Analysis/Synthesis Filter Bank Design Based on Time Domain Aliasing Cancellation," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-34, No. 5, October 1986.

[2] Princen, J. P., Johnson, A. W., Bradley, A. B., "Subband/Transform Coding Using Filter Bank Designs Based on Time Domain Aliasing Cancellation," IEEE International Conference on Acoustics, Speech, and Signal Processing, 1987, Dallas, USA, pp. 2161-2164.

[3] Rothweiler, J.H., "Polyphase Quadrature Filters – A New Subband Coding Technique," IEEE International Conference on Acoustics, Speech, and Signal Processing, 1983, Boston, USA, pp. 1280-1283.

[4] Edler, B., "Coding of Audio Signals with Overlapping Block Transform and Adaptive Window Functions," (in German), Frequenz, vol.43, pp.252-256, 1989

[5] Wang, Y., "Selected Advances in Audio Compression and Compressed Domain Processing," Ph.D. thesis, Tampere University of Technology, 2001.

[6] Ferreira, A., "Spectral Coding and Post-Processing of High Quality Audio," Ph.D. thesis, University of Proto, http://telecom.inescn.pt/doc/phd_en.html

[7] Malvar, H., "A Modulated Complex Lapped

Transform and Its Apllications to Audio Processing," IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999, Phoenix, USA.

[8] Malvar, H., "Signal Processing with Lapped Transforms," Artech House, Inc., 1992.

[9] ISO/IEC JTC1/SC29/WG11, "Coding of moving pictures and audio - MPEG-2 Advanced Audio Coding AAC," ISO/IEC 13818-7 International Standard, 1997.

[10] Wang, Y., Yaroslavsky, L., Vilermo, M., Väänänen, M. "Restructured Audio Encoder for Improved Computational Efficiency," AES 108th International Convention, February 19-22, 2000, Paris, France.

[11] Wang, Y., Yaroslavsky, L., Vilermo, M., "On the Relationship between MDCT, SDFT and DFT," 16th IFIP World Computer Congress (WCC2000)/5th International Conference on Signal Processing (ICSP2000), August 21-25, 2000, Beijing, China.

[12] Wang, Y., Yaroslavsky, L., Vilermo, M., Väänänen, M. "Some Peculiar Properties of the MDCT," 16th IFIP World Computer Congress (WCC2000)/5th International Conference on Signal Processing (ICSP2000), August 21-25, 2000, Beijing, China.

[13] Yaroslavsky, L., Eden, M., "Fundamentals of Digital Optics," Birkhauser, Boston, 1996.

[14] Bosi, M., Brandenburg, K., Quackenbush, S., Fielder, L., Akagiri, K., Fuchs, H., Dietz, M., Herre, J., Davidson, G., Oikawa, Y., "ISO/IEC MPEG-2 Advanced Audio Coding," Journal of Audio Engineering Society, vol. 45, no. 10, 1997.

[15] Perkins, C., Hodson, O., Hardman, V., "A Survey of Packet-loss Recovery Techniques for Streaming Audio," IEEE Network, Sept/Oct 1998

[16] Wang, Y., Streich, S., "A Drumbeat-Pattern based Error Concealment Method for Music Streaming Applications," Accepted by IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2002), May 13-17, 2002, Orlando, Florida, USA.

[17] Moore, B.C.J, "An Introduction to the Psychology of Hearing," 4. Edition, Academic Press, London, 1997.