Modified Discrete Cosine Transform—Its Implications for Audio Coding and Error Concealment*

YE WANG, *AES Member,* **AND MIIKKA VILERMO**

Nokia Research Center, FI-33721 Tampere, Finland

A study of the modified discrete cosine transform (MDCT) and its implications for audio coding and error concealment is presented from the perspective of Fourier frequency analysis. A relationship between MDCT and DFT via shifted discrete fourier transform (SDFT) is established, which provides a possible fast implementation of MDCT employing a fast Fourier transform (FFT) routine. The concept of time-domain alias cancellation (TDAC), the symmetric and nonorthogonal properties of MDCT, is analyzed and illustrated with intuitive examples. New insights are given for innovative solutions in audio codec design and MDCT-domain audio processing such as error concealment.

0 INTRODUCTION

With the 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.

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 an energy compaction capability similar to DCT, MDCT simultaneously achieves critical sampling, a reduction of the block effect, and flexible window switching.

In applications such as streaming audio to handheld devices, it is often necessary to have fast implementations and optimized 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 were motivations for us to conduct this study.

MDCT uses the concept of time-domain alias cancellation (TDAC) [1], [2], whereas the quadrature mirror filter bank (QMF) uses the concept of frequency-domain alias cancellation [3]. This can be viewed as a duality of MDCT and QMF. However, it should be noted that MDCT also cancels frequency-domain aliasing, whereas QMF does not cancel time-domain aliasing. In other words, MDCT is designed to achieve perfect reconstruction, QMF is not.

Prior to the introduction of MDCT, transform-based audio coding techniques used the discrete Fourier transform (DFT) and the discrete cosine transform (DCT) with window functions such as rectangular and sine-taper functions. However, these early coding techniques have failed to fulfill the contradictory requirements imposed by high-quality audio coding. For example, with a rectangular window the analysis/synthesis system is critically sampled, that is, the overall number of 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 these transformers are not critically sampled. MDCT has solved the paradox satisfactorily and is currently the best solution. The concept of 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 models of human auditory systems could still be the cause of certain coding artifacts at low bit rates [5].

A complex version of MDCT has been investigated in [6]–[8] in terms of filter-bank theory. Our research has approached the problem from a different perspective—Fourier spectrum analysis. We hope that this study cannot only provide an intuitive tutorial of the concept of MDCT and TDAC, but also some stimulation for innovative solutions in applications such as MDCT-domain audio processing and error concealment.

* Presented at the AES 22nd International Conference, Espoo, Finland, 2002 June 15–17; revised 2002 September 20.

This paper is organized as follows. A relationship between MDCT and DFT via shifted discrete Fourier transform (SDFT) is established in Section 1. The symmetric properties of the MDCT and TDAC concepts are illustrated in Section 2. The nonorthogonal property of MDCT is then discussed in Section 3. The implications for audio coding and error concealment are outlined in Section 4. Section 5 concludes with some discussions.

1 INTERCONNECTION BETWEEN MDCT, SDFT, AND DFT

The direct and inverse MDCT and its inverse modified discrete cosine transform (IMDCT) are defined as [1], [2]

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

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

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

$$h_k = h_{2N-1-k} \tag{3}$$

$$h_k^2 + h_{k+N}^2 = 1 . \tag{4}$$

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

$$h_k = \sin\left(\pi \frac{k + 1/2}{2N}\right), \qquad k = 0, \ldots, 2N-1 . \tag{5}$$

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

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

The relationship between MDCT and DFT can be established via SDFT. The direct and inverse SDFTs are defined as [9]

$$\mathrm{SDFT}_{u,v} = \alpha_r^{u,v} = \sum_{k=0}^{2N-1} a_k \exp\left[\mathrm{i}2\pi \frac{(k+u)(r+v)}{2N}\right] \tag{7}$$

$$\mathrm{ISDFT}_{u,v} = a_k^{u,v} = \frac{1}{2N} \sum_{r=0}^{2N-1} \alpha_r^{u,v} \exp\left[-\mathrm{i}2\pi \frac{(k+u)(r+v)}{2N}\right] \tag{8}$$

where u and v represent arbitrary time- and frequency-domain shifts, respectively. SDFT is a generalization of DFT, which allows a possible arbitrary shift in position of the samples in the time and frequency domains 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],

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

The righthand side of Eq. (9) is the $\mathrm{SDFT}_{(N+1)/2,1/2}$ of the signal $\hat{a}_k$ formed from the initial windowed signal $\tilde{a}_k$ according to Eq. (6). Physical interpretation of Eq. (6) is straightforward. MDCT coefficients can be obtained by adding the $\mathrm{SDFT}_{(N+1)/2,1/2}$ coefficients of the initial windowed signal and the alias.

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

$$\alpha_r = \mathrm{real}\left\{\mathrm{SDFT}_{(N+1)/2,\,1/2}\left(\tilde{a}_k\right)\right\} . \tag{10}$$

With reference to Eqs. (6) and (9) and Fig. 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 in Fig. 1(a)] is mirrored in the time domain and then inverted, before being subsequently added to the original signal. The second half of the window (the signal portion between points B and C) is also mirrored in the time domain and added to the original signal.

From Eqs. (1), (2), (6), and (9) and Fig. 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 $\mathrm{SDFT}_{(N+1)/2,1/2}$ are lost in the MDCT transform, which is equivalent to a decimation operation. Applying an MDCT and then an IMDCT converts the input signal into one that contains a certain twofold symmetric alias [see Eq. (6) and Fig. 1(f)]. The introduced alias is canceled in the overlap–add process to achieve perfect reconstruction (see Fig. 2).

The formulation in Eq. (9) is different when compared with the odd-DFT concept discussed in [6]. The odd-DFT is the $\mathrm{SDFT}_{0,1/2}$ of the initial windowed signal $\tilde{a}_k$.

The $\mathrm{SDFT}_{(N+1)/2,1/2}$ can be expressed by means of the conventional DFT as

$$\begin{aligned} &\sum_{k=0}^{2N-1} \hat{a}_k \exp\left\{\mathrm{i}2\pi \frac{\left[k + (\mathrm{N}+1)/2\right](r + 1/2)}{2N}\right\} \\ &\quad = \left\{\sum_{k=0}^{2N-1}\left[\hat{a}_k \exp\left(\mathrm{i}2\pi \frac{k}{4N}\right)\right] \exp\left(\mathrm{i}2\pi \frac{kr}{2N}\right)\right\} \\ &\qquad \times \exp\left[\mathrm{i}2\pi \frac{(N+1)r}{4N}\right] \exp\left(\mathrm{i}\pi \frac{N+1}{4N}\right) . \end{aligned} \tag{11}$$

On the right-hand side of Eq. (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 one-half 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 $\mathrm{SDFT}_{(N+1)/2,1/2}$ 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 one-half the frequency-sampling interval. This for-

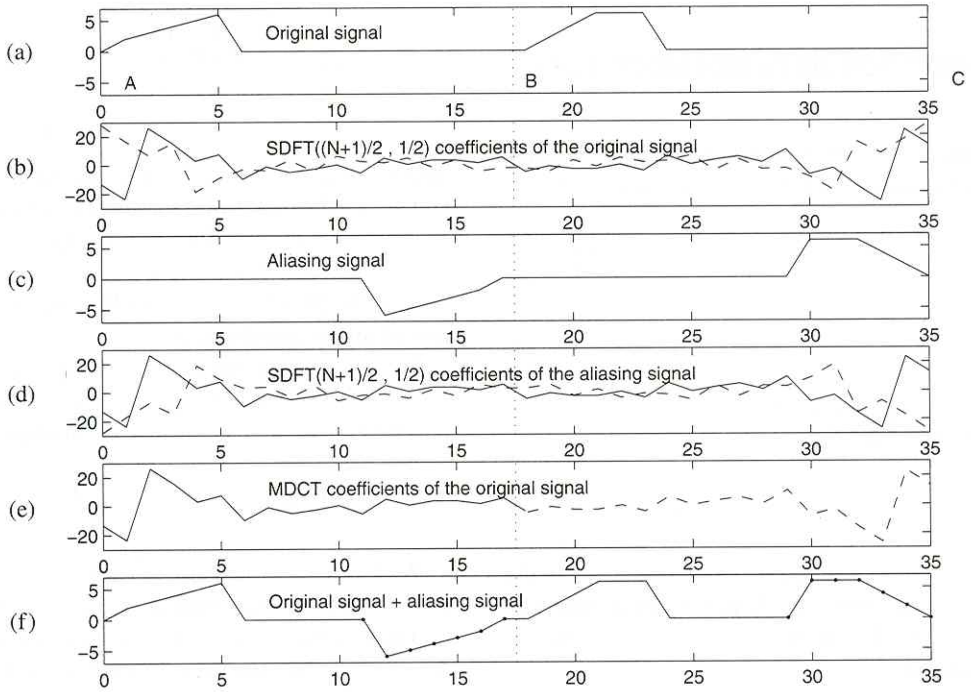


Fig. 1. Relationship between MDCT and $\mathrm{SDFT}_{(N+1)/2,1/2}$. N is even. (a) Artificial time-domain signal of 36 samples. (b) $\mathrm{SDFT}_{(N+1)/2,1/2}$ coefficients of signal in (a). (c) Time-domain alias. (d) $\mathrm{SDFT}_{(N+1)/2,1/2}$ coefficients of alias. —— real parts; – – – imaginary parts in (b) and (d). (e) MDCT coefficients of time signal in (a). – – – odd symmetric to solid line, thus is redundant. (f) Alias embedded time signal.

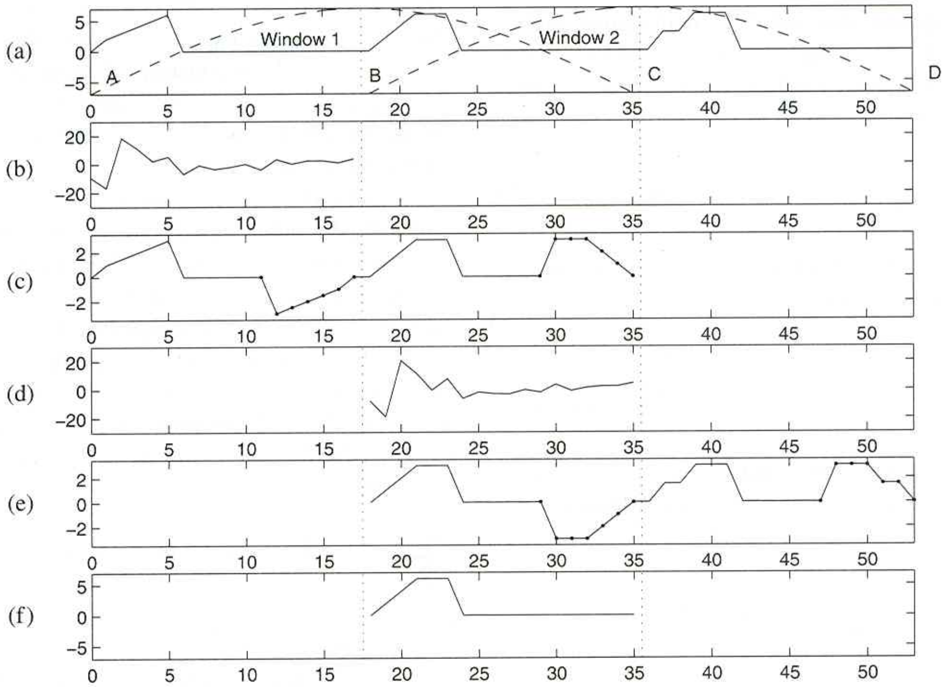


Fig. 2. Illustration of MDCT, overlap–add procedure, and time-domain alias cancellation (TDAC). (a) Artificial time signal. – – – 50% overlapped windows. (b) MDCT coefficients of signal in window 1. (c) IMDCT coefficients of signal in (b); alias shown by markers on line. (d) MDCT coefficients of signal in window 2. (e) IMDCT coefficients of signal in (d); alias shown by markers on line. (f) Reconstructed time-domain signal after overlap–add procedure. Original signal in overlapped part (between points B and C) is perfectly reconstructed.

mulation provides one possible fast implementation using an FFT routine.

2 SYMMETRIC PROPERTIES OF MDCT AND TDAC CONCEPTS

2.1 Symmetric Property of MDCT

The $\mathrm{SDFT}_{(N+1)/2,1/2}$ coefficients exhibit symmetric properties,

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

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

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

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

We have proven that

$$\mathrm{IMDCT}(\alpha_r) = \mathrm{ISDFT}_{(N+1)/2,\,1/2}\,(\alpha_r)\,, \qquad r = 0\,,\ldots\,,2N-1\,. \tag{14}$$

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, in order to have $2N$ dependent frequency components, to apply the symmetric property of MDCT to the ISDFT routine, as shown in Fig. 1(e).

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

2.2 Intuitive Illustration of TDAC Concept

Based on Eqs. (1), (2), (6), and (9), we have used a similar artificial time-domain signal as in Fig. 1(a) to illustrate the TDAC concept in an intuitive way. The artificial signal of 54 samples is shown in Fig. 2(a). The MDCT coefficients of the signal in window 1 are shown in Fig. 2(b). To illustrate the concept, a rectangular window is used. Due to the 50% decimation in MDCT [from $2N$ time-domain samples in Fig. 2(a) to N independent frequency-domain coefficients in Fig. 2(b)], the alias is introduced. This is illustrated in Fig. 2(c). The IMDCT introduces redundancy [from N frequency-domain coefficients in Fig. 2(b) to $2N$ time-domain samples in Fig. 2(c)]. The MDCT coefficients of the signal in window 2 are presented in Fig. 2(d). The corresponding IMDCT time-domain signal is shown in Fig. 2(e). If the overlap–add procedure is performed with Fig. 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 for the first half of the first window and the second half of the last window, as indicated in Fig. 2.

In order to illustrate the TDAC concept during the window switching specified in the MPEG AAC ISO/IEC standard [12], 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 Eq. (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]. Window switching is an important concept to reduce pre-echo in an MDCT-based audio codec such as MPEG-2 AAC. The TDAC concept during window switching in AAC is illustrated in Fig. 3.

3 NONORTHOGONAL PROPERTY OF MDCT

3.1 Observation from a Single Transform Block

If a signal exhibits local symmetry such that

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

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

If a signal exhibits local symmetry such that

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

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

To illustrate in an intuitive way that MDCT does not fulfill Parseval's theorem, we have designed a phase/ frequency-modulated time signal in Fig. 4(a), which has two different frequency elements with a duration of half a frame size (frame size = 512 samples). The dashed lines in Fig. 4(a) illustrate the 50% window overlap. However, the MDCT spectra of different time slots in Fig. 4(b), (d), and (f) are calculated with rectangular windows for illustrative purposes. The IMDCT time-domain samples of frames 1, 2, and 3 are shown in Fig. 4(c), (e), and (g), respectively. The reconstructed time-domain samples after the overlap–add procedure are shown in Fig. 4(h). With frame 2 the condition in Eq. (17) holds, and the MDCT coefficients are all zero. Nevertheless the time-domain samples in frame 2 can still be reconstructed perfectly after the overlap–add procedure. With frame 3 the condition in Eq. (18) holds, and the original time samples are reconstructed perfectly even without the overlap–add procedure. These are, of course, very special occurrences, which are rare in real-life audio signals, especially after

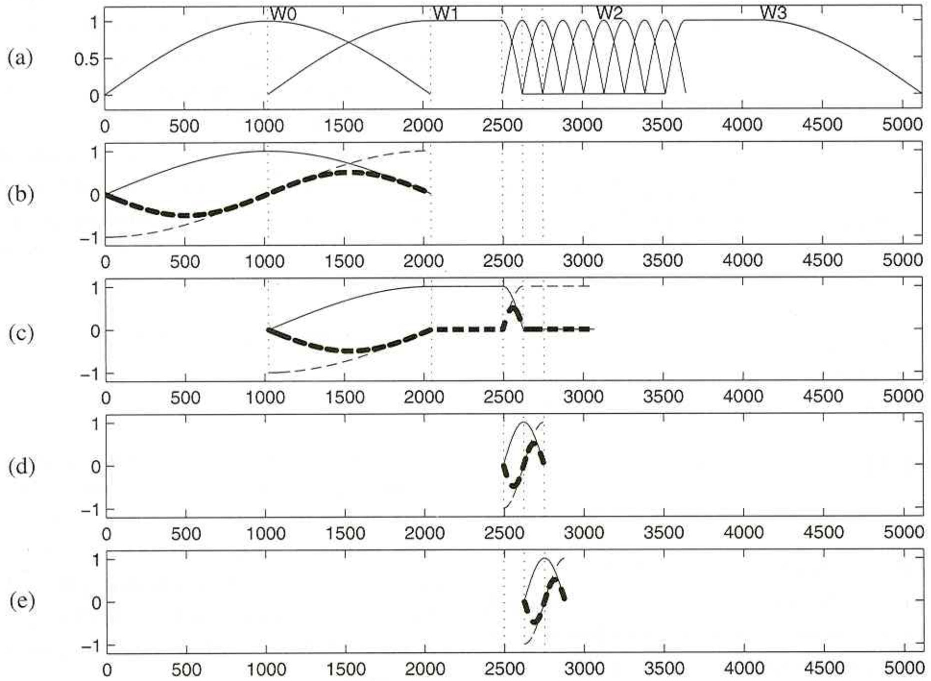


Fig. 3. TDAC in case of window switching. (a) Four types of window shape in MPEG-2 AAC indicated by W0, … , W3. (b) Window function in long window (——), time-domain alias (– – –), and time-domain alias after weighting with window function (**– – –**). (c) Window function in transition window (——), time-domain alias (– – –), and time-domain alias after weighting with window function (**– – –**). (d), (e) Window function in short window (——), time-domain alias (– – –), and time-domain alias after weighting with window function (**– – –**).

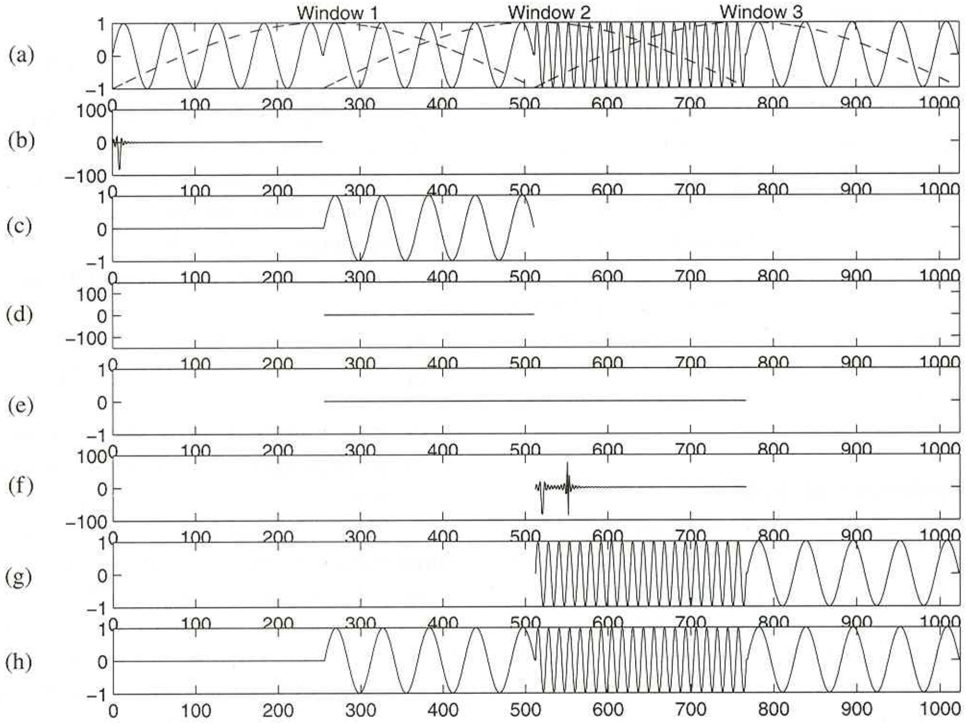


Fig. 4. Signal analysis/synthesis with MDCT, overlap–add procedure, and perfect reconstruction of time-domain samples. (a) Phase/frequency-modulated time signal. (b), (d), (f) MDCT spectra in different time slots, indicated in (a) as frames 1, 2, 3. (c), (e), (g) reconstructed time-domain samples (with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

proper windowing such as a sine window. If the signal is close to the condition in Eq. (17), however, the MDCT spectrum will be very unstable in comparison with the DFT spectrum. In this case, using the output of the DFT-based psychoacoustic model to quantize the MDCT coefficients could cause certain coding artifacts. This is a limitation of MDCT.

3.2 Observation from Multiple Transform Blocks

As shown in Eq. (19),

$$P = \begin{bmatrix} p_{1,1} & p_{1,2} & p_{1,3} & \cdots & p_{1,2N} \\ p_{2,1} & p_{2,2} & p_{2,3} & \cdots & p_{2,2N} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ p_{N,1} & p_{N,2} & p_{N,3} & \cdots & p_{N,2N} \end{bmatrix} \tag{19}$$

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.

In the case of a continuous input stream x, a block-diagonal matrix T can be made with the MDCT matrices P on the diagonal and zeros elsewhere,

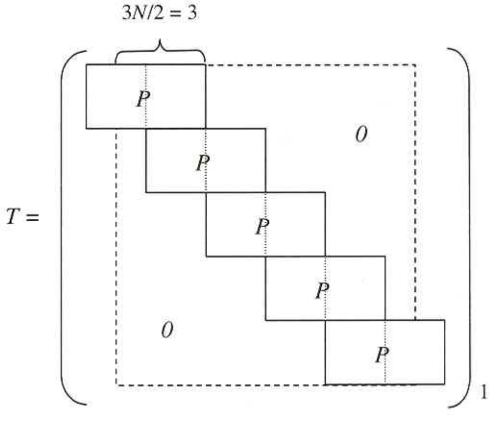
$$X_{nN} = \begin{bmatrix} P & & & & 0 \\ & P & & & \\ & & \ddots & & \\ & & & P & \\ 0 & & & & P \end{bmatrix}_{(nN)\times[(n+1)N]} \cdot x_{(n+1)N} = T \cdot x_{(n+1)N} \tag{20}$$

where x is the input vector of the signal and X is the output vector of the MDCT coefficients. This block-diagonal matrix T for transforming $(n + 1)\,N$ input samples to nN spectral components is of size $(nN) \times [(n + 1)\,N]$. T becomes an orthogonal and square matrix if $n \rightarrow \infty$.

The orthogonality of T implies

$$T^{\mathrm{T}} \cdot T = T \cdot T^{\mathrm{T}} = I\,. \tag{21}$$

However, in the case of finite-length input signals, T is no longer 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,

$$T \cdot T^{\mathrm{T}} = \begin{bmatrix} \begin{bmatrix}1&0\\0&1\end{bmatrix} & & & & 0 \\ & \begin{bmatrix}1&0\\0&1\end{bmatrix} & & & \\ & & \begin{bmatrix}1&0\\0&1\end{bmatrix} & & \\ & & & \begin{bmatrix}1&0\\0&1\end{bmatrix} & \\ 0 & & & & \begin{bmatrix}1&0\\0&1\end{bmatrix} \end{bmatrix} = I \tag{23}$$

$$T^{\mathrm{T}} \cdot T = \begin{bmatrix} \begin{bmatrix}0.5&0.5\\0.5&0.5\end{bmatrix} & & & & & 0 \\ & \begin{bmatrix}1&0\\0&1\end{bmatrix} & & & & \\ & & \begin{bmatrix}1&0\\0&1\end{bmatrix} & & & \\ & & & \begin{bmatrix}1&0\\0&1\end{bmatrix} & & \\ & & & & \begin{bmatrix}1&0\\0&1\end{bmatrix} & \\ 0 & & & & & \begin{bmatrix}0.5&0.5\\0.5&0.5\end{bmatrix} \end{bmatrix}. \tag{24}$$

It is clear from Eq. (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 MDCT domain, such as editing and error concealment. Two applications of the theoretical background are discussed in the following section.

4 IMPLICATIONS FOR AUDIO CODING AND ERROR CONCEALMENT

4.1 MDCT-Based Perceptual Audio Coding

Modern perceptual audio encoders are conceptually similar in the sense that they consist of four basic building blocks: a transform or filter bank (such as MDCT), a perceptual model, requantization and coding, and bit-stream formatting. The basic structure of an MDCT-based audio encoder is shown in the block diagram of Fig. 5, where the bit-stream formatting is omitted.

The concept of perceptual audio coding (bit-rate reduction) described from the viewpoint of quantization-noise shaping is as follows. Initially a PCM signal, such as music on a commercial CD, has the quantization noise distributed uniformly across the whole frequency band. A transform or filter bank creates a frequency-domain representation of this signal. A perceptual model usually uses the original signal to estimate a time- and frequency-dependent masking threshold, indicating the maximum quantization noise inaudible in the presence of this audio signal. By requantization a quantizer then reduces the number of bits used to represent this signal, which will result in an increase and shaping of quantization noise to the limit of the masking threshold. This explains the significance of masking in perceptual audio coding technologies.

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

$$\mathrm{EN} = \frac{\mathrm{ES}}{\mathrm{SMR}} \tag{25}$$

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, the 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, which does not seem to pose a serious problem in coding applications if a proper window function is used.

Another important issue in audio encoder design is computational simplicity. In an MDCT-based audio encoder a complex transform such as the FFT is a necessary step for the psychoacoustic model (see Fig. 5). To reduce the computational complexity of the encoder, it is desired that the MDCT and the complex frequency-domain values required in the psychoacoustic model may be calculated from the same set of computations. Luckily this desire can be fulfilled via an SDFT. The simplified encoder structure is illustrated in Fig. 6.

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 to 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 [13]. 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 neighboring packet, the drumbeat is simply eliminated. On the other hand, if the drumbeat is copied to the 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. It also depends on the distance between the original drumbeat and that generated due to packet repetition [14].

Due to the nonorthogonal property of MDCT, the repetition violates the TDAC conditions. Consequently the alias distortions in the overlapped parts cannot cancel each other out (Fig. 7). However, the MDCT window functions enable a natural fade-in and fade-out in the overlap–add

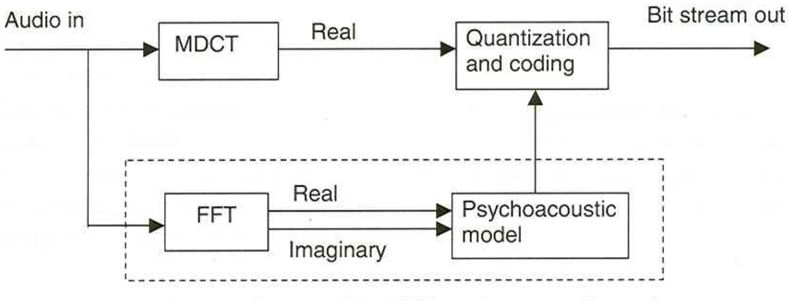


Fig. 5. Block diagram of MDCT-based perceptual encoder.

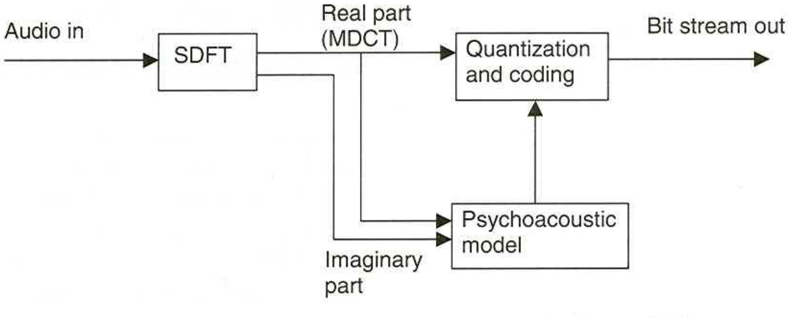


Fig. 6. Modified structure of MDCT-based perceptual encoder in Fig. 5 with less computation.

operation in the time domain. The uncancelled alias is normally not perceptible if the signal is stationary and the lost data unit is short enough.

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 Fig. 8.

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, and 3, respectively. The short window is introduced to tackle transient signal better; 50% window overlap is used with MDCT.

If two consecutive short window frames indexed as 22 in a window-switching sequence 1223 are lost in a transmission channel, it is easy to deduce their window types from their neighboring frames. This information could be used in error concealment [14]. However, if we disregard the window-switching information available from the audio bit stream and perform simple repetition, it could result in window-switching patterns of 1113 (see Fig. 8). In this case not only are the TDAC conditions violated in the window overlapped areas, but we also 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, state-of-the-art audio coding techniques tend to use longer transform block lengths than their predecessors, for example, 1024 MDCT coefficients, which correspond to 2048 PCM samples in AAC. 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 a long window in an AAC encoder according to our examinations of AAC bit streams. The reduced time resolution increases the effect of double-drumbeat problems if simple repetition or drumbeat replacement is used [14]. Fig. 9 illustrates potential problems with our previous method, described in [14], if the locations of the original beat and the replacement beat are not consistent.

It is impossible to solve the problem with the time resolution of the AAC frame length. However, if the beat detection is performed with an increased time resolution, as illustrated in Fig. 10, we will have a better chance to tackle the double/quadruple drumbeat problem.

To increase the time resolution of the beat detector, we perform a parallel signal analysis with the short windows, which improves the time resolution by a factor of 8, as shown in Fig. 10. 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

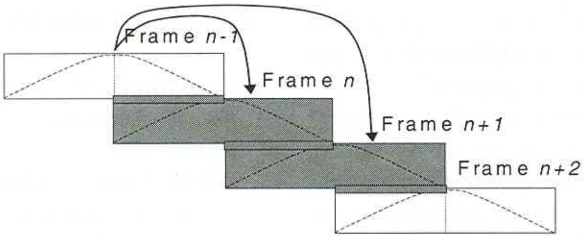


Fig. 7. Illustration of a special problem with repetition scheme in MDCT domain. Shaded rectangles—corrupted data units; blank rectangles—error-free ones; heavily shaded rectangles— uncancelled alias; – – – window shape. Arrows indicate packet repetition operations. *n* is an integer number representing data unit index.

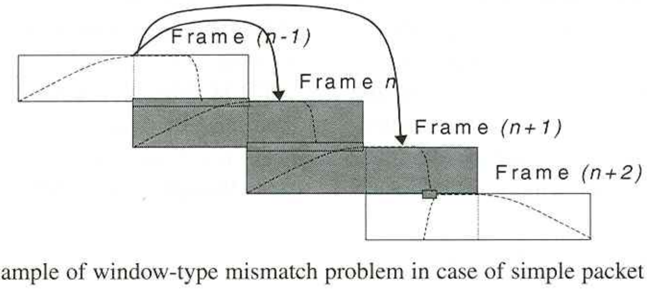


Fig. 8. Example of window-type mismatch problem in case of simple packet repetition.

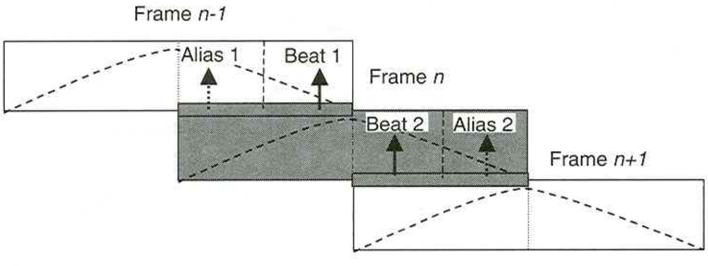


Fig. 9. Possible quadruple-drumbeat problem in case of beat replacement when using a long MDCT transform block. Original (beat 1) and inserted (beat 2) beats are not aligned in time, and their aliases (alias 1 and alias 2) do not cancel each other.

about 23 ms, and the improved time resolution is about 3 ms, which is close to the time resolution of the human ear [15]. 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 in order to improve the performance of the error concealment method described in [14]. A detailed description of the new method will be published elsewhere.

5 DISCUSSION

A study of the modified discrete cosine transform (MDCT) and its implications for audio coding and error concealment has been presented from the perspective of Fourier frequency analysis. Some remarks on MDCT are 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 analyzer and leads to possible mismatch problems between MDCT- and DFT-based perceptual models. Nevertheless MDCT has been applied successfully to perceptual audio compression without major problems if a proper window, such as a sine window, is employed.
- The TDAC of an MDCT filter bank can only be achieved with the overlap–add process in the time domain. Although MDCT coefficients are quantized in an individual data block, MDCT 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 out.
- 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 canceled, 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 increase 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) [16].
- In very low bit-rate coding the high-frequency components are often removed. This corresponds to a very steep low-pass filter. Due to the increased window size, the ringing effect caused by high-frequency cutting is longer.

Two application types are studied—MDCT-domain audio coding and error concealment. Some challenges are presented with possible solutions.

6 REFERENCES

[1] J. P. Prince and A. B. Bradley, "Analysis/Synthesis Filter Bank Design Based on Time Domain Aliasing Cancellation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34 (1986 Oct.).

[2] J. P. Prince, A. W. Johnson, and A. B. Bradley, "Subband/Transform Coding Using Filter Bank Designs Based on Time Domain Aliasing Cancellation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (Dallas, TX, 1987), pp. 2161–2164.

[3] J. H. Rothweiler, "Polyphase Quadrature Filters—A New Subband Coding Technique," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (Boston, MA, 1983), pp. 1280–1283.

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

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

[6] A. Ferreira, "Spectral Coding and Post-Processing of High Quality Audio," Ph.D. thesis, University of Proto, Finland (1998).

[7] H. Malvar, "A Modulated Complex Lapped Transform and Its Applications to Audio Processing," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (Phoenix, AZ, 1999), pp. 1421–1424.

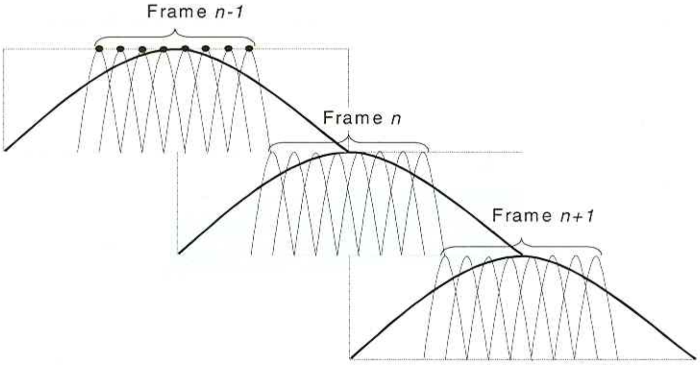


Fig. 10. Improved time resolution of beat detector. • central position of each short window, indicating finer time grids.

[8] H. Malvar, *Signal Processing with Lapped Transforms* (Artech House, Boston, MA, 1992).

[9] L. Yaroslavsky and M. Eden, *Fundamentals of Digital Optics* (Birkhauser, Boston, MA, 1996).

[10] Y. Wang, L. Yaroslavsky, M. Vilermo, and M. Väänänen, "Restructured Audio Encoder for Improved Computational Efficiency," presented at the 108th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 48, p. 352 (2000 Apr.), preprint 5103.

[11] Y. Wang, L. Yaroslavsky, and M. Vilermo, "On the Relationship between MDCT, SDFT, and DFT," presented at the 16th IFIP World Computer Congr. (WCC2000)/5th Int. Conf. on Signal Processing (ICSP2000), Beijing, China, 2000 Aug. 21–25.

[12] ISO/IEC 13818-7, "Coding of Moving Pictures and Audio—MPEG-2 Advanced Audio Coding AAC," ISO/IEC JTC1/SC29/WG112, International Standards Organization, Geneva, Switzerland (1997).

[13] C. Perkins, O. Hodson, and V. Hardman, "A Survey of Packet-Loss Recovery Techniques for Streaming Audio," *IEEE Network* (1998 Sept./Oct.).

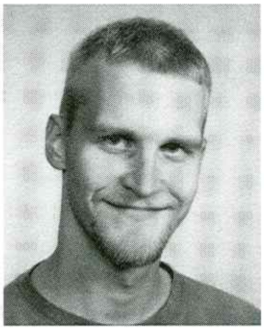
[14] Y. Wang and S. Streich, "A Drumbeat-Pattern-Based Error Concealment Method for Music Streaming Applications," presented at the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP2002), Orlando, FL, 2002 May 13–17.

[15] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 4th ed. (Academic Press, London, 1997).

[16] J. Herre and J. D. Johnston, "Enhancing the Performance of Perceptual Audio Coders by Using Temporal Noise Shaping (TNS)," presented at the 101st Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 44, p. 1175 (1996 Dec.), preprint 4384.

THE AUTHORS

Y. Wang



M. Vilermo

Ye Wang received a B.Sc. degree in wireless communications from the Southern China University of Technology, China in 1983, a Diplom-Ingenieur in telecommunications from the Technische Universität Braunschweig, Germany in 1993, and Licentiate and Doctoral degrees in information technology from the Tampere University of Technology, Finland, in 2000 and 2002, respectively.

He received a Nokia Foundation award and a scholarship from the Academy of Finland, where he worked as a visiting scholar at the Experimental Psychology Department, University of Cambridge, UK, during the spring of 2001. In 1994 he worked as a research engineer in the Speech and Audio Systems Laboratory, Nokia Research Center, Tampere, Finland, and in 2000 became a senior research engineer. He was appointed assistant professor at the National University of Singapore in 2002.

Dr. Wang is a member of the Audio Engineering Society, the AES Technical Committee on Coding of Audio Signals, and the Institute of Electrical and Electronics Engineers. His current research interests include parametric audio compression, compressed domain processing, and error resilient audio content delivery in wireless packet networks. He speaks fluent Chinese, English, and German; and Finnish fairly well. He enjoys working in an international environment with people from different cultural backgrounds.

●

Miikka Vilermo has worked at Nokia Research Center since 1997, first as a trainee and then as assistant research engineer in audio signal processing. His experience includes design of digital signal processing algorithms for high-quality audio applications. His main research topics have been high-quality audio coding and psychoacoustic models.

He is now working toward the completion of his M.Sc. thesis at the Tampere University of Technology, Finland. He has played the violin since he was seven years old and is currently studying at the Tampere Conservatory.