Multimed Tools Appl
DOI 10.1007/s11042-009-0326-4

Spatial parameters for audio coding: MDCT domain analysis and synthesis

Shuixian Chen · Naixue Xiong · Jong Hyuk Park · Min Chen · Ruimin Hu

Abstract We use Modified Discrete Cosine Transform (MDCT) to analyze and synthesize spatial parameters. MDCT in itself lacks phase information and energy conservation, which are needed by spatial parameters representation. Completing MDCT with Modified Discrete Sine Transform (MDST) into "MDCT-*j**MDST" overcomes this and enables the representation in a form similar to that of DFT. And due to overlap-add in time domain, a MDST spectrum can be built perfectly from MDCT spectra of neighboring frames through matrix-vector multiplication. The matrix is heavily diagonal and keeping only a small number of its sub-diagonals is sufficient for approximation. When using MDCT based core coder in spatial audio coding, like Advanced Audio Coding (AAC), we need no separate transforming for spatial processing, cutting down significantly the computational complexity. Subjective listening tests also show that MDCT domain spatial processing has no quality impairment.

Keywords Audio coding · MDCT · MDST · Singular value · Spatial parameter

S. Chen · R. Hu
Computer School, Wuhan University, Wuhan, China

S. Chen
e-mail: csx792@163.com

N. Xiong
Department of Computer Science, Georgia State University, Atlanta, GA, USA
e-mail: nxiong@cs.gsu.edu

J. Hyuk Park (✉)
Department of Computer Science and Engineering, Kyungnam University, Masan, Korea
e-mail: parkjonghyuk1@hotmail.com

M. Chen
School of Computer Science & Engineering, Seoul National University, Seoul 151-744, Korea
e-mail: mchen@mmlab.snu.ac.kr

Published online: 2009

1 Introduction

Stereo and multichannel audio coding based on spatial parameters leads to sharp reduction in bitrate, this coding scheme, called Spatial Audio Coding (SAC), first emerged in Binaural Cue Coding (BCC) [4–6, 19, 21–25], then was enhanced and specialized in Parametric Stereo (PS) [2, 13, 28, 51, 52] and MPEG Surround (MPS) [10, 12, 14, 15, 17, 29–31, 33, 46, 49, 50], both of which have been standardized in MPEG-4 recently. In case of PS, almost half of the bitrate is saved—a little overhead of the spatial parameters added to a mono audio versus "left + right" as in MPEG-1 Layer III (MP3) and MPEG-2 Advanced Audio Coding (AAC) [9, 32]—with little quality impairment [13]. And MPEG Surround has only more startling bitrate saving for multichannel audios [12, 29, 30].

The workhorse of SAC is analysis and synthesis of the spatial parameters in stereo and multichannel audio signals, including Interchannel Level Difference (ILD), Interchannel Time Difference (ITD), and Interchannel Coherence (IC) [7, 27, 53]. Recording an audio signal with more than one channel is in fact discrete spatial domain sampling, which provides sound source localization and size cues, concreting to the above spatial parameters. Binaural hearing studies also find that we have the physiological facilities sensing the parameters [11, 35, 43, 45], alongside sensing amplitude and frequency. Instead of individual channels, we can represent the signal as a downmixed channel for all the time-frequency information and a sequence of the parameters for the spatial information. And the latter is in much lower quantity. It is shown that 8 kbps is sufficient for perceptually transparent coding of stereo spatial parameters [13], but not until above 64 kbps for an individual channel [8]. This turns out to be a more economic way of audio coding.

Location and size are well defined only for an individual sound source during a short time period, so are the ILD, ITD, and IC. Separating sound sources is a very hard problem, it is circumvented by applying short-time time-to-frequency (T/F) transform and taking signal components within each time-frequency zone as a virtual source. DFT and Quadrature Mirror Filterbank (QMF) [13, 52] are two most usual T/F tools for this purpose. PS and MPS achieve good results using them [12, 13, 30]. But both the transforms introduce significant computational load [52] and algorithmic delay [2]. Figure 1 shows the normal practice of SAC: transforming all channels, extracting the spatial parameters, then downmixing to one core channel and transforming it back to time. We can see Fig. 1 that both forward band backward transforms must be performed and buffers are needed for these block operations.

However, if the core encoding uses directly the downmixed T/F spectra, no F/T and accompanied buffering are needed. Both PS and MPS choose AAC as their core encoder, working in Modified Discrete Cosine Transform (MDCT) [37–39, 42, 44, 47, 48] domain, for its quality and establishment. Practically it is impossible to force AAC to use other transforms. Instead we may force spatial analysis and synthesis to use MDCT. But MDCT is a real transform. Phase information is not as explicit as in DFT or QMF—ITD cannot be readily evaluated; energy is not conserved—ILD cannot be readily evaluated too.

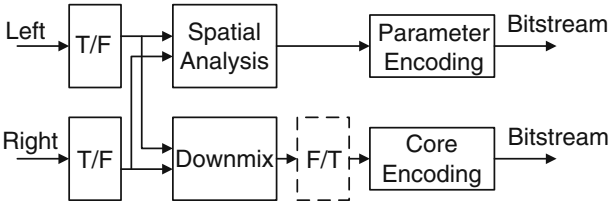


Fig. 1 SAC encoding structure, stereo case

Phase and energy conservation will come back if we combine MDCT with Modified Discrete Sine Transform (MDST). They form a Modified Discrete Fourier Transform (MDFT) [26, 36] in the way of "MDCT-*j**MDST", which is also named Modulated Complex Lapped Transform (MCLT) [40, 41]. But it will be more pleasing if we can avoid explicit MDST computation. MDCT does provide this possibility. Both MDCT and MDST are invertible. Therefore, either can be deduced from the other, by transforming back to time and applying the other transform. So finding MDST from MDCT is theoretically viable, but practically viable only if it involves less computation than *inverse* MDCT (IMDCT) with *forward* MDST.

In this paper, we first derive the spatial parameters in MDFT domain in Section 2. Then we develop a low complexity approximation of MDST from MDCT in Section 3. Section 4 discusses how to apply MDCT domain spatial analysis and synthesis in SAC. Section 5 gives testing results of a stereo coder using the proposed method. We conclude the paper in Section 6.

2 MDFT domain representation

2.1 Time domain spatial parameters

A sound wave travels from source to microphones or ears through different paths. We can model it as a linear time-invariant process with sound source signal $S(t)$, path impulse response $h_l(t)$ and $h_r(t)$, interference signals $n_l(t)$ and $n_r(t)$ which are noises and/or other sound source signals, and received signals $x_l(t)$ and $x_r(t)$ (Fig. 2), or

$$\begin{cases} x_l(t) = S(t)*h_l(t) + n_l(t) \\ x_r(t) = S(t)*h_r(t) + n_r(t) \end{cases}, \tag{1}$$

where "*" indicates convolution. In spatial hearing, $h_l(t)$ and $h_r(t)$ are called Head Related Impulse Response (HRIR) [3], relying on the source location. Normally HRIR is an irregular curve with a time span less than 2 ms. But what is more relevant to sound source localization is their difference, Head Related Difference Impulse Response (HRDIR) $h\Delta(t) = h_r(t) * h_l^{-1}(t)$, since we cannot know $S(t)$ directly. HRDIR is mainly a result of relative level and time differences, or ILD and ITD. Ignoring the interference signals, they can be represented through the observed signals $x_l(t)$ and $x_r(t)$ as

$$\begin{cases} \mathrm{ILD} = 10\log_{10}\left(\int x_l^2(t)dt / \int x_r^2(t)dt\right) \\ \mathrm{ITD} = \arg\max_{\tau}\left\{\int x_l(t)x_r(t+\tau)dt\right\} \end{cases}. \tag{2.a}$$

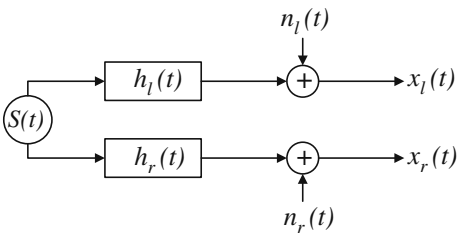


Fig. 2 Sound waves traveling as filtering

Then ILD is the energy ratio, and ITD is the time shift of the maximum inner product. Interference signals de-correlate the observed signals. Then IC takes the normalized correlation to measure the disruption of the noises,

$$\mathrm{IC} = \max_{\tau}\left\{\frac{\int x_l(t)x_r(t+\tau)dt}{\left(\int x_l^2(t)dt\right)^{1/2}\left(\int x_r^2(t)dt\right)^{1/2}}\right\} \tag{2.b}$$

For discrete time signals, we need only substitute the integral with summing.

2.2 Frequency domain representation

For energy conserving complex transforms such as DFT, ILD is readily represented as spectral energy ratio, ITD the group delay, and IC the real part of normalized spectral correlation. But MDCT in itself lacks those advantages. Let $x(n)$, $n = 0, 1, \ldots, 2N-1$ be a $2N$-point time signal. Its MDCT spectrum is

$$X(k) = \langle x, c_k\rangle = \sum_{n=0}^{2N-1} x(n)\cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}+\frac{N}{2}\right)\left(k+\frac{1}{2}\right)\right], \tag{3.a}$$

where $X(k)$ is the kth MDCT spectral line, $k = 0, 1, \ldots, N-1$, and $c_k(n)$ the kth MDCT basis vector. Note (3.a) is a mapping of $2N$-real time points to N-real MDCT lines. This is the root of the trouble—energy cannot be conserved and no phase information is available. Its conjugate transform, MDST, provides just what MDCT lacks. MDST is defined as

$$Y(k) = \langle x, s_k\rangle = \sum_{n=0}^{2N-1} x(n)\sin\left[\frac{\pi}{N}\left(n+\frac{1}{2}+\frac{N}{2}\right)\left(k+\frac{1}{2}\right)\right], \tag{3.b}$$

in which $Y(k)$ is kth MDST spectral line, and $s_k(n)$ the kth MDST basis vector. Taking MDCT as real part and MDST as imaginary part, we have MDFT as

$$Z(k) = X(k) - jY(k) = \sum_{n=0}^{2N-1} x(n)e^{-j\frac{\pi}{N}\left(n+\frac{1}{2}+\frac{N}{2}\right)\left(k+\frac{1}{2}\right)}. \tag{3.c}$$

This is also Shifted Discrete Fourier Transform (SDFT) [54]—conserving energy except for a constant scaling (Appendix A), and converting time shift d into phase shift $-\pi d(k+1/2)/N$ (Appendix B). Then in MDFT domain, the equivalent of (2.a) is

$$\begin{cases}\mathrm{ILD} = 10\log_{10}\dfrac{\sum\|Z_l(k)\|^2}{\sum\|Z_r(k)\|^2}\\ \mathrm{ITD} = \dfrac{d}{dk}\{\arg(Z_l(k)Z_r{}^*(k))\}\ ,\end{cases} \tag{4.a}$$

where $Z_l(k)$ and $Z_r(k)$ are MDFT spectra for the received time signals $x_l(n)$ and $x_r(n)$, the right side of the ITD equation means group delay for the interested MDFT spectrum range or the slope of the regression line. The equivalent of (2.b) is

$$\mathrm{IC} = \frac{\sum\left|Z_l(k)Z_r^*(k)\right|}{\left(\sum\|Z_l(k)\|^2\right)^{1/2}\left(\sum\|Z_r(k)\|^2\right)^{1/2}}. \tag{4.b}$$

In practice, summation in (4.a) and (4.b) is over a subband of a perceptually partitioned band table [13, 25]. And due to the duplex theory, ITD below 1.5 kHz is replaced by the more sensitive average phase difference [25]. For robust estimation, ITD is even replaced by Interchannel Phase Difference (IPD) for all subbands [13].

2.3 Representation in windowed transform

Windowing is a usual procedure for MDCT in audio coding. It scales time signals by a window function. The most often used one is the sine window [37] $w_s(n) = \sin[\pi(n+0.5)/(2N)]$, $n = 0, 1, \ldots, 2N-1$ (Fig. 3). To have similar spatial representation as (4.a) and (4.b), windowed MDFT must preserve DFT-like temporal-spectral correspondence—energy conserving and time shift becomes linear phase shift. Intuitively, MDST should have the same window function. But this will ruin the correspondence.

We shall see that MDST with the cosine function $w_c(n) = \cos[\pi(n+0.5)/(2N)]$, $n = 0, 1, \ldots, 2N-1$ (Fig. 3) does satisfy those requirements. Let $X(k) = \langle w_s x, c_k \rangle$ be the sine-windowed MDCT spectrum, and $Y(k) = \langle w_c x, s_k \rangle$ be the cosine-windowed MDST spectrum. The new MDFT is defined as

$$Z(k) = \begin{cases} Z_-(0) & -jZ_-(0), & k = 0 \\ -Z_+(k-1) & -jZ_-(k), & k = 1, ..., N-1 \\ -Z_+(N-1) & -jZ_+(N-1), & k = N \end{cases} \tag{5.a}$$

where $Z_-(k) = Y(k) - X(k)$ and $Z_+(k) = Y(k) + X(k)$. A very interesting property of (5.a) is that it is equivalent to DFT with a linear phase shift (Appendix C), or

$$\begin{cases} Z(k) = e^{-j\varphi(k)} \sum_{n=0}^{2N-1} x(n) e^{-j\frac{\pi}{N}nk} \\ \varphi(k) = \frac{\pi}{N} k \left(\frac{1}{2} + \frac{N}{2} \right) + \frac{\pi}{4}. \end{cases} \tag{5.b}$$

This way all time-frequency relations of DFT automatically go to the MDFT. Then (4.a) and (4.b) still apply.

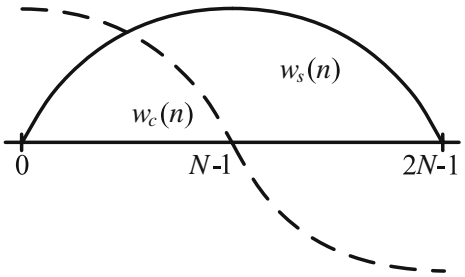


Fig. 3 Window functions $w_s(n)$ and $w_c(n)$ for MDCT and MDST respectively. The shape of $w_c(n)$, odd symmetric about the center, is contrary to intuition on window functions, but essential for DFT-like temporal-spectral correspondence

3 Approximation in MDCT domain

3.1 Deriving MDST from MDCT

AAC performs only MDCT during encoding. So the obvious way to obtain MDFT spectra then ILD, ITD, and IC is to use a separate MDST on incoming audio signals. But it is possible to derive MDST from MDCT. MDCT is invertible, as a row of 50% over-lapped block transform (Fig. 4), its property of Time Domain Aliasing Cancellation (TDAC) [47, 48] enables perfectly reconstructing time points from the same number of MDCT spectral lines. Thus we can find MDST spectra by transforming back to time domain and then to MDST domain. This roundtrip is meaningful only during decoding, where audio time samples are to be found out. During encoding, audio time samples are readily available.

But this roundtrip may have a shortcut, without transforming back and forth. MDCT basis vector $c_k(n)$ has a special structure: odd-symmetric on the first half, time span $0, 1, \ldots, N-1$; and even-symmetric on the second half, time span $N, N+1, \ldots, 2N-1$. MDST basis vector $s_k(n)$ has a similar structure, but even-symmetric on the first half, and odd-symmetric on the second half. In Fig. 4, MDCT on frame i−1 and MDCT on frame i+1, due to the 50% overlap ratio, just provide the even part of the first half and odd part of the second half of frame i respectively, sufficient for MDST on that frame.

3.2 The near-diagonal conversion matrix

We shall see the conversion matrix of frame i−1 MDCT spectrum X^{i-1} and frame i+1 MDCT spectrum X^{i+1} to frame i MDST spectrum Y^i is near-diagonal for typical N used in audio coding. Thus the conversion matrix can be approximated by a small number of sub-diagonals, instead of the full one. This reduces the computational complexity from $O(N^2)$ to $O(mN)$, where m is the number of sub-diagonals for approximation. It is in this sense that the near-diagonal conversion matrix is just the short cut.

We first partition the $2N$-dimensional basis vectors c_k and s_k evenly into N-dimensional column sub-vectors as $(c_k)^{\mathrm{T}} = \left((c_k^0)^{\mathrm{T}} (c_k^1)^{\mathrm{T}}\right)$, and $(s_k)^{\mathrm{T}} = \left(c(s_k^0)^{\mathrm{T}} (s_k^1)^{\mathrm{T}}\right)$, where $k = 0, 1, \ldots, N-1$. The superscript "0" indicates the first half sub-vectors, "1" indicates the second half sub-vectors, and "T" for transpose. Then we have four $N \times N$ matrices

$$\begin{cases} \mathbf{C_0} = \begin{pmatrix} c_0^0 & c_1^0 & \ldots & c_{N-1}^0 \end{pmatrix}, & \mathbf{C_1} = \begin{pmatrix} c_0^1 & c_1^1 & \ldots & c_{N-1}^1 \end{pmatrix} \\ \mathbf{S_0} = \begin{pmatrix} s_0^0 & s_1^0 & \ldots & s_{N-1}^0 \end{pmatrix}, & \mathbf{S_1} = \begin{pmatrix} s_0^1 & s_1^1 & \ldots & s_{N-1}^1 \end{pmatrix} \end{cases}. \tag{6}$$

The definition of MDCT in (3.a) implies that each column vector c_k^0 of $\mathbf{C}_0$ is odd-symmetric while each column vector c_k^1 of $\mathbf{C}_1$ is even symmetric. Also MDCT basis vectors

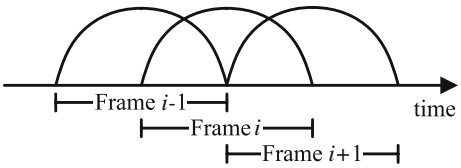


Fig. 4 MDCT with 50% overlap. TDAC ensures perfect reconstruction without redundancy

$c_0, c_1, \ldots, c_{N-1}$ are orthogonal. Then in matrix form, the symmetry and orthogonality of $\mathbf{C}_0$ and $\mathbf{C}_1$ are

$$\begin{cases} \mathbf{J}\mathbf{C}_0 = -\mathbf{C}_0, \mathbf{J}\mathbf{C}_1 = \mathbf{C}_1 \\ \mathbf{C}_0^{\mathrm{T}}\mathbf{C}_0 + \mathbf{C}_1^{\mathrm{T}}\mathbf{C}_1 = N\mathbf{I}, \mathbf{C}_1\mathbf{C}_0^{\mathrm{T}} = \mathbf{C}_0\mathbf{C}_1^{\mathrm{T}} = \mathbf{0}\ ., \end{cases} \tag{7.a}$$

where $\mathbf{J}$ is the $N\times N$ anti-diagonal matrix having only 1 on its anti-diagonal, and $\mathbf{I}$ is the $N\times N$ identity matrix. The definition of MDST in (3.b) implies s_k^0 and s_k^1 can be obtained from c_k^1 and c_k^0 with sign changing, or

$$\mathbf{S}_0 = -\mathbf{C}_1\mathbf{P}, \mathbf{S}_1 = \mathbf{C}_0\mathbf{P}, \tag{7.b}$$

where $\mathbf{P}$ is the $N\times N$ sign changing matrix with only $+1, -1, +1, -1, \ldots$, on its diagonal.

Note N-dimensional column vectors x_0^i and x_1^i the first half and second half of the frame i. They can be reconstructed perfectly from the three MDCT spectra X^{i-1}, X^i, and X^{i+1} for frame $i-1$, i, and $i+1$ respectively, by the following matrix multiplication,

$$\begin{pmatrix} x_0^i \\ x_1^i \end{pmatrix} = \frac{1}{N}\begin{pmatrix} \mathbf{C_1}X^{i-1} + \mathbf{C_0}X^i \\ \mathbf{C_1}X^i + \mathbf{C_0}X^{i+1} \end{pmatrix}. \tag{8}$$

It can be verified from (7.a) that $\mathbf{S}_0^{\mathrm{T}}\mathbf{C}_0 = \mathbf{S}_1^{\mathrm{T}}\mathbf{C}_1 = \mathbf{0}$ (Appendix D). Then substituting (8) into the MDST definition in matrix form, we have Y^i as the following,

$$\begin{aligned} Y^i &= \begin{pmatrix} \mathbf{S_0^T} & \mathbf{S_1^T} \end{pmatrix}\begin{pmatrix} x_0^i \\ x_1^i \end{pmatrix} \\ &= \frac{1}{N}\left(\mathbf{S_0^T}\mathbf{C_1}X^{i-1} + \left(\mathbf{S_0^T}\mathbf{C_0} + \mathbf{S_1^T}\mathbf{C_1}\right)X^i + \mathbf{S_1^T}\mathbf{C_0}X^{i+1}\right) \\ &= \frac{1}{N}\left(\mathbf{S_1^T}\mathbf{C_0} - \mathbf{S_0^T}\mathbf{C_1}\right)X_-^i + \frac{1}{N}\left(\mathbf{S_1^T}\mathbf{C_0} + \mathbf{S_0^T}\mathbf{C_1}\right)X_+^i \end{aligned} \tag{9}$$

where $X_-^i = (X^{i+1} - X^{i-1})/2$, and $X_+^i = (X^{i+1} + X^{i-1})/2$. Since X^i is annihilated, the MDST spectrum Y^i for frame i can be built perfectly from neighboring MDCT spectra X^{i-1} and X^{i+1} or equivalently X_-^i and X_+^i.

Each column vector of $\mathbf{C_0}$, $\mathbf{C_1}$, $\mathbf{S_0}$, and $\mathbf{S_1}$ is part of a cosine or sine sequence with different frequencies. As N increases, these vectors tend to be orthogonal, and the matrices $\mathbf{S_0^T}\mathbf{C_1}$ and $\mathbf{S_1^T}\mathbf{C_0}$ become more diagonal then, there is however more diagonality to explore when considering their sum and difference. Through (7.b) and (7.a), we have the difference matrix

$$\frac{1}{N}\left(\mathbf{S_1^T}\mathbf{C_0} - \mathbf{S_0^T}\mathbf{C_1}\right) = \frac{1}{N}\mathbf{P}^{\mathrm{T}}\left(\mathbf{C_0^T}\mathbf{C_0} + \mathbf{C_1^T}\mathbf{C_1}\right) = \mathbf{P}^{\mathrm{T}}, \tag{10.a}$$

which has only $(-1)^k$ entries on the diagonal; and noting $\theta = \pi/(2N)$, the sum matrix $\mathbf{T}$

$$\begin{aligned} (\mathbf{T})_{k,l} &= \frac{1}{N}\left(\mathbf{S_1^T}\mathbf{C_0} + \mathbf{S_0^T}\mathbf{C_1}\right)_{k,l} \\ &= \begin{cases} \dfrac{\mathrm{Re}\{j^{k+l-1}\}}{N\sin[\theta(k-l)]}, & k-l = odd \\ \dfrac{\mathrm{Re}\{j^{k-l+2}\}}{N\sin[\theta(k+l+1)]}, & k-l = even \end{cases}. \end{aligned} \tag{10.b}$$

whose entry $(\mathbf{T})_{k,l}$ vanishes quickly away from the diagonal (Fig. 5).

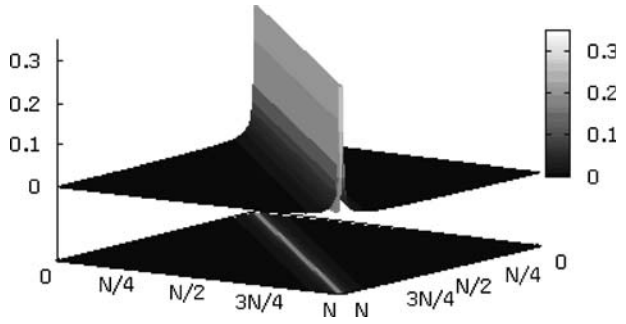


Fig. 5 $\left|(\mathbf{T})_{k,l}\right|$ of the sum matrix $\mathbf{T} = \left(\mathbf{S_0^T C_1} + \mathbf{S_1^T C_0}\right)/N$. Only entries about the diagonal are significant

The near-diagonality of $\mathbf{T}$ provides a computationally effective way to derive approximately MDST from MDCT through (9). Omitting orders and signs, we can find from (10.b) that each row or column of $\mathbf{T}$ shares the same set of $N/2$ terms $1/(N\sin[\theta(2s+1)])$, $s = 0, 1, \ldots, N/2-1$, and each unique term appears exactly twice. Let $\mathbf{T}_m$ be the matrix having in common with $\mathbf{T}$ the most significant $2mN$ entries, all the rest zero. Then nonzero entries of $\mathbf{T}_m$ have their absolute values no less than $1/(N\sin[\theta(2m-1)])$ and center heavily around the diagonal. We substitute $\mathbf{T}$ with $\mathbf{T}_m$ in (9). And with (10.a), (9) is approximated in a FIR-like fashion as

$$\begin{aligned} Y^i(k) &\approx (-1)^k X_-^i(k) + \sum_{l=0}^{N-1} (\mathbf{T}_m)_{k,l} X_+^i(l) \\ &= (-1)^k X_-^i(k) + (-1)^k \sum_{s=-m}^{m-1} \frac{X_+^i(\mathrm{ind}(k-2s-1))}{(-1)^s N \sin[\theta(2s+1)]} \end{aligned} \tag{11}$$

where ind(•) is a spectral line index mapping function, $\mathrm{ind}(l) = -l-1$, if $l < 0$; $\mathrm{ind}(l) = 2N-l-1$, if $l > N-1$; and ind(l)=l else. This way we avoid explicit matrix-vector multiplication.

3.3 Windowed MDCT-MDST conversion

We deal with here the windowing scenario mentioned in the part *C* of Section 2. Note $\mathbf{W}_0$ and $\mathbf{W}_1$ two $N\times N$ diagonal matrices whose diagonals are the sine window function $w_s(n)$ spanning from 0 to $N-1$ and N to $2N-1$ respectively. They satisfy the general requirements for TDAC on MDCT with windowing,

$$\begin{cases} \mathbf{W}_0\mathbf{W}_0 + \mathbf{W}_1\mathbf{W}_1 = \mathbf{I} \\ \mathbf{W}_1 = \mathbf{J}\mathbf{W}_0\mathbf{J}\ . \end{cases} \tag{12}$$

By the sine-cosine duality, the matrices of the cosine window function $w_c(n)$ are $\mathbf{W}_1$ and $-\mathbf{W}_0$ for $n = 0, \ldots, N-1$ and $n = N, \ldots, 2N-1$ respectively. Then (8) becomes

$$\begin{pmatrix} x_0^i \\ x_1^i \end{pmatrix} = \frac{2}{N} \begin{pmatrix} \mathbf{W}_1\mathbf{C}_1 X^{i-1} + \mathbf{W}_0\mathbf{C}_0 X^i \\ \mathbf{W}_1\mathbf{C}_1 X^i + \mathbf{W}_0\mathbf{C}_0 X^{i+1} \end{pmatrix}, \tag{13}$$

and with $\mathbf{S}_0^{\mathrm{T}}\mathbf{W}_1\mathbf{W}_0\mathbf{C}_0 = \mathbf{S}_1^{\mathrm{T}}\mathbf{W}_0\mathbf{W}_1\mathbf{C}_1 = 0$, $\mathbf{S}_0^{\mathrm{T}}\mathbf{W}_1\mathbf{W}_1\mathbf{C}_1 = \mathbf{S}_0^{\mathrm{T}}\mathbf{C}_1/2$, and $\mathbf{S}_1^{\mathrm{T}}\mathbf{W}_0\mathbf{W}_0\mathbf{C}_0 = \mathbf{S}_1^{\mathrm{T}}\mathbf{C}_0/2$ (Appendix D) deduced from (7.a) and (14), (9) becomes

$$\begin{aligned} Y^i &= \begin{pmatrix}\mathbf{S}_0^{\mathrm{T}} & \mathbf{S}_1^{\mathrm{T}}\end{pmatrix}\begin{pmatrix}\mathbf{W}_1 & \mathbf{0}\\ \mathbf{0} & -\mathbf{W}_0\end{pmatrix}\begin{pmatrix}x_0^i\\ x_1^i\end{pmatrix} \\ &= \frac{2}{N}\left(\mathbf{S}_0^{\mathrm{T}}\mathbf{W}_1\mathbf{W}_0\mathbf{C}_0 - \mathbf{S}_1^{\mathrm{T}}\mathbf{W}_0\mathbf{W}_1\mathbf{C}_1\right)X^i \\ &\quad + \frac{2}{N}\mathbf{S}_0^{\mathrm{T}}\mathbf{W}_1\mathbf{W}_1\mathbf{C}_1X^{i-1} - \frac{2}{N}\mathbf{S}_1^{\mathrm{T}}\mathbf{W}_0\mathbf{W}_0\mathbf{C}_0X^{i+1} \\ &= \frac{1}{N}\mathbf{S}_0^{\mathrm{T}}\mathbf{C}_1X^{i-1} - \frac{1}{N}\mathbf{S}_1^{\mathrm{T}}\mathbf{C}_0X^{i+1} \\ &= \frac{1}{N}\left(\mathbf{S}_1^{\mathrm{T}}\mathbf{C}_0 - \mathbf{S}_0^{\mathrm{T}}\mathbf{C}_1\right)X_-^i + \frac{1}{N}\left(\mathbf{S}_1^{\mathrm{T}}\mathbf{C}_0 + \mathbf{S}_0^{\mathrm{T}}\mathbf{C}_1\right)X_+^i \end{aligned} \tag{14}$$

where $X_-^i = (-X^{i+1} - X^{i-1})/2$ and $X_+^i = (-X^{i+1} + X^{i-1})/2$. Therefore, (10.a) and (12) still apply only with different definitions for X_-^i and X_+^i.

In [16], a method was proposed to estimate MDST from MDCT based on trigonometry manipulation. However it is restricted to the non-windowed and sine-windowed case. Our method is more general, all window functions satisfying (12) applies.

To clarify the algorithm in the last two subsections, we list steps to compute MDST spectrum from MDCT spectrum in Table 1.

3.4 Approximation error analysis

The Mean Square Error (MSE) of this approximation can be estimated from the square sum of the singular values of $\mathbf{T}-\mathbf{T}_m$. It is equal to the trace of $(\mathbf{T}-\mathbf{T}_m)^{\mathrm{T}}(\mathbf{T}-\mathbf{T}_m)$ or the energy sum of all column vectors of $\mathbf{T}-\mathbf{T}_m$. But from (7.a), we can find that $\mathbf{T}$ is orthogonal and energy of each column vector of $\mathbf{T}$ is 1 (Appendix E). Therefore the MSE is also equal to the relative MSE or the ratio of MSE to the energy of $\mathbf{T}X_+^i$. By the element pattern of $\mathbf{T}$ as discussed above, this relative MSE can be found by analyzing any single column vector of $\mathbf{T}$. From the orthogonality of $\mathbf{T}$, we have the equality

$$\frac{\theta^2}{\sin^2\theta} + \ldots + \frac{\theta^2}{\sin^2[\theta(2m-1)]} + \ldots + \frac{\theta^2}{\sin^2[\theta(N-1)]} = \frac{\pi^2}{8}. \tag{15}$$

The first m terms in (12) appear in both $\mathbf{T}$ and $\mathbf{T}_m$, and the rest terms relates to the MSE. As N increases, $\theta=\pi/(2N)$ decreases and its mth term $\theta^2/\sin^2[\theta(2m-1)] \to 1/(2m-1)^2$ decreasing monotonically. But the sum of the infinite sequence $1/1^2, 1/3^2, \ldots, 1/(2m-1)^2, \ldots$ is also $\pi^2/8$. Therefore, given any $\varepsilon \in (0,1)$, we have a positive integer m such that the sum of the first m terms of the infinite sequence is larger than $(1-\varepsilon)\pi^2/8$, and then the sum of the first m terms of (12) is also larger than $(1-\varepsilon)\pi^2/8$. Thus we can find m independent of N satisfying a given relative MSE upper limit ε, or

$$\begin{aligned} \mathrm{MSE}(m) &= 1 - \left(\frac{\theta^2}{\sin^2\theta} + \ldots + \frac{\theta^2}{\sin^2[\theta(2m-1)]}\right)\frac{8}{\pi^2} \\ &< 1 - \left(\frac{1}{1^2} + \ldots + \frac{1}{(2m-1)^2}\right)\frac{8}{\pi^2} < \varepsilon\ . \end{aligned} \tag{16}$$

For example, if requiring relative MSE less than 0.1, m = 2 is sufficient for all N.

Table 1 MDST spectrum computation

Before Start: A constant table holding the absolute values of the m most significant values, i.e. `t(1:m)=1/(N*sin(pi/(2*N)*(2*[1:m]-1))` in Matlab language. This needs to be calculated only once. An example for `N=128` and `m=5` is

```
t=[0.6366636, 0.212255, 0.127404, 0.091058, 0.070880];
```

then setup a FIR filter

```
tf=t([m:-1:1,1:m]).*(-1).^[-m:m-1];
```

Step 1: given 3 MDCT spectra `X0`, `X1`, `X2`, each `Nx1` column vector, of the last, current, and next frame, find the difference vector `Xm` with and the sum vector `Xp` without window,

```
Xm=(X2-X0)/2; Xp=(X2+X0)/2;
```

With window

```
Xm=(-X2-X0)/2; Xp=(-X2+X0)/2;
```

Step 2: find `Xpf` by filter `Xp` with `t(1:m)`, here `ind()` is the index mapping function. In fact, except low frequency end (`<2*m`) and high frequency end (`>N-2*m+1`), the function acts as an identity operator:

```
for k=1:N
    Xpf(k)=tf*Xp(ind([k+2m-1:-2:k-2m+1]));
end
```

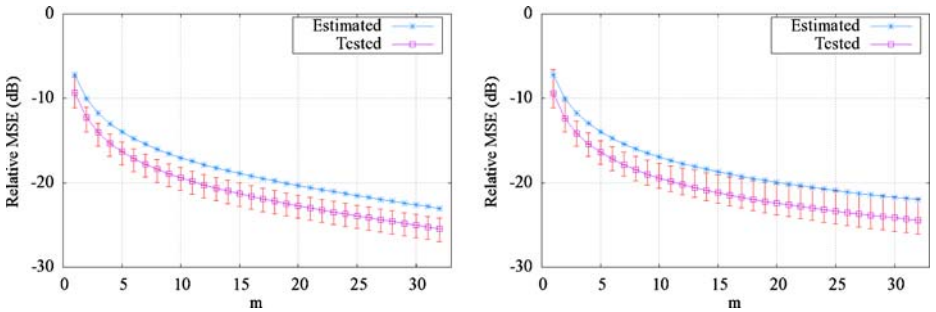
Step 3: find MDST spectrum `Y1`

```
Y1=(Xm+Xpf).*(-1).^[0:N-1]';
```

Step 4: Shft one frame forward before goto **step 1**

```
X1=X0;X1=X2;X2=new MDCT spectrum;
```

There are two terms in (9). The above error analysis is only for the second term while the first one involves only sign changes. When the two terms annihilate each other and the resulting Y^i is zero or very small, any small error in the second term transfers to very large relative error. Fortunately, it is a very rare case that time signal x_i^0 is odd-symmetric and x_i^1 is even-symmetric. For real audio signals, their relative MSEs are normally smaller given m (Fig. 6).



(a) N=128 (b) N=1024

Fig. 6 Relative MSE of approximating MDST from MDCT. The conversion matrix $\mathbf{T}$ is substituted by $\mathbf{T}_m$ which shares its most significant $2mN$ entries with the rest 0. The vertical bars indicate maximum, mean, and minimum relative MSE over 12 audio sequences of different types, 48,000 points each. **a** N=128, **b** N=1024

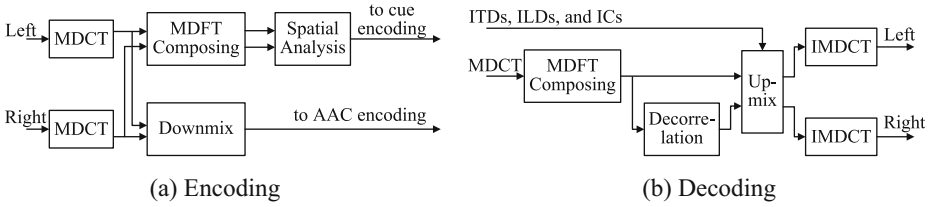


Fig. 7 Spatial stereo coding using only MDCT. The encoder has only forward transforms, and the decoder has only backward transforms. MDCT spectra are shared by core coding and spatial coding. **a** Encoding, **b** Decoding

4 Spatial encoding and decoding

We have developed the necessary tools for spatial analysis and synthesis using MDCT spectrum. Without the structural burden of separate complex transforms, spatial coding modules will integrate more closely into core coders, provided that they use the same transform. One of the most used core coders AAC, based on MDCT, enables this possibility.

Figure 7(a) shows our proposed encoding structure. Incoming audio signals go though MDCT on every channel. From (11) we can find their MDST spectra. By (5.a) and (5.b), MDFT spectra are constructed. Then ILD, ITD, and IC are given by (4.a) and (4.b). This process involves no explicit MDST transforming or separate complex transforming. PS encoder based on FFT, uses two FFT, one IFFT, and one MDCT for core coding [13]. In stereo case, the proposed scheme only uses two MDCT. Suppose the frame length is N. Since there are always 50% overlap between adjacent frames in both cases, the length of FFT/IFFT and MDCT are $2N$. The time inputs are real samples, so FFT/IFFT can be efficiently computed through N-point complex FFT. Also MDCT can be efficiently computed through $N/2$-point complex-FFT (CFFT). So the proposed coder saves about 5/7 of the transform complexity (Table 2). This is similar at decoder end. Since MDST spectrum can be computed using low-order FIR filtering from MDCT spectrum, for this computational overhead is low compared to the transform itself. But a main source of complexity of spatial coding is transforming [52]. It significantly cuts down CPU cycles.

Decoding is roughly an inverse process of encoding, which is shown in Fig. 7(b). A core AAC decoder only needs to output reconstructed MDCT spectra, on which MDST spectra are built by (11) and then MDFT spectra by (5.a) and (5.b). A decorrelation procedure [13, 18, 20] generates counterpart MDFT spectra, having similar envelops but different fine

Table 2 Transform complexity comparison

	$2N$-FFT	$2N$-IFFT	$2N$-MDCT	$2N$-IMDCT	complexity
FFT-based encoder	2 times	1 time	1 time	0 time	~7 $N/2$-CFFT
FFT-based decoder	1 time	2 times	0 time	1 time	~7 $N/2$-CFFT
MDCT-based decoder	0 time	0 time	2 times	0 time	~2 $N/2$-CFFT
MDCT-based decoder	0 time	0 times	0 time	2 times	~2 $N/2$-CFFT

Table 3 Testing sequences description

Type	File Name	Description
Speech	es01.wav	Vocal (Suzan Vega)
	es02.wav	German speech
	es03.wav	English speech
Single Instrument	si01.wav	harpsichord
	si01.wav	castanets
	si01.wav	pitch pipe
Simple Sound Mixture	sm01.wav	bagpipes
	sm01.wav	glockenspiel
	sm01.wav	plucked strings
Complex Sound Mixture	sc01.wav	trumpet solo and orchestra
	sc01.wav	orchestral piece
	sc01.wav	contemporary pop music

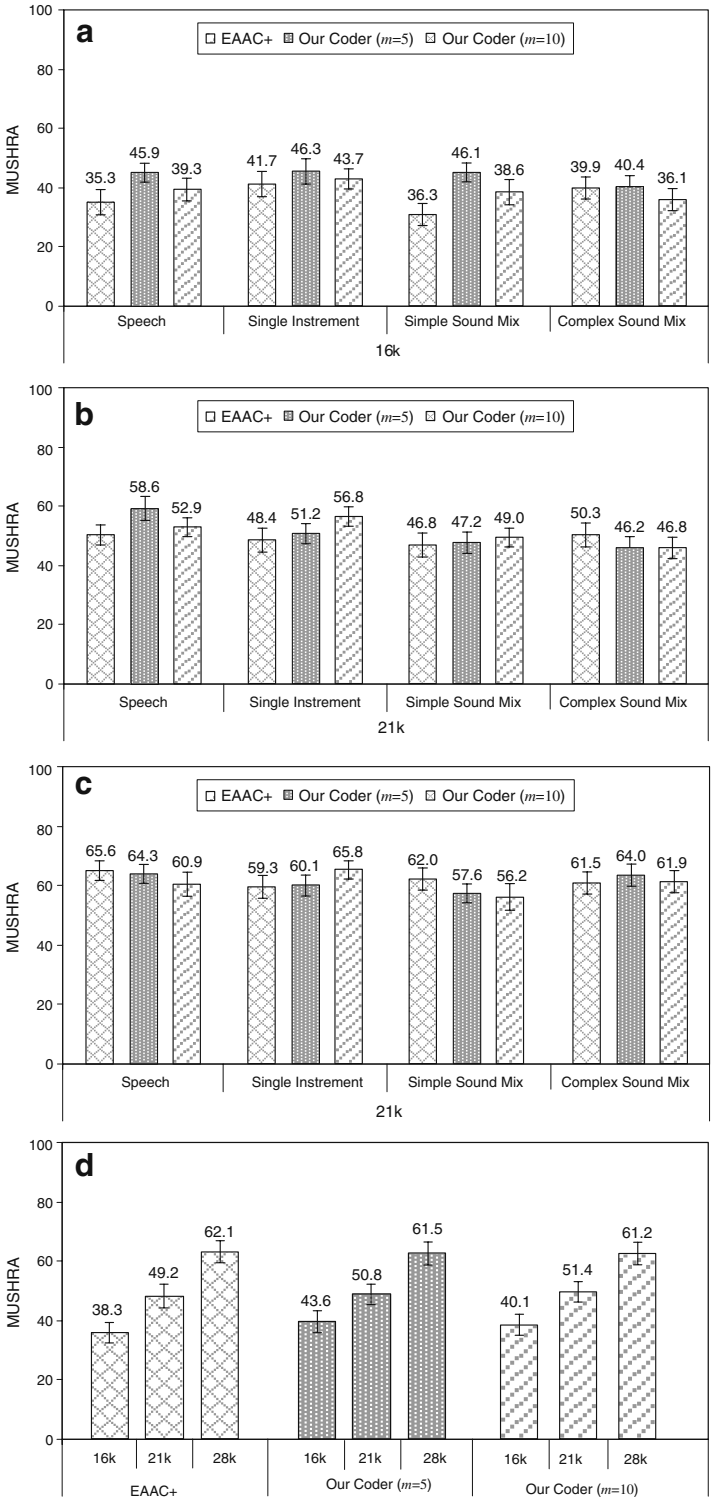
structures both in time and frequency domain. With a pair of a decoded MDFT spectrum and its decorrelated version, we can build two MDFT spectra through linear combination and phase adjustment matching the given ILDs, ITDs, and ICs on all bands. Time signals follow readily from MDCT part of the reconstructed MDFT through IMDCT and overlap-and-add (Fig. 4). Apart from MDCT and MDFT, it is very similar to PS. The goal is not waveform recovery but time-frequency information and spatial information recovery to the extent of human hearing perception.

5 Performance

To test the viability of MDCT domain spatial analysis and synthesis, we implement a stereo audio coder, it is based on a "PS + AAC" stereo coder EAAC+ from 3GPP [1], a state-of-the-art low bitrate audio coder used in 3G mobile audio services. The main modification is to replace QMF in PS by MDFT. Other modifications are applied only if necessary. Specifically, we use same number of subbands with as close as possible bandwidth between MDFT and QMF. The spatial parameters are quantized using the same quantization tables as EAAC+, resulting in close parameter bit consumption on each subband. However, the average parameter bitrate is essentially the same. The decorrelation module uses the same algorithm, but there are major code level modifications the discrepancy between QMF and MDFT.

The test is to find the subjective quality impairment of our coder at bitrate below 36 kbps [1], which is the sweet spot for stereo coding using the PS scheme. The original EAAC+ is also tested for reference and comparison. Other popular audio coders such as

Fig. 8 MUSHRAM (ITU-R BS1534-1) test scores, mean and 95% confidence interval. The 0–100 score ranges is interpreted as bad (0–20), poor (20–40), fair (40–60), good (60–80), and excellent (80–100). Bitrate used is 16 kbps in (**a**), 21 kbps in (**b**), and 28 kbps in (**c**), each split into four categories according to the four sequence types. The overall test results are shown in (**d**). Here m is the order of MDST approximation ▶


a
EAAC+
Our Coder (m=5)
Our Coder (m=10)
MUSHRA
100
80
60
40
20
0
35.3
45.9
39.3
41.7
46.3
43.7
36.3
46.1
38.6
39.9
40.4
36.1
Speech
Single Instrement
Simple Sound Mix
Complex Sound Mix
16k
b
EAAC+
Our Coder (m=5)
Our Coder (m=10)
MUSHRA
58.6
52.9
48.4
51.2
56.8
46.8
47.2
49.0
50.3
46.2
46.8
Speech
Single Instrement
Simple Sound Mix
Complex Sound Mix
21k
c
EAAC+
Our Coder (m=5)
Our Coder (m=10)
MUSHRA
65.6
64.3
60.9
59.3
60.1
65.8
62.0
57.6
56.2
61.5
64.0
61.9
Speech
Single Instrement
Simple Sound Mix
Complex Sound Mix
21k
d
MUSHRA
38.3
49.2
62.1
43.6
50.8
61.5
40.1
51.4
61.2
16k
21k
28k
EAAC+
Our Coder (m=5)
Our Coder (m=10)

MP3, MPEG-2 AAC, and WMA do not use PS but code stereo signals essentially as separate channels, typically operating at much higher bitrates (96–128 kbps). So they are not used in this test.

There are 12 stereo sequences used in this test, each 10~20 s long, 48 kHz sampled, and 16-bit PCM width, all from MPEG standard audio test library. The signal types are listed in Table 3.

The test method complies ITU-R BS.1534-1 standards [34], or "MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA)". It is specifically designed for intermediate audio quality evaluation requiring relatively small listening panel size. We enrolled 15 subjects. Each was asked to evaluate the randomly ordered test sequences with regard to their corresponding reference sequences, on a scale from 0 (worst)—100 (best). The reference sequences are those listed in Table 3, without any processing. The test sequences are those processed by our coder in two configurations: m=5 and m=10, and EAAC+, as well as hidden references, 3.5 kHz and 7 kHz low-pass filtered anchors. The last three are used to screen out invalid scores. Test results classified according to signal types are given in Fig. 8(a), (b) and (c) for 3 bitrates 16 kbps, 21 kbps, and 28 kbps respectively. Figure 8(d) gives the overall results. Scores are shown as mean and 95% confidence interval.

The test indicates that our coder has same quality as EAAC+ statistically. This confirms that the replacing of QMF by MDFT in PS does not have negative impact on overall audio quality. More interestingly, even the MDST approximation errors decrease from 0.023 using m=5 to 0.011 using m=10, the mean scores sees no quality increasing. This agrees our observation that spatial parameter quantization errors in PS are much larger than those introduced by MDST approximation. All those errors become irrelevant to human hearing if they are below the just noticeable differences or hearing thresholds of the spatial parameters. This is the design principle of the spatial parameter quantization tables. And output time signals are built only by MDCT, which is not directly related to MDST error. We can expect beyond a certain value, higher m will not bring higher quality.

6 Conclusions

We give two types of MDCT-MDST combination for MDFT, non-windowed and windowed, sharing the same temporal-spectral correspondence as FFT. It enables spatial parameters analysis and synthesis in MDFT domain. We also find that instead of direct transforming, MDST spectrum can be converted from neighboring MDCT spectra. And the conversion matrix is heavily diagonal so a small number of its sub-diagonals are sufficient for approximating MDST in spatial coding. Our spatial stereo coder using approximated MDFT shows statistically equal audio quality as PS, but saving half transforms at both the encoding and decoding ends.

This coding scheme makes spatial coding more a tool inside the core AAC coder than outside it. But synthesizing spatial parameters for the current MDCT spectrum uses indirectly the next MDCT spectrum, adding a frame of delay. If avoiding this, spatial coding will have no additional delay other than those introduced in AAC coding. Besides, the finding that sine-windowed MDFT is in fact FFT with a linear phase shift will also make separate T/F tools for psychoacoustics analysis and spectral processing unnecessary—the real part of the sine-windowed MDFT is exactly the sine-windowed MDCT.

Acknowledgement This research was supported by National Science Foundation of China (grant 60832002) and MKE(Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) Support program supervised by the IITA(Institute of Information Technology Advancement) (IITA-2009-C1090-0902-0020)

Appendix

A. MDFT energy conservation

As in (3.a) and (3.b), $c_0,\ldots,c_{N-1}$ and $s_0,\ldots,s_{N-1}$ are $2N$-dimensional basis vectors for MDCT and MDST respectively. The inner products between them are

$$\begin{cases} \langle c_k, c_l\rangle = N\delta(k-l), & k,l = 0,\ldots,N-1 \\ \langle s_k, s_l\rangle = N\delta(k-l), & k,l = 0,\ldots,N-1, \\ \langle c_k, s_l\rangle = 0, & k,l = 0,\ldots,N-1 \end{cases} \tag{A.1}$$

where δ(•) is the unit impulse function. They compose an orthogonal basis for $2N$-dimensional real vector space. Then for a time signal $x(n)$, $n = 0,\ldots,2N-1$, and its MDCT spectrum $X(k)$ and MDST spectrum $Y(k)$, $k = 0, ..., 2N-1$, their energy satisfies

$$\begin{aligned} & N\langle x,x\rangle \\ &= \frac{1}{N}\left\langle \sum_{k=0}^{N-1}(X(k)c_k + Y(k)s_k), \sum_{k=0}^{N-1}(X(k)c_k + Y(k)s_k)\right\rangle \\ &= \langle X,X\rangle + \langle Y,Y\rangle = \langle X+jY, X+jY\rangle\ . \end{aligned} \tag{A.2}$$

This verifies that MDFT spectral energy is N times of temporal energy.

B. MDFT time shift and phase shift

From MDFT definition in (3.c), we have when time signal $x(n)$ has a shift d and satisfies $x(n-2N) = -x(n)$, its MDFT spectrum as

$$\begin{aligned} \widetilde{Z}(k) &= \sum_{n=0}^{2N-1} x(n-d)\exp\left[-j\frac{\pi}{N}\left(n+\frac{1}{2}+\frac{N}{2}\right)\left(k+\frac{1}{2}\right)\right] \\ &= \sum_{n=-d}^{2N-1-d} x(n)\exp\left[-j\frac{\pi}{N}\left(n+d+\frac{1}{2}+\frac{N}{2}\right)\left(k+\frac{1}{2}\right)\right] \\ &= \sum_{n=0}^{2N-1-d} x(n)\exp\left[-j\frac{\pi}{N}\left(n+d+\frac{1}{2}+\frac{N}{2}\right)\left(k+\frac{1}{2}\right)\right] \\ &\quad - \sum_{n=2N-d}^{2N-1} x(n-2N)\exp\left[-j\frac{\pi}{N}\left(n+d+\frac{1}{2}+\frac{N}{2}\right)\left(k+\frac{1}{2}\right)\right] \\ &= Z(k)\exp\left[-j\frac{\pi}{N}d\left(k+\frac{1}{2}\right)\right] \end{aligned} \tag{A.3}$$

where $Z(k)$ is MDFT spectrum of $x(n)$ without shift. The condition $x(n-2N) = -x(n)$ parallels DFT's requirement of periodicity but with a negative sign. For real signals and $d<<2N$, (A.4) is an approximation.

C. Windowed MDFT

Note $X(k)$ and $Y(k)$ are sine-windowed MDCT spectrum and cosine-windowed MDST spectrum respectively. Then we have

$$\begin{aligned} Z_+(k) = Y(k) + X(k) = & \sum_{n=0}^{2N-1} x(n) \cos\left[\frac{\pi}{2N}\left(n+\frac{1}{2}\right)\right] \sin\left[\frac{\pi}{N}\left(n+\frac{1}{2}+\frac{N}{2}\right)\left(k+\frac{1}{2}\right)\right] \\ & + \sum_{n=0}^{2N-1} x(n) \sin\left[\frac{\pi}{2N}\left(n+\frac{1}{2}\right)\right] \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}+\frac{N}{2}\right)\left(k+\frac{1}{2}\right)\right] \\ = & -\sum_{n=0}^{2N-1} x(n) \cos\left[\frac{\pi}{N} n(k+1) + \frac{\pi}{N}(k+1)\left(\frac{1}{2}+\frac{N}{2}\right) + \frac{\pi}{4}\right] \end{aligned} \tag{A.4}$$

and

$$\begin{aligned} Z_-(k) &= Y(k) - X(k) \\ &= \sum_{n=0}^{2N-1} x(n) \cos\left[\frac{\pi}{2N}\left(n+\frac{1}{2}\right)\right] \sin\left[\frac{\pi}{N}\left(n+\frac{1}{2}+\frac{N}{2}\right)\left(k+\frac{1}{2}\right)\right] \\ &\quad - \sum_{n=0}^{2N-1} x(n) \sin\left[\frac{\pi}{2N}\left(n+\frac{1}{2}\right)\right] \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}+\frac{N}{2}\right)\left(k+\frac{1}{2}\right)\right] \\ &= \sum_{n=0}^{2N-1} x(n) \sin\left[\frac{\pi}{N} nk + \frac{\pi}{N} k\left(\frac{1}{2}+\frac{N}{2}\right) + \frac{\pi}{4}\right] . \end{aligned} \tag{A.5}$$

Take (A.4) and (A.5) as real part and imaginary part respectively,

$$\begin{aligned} & -Z_+(k-1) - jZ_-(k) \\ &= \sum_{n=0}^{2N-1} x(n) \cos\left[\frac{\pi}{N} nk + \frac{\pi}{N} k\left(\frac{1}{2}+\frac{N}{2}\right) + \frac{\pi}{4}\right] \\ &\quad -j \sum_{n=0}^{2N-1} x(n) \sin\left[\frac{\pi}{N} nk + \frac{\pi}{N} k\left(\frac{1}{2}+\frac{N}{2}\right) + \frac{\pi}{4}\right] \\ &= \exp\left\{-j\left[\frac{\pi}{N} k\left(\frac{1}{2}+\frac{N}{2}\right) + \frac{\pi}{4}\right]\right\} \sum_{n=0}^{2N-1} x(n) \exp\left[-j\frac{\pi}{N} nk\right] \end{aligned} \tag{A.6}$$

which is $2N$-point DFT with a phase shift. Moreover with $Z_+(-1) = -Z_-(0)$ and $Z_-(N) = Z_+(N-1)$, (A.6) leads to (5.a).

D. Properties of MDCT and MDST transform matrices

From (6), we can see each column vector of $\mathbf{C}_0$ and $\mathbf{S}_1$ are odd-symmetric, and each column vector of $\mathbf{C}_1$ and $\mathbf{S}_0$ are even-symmetric. With the help of anti-diagonal matrix $\mathbf{J}$ having only 1 on its anti-diagonal, the symmetries are equivalent to $\mathbf{JC}_0 = -\mathbf{C}_0, \mathbf{JS}_1 = -\mathbf{S}_1$ and $\mathbf{JC}_1 = \mathbf{C}_1, \mathbf{JS}_0 = \mathbf{S}_0$ respectively. From this and $\mathbf{J}^{\mathrm{T}}\mathbf{J} = \mathbf{JJ} = \mathbf{I}$, we have

$$\mathbf{S}_0^{\mathrm{T}}\mathbf{C}_0 = \mathbf{S}_0^{\mathrm{T}}\mathbf{J}^{\mathrm{T}}\mathbf{JC}_0 = (\mathbf{JS}_0)^{\mathrm{T}}(\mathbf{JC}_0) = -\mathbf{S}_0^{\mathrm{T}}\mathbf{C}_0, \tag{A.7}$$

which implies $\mathbf{S}_0^{\mathrm{T}}\mathbf{C}_0 = \mathbf{0}$. And for the same reason, $\mathbf{S}_1^{\mathrm{T}}\mathbf{C}_1 = \mathbf{0}$. For the windowed case, from the second equation of (14) $\mathbf{W}_1$=$\mathbf{J}\mathbf{W}_0\mathbf{J}$ and that $\mathbf{W}_0$ and $\mathbf{W}_1$ are diagonal matrices then $\mathbf{W}_0\mathbf{W}_1$=$\mathbf{W}_1\mathbf{W}_0$, we have

$$\begin{aligned}\mathbf{S}_0^{\mathrm{T}}\mathbf{W}_1\mathbf{W}_0\mathbf{C}_0 &= \mathbf{S}_0^{\mathrm{T}}\mathbf{J}^{\mathrm{T}}\mathbf{J}\mathbf{W}_1\mathbf{J}\mathbf{J}\mathbf{W}_0\mathbf{J}\mathbf{J}\mathbf{C}_0 \\ &= (\mathbf{J}\mathbf{S}_0)^{\mathrm{T}}(\mathbf{J}\mathbf{W}_1\mathbf{J})(\mathbf{J}\mathbf{W}_0\mathbf{J})(\mathbf{J}\mathbf{C}_0) \\ &= -\mathbf{S}_0^{\mathrm{T}}\mathbf{W}_0\mathbf{W}_1\mathbf{C}_0 = -\mathbf{S}_0^{\mathrm{T}}\mathbf{W}_1\mathbf{W}_0\mathbf{C}_0\end{aligned} \tag{A.8}$$

which implies $\mathbf{S}_0^{\mathrm{T}}\mathbf{W}_1\mathbf{W}_0\mathbf{C}_0 = \mathbf{0}$. And for the same reason, $\mathbf{S}_1^{\mathrm{T}}\mathbf{W}_0\mathbf{W}_1\mathbf{C}_1 = \mathbf{0}$. Also by similar procedure as (A.8), we have $\mathbf{S}_0^{\mathrm{T}}\mathbf{W}_1\mathbf{W}_1\mathbf{C}_1 = \mathbf{S}_0^{\mathrm{T}}\mathbf{W}_0\mathbf{W}_0\mathbf{C}_1$. From this and with the help of the first equation of (14) $\mathbf{W}_0\mathbf{W}_0 + \mathbf{W}_1\mathbf{W}_1 = \mathbf{I}$, we can see

$$\begin{aligned}\mathbf{S}_0^{\mathrm{T}}\mathbf{W}_1\mathbf{W}_1\mathbf{C}_1 &= \frac{1}{2}\left(\mathbf{S}_0^{\mathrm{T}}\mathbf{W}_0\mathbf{W}_0\mathbf{C}_1 + \mathbf{S}_0^{\mathrm{T}}\mathbf{W}_1\mathbf{W}_1\mathbf{C}_1\right) \\ &= \frac{1}{2}\mathbf{S}_0^{\mathrm{T}}(\mathbf{W}_0\mathbf{W}_0 + \mathbf{W}_1\mathbf{W}_1)\mathbf{C}_1 \\ &= \frac{1}{2}\mathbf{S}_0^{\mathrm{T}}\mathbf{C}_1\ .\end{aligned} \tag{A.9}$$

And for the same reason, $\mathbf{S}_1^{\mathrm{T}}\mathbf{W}_0\mathbf{W}_0\mathbf{C}_0 = \mathbf{S}_1^{\mathrm{T}}\mathbf{C}_0/2$.

E. Properties of the conversion matrix $\mathbf{T}$

As in (7.b), $\mathbf{P}$ is a matrix having only $+1, -1, +1, -1, \ldots,$ on its diagonal, implying $\mathbf{P}\mathbf{P}^{\mathrm{T}} = \mathbf{I}$. And with $\mathbf{S}_0 = -\mathbf{C}_1\mathbf{P}, \mathbf{S}_1 = \mathbf{C}_0\mathbf{P}$ in (7.b), we have $\mathbf{S}_1\mathbf{S}_1^{\mathrm{T}} = \mathbf{C}_0\mathbf{C}_0^{\mathrm{T}}$, $\mathbf{S}_0\mathbf{S}_0^{\mathrm{T}} = \mathbf{C}_1\mathbf{C}_1^{\mathrm{T}}$, $\mathbf{S}_0\mathbf{S}_1^{\mathrm{T}} = -\mathbf{C}_1\mathbf{C}_0^{\mathrm{T}}$, $\mathbf{S}_1\mathbf{S}_0^{\mathrm{T}} = -\mathbf{C}_0\mathbf{C}_1^{\mathrm{T}}$. With the help of $\mathbf{C}_0^{\mathrm{T}}\mathbf{C}_0 + \mathbf{C}_1^{\mathrm{T}}\mathbf{C}_1 = N\mathbf{I}$ and $\mathbf{C}_1\mathbf{C}_0^{T} = \mathbf{C}_0\mathbf{C}_1^{T} = \mathbf{0}$ in (7.a), the conversion matrix defined in (10.b) is orthogonal, or

$$\begin{aligned}\mathbf{T}^{\mathrm{T}}\mathbf{T} &= \frac{1}{N^2}\left(\mathbf{S}_1^{\mathrm{T}}\mathbf{C}_0 + \mathbf{S}_0^{\mathrm{T}}\mathbf{C}_1\right)^{\mathrm{T}}\left(\mathbf{S}_1^{\mathrm{T}}\mathbf{C}_0 + \mathbf{S}_0^{\mathrm{T}}\mathbf{C}_1\right) \\ &= \frac{1}{N^2}\left(\mathbf{C}_0^{\mathrm{T}}\mathbf{S}_1\mathbf{S}_1^{\mathrm{T}}\mathbf{C}_0 + \mathbf{C}_1^{\mathrm{T}}\mathbf{S}_0\mathbf{S}_0^{\mathrm{T}}\mathbf{C}_1 + \mathbf{C}_1^{\mathrm{T}}\mathbf{S}_0\mathbf{S}_1^{\mathrm{T}}\mathbf{C}_0 + \mathbf{C}_0^{\mathrm{T}}\mathbf{S}_1\mathbf{S}_0^{\mathrm{T}}\mathbf{C}_1\right) \\ &= \frac{1}{N^2}\left(\mathbf{C}_0^{\mathrm{T}}\mathbf{C}_0\mathbf{C}_0^{\mathrm{T}}\mathbf{C}_0 + \mathbf{C}_1^{\mathrm{T}}\mathbf{C}_1\mathbf{C}_1^{\mathrm{T}}\mathbf{C}_1 - \mathbf{C}_1^{\mathrm{T}}\mathbf{C}_1\mathbf{C}_0^{\mathrm{T}}\mathbf{C}_0 - \mathbf{C}_0^{\mathrm{T}}\mathbf{C}_0\mathbf{C}_1^{\mathrm{T}}\mathbf{C}_1\right) \\ &= \frac{1}{N^2}\left(\mathbf{C}_0^{\mathrm{T}}\mathbf{C}_0\mathbf{C}_0^{\mathrm{T}}\mathbf{C}_0 + \mathbf{C}_1^{\mathrm{T}}\mathbf{C}_1\mathbf{C}_1^{\mathrm{T}}\mathbf{C}_1 + \mathbf{C}_1^{\mathrm{T}}\mathbf{C}_1\mathbf{C}_0^{\mathrm{T}}\mathbf{C}_0 + \mathbf{C}_0^{\mathrm{T}}\mathbf{C}_0\mathbf{C}_1^{\mathrm{T}}\mathbf{C}_1\right) \\ &= \frac{1}{N^2}\left(\mathbf{C}_0^{\mathrm{T}}\mathbf{C}_0 + \mathbf{C}_1^{\mathrm{T}}\mathbf{C}_1\right)\left(\mathbf{C}_0^{\mathrm{T}}\mathbf{C}_0 + \mathbf{C}_1^{\mathrm{T}}\mathbf{C}_1\right) = \mathbf{I}\ .\end{aligned} \tag{A.10}$$

References

1. 3GPP specification Series TS 26.410 (2005) General audio codec audio processing functions; enhanced aacPlus general audio codec; floating-point ANSI-C code, http://www.3gpp.org/ftp/Specs/html-info/26-series.htm, Apr. 2005
2. 3GPP Specification Series TS26.405 (2005) General audio codec audio processing functions; enhanced aacPlus general audio codec; encoder specification; parametric stereo part, http://www.3gpp.org/ftp/Specs/html-info/26-series.htm, Apr. 2005

3. Algazi VR, Duda RO, Thompson DM, Avendano C (2001) The CIPIC HRTF database. Presented at IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics
4. Baumgarte F, Faller C (2002a) Estimation of auditory spatial cues for binaural cue coding. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp 1801–1804
5. Baumgarte F, Faller C (2002b) Why binaural cue coding is better than intensity stereo coding. Presented at the 112th AES Convention, Munich, Germany
6. Baumgarte F, Faller C (2003) Binaural cue coding—part I: psychoacoustic fundamentals and design principles. IEEE Trans Speech Audio Process 11:509–519. doi:10.1109/TSA.2003.818109
7. Blauert J (1983) Spatial hearing: the psychophysics of human sound localization. MIT, USA
8. Bosi M, Goldberg R (2003) MPEG-2 AAC. In: Introduction to digital audio coding and standards, chap. 13. Kluwer Academic, USA, pp 333–367
9. Bosi M, Brandenburg K, Quackenbush S, Fielder L, Akagiri K, Fuchs H, Dietz M (1997) ISO/IEC MPEG-2 advanced audio coding. J Audio Eng Soc 45(10):789–814
10. Breebaart J (2007) Analysis and synthesis of binaural parameters for efficient 3D audio rendering in MPEG Surround. In: IEEE International Conference on Multimedia and Expo, Beijing, China, pp 1878–1881
11. Breebaart J, van de Par S, Kohlrausch A (2001) Binaural processing model based on contralateral inhibition. I. Model structure. J Acoust Soc Am 110:1074–1088. doi:10.1121/1.1383297
12. Breebaart J, Disch S, Faller C, Herre J, Hotho G, Kjörling K, Myburg F, Neusinger M, Oomen W, Purnhagen H, Rödén J (2005a) MPEG spatial audio coding / MPEG Surround: overview and current status. Presented at the 119th AES Convention, New York
13. Breebaart J, van de Par S, Kohlrausch A, Schuijers E (2005b) Parametric coding of stereo audio. EURASIP J Appl Signal Process 9:1305–1322. doi:10.1155/ASP.2005.1305
14. Breebaart J, Hotho G, Koppens J, Schuijers E, Oomen W, van de Par S (2007) Background, concept and architecture for the recent MPEG Surround standard on multi-channel audio compression. J Audio Eng Soc 55:331–351
15. Breebaart J, Villemoes L, Köjrling K (2008) Binaural rendering in MPEG Surround. EURASIP J. Advances in Signal Processing, Article ID 732895
16. Cheng CI (2004) Method for estimating magnitude and phase in the MDCT domain. Presented at the 116th AES Convention, Berlin, Germany
17. Disch S, Ertel C, Faller C, Herre J, Hilpert J, Hoelzer A, Kroon P, Linzmeier K, Spenger C (2004) Spatial audio coding: next-generation efficient and compatible coding of multi-channel audio. Presented at the 117th AES Convention, San Francisco, USA
18. Engdegård J, Purnhagen H, Rödén J, Liljeryd L (2004) Synthetic ambience in parametric stereo coding. Presented at 116th AES Convention, Berlin, Germany
19. Faller C (2004) Parametric coding of spatial audio. Ph.D. Dissertation, Institut de systèmes de communication, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
20. Faller C (2006) Parametric multichannel audio coding: synthesis of coherence cues. IEEE Trans Audio Speech Lang Process 14:299–310. doi:10.1109/TSA.2005.854105
21. Faller C, Baumgarte F (2001) Efficient representation of spatial audio using perceptual parameterization. Presented at IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York
22. Faller C, Baumgarte F (2002a) Binaural cue coding: a novel and efficient representation of spatial audio. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp 1841–1844
23. Faller C, Baumgarte F (2002b) Binaural cue coding applied to stereo and multi-channel audio compression. Presented at the 11th AES Convention, Munich, Germany
24. Faller C, Baumgarte F (2002c) Binaural cue coding applied to audio compression with flexible rendering. Presented at the 113th AES Convention, Los Angeles, USA
25. Faller C, Baumgarte F (2003) Binaural cue coding—part II: schemes and applications. IEEE Trans Speech Audio Process 11:520–531. doi:10.1109/TSA.2003.818108
26. Fliege NJ (1994) Modified DFT Polyphase SBC filter banks with almost perfect reconstruction. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp 149–152
27. Gilkey R, Anderson TR (eds) (1997) Binaural and spatial hearing in real and virtual environments. Erlbaum, Mahwah, NJ
28. Herre J (2004) From joint stereo to spatial audio coding-recent progress and standardization. In: Proc. of the 7th Int. Conference on Digital Audio Effects, Naples, Italy, Oct. 2004, pp. 157–162

29. Herre J, Purnhagen H, Breebaart J, Faller C, Disch S, Kjörling K (2005) The reference model architecture for MPEG spatial audio coding. Presented at the 118th AES Convention, Barcelona, Spain
30. Herre J, Köjrling K, Breebaart J, Faller C, Disch S, Purnhagen H, Koppens J, Hilpert J, Rödén J, Oomen W, Linzmeier K, Chong KS (2008) MPEG Surround—the ISO/MPEG standard for efficient and compatible multi-channel audio coding. J Audio Eng Soc 56:932–955
31. Hotho G, Villemoes LF, Breebaart J (2008) A backward-compatible multichannel audio codec. IEEE Trans Audio Speech Lang Process 16:83–93. doi:10.1109/TASL.2007.910768
32. ISO/IEC JTC1/SC29/WG11 (2005) Information technology—generic coding of moving pictures and associated audio information—part 7: advanced audio coding (AAC), ISO/IEC 13818-7: 2005(E)
33. ISO/IEC JTC1/SC 29/WG11 (2006) MPEG Audio sub-group, Text of ISO/IEC 23003-1:2006/FCD, MPEG Surround
34. ITU (2003) Method for the subjective assessment of intermediate quality level of coding systems, ITU-R BS.1534-1
35. Joris P, Yin TCT (2006) A matter of time: internal delays in binaural processing. Trends Neurosci 30:70–78. doi:10.1016/j.tins.2006.12.004
36. Karp T, Fliege NJ (1995) MDFT filter banks with perfect reconstruction. Presented at IEEE International Symposium on Circuits and Systems
37. Malvar HS (1990) Lapped transforms for efficient transform/subband coding. IEEE Trans Acoust Speech Signal Process 38:969–978. doi:10.1109/29.56057
38. Malvar HS (1991) Fast algorithm for modulated lapped transform. Electron Lett 27(9):775–776. doi:10.1049/el:19910482
39. Malvar HS (1992) Signal processing with lapped transforms. Artech House, Norwood, MA
40. Malvar H (1999) A modulated complex lapped transform and its applications to audio processing. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp 1421–1424
41. Malvar HS (2003) Fast algorithm for the modulated complex lapped transform. IEEE Signal Process Lett 10:8–10. doi:10.1109/LSP.2002.806700
42. Malvar HS, Staelin DH (1989) The LOT: transform coding without blocking effects. IEEE Trans Acoust Speech Signal Process 37:553–559. doi:10.1109/29.17536
43. McAlpine D, Jiang D, Palmer AR (2001) A neural code for low-frequency sound localization in mammals. Nat Neurosci 4:396–401. doi:10.1038/86049
44. Mu-Huo C, Yu-Hsin H (2003) Fast IMDCT and MDCT algorithms—a matrix approach. IEEE Trans Signal Process 51:221–229. doi:10.1109/TSP.2002.806566
45. Munkong R, Biing-Hwang J (2008) Auditory perception and cognition. IEEE Signal Process Mag 25:98–117. doi:10.1109/MSP.2008.918418
46. Plogsties J, Breebaart J, Herre J, Villemoes L, Jin C, Kjörling K, Koppens J (2006) MPEG Surround binaural rendering—surround sound for mobile devices. Presented at 24th Tonmeistertagung—VDT International Convention, Leipzig, Germany
47. Princen J, Bradley A (1986) Analysis/synthesis filter bank design based on time domain aliasing cancellation. IEEE Trans Acoust Speech Signal Process 34:1153–1161. doi:10.1109/TASSP.1986.1164954
48. Princen JP, Johnson AW, Bradley AB (1987) Subband/transform coding using filter bank designs based on time domain aliasing cancellation. In: Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing, pp 2161–2164
49. Quackenbush S, Herre J (2005) MPEG Surround. IEEE Multimedia 12:18–23. doi:10.1109/MMUL.2005.76
50. Roden J, Breebaart J, Hilpert J, Purnhagen H, Schuijers E, Koppens J, Linzmeier K, Holzer A (2007) A study of the MPEG Surround quality versus bit-rate curve. Presented at the 123rd AES Convention, New York, USA
51. Schuijers EGP, Oomen AWJ, den Brinker AC, Gerrits AJ (2003) Advances in parametric coding for high-quality audio. Presented at the 114th AES Convention, Amsterdam, The Netherlands
52. Schuijers E, Breebaart J, Purnhagen H, Engdegard J (2004) Low complexity parametric stereo coding. Presented at 116th AES Convention, Berlin, Germany
53. Strutt JW (1907) (Lord Rayleigh), on our perception of sound direction. Philos Mag 13:214–232
54. Wang Y, Vilermo M (2003) Modified discrete cosine transform—its implications for audio coding and error concealment. J Audio Eng Soc 51:52–61

Shuixian Chen is currently pursuing the Ph.D. in computer science at the National Key Lab of Multimedia Communications, Wuhan University, China. Her researches focus on spatial audio perception, parametric representation of spatial audio scenes, parametric and transform audio coding, and audio communication system design and implementation. She is a member of the audio expert subgroup of China Audio Video Standard (AVS). She also patents five audio related technologies.

Naixue Xiong received his both PhD degrees in Wuhan University, and Japan Advanced Institute of Science and Technology, respectively. Both are on computer science. Now he is a research scientist in the Department of Computer Science, Georgia State University, Atlanta, USA. His research interests include Communication Protocols, Network Architecture and Design, and Optimization Theory. Until now, he published about 140 research articles (including 54 journal articles). Some of his works were published or submitted in IEEE JSAC, IEEE or ACM transactions, and IEEE INFOCOM. He has been a General Chair, Program Chair, Publicity Chair, PC member and OC member of about 73 international conferences, and as a reviewer of about 56 international journals, including IEEE JSAC, IEEE Transactions on Communications, IEEE Transactions on Mobile Computing, IEEE Trans. on Parallel and Distributed Systems. He is serving as an editor for 9 international journals, and a guest editor for 9 international journals, including Information Science (Springer). He has received the Best Paper Awards in the 10th IEEE International Conference on High Performance Computing and Communications (HPCC-08) and the 28th North American Fuzzy Information Processing Society Annual Conference (NAFIPS2009). He is a member of IEEE and IET. E-mail: nxiong@cs.gsu.edu.

Dr. Jong Hyuk Park received his Ph.D. degree in the Graduate School of Information Security from Korea University, Korea. He is now a professor at the Department of Computer Science and Engineering, Kyungnam University, Korea. He has published about 100 research papers in international journals and conferences. He has been serving as chairs, program committee, or organizing committee chair for many international conferences and workshops. He was editor-in-chief of the International Journal of Multimedia and Ubiquitous Engineering (IJMUE), the managing editor of the International Journal of Smart Home (IJSH). He is Associate Editor/Editor of 14 international journals including 8 journals indexed by SCI(E). In addition, he has been serving as a Guest Editor for international journals by some publishers: Springer, Elsevier, John Wiley, Oxford Univ. press, Hindawi, Emerald, Inderscience. His research interests include security and digital forensics, ubiquitous and pervasive computing, context awareness, multimedia services, etc. He got the best paper award in ISA-08 conference, April, 2008. And he got the outstanding leadership awards from IEEE HPCC-09 and ISA-09, June, 2009.

Min Chen received the Ph.D degree in Electrical Engineering from South China University of Technology in 2004, when he was 23 years old. He is an assistant professor in School of Computer Science and Engineering at Seoul National University (SNU). He has been a Research Associate in the Department of Computer Science at University of British Columbia (UBC) for half year, and worked as a Post-Doctoral Fellow in Department of Electrical and Computer Engineering at UBC for three years. Before joining UBC, he was a Post-Doctoral Fellow at the SNU for one and half years. Dr. Chen has published more than 50 technical papers. He received the Best Paper Runner-up Award from QShine 2008. He was interviewed by Chinese Canadian Times where he appeared on the celebrity column in 2007. Dr. Chen is the author of OPNET Network Simulation (Tsinghua University Press, 2004). He has served as session chairs for several conferences, including VTC'08, QShine'08, ICACT'09, Trientcom'09, and ICC'09. He was the TPC chair of ASIT'09 and TPC co-chair of PCSI'09 and PCSI'10.

Ruimin Hu received his Ph.D. degree in Electronic and Information Engineering from Huzhong University of Science and Technology, Wuhan, China. He is now a professor at the Department of Computer Science, Wuhan University, China. Dr. Hu has published many research papers in international journals and conferences. He is director of National Engineering Research Center for Multimedia Software, P. R. China and Key Lab of Multimedia & Network Communications Engineering, P. R. China. Dr. Hu's research interests include audio & video coding, Security & Surveillance services, multimedia signal processing, etc.