LETTER

# Efficient FFT Algorithm for Psychoacoustic Model of the MPEG-4 AAC

**Jae-Seong LEE**[†a)], **Chang-Joon LEE**[†], ***Nonmembers***, **Young-Cheol PARK**[††], ***Member***, ***and*** **Dae-Hee YOUN**[†], ***Nonmember***

**SUMMARY** This paper proposes an efficient FFT algorithm for the Psycho-Acoustic Model (PAM) of MPEG-4 AAC. The proposed algorithm synthesizes FFT coefficients using MDCT and MDST coefficients through circular convolution. The complexity of the MDCT and MDST coefficients is approximately half of the original FFT. We also design a new PAM based on the proposed FFT algorithm, which has 15% lower computational complexity than the original PAM without degradation of sound quality. Subjective as well as objective test results are presented to confirm the efficiency of the proposed FFT computation algorithm and the PAM.

***key words:*** *FFT, MDCT, MDST, psychoacoustic model, MPEG-4 AAC*

## 1. Introduction

MPEG-4 AAC encoder [1] achieves high coding gain by exploiting both perceptual irrelevancies and statistical redundancies. MPEG-4 AAC encoder uses two different time-frequency mappings. The coders use a Modified Discrete Cosine Transform (MDCT) for the subband analysis of the input signal [2]–[4]. At the same time, for the psychoacoustic modeling (PAM) of human hearing, the Fast Fourier Transform (FFT) is used [2]–[4]. The MDCT filterbank dynamically changes the frequency resolution to prevent the spread of so-called pre-echo. In the steady-state input condition, a high resolution (long-window) MDCT filterbank is used to increase coding gain. However, whenever signal attack is detected, the long-window switches to a shorter window. PAM, on the other hand, simultaneously computes separate FFT filter-banks of two sizes (2048 and 256 samples) for the long- and short-windows. Since approximately 35% of the computational complexity of MPEG-4 AAC encoder is caused by the PAM process [5], simplification of the PAM is a determining factor in the successful implementation of the AAC encoder. The FFT is one of the most crucial parts of the PAM.

In this paper, we propose an efficient FFT algorithm for the PAM in MPEG-4 AAC encoder [1]. The proposed algorithm obtains exact FFT spectra by combining the MDCT coefficients from the analysis filterbank with additional Modified Discrete Sine Transform (MDST) coefficients. Major advantage of the proposed algorithm is that, by removing redundant operations, it can significantly reduce the computational complexity of the FFT. In this paper, we also present a new PAM based on the proposed FFT algorithm. As a result, the new PAM is 15% less complex than the original. In addition, by obtaining the FFT spectrum directly from the windowed input to the MDCT filterbank, the new PAM simulates energy values of MDCT coefficients more precisely than the PAM in MPEG-4 AAC encoder [1]. The rest of this paper is organized as follows: Section 2 addresses the modified FFT computation, experimental results are presented in Sect. 3, and brief conclusions are given in Sect. 4.

Manuscript received July 1, 2009.

[†]The authors are with Yonsei University, Seoul, Korea.

[††]The author is with Yonsei University, 234 Heungup, maeji Wonju-city, Kwangwon, Korea.

a) E-mail: dream7070@dsp.yonsei.ac.kr

DOI: 10.1587/transinf.E92.D.2535

## 2. Low-Complexity FFT Algorithm

In this section, an efficient FFT algorithm for PAM of MPEG-4 AAC is presented. The new algorithm utilizes the relationship between MDCT/MDST and DFT [7]. The MDCT coefficient $X_c(k)$ of the input sequence $x(n)$ is computed as [2]

$$X_c(k) = \sum_{n=0}^{N-1} h(n)x(n)\cos\left[\frac{2\pi}{N}\left(n+\frac{N}{4}+\frac{1}{2}\right)\left(k+\frac{1}{2}\right)\right] \quad (1)$$

where $h(n)$ is a window function and $N$ is the size of the transform block. In MPEG-4 AAC, $N = 2048$ is used. Defining $n_0 = N/4 + 1/2, k_0 = 1/2$ and denoting a windowed input signal as $z(n) = h(n)x(n), 0 \le n \le N-1$, the MDCT equation is rewritten as

$$\begin{aligned} X_c(k) &= \sum_{n=0}^{N-1} z(n)\cos\left[\frac{2\pi}{N}(k+k_0)(n+n_0)\right] \\ &= real\left\{\sum_{n=0}^{N-1} z(n)\exp\left[-j\frac{2\pi}{N}(k+k_0)(n+n_0)\right]\right\} \end{aligned} \quad (2)$$

where $real\{\cdot\}$ denotes the real part of a complex function. By splitting the complex exponent in Eq. (2), we have $X_c(k) = real\{\tilde{X}(k)\}$ [7] where

$$\begin{aligned} \tilde{X}(k) = &\left[\sum_{n=0}^{N-1}\left(z(n)\exp\left(-j\frac{2\pi}{N}k_0 n\right)\right)\exp\left(-j\frac{2\pi}{N}kn\right)\right] \\ &\cdot\exp\left(-j\frac{2\pi}{N}(k_0+k)n_0\right) \end{aligned} \quad (3)$$

The summation inside the square brackets corresponds to the conventional DFT of $z(n)$ with a spectral shift by $k_0$ bin,

and the last exponent represents a time shift by $n_0$ samples. Then, $\tilde{X}(k)$ can be rewritten as

$$\tilde{X}(k) = DFT\left\{z(n)\exp\left(-j\frac{2\pi}{N}k_0 n\right)\right\}\cdot\exp\left(-j\frac{2\pi}{N}(k_0+k)n_0\right) \tag{4}$$

If the imaginary part of $\tilde{X}(k)$ in Eq. (3) is taken, it becomes a Modified Discrete Sine Transform (MDST). Thus by denoting the MDST coefficients as $X_s(k)$, $\tilde{X}(k)$ is simply expressed as $\tilde{X}(k) = X_c(k) - jX_s(k)$. After transposing the exponent on the right side of Eq. (4) and taking IDFT on both sides, we have

$$IDFT\left\{\tilde{X}(k)\cdot\exp\left(j\frac{2\pi}{N}(k_0+k)n_0\right)\right\} = z(n)\exp\left(-j\frac{2\pi}{N}k_0 n\right)$$

*or*

$$z(n) = IDFT\left\{\tilde{X}(k)\cdot\exp\left(j\frac{2\pi}{N}(k_0+k)n_0\right)\right\}\cdot\exp\left(j\frac{2\pi}{N}k_0 n\right) \tag{5}$$

Now, Eq. (5) is readily followed by

$$DFT\{z(n)\} = \left\{\tilde{X}(k)\cdot\exp\left(j\frac{2\pi}{N}(k_0+k)n_0\right)\right\} * DFT\left(\exp\left(j\frac{2\pi}{N}k_0 n\right)\right) \tag{6}$$

where $*$ represents circular convolution. Eq. (6) indicates that the DFT spectrum of the windowed input $z(n)$ can be synthesized using $\tilde{X}(k)$ or MDCT and MDST coefficients $X_c(k)$, $X_s(k)$. Since the MDCT coefficients are obtained through the analysis filterbank, MDST, instead of DFT, can be computed. Furthermore fast algorithms developed for MDCT ([9], for example) can be used for MDST with minor modifications.

### 2.1 Window Translation

It is important to note that the PAM in MPEG-4 AAC encoder [1] uses the Hann window for the DFT computation, not the sine window [1]. Therefore, we may need to translate Eq. (6) for the Hann window. This can be done using a simple translation. For the long sequence, the translation is given by

$$DFT\{x(n)h_{H\_long}(n)\} = DFT\left\{z(n)\cdot\frac{h_{H\_long}(n)}{h_{S\_long}(n)}\right\} = \left[\tilde{X}(k)\cdot\exp\left(j\frac{2\pi}{N}(k_0+k)n_0\right)\right] * C_l(k) \tag{7}$$

where

$$C_l(k) = DFT\left\{\exp\left(j\frac{2\pi}{N}k_0 n\right)\cdot\frac{h_{H\_long}(n)}{h_{S\_long}(n)}\right\} \tag{8}$$

Observing the relationship between Hann and sine windows, $h_H(n) = h_S(n)\cdot h_S(n)$, obtained by the cosine double-angle formula, we can see that $C_l(k)$ is a simple DFT of an exponential function being multiplied by the long sine window. Since $k_0 = 0.5$, spectral leakage may occur. However, the windowing operation in Eq. (8) reduces the leakage. Most of the energy of $C_l(k)$ is concentrated on first and second coefficients. If so, the circular convolution in Eq. (7) can be conveniently substituted with a first-order FIR filtering.

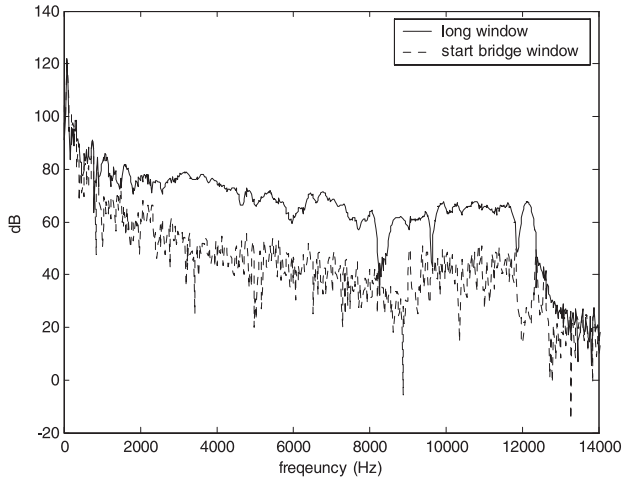


**Fig. 1** Power spectra comparison between long Hann and start sine windows.

It should be noted that we can avoid the window translation $h_H(n)/h_S(n)$ in Eqs. (8) and (10) by simply applying the sine window to the FFT. Then, the circular convolution is performed with a function:

$$C_l'(k) = DFT\left\{\exp\left(j\frac{2\pi}{N}k_0 n\right)\right\}. \tag{9}$$

However, since no window function is involved, spectral leakage in $C_l'(k)$ will be more visible than $C_l(k)$, which makes it difficult to approximate the circular convolution to a first-order FIR filtering without sacrificing performance. Thus, a window translation is beneficial for MDCT/MDST-based FFT computation. The current PAM uses the Hann window-based FFT spectrum in MPEG-4 AAC encoder [8].

For window switching, the PAM in MPEG-4 AAC encoder [1] uses only short (256 samples) and long (2048 samples) windows. It doesn't define the start and stop bridge windows. However, since the MDCT filterbank uses start and stop windows, it is desirable to use the same type of window for both FFT and MDCT. In such a way, it is possible to simulate the energy values in MDCT coefficients more precisely.

In other words, the FFT input is likely to include transient components appearing at the end of a block, while the MDCT input will exclude those components due to the shape of the start bridge window which has zeros in the 1600th~2048th sample period. The resulting power spectra in Fig. 1 show that there are noticeable differences between the two. Therefore, when the start or stop bridge window is selected for the MDCT, PAM can not properly reflect spectral contents of the MDCT input.

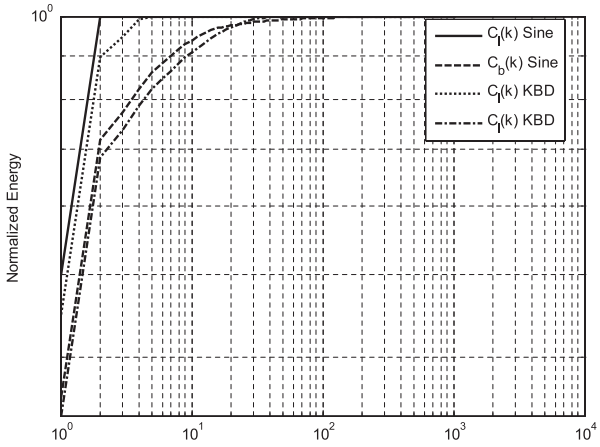


**Fig. 2** Cumulative energy of coefficients $C_l(k)$ and $C_b(K)$.

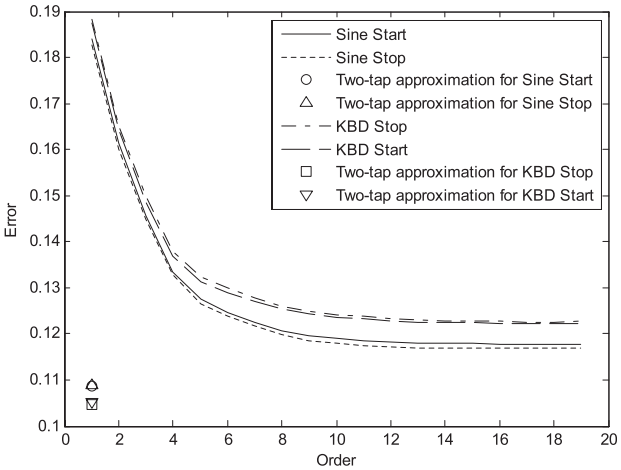


**Fig. 3** Errors in the FFT spectrum caused by the approximation of circular convolution at bridge window.

If we use the start and stop bridge windows, the same translation can be applied. Again, we can rewrite the DFT of the input block being windowed using the bridge Hann window as

$$\begin{aligned} &DFT\{x(n)h_{H\_start\|stop}(n)\} \\ &\quad = DFT\{z(n)\cdot h_{S\_start\|stop}(n)\} \\ &\quad = \left[\tilde{X}(k)\cdot \exp\left(j\frac{2\pi}{N}(k_0+k)n_0\right)\right] * C_b(k) \end{aligned} \tag{10}$$

where

$$C_b(k) = DFT\left\{\exp\left(j\frac{2\pi}{N}k_0 n\right)\cdot h_{S\_start\|stop}(n)\right\}. \tag{11}$$

In Eq. (10),$h_{H\_start\|stop}$ and $h_{S\_start\|stop}$ are bridge (start or stop) Hann window and bridge Sine window respectively. Now, $C_b(k)$ is a DFT of the exponential function being windowed using the bridge Hann window. However unlike Eq. (8), the windowing operation comprised in Eq. (11) cannot effectively prevent spectral leakage, which is due to the asymmetrical shape of the bridge Hann window.

To assess the spectral leakage, we first measured the cumulative energies of the coefficients $C_l(k)$ and $C_b(k)$. Results are shown in Fig. 2. Unlike $C_l(k)$ where the energy is concentrated on the first two coefficients, the energy of $C_b(k)$ is distributed widely over frequency bins. Thus, in the case of the start/stop windows, the low-order approximation of the circular convolution is not possible without introducing substantial errors. But the order of circular convolution should be kept to minimum, since it affects the overall complexity of the FFT computation. To search for the minimum order of the circular convolution, we measured the error of the FFT spectrum. Figure 3, shows the normalized spectrum error. The input was white Gaussian noise and the spectral error was computed with $\zeta = E\left\{\frac{\sum_k |s_o(k)-s_T(k)|^2}{\sum_k |s_o(k)|^2}\right\}$ where $s_o(k)$ were the true FFT coefficients and $s_T(k)$ were the estimate through the circular convolution with the truncated $C_b(k)$. As shown in Fig. 2, the first 20 coefficients of $C_b(k)$ include about 98% of the total energy. The error in Fig. 3 is saturated at the order around 12, and slowly decays afterward, which indicates that the low-order approximation is problematic in terms of having a reasonable accuracy in circular convolution in Eq. (10).

Another possible approximation of the circular convolution in Eq. (10) is the use of $C_l(k)$ instead of $C_b(k)$. As explained, the low-order approximation of $C_b(k)$ will introduce a loss of energy, and, in turn, the circular convolution using the approximated $C_b(k)$ may cause an underestimation of the spectral level of the input. Since the lowered spectral level can produce a lowered masking threshold, the approximation affects the quality of the output sound. On the other hand, the use of the two-tap approximation of $C_l(k)$ in place of $C_b(k)$ can preserve most of the input energy because no significant energy loss is introduced during this approximation. The only problem is that we may lose the fine structure of the spectrum of the input block. But the PAM calculates one representative masking level for each scalefactor band. Fine structure of the FFT spectrum is less important than the energy level in each scalefactor band. Thus, the two-tap approximation of $C_l(k)$ is another choice for the circular convolution in Eq. (10). In Fig. 3, we also show the spectral error introduced by the use of the two-tap approximation of $C_l(k)$, instead of $C_b(k)$ in Eq. (11), for the circular convolution in Eq. (10). The two-tap approximation of $C_l(k)$ yields lower spectral errors than any other case using the truncated $C_b(k)$. Thus, in the proposed algorithm, the two-tap approximation of $C_l(k)$ is used instead of $C_b(k)$. Unlike the case of long sine to Hann window translation, there is no explicit relationship between KBD and Hann windows. Thus, the spectral leakage occurs in the process of window translation. But then again, the minimum order for approximating the KBD to Hann window translation should be determined to minimize the complexity. For this, we measured the cumulative energy of $C_l(k)$ and $C_b(k)$, and normalized spectral errors in the case of KBD window. For ease of comparison, results are overlaid on Figs. 2 and 3. After experimenting, the minimum spectral error was obtained by approximating the convolution process in Eq. (8) using the

first two taps of $C_l(k)$ given by Eq. (8), and the error was, on average, 8.01%. The spectral error, however, was translated down to less than 0.03% of energy difference in scalefactor bands, which resulted in a negligible difference in masking threshold. The same approximation used in the case of sine bridge window was used for the KBD bridge window, and the spectral error induced by the approximation was slightly lower than for the sine bridge window.

Base on the proposed FFT algorithm, we design a new PAM. Figure 4 shows a signal flow for a new PAM employing the proposed MDCT/MDST-based FFT algorithm. The new PAM uses a time-domain block-switching algorithm. The conventional transient detection based on Perceptual Entropy (PE) determines the window type for the current input frame at the end of the PAM process. Thus, if the window switching is determined, we should restart the PAM process by calculating FFTs using all window types, which results in a significant computational increase. To avoid this, the block type has to be determined prior to the FFT computation. To this end, we use the time-domain transient detection algorithm adopted in AC3 [11]. The algorithm examines the time segments to detect an increase of energy. Segments are also examined in different time scales. If a significant increase in signal energy is detected, a transient is indicated. By using the time-domain algorithm, transient detection can be performed prior to the PAM process. Thus, the MDCT/MDST-based FFT algorithm can be implemented without an increase in computational cost.

### 2.2 Complexity

There are many varieties of fast algorithms for MDCT computation. In particular, the algorithm in [9] requires $N\times(\log_2 N+1)/4$ real multiplications and $N\times(3\log_2 N-1)/4$ real additions for an N-point MDCT. The same algorithm can be used for MDST at the same complexity. The $n_0$ spectral shifts in Eqs. (7) and (10) require $3N/2$ real multiplications and $3N/2$ real additions, and they require that the circular convolutions in Eqs. (8) and (11) be approximated using a two-tap FIR filtering with $3N$ real multiplications and $7N/2$ additions. Therefore, the proposed algorithm requires $N\times\log_2 N+19N/2$ real operations for the FFT computation, which can be compared with the conventional split radix FFT [10] requiring $4N\times(\log_2 N-1)+8$ real operations. Since PAM in MPEG-4 AAC adopts a 2048-point FFT ($N=2048$), it is possible to save about 51% of the computational cost of the standard FFT computation using the new algorithm. The computing time for new PAM was measured using a C language model on a PC; we measured the complexity using the CPUs internal clock cycle counter. As a result of modified FFT, we confirmed the new PAM requires 12% less execution time than the PAM in MPEG-4 AAC encoder which is ISO/IEC 14496-5 reference software [8] at a PC.

## 3. Experimental Results

The performance of the proposed algorithm was measured through both the objective difference grade (ODG) [6] as well as the subjective tests. The tests were conducted with audio signals listed in Table 1. The sampling frequency was 44.1 kHz and all signals were encoded with MPEG-4 AAC at 64 kbps. Output score of ODG ranges from 0 to −4, where the value corresponds to 'imperceptible degradation' and −4 corresponds to 'very annoying'. ODG scores in Fig. 5 show that the sound quality of the proposed algorithm is equivalent to MPEG-4 AAC encoder which was ISO/IEC 14496-5 reference software [8] for all test materials. Figure 6 presents the results of the MUSHRA listening test, including 95% confidence intervals as bars. The important point to note is that there was no sound item where the modified FFT lead to reduced performance. Comparing the performances over the individual audio sequences, we see that relative performances are similar.

**Table 1** Test materials for objective tests.

| Item | Description | | sec |
|---|---|---|---|
| es01 | German speech | Speech | 8 |
| es02 | English speech | | 7 |
| es03 | Korean speech | | 8 |
| si01 | Glockenspiel | Single instruments | 10 |
| si02 | Castanets | | 7 |
| si03 | Hihat | | 3 |
| sc01 | Jan Rafaj's "test signal2" | Complex sound mixtures | 4 |
| sc02 | Sarah McLachlan's "Elsewhere" | | 6 |
| sc03 | Fool's Garden's "Lemon Tree" | | 10 |

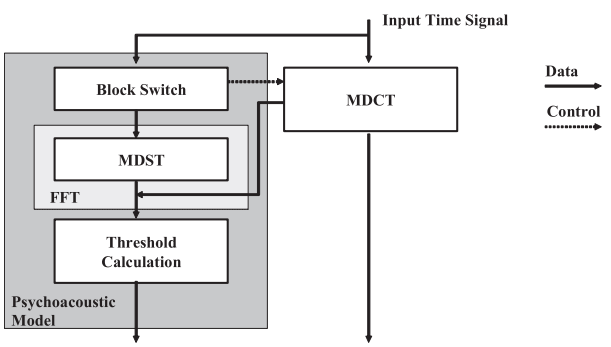


**Fig. 4** New PAM employing MDCT/MDST-based FFT computation.

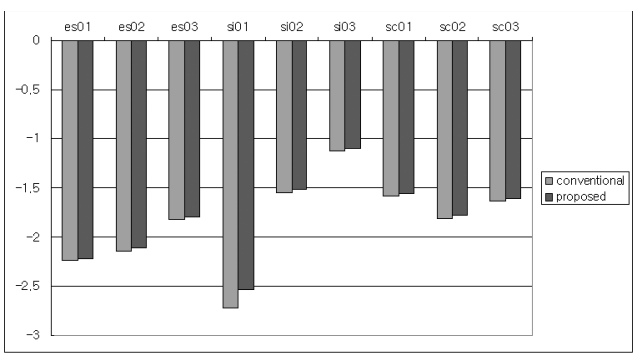


**Fig. 5** Listening test through ODG.

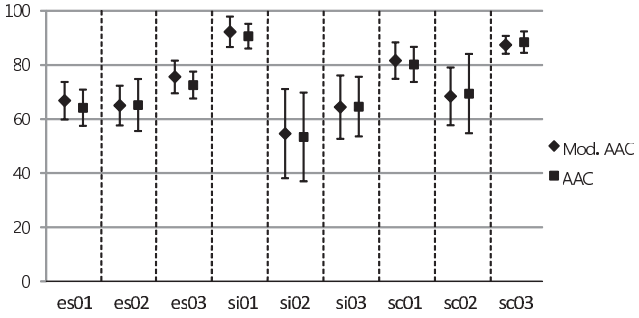


**Fig. 6** Listening test through MUSHRA.

## 4. Conclusions

In this paper, we proposed a low-complexity FFT computation method using MDCT and MDST coefficients for MPEG-4 AAC. The proposed FFT method uses start/stop Hann windows for the FFT computation in the PAM when the input block is judged to be a transient one. For the implementation of the proposed method, we adopted time-domain transient detection. Since MDCT coefficients are obtained directly from the analysis filterbank, the proposed algorithm being compared with the conventional split radix FFT can save roughly 51% of the computational cost. Through subjective evaluations, the proposed method was verified to provide audio quality as good as the conventional method even though it offers lower computation complexity.

**References**

[1] ISO/IEC 14496-3:1999. Information Technology: Coding of Audio-Visual Objects—Part 3: Audio, 1999.
[2] J.P. Princen, A.W. Johnson, and A.B. Bradley, "Subband/transform coding using filter bank designs based on time domain aliasing cancellation," Proc. IEEE Int. Conf. Acoust. Speech Signal Process., pp.2161–2164, 1987.
[3] ISO/IEC 11172-3:1992, Information technology: Coding of Moving Pictures and Associated Audio for Digital Storage Media up to about 1.5 Mbit/s—Part3: Audio, 1992.
[4] ISO/IEC 13818-7:1997, Information technology: Generic Coding of Moving Pictures and Associated Audio Information—Part 7: Advanced Audio Coding (AAC), 1997.
[5] T.-H. Tsai, S.-W. Huang, and L.-G. Chen, "Design of a low power psychoacoustic model co-processor for MPEG-2/4 AAC LC stereo encoder," Proc. 2003 IEEE International Symposium on Circuits and Systems, vol.2, pp.552–555, 2003.
[6] ITU Radiocommunication Study Group 6, "Draft revisoin to recommendation ITU-R BS.1387—Method for objective measurements of perceived audio quality."
[7] Y. Wang, L. Yaroslavsky, and M. Vilermo, "On the relationship between MDCT, SDFT and DFT," Proc. ICSP, pp.44–47, 2000.
[8] ISO/IEC 14496-5:2000. Information Technology: Coding of Audio-Visual Objects—Part 5: Reference Software, 2000.
[9] P. Duhamel, Y. Mahieux, and J.P. Petit, "A fast algorithm for the implementation of filter banks based on time domain aliasing cancellation," Proc. IEEE Int. Conf. Acoust. Speech Signal Process., pp.2209–2212, 1991.
[10] P. Duhamel, "Implementation of split radix FFT algorithms for complex, real and real symmetric data," IEEE Trans. Acoust. Speech Signal Process., vol.34, no.2, pp.285–295, April 1986.
[11] M.J. Smithers and M.C. Fellers, "Increased efficiency MPEG-2 AAC encoding," Proc. 111th Convention of AES, Preprint no.5490, 2001.