IMPROVED LOW-DELAY MDCT-BASED CODING OF BOTH STATIONARY AND TRANSIENT AUDIO SIGNALS

Christian R. Helmrich[1], *Goran Marković*[2], *and Bernd Edler*[1]

[1] International Audio Laboratories Erlangen [2] Fraunhofer Institut für Integrierte Schaltungen (IIS)
Am Wolfsmantel 33, 91058 Erlangen, Germany
christian.helmrich@audiolabs-erlangen.de

ABSTRACT

General-purpose MDCT-based audio coders like MP3 or HE-AAC utilize long inter-transform overlap and lookahead-based transform length switching to provide good coding quality for both stationary and non-stationary, i. e. transient, input signals even at low bitrates. In low-delay communication scenarios such as Voice over IP, however, algorithmic delay due to framing and overlap typically needs to be reduced and additional lookahead must be avoided. We show that these restrictions limit the performance of contemporary low-delay transform coders on either stationary or transient material and propose 3 modifications: an improved noise substitution technique and increased overlap between "long" transforms for stationary, and "long to short" transform length switching without lookahead and directly from the long overlap for transient frames. A listening test indicates the merit of these changes when integrated into AAC-LD.

Index Terms— Audio coding, delay, LPC, MDCT, parametric

1. INTRODUCTION

Following a need for combining the previously separate paradigms of speech and transform coding into a single general-purpose audio codec (coder/decoder), three new audio coding standards evolved over the last decade. Building upon the principle used in AMR-WB Plus [1], two new codecs were proposed in 2012: Extended High-Efficiency AAC based on AMR-WB and HE-AACv2 [2] and Opus based on SILK and the newly developed CELT [3]. The respective transform-coder parts of these two standards, namely the improved HE-AACv2 and CELT, both employ the modified discrete cosine transform (MDCT) [4] to obtain frequency-domain representations for quantization and coding. However, they differ in some details:

- *inter-transform overlap*: HE-AAC uses a maximum overlap of 43 ms (100 % of the frame length), while CELT employs a fixed 2.5 ms overlap regardless of frame or transform length.
- *block switching* for transform length adaptation: both codecs allow switching between frames of either one "long" or eight "short" transforms, but HE-AAC requires transition frames.
- *algorithmic delay*: unlike in CELT, which only needs 2.5 ms of lookahead for the transform overlap, further delay sources exist in HE-AAC: QMF banks and block-switch lookahead.

The reason for these design differences is that CELT/Opus is a codec supporting low-delay applications like Voice over IP (VoIP), whereas HE-AAC is targeted at offline usage (e. g. file storage) and broadcast scenarios, where algorithmic delay is mostly irrelevant. Utilizing (Extended) HE-AAC for real-time communication is impractical due to its inherent delay of far more than 100 ms. For this reason, two low-delay variants of HE-AAC, namely AAC-LD and AAC-ELD, were standardized in 2000 and 2008, respectively [5, 6] and, like Opus, allow for coding with less than 30 ms delay. It must be noted that, in principle, Opus and AAC-(E)LD may also be used in offline, non-realtime situations. However, as will be described in this paper, they have certain drawbacks which limit the achievable audio quality on certain audio material to a level below that of the general-purpose (Extended) HE-AAC. Aside from efficient stereo coding and bandwidth extension, which will not be discussed here, CELT's performance falls short on very tonal stationary signals due to the short transform overlap, and AAC-(E)LD struggles with musical attacks and other very strong non-stationarities due to the lack of block switching functionality. In addition, at low bitrates (E)LD occasionally exhibits audible "musical" narrow spectral holes, and CELT sometimes sounds noisy. It is therefore desirable to develop a low-delay codec which avoids all these weakpoints and, as a result, offers a level of audio quality matching that of HE-AAC. The design of such an improved codec is the subject of this paper.

The remainder of the document is organized as follows. Section 2 examines the MDCT architectures of (E)LD and CELT in greater detail and illustrates why the aforementioned artifacts occur at low bitrates. Then, a modified codec architecture is proposed in Section 3, with its three key components described in Sections 4, 5, and 6, respectively. Section 7 presents and discusses the results of a blind listening test conducted to evaluate the subjective performance of a (E)LD-based coding system improved by integrating the techniques of the previous sections. Finally, Section 8 concludes the paper.

2. LOW-DELAY TRANSFORM CODING

As indicated in the introduction, CELT as well as AAC-(E)LD, and in fact most modern audio coders, utilize the MDCT – a real-valued and critically sampled, lapped transform – to convert each frame of time-samples into blocks of frequency-domain coefficients, which can be quantized and coded efficiently. A general signal diagram of both coding schemes, separated into encoder and decoder, is shown in Figure 1. After the time-frequency transformation of the MDCT, the resulting spectral coefficients, or lines, are grouped into bands whose widths are modeled after the perceptual Bark or ERB scales [7]. Before or during a quantization loop over these bands, a scale or energy factor is then computed for each band, and its inverse is applied to all lines of the respective band. Due to this process, the bands are known as scalefactor bands in the AAC family of coders. The quantization, controlled by a per-band bit allocation algorithm, serves the reduction of perceptual irrelevancy present in the signal.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

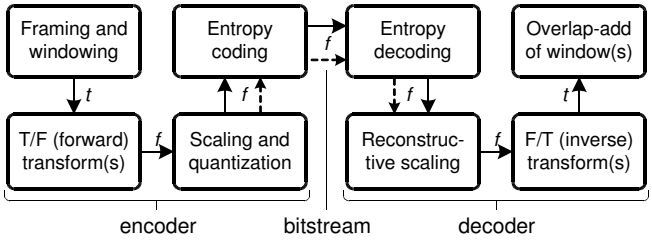

Fig. 1. *Block diagram of a typical low-delay encoder and decoder.*

Reduction of redundancy is attained by the energy-compacting MDCT and lossless entropy coding of the quantized lines and scale or energy factors. Moving on to the decoder in the right half of Fig. 1, where the entropy coded bitstream multiplex is received, demultiplexing and entropy decoding are performed to obtain the factors and quantized line representations. The lines are then reconstructed via multiplication of each band's line-quantization indices with the corresponding reconstructed energy or scale factor. The final steps are an inverse MDCT, transforming the reconstructed spectra back into time-domain samples, and an overlap-add (OLA) procedure to cancel time-domain aliasing (TDA) due to the MDCT and to build a gapless output waveform from the individually coded frames.

Beside the bitrate used for coding, the outer blocks depicted in Fig. 1 have the largest influence on the quality of the reconstructed audio signals. A graphical comparison of the framing, MDCT, and time-resolution optimization blocks of a CELT and (E)LD encoder is shown in Figure 2. It can be seen that the shapes of the windows processed by the MDCT differ significantly. A CELT encoder forms nearly rectangular Tukey-like windows, whereas (E)LD uses much smoother bell-shape windows. Obviously, CELT's window exhibits more spectral leakage than the (E)LD windows, which explains (at least partially) the fidelity and efficiency problems CELT shows on very tonal instruments such as trumpets. However, it has a notable advantage over the (E)LD windows: it enables switching between a *long* and eight *short* transforms like HE-AAC, but without intermediate *start* or *stop* transition windows, thus avoiding additional block-switch lookahead. In (E)LD, block switching is very difficult to achieve without violating the Princen-Bradley condition [4] for perfect reconstruction in the absence of quantization, which is why the latter codecs only offer one *long* transform length. This, in turn, can be taken as the reason why (E)LD's performance on transients like those in e. g. castanets or electronic music is sub-par. Although AAC-LD can reduce the transform overlap upon detecting a transient, the shortest possible time span of its windows still is about 2.5 times longer (13.3 ms) than that of CELT's *short* windows (5.3 ms), assuming both codecs operate at 48 kHz sample rate. Since coding error due to line quantization extends over the entire duration of a reconstructed window [4], pre-echo artifacts, i. e. temporal unmaskings of coding noise, are more likely to arise in (E)LD than CELT.

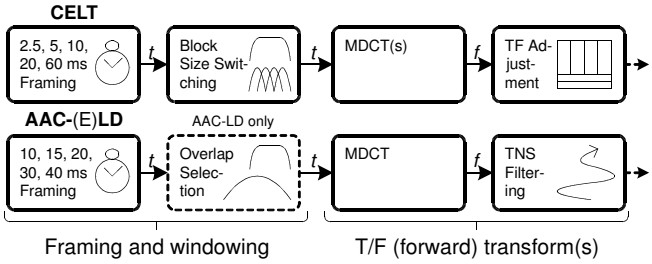

Fig. 2. *Zoomed view of the encoder in Fig. 1 for CELT and (E)LD.*

It must be noted in this regard that, as illustrated in Fig. 2, both coding systems provide means for improving the time resolution of a frame *after* the MDCT stage. In (E)LD, frequency-domain linear predictive filtering, known as temporal noise shaping (TNS), can be applied to each spectrum [5, 8] while in CELT, adjacent lines are optionally subjected to TF adjustment by means of Hadamard transformation [3]. As both tools should yield similar levels of pre-echo reduction, CELT's advantage on transients due to block switching remains since TF adjustment can also be used on *short* windows.

Furthermore, it is worth noting that none of the two examined low-delay coders feature a sophisticated noise insertion technique such as the noise filling methods employed in Extended HE-AAC [9]. This has two implications. First, in the case of (E)LD, where the scalefactor-band-wise perceptual noise substitution (PNS) is used, bands having at least one line not quantized to zero are not filled up with noise at all. At low bitrates, where only a few lines "survive" the quantization in each band, this causes narrow spectral holes of audibly tonal character, often referred to as musical noise. Second, the flexibility of noise insertion is limited, particularly in CELT, as the latter only offers "anti-collapse" filling of zero-quantized bands in *short* windows. In *long* windows, CELT aims at preventing musical noise by means of "spreading", a line-wise rotation pre- and post-processing around the quantizer [3]. The activation of this tool is based on a frequency-domain measure of frame tonality and a bit sensitive to mis-detection, occasionally leading to audible noise in the decoded signal. Given the low frequency selectivity of the near-rectangular *long* windows, this is not surprising: as shown in Figure 3, a MDCT of closely spaced harmonics resembles one of noise. In CELT, due to the low selectivity, even soft application of spreading to a tonal signal hence is likely to result in audible quality loss.

3. A MODIFIED LOW-DELAY CODING SCHEME

Summarizing the findings of the previous section, we can state that

- low-overlap near-rectangular MDCT windows exhibit little frequency selectivity, and thus efficiency, on stationary input.
- high-overlap AAC-like windows complicate block switching without extra encoder lookahead, which is therefore avoided in (E)LD; the implication is low efficiency on transient input.
- when using noise insertion methods, low application flexibility and MDCT selectivity limit their benefit or cause misuse.

To reduce or even avoid these three drawbacks and to combine the relative strengths of (E)LD and CELT, namely stationary and transient performance, respectively, into a single low-delay codec, we propose a modified coding architecture, depicted in Figure 4, with three improved components which will be examined hereafter.

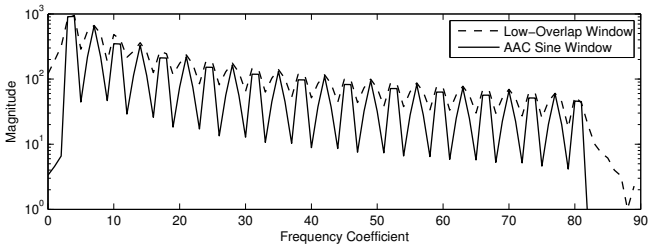

Fig. 3. *MDCT of low-pitch harmonic signal with different windows.*

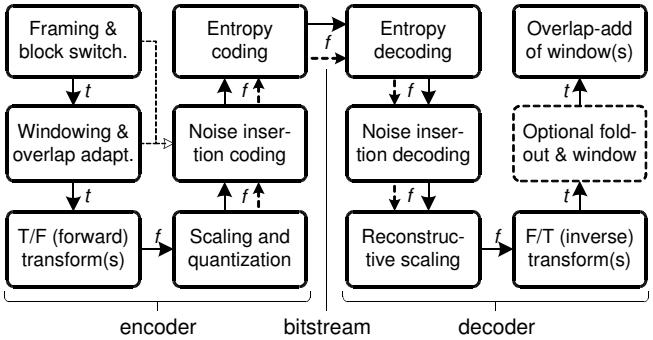

Fig. 4. *Block diagram of the proposed improved low-delay scheme.*

4. SYMMETRIC WINDOWS WITH ADAPTIVE OVERLAP

To ameliorate the issue of low frequency selectivity exhibited by a low-overlap window, we increase the overlap of the *long* windows to 50 % of the frame length, which was found to be a good tradeoff between performance and windowing delay. In addition, we employ the overlap adaptation used in AAC-LD [5] to isolate transients in a frame-overlap region into only one of the windows (Figure 5 a).

Regarding the choice of *long* window shape and symmetry, we have two options: symmetric windows as in AAC-LD or CELT, or special asymmetric windows for reduced delay as in AAC-ELD [6] or G.718 [10]. Here it is helpful to re-examine [4], which relates the analysis w_a and synthesis w_s window for a coded frame i as follows:

$$\hat{x}_i(t) = w_s(t+F)\, y_{i-1}(t+F) + w_s(t)\, y_i(t), \quad t = 0 \ldots F-1, \qquad (1)$$

with

$$y_i(t) = \frac{1}{F} \sum_{f=0}^{F-1} X_i(f) \cos\left(\frac{\pi}{F}\left(f+\frac{1}{2}\right)(t+T)\right), \quad t = 0 \ldots 2F-1, \qquad (2)$$

and

$$f = 0 \ldots F-1,$$

$$X_i(f) = \sum_{t=0}^{2F-1} x_i(t)\, w_a(2F-1-t) \cos\left(\frac{\pi}{F}\left(f+\frac{1}{2}\right)(t+T)\right), \qquad (3)$$

where x and $\hat{x}$ are the input and decoded signals, respectively, F is the frame length and T is a constant. Quantization errors and noise-filled lines, however, which are generally uncorrelated between the frames and which form in the MDCT domain of X, aren't subjected to (3) and hence only windowed by w_s. Due to (1) such components exhibit slight amplitude modulation in $\hat{x}$ after OLA for asymmetric windows as in [6, 10], illustrated by the dashed line in Fig. 5 b. Thus we use symmetric windows with w_s=w_a, which don't have this issue.

5. IMPROVED DELAYLESS BLOCK SWITCHING

Based on the analysis of a time-domain (TD) transient detector, we propose a block switching scheme allowing a direct transition from long to short overlap while retaining the perfect reconstruction property. It is similar to the one in [11] but avoids signal amplification [w_1 in 11] and is simpler. The key part is the order of operations and an extra folding process (dashed block in Fig. 4), as shown below.

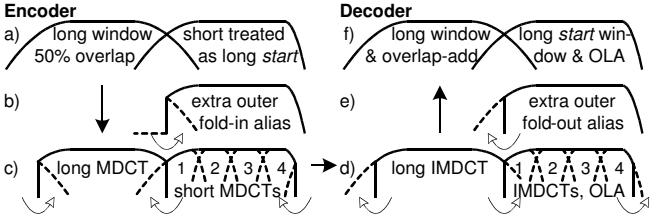

Fig. 6. *Order of operations in proposed delayless block switching.*

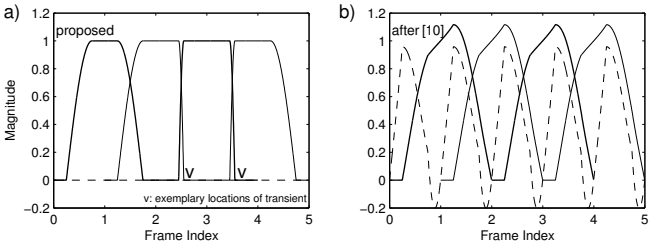

Fig. 5. *Noise shaping via* a) *symmetric,* b) *asymmetric windows* w_s. *dashed: temporal envelope in dB of noise signal created in MDCT.*

Fig. 6 suggests, simply speaking, to treat the frame containing short transforms as a classic long *start* transition window regarding overlap with windows of adjacent frames. In the encoder, outer fold-in aliasing must be applied prior to exertion of short inner w_a windows and MDCTs. Likewise, on the decoder side, the short-transform TD signals first need to be fully reconstructed via short IMDCTs, inner w_s windows and OLA between the short windows before outer fold-out aliasing and long *start* synthesis windowing can be carried out.

6. PERCEPTUAL, TONALITY-BASED NOISE INSERTION

After the scaling and quantization process, a noise insertion block is added in the encoder which computes two parameters for a noise filling procedure similar to the one used in Extended HE-AAC [2], operating before the reconstructive scaling in the decoder. The first parameter – as in [2, 9] – is a noise level defining the magnitude or energy of inserted pseudo-random lines, replacing zero-quantized lines, prior to reconstructive scaling. For best quality, harmonic signals should have less noise filling applied than noisy input (see also the end of section 2). Since in low-overlap windows, it is difficult to distinguish noisy spectra from low-pitched tonal ones in which harmonic components leak into neighboring lines, as shown in Fig. 3, we add to MDCT-domain tonality measures time-domain stationarity and zero-crossing data obtained in the transient detector (thin dashed arrow in Fig. 4) to control the noise level attenuation.

For a scalefactor band which is completely quantized to zero in [2], the corresponding scale factor controls the energy of that band after the substitution by the pseudo-random lines, weighted by the noise level. In low-bitrate coding, though, it is sometimes desirable to combine all bands into a single one and to perform quantization noise shaping by means of an LPC filter's transfer function [1]. In that case, the magnitudes of inserted noise coefficients will follow the shape of the LPC transfer function. It is, however, well known that due to psychoacoustic effects such as phase locking or reduced masking [7], fewer lines should be quantized to zero – or put differently, a higher coding SNR should be reached – at low than at high frequencies. This is e. g. reflected by the downward slope of CELT's fixed bit allocation prototype [12] and is illustrated in Figure 7.

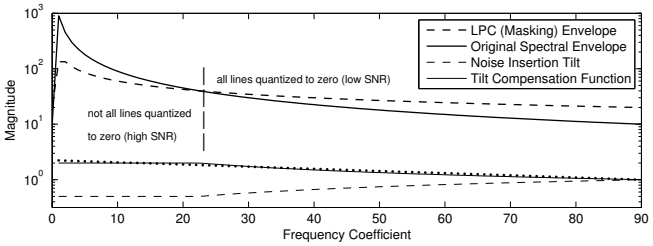

Fig. 7. *Example of tilt compensation for proposed noise insertion.*

If as noted, an LPC filter is to be utilized for quantization noise shaping, that filter will therefore exhibit a spectral tilt relative to the signal's frequency envelope, especially when it is derived from the pre-emphasized TD input [9]. The substituted noise lines, however, should follow the (perceptual) spectral envelope, not the LPC filter (quantization noise masking) envelope, in order to avoid a tilt in the noise-fill contribution when large parts of the spectrum are quantized to zero. To avoid this situation, depicted by the lowest line in Fig. 7, we propose a line-wise factor to compensate for the tilt, i. e. for the ratio between masking envelope and perceptual envelope. A line on a logarithmic scale (dotted) serves well to model the factors.

7. IMPLEMENTATION AND EVALUATION

To assess the performance of the proposed low-delay codec system of Fig. 4, a low-delay variant of Extended HE-AAC, architecturally similar to AAC-LD, was enhanced by the 3 improved components described in the last sections. The integration was done as follows:

- framing and windowing: each 20-ms frame contains either 1 *long* or 4 *short* windows. The maximum overlap is 50 % (10 ms), the minimum is 3.125 % of the frame length (0.625 ms). The windowing-induced lookahead thus amounts to 10 ms.
- delayless block switching: a 4-*short* frame is chosen whenever a non-stationarity measure obtained in the TD transient detector exceeds a threshold. Moreover, low window overlap is used if it allows to exclude transients from the frame.
- perceptual noise insertion: after uniform quantization of the lines, the noise energy and tilt parameters are determined as in [1,2] and sec. 6, guided by time-domain tonality measures.

Further notable components of the codec framework include a long term prediction (LTP) pre- and post-filter, similar to CELT's pitch filter [12], noise shaping via an LPC filter instead of scale factors, as in sec. 6, and a spectral band replication (SBR) tool as in ELD [6], but with even lower delay (2 ms), extending the coded bandwidth from 12.8 to 16 kHz. The pitch lag and gain computed for the LTP are re-used as one of the frame-tonality measures for the derivation of the noise insertion parameters. As the proposed block switching does not require additional lookahead, the total end-to-end delay of the coder framework is given by the sum of the framing delay, SBR delay and lookahead, and windowing lookahead, i. e. 32 ms.

For subjective evaluation of our implementation a double-blind listening test following the MUSHRA (multi-stimulus with hidden reference and anchor) methodology [13] was conducted at a bitrate of 48 kb/s mono. For comparison, the latest HE-AAC (Winamp 5.7) and Opus (1.1) coder versions at the time of writing were included as well. 10 trained, mostly expert listeners, all under the age of 37, participated in the test. They were asked to evaluate the basic audio qualiy [13] of all six conditions on the 11 signals listed in Table 1.

Applause	downmix of rear L+R of EBU 5.1 item	tech.ebu.ch/docs/tech/tech3339.pdf
Castanets	part 1 of EBU SQAM CD, track 27	tech.ebu.ch/publications/sqamcd
Fatboy	beginning of "Kalifornia" by Fatboy Slim	CD "You've come a long way, Baby"
Flamenco	excerpt: castanets and flamenco guitar	"castanets" from hydrogenaudio.org
Glockenspiel	part 2 of EBU SQAM CD, track 35	tech.ebu.ch/publications/sqamcd
Harpsichord	part 1 of EBU SQAM CD, track 40	tech.ebu.ch/publications/sqamcd
Voleurs	excerpt: "Les Voleurs de la République"	CD "Elohim" by Alpha Blondy
Robots	excerpt: "Die Roboter (The Robots)"	CD "the Man Machine" by Kraftwerk
RockYou	beginning of "Rock You Gently"	CD "the Hunter" by Jennifer Warnes
Te15	beginning of "Aragonaise IV" by Bizet	CD "Carmen Suite, Symphony no. 1"
Velvet	beginning of "Coitus" by Green Velvet	"velvet" from hydrogenaudio.org

Table 1. *Details on the 11 audio signals used in the MUSHRA test.*

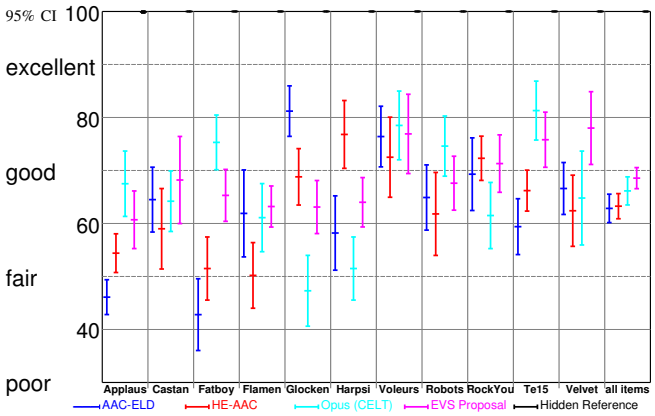

Fig. 8. *Result of the 48-kb/s listening test. CI: confidence interval.*

The results of the listening test are shown in Figure 8. To allow for better comparison, the vertical quality scale is magnified to the range from "poor" to "excellent", hence the results for the 3.5-kHz anchor (average grade of 21) are not displayed. It can be observed that overall, the proposed codec not only matches but exceeds the quality offered by AAC-ELD (31.3 ms delay) and HE-AAC (more than 120 ms delay) and, like CELT, performs relatively well on the transient signals while beating the latter on the tonal Glockenspiel, Harpsichord, and RockYou items. This proposed low-delay system was selected part of the EVS coding standard currently being finalized in 3GPP. The still quite low ratings on the Fatboy (vocoder) and Glockenspiel signals remain a topic for future investigation.

8. CONCLUSION

It was shown that MDCT-based low-delay audio coders face quality problems when encoding either stationary or transient signals, and a solution was presented which largely ameliorates this issue. The proposal has three components: a longer maximum window overlap and overlap adaptation, delayless block switching directly from the long overlap and a modified noise insertion method. A listening test of an implementation of these revealed that the individual strengths of AAC-ELD and Opus/CELT can be combined into a single codec.

9. ACKNOWLEDGMENT

The authors thank Ralf Geiger for helpful discussions and advice.

10. REFERENCES

[1] R. Salami, R. Lefebvre, A. Lakaniemi, K. Kontola, S. Bruhn, and A. Taleb, "Extended AMR-WB for high-quality audio on mobile devices," *IEEE Comm. Mag.*, vol. 44, no. 5, May 2006.

[2] ISO/IEC 23003-3:2012, "MPEG audio technologies – Part 3: Unified speech and audio coding," Geneva, January 2012.

[3] J.-M. Valin, K. Vos, and T. Terriberry, "Definition of the Opus Audio Codec," IETF proposed standard RFC-6716, September 2012. Available online at http://tools.ietf.org/html/rfc6716.

[4] J.P. Princen, A.W. Johnson, and A.B. Bradley, "Subband/transform coding using filter bank design based on time domain aliasing cancellation," *Proc. IEEE ICASSP*, pp. 2161–2164, 1987.

[5] E. Allamanche, R. Geiger, J. Herre, and T. Sporer, "MPEG-4 Low Delay Audio Coding based on the AAC Codec," in *Proc. AES 106th Convention*, Munich, no. 4929, May 1999.

[6] M. Schnell, et al., "MPEG-4 Enhanced Low Delay AAC – a new standard for high quality communication," in *Proc. AES 125th Convention*, San Francisco, no. 7503, October 2008.

[7] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 6th edition, Bingley: Emerald Group Publishing, 2012.

[8] J. Herre and J. D. Johnston, "Continuously signal-adaptive filterbank for high-quality perceptual audio coding," in *Proc. IEEE ASSP WASPAA*, New Paltz, October 1997.

[9] M. Neuendorf, et al., "A novel scheme for low bitrate unified speech and audio coding – MPEG RM0," in *Proc. AES 126th Convention*, Munich, no. 7713, May 2009.

[10] M. Jelínek, T. Vaillancourt, and J. Gibbs, "G.718: A new embedded speech and audio coding standard with high resilience," *IEEE Comm. Mag.*, vol. 47, no. 10, October 2009.

[11] D. Virette, B. Kövesi, and P. Philippe, "Adaptive time-frequency resolution in modulated transform at reduced delay," *Proc. IEEE ICASSP*, Las Vegas, pp. 3781–3784, April 2008.

[12] J.-M. Valin, G. Maxwell, T. Terriberry, and K. Vos, "High-quality, low-delay music coding in the Opus codec," in *Proc. AES 135th Convention*, New York, no. 8942, October 2013.

[13] International Telecommunication Union, Radiocommunication Assembly, "Recommendation ITU-R BS.1534-1: Method for the subjective assessment of intermediate quality level of coding systems (MUSHRA)," Geneva, January 2003.