# SPECTRAL ENVELOPE RECONSTRUCTION VIA IGF FOR AUDIO TRANSFORM CODING

*Christian R. Helmrich*[1]*, Andreas Niedermeier*[2]*, Sascha Disch*[2]*, and Florin Ghido*[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

In low-bitrate audio coding, modern coders often rely on efficient parametric techniques to enhance the performance of the waveform preserving transform coder core. While the latter features well-known perceptually adapted quantization of spectral coefficients, parametric techniques reconstruct the signal parts that have been quantized to zero by the encoder to meet the low-bitrate constraint. Large numbers of zeroed spectral values and especially consecutive zeros constituting gaps often lead to audible artifacts at the decoder. To avoid such artifacts the new 3GPP Enhanced Voice Services (EVS) coding standard utilizes noise filling and intelligent gap filling (IGF) techniques, guided by spectral envelope information. In this paper the underlying considerations of the parametric energy adjustment and transmission in EVS and its relation to noise filling, IGF, and tonality preservation are presented. It is further shown that complex-valued IGF envelope calculation in the encoder improves the temporal energy stability of some signals while retaining real-valued decoder-side processing.

***Index Terms***— Audio coding, noise shaping, parametric

## 1. INTRODUCTION

The recently standardized 3GPP EVS codec, release 12 [1], integrates parametric extensions to enhance the performance of the waveform preserving full-band transform coded excitation (TCX) core at low bitrates by efficiently combining noise filling [2] and intelligent gap filling (IGF) [3] directly within the modified discrete cosine transform (MDCT) domain. IGF harvests spectral tiles from lower-frequency (LF) non-zero MDCT bins to fill spectral gaps in the decoded signal. Several existing audio coding standards such as Extended HE-AAC [4] implement predecessors of these techniques. In particular, Extended HE-AAC supports MDCT-domain noise filling but, being restricted to pseudo-random noise insertion, it fails to preserve the harmonic fine-structure of a quasi-stationary tonal signal and relies on Temporal Noise Shaping (TNS) [5] to reconstruct the temporal fine-structure of a transient signal. However, like the TCX core of AMR-WB+ [6], the MDCT-based TCX part of Extended HE-AAC does not support TNS.

Extended HE-AAC also comprises an audio bandwidth extension toolset called enhanced Spectral Band Replication (eSBR). With eSBR the audio signal is downsampled at least by a factor of two after zeroing out the high-frequency (HF) part of the spectrum [7, 8]. The HF part is replicated through eSBR at the receiver, restricting waveform preservation to solely the low frequencies and parameter controlled signal replacement to high frequencies only. Further, eSBR requires the usage of an additional filterbank pair, an analysis and synthesis QMF transform, which is utilized to replace the empty HF part and to resample the audio signal [4]. This increases the computational complexity and memory consumption as well as algorithmic delay of the audio codec. Finally, eSBR uses transmitted energy values on a QMF time-frequency grid to reshape the spectral envelope. Designed for QMF-domain replication of empty HF bands using MDCT-domain coded LF portions, eSBR envelope coding and HF reconstruction cannot handle partially zeroed HF spectra. Since all of these shortcomings should be avoided in a mobile communication codec such as EVS, a low-delay version of eSBR which was used in an early version of EVS [2] was replaced by IGF.

An overview of the processing chain in the MDCT-based TCX encoder of the EVS codec is depicted in Figure 1. IGF is operated on the filtering residuals of TNS and its extension for the envelope-coded region, Temporal Tile Shaping (TTS). TTS is described in detail in an accompanying paper [9].

It is crucial for best perceptual quality of a decoded audio signal that the spectral envelope of the input signal, and thereby the energy distribution among spectral coefficients, is preserved as closely as possible [10]. The following two sections therefore focus on a proper calculation and application of quantized IGF spectral energy representations especially in frequency bands not fully quantized to zero by the encoder.

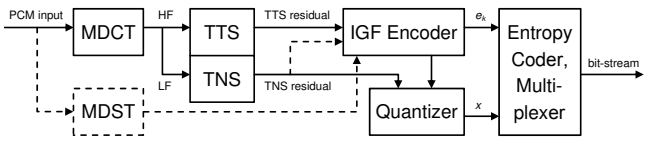


**Fig. 1**. *Audio coder with IGF and TTS coding as used in EVS*

## 2. REAL-VALUED ENVELOPE CALCULATION

IGF is employed for parametric signal reconstruction of spectral bands comprising multiple MDCT bins set to zero by the encoder, either purposely a-priori or by hitting the dead-zone of the MDCT quantizer. For each IGF band an envelope value in the form of an energy scale-factor is coded. A real-valued IGF envelope calculation is utilized whenever a TNS or TTS filter is applied, a choice which is explained in section 3.1.

### 2.1. Terms and Definitions

Let $X \in \mathbb{R}^N$ be the real-valued MDCT spectral representation of a windowed audio signal of window-length $2N$. IGF requires the definition of a start bin $b_s$ and an end bin $b_e \leq N$ together with a band-partitioning of the IGF range $[b_s, b_e[$:

$$\omega_k = \left[ b_s + \sum_{i=1}^{k-1} w_i,\ b_s + \sum_{i=1}^{k} w_i \right[,\ 1 \leq k \leq n, \tag{1}$$

where $w_1, w_2, ..., w_n$ specifies the width of each interval, or band, in the grid $\omega$. This width sequence is usually isotone,

$$0 < w_1 \leq w_2 \leq ... \leq w_n, \tag{2}$$

reflecting the critical-band behavior of the human ear, where auditory filter bandwidth increases with center frequency, and

$$b_s + \sum_{i=1}^{n} w_i = b_e. \tag{3}$$

For each band index $k$ an energy scale-factor $e_k$ is computed.

### 2.2. Encoder – Transmitter Side

With $X$ and $\omega$ the encoder calculates the spectral envelope $E$:

$$E_k = \sqrt{\frac{1}{w_k} \sum_{i \in \omega_k} X_i^2},\ 1 \leq k \leq n. \tag{4}$$

To prepare $E_k$, which represents the root mean square (RMS) of band $k$, for transmission, quantization is applied, yielding

$$e_k = \lfloor 4 \cdot \log_2(s \cdot E_k) \rceil,\ 1 \leq k \leq n, \tag{5}$$

which parameterizes the target energy in that band. A scalar $s$ serves to scale all $E_k$ to a convenient range for the binary-logarithmic quantization. An equivalent quantization method is used for the scale-factors in the family of AAC encoders [4, 7]; there $s$ is applied additively in the logarithmic domain.

After the determination of $e_k$, the MDCT coefficients of $X$ are quantized as well. In EVS and other constant-bitrate (CBR) coders this is done in a rate-distortion loop to match the frame-wise bit budget constraint. The quantization of $X$ results in levels $x$ which are entropy coded, usually by way of Huffman coding as in AAC or various arithmetic coding methods as in CELT [10], Extended HE-AAC [4], or EVS [1]. $x$ and the IGF side-information, the noiselessly coded $e_k$, are multiplexed into a bit-stream for transmission to the receiver.

### 2.3. Decoder – Receiver Side

A block diagram of the decoder corresponding to the encoder of Figure 1 discussed above is illustrated in Figure 2. It can be seen that the IGF technique is applied just before the TTS and/or TNS filters reshape the spectrum for inverse MDCT.

At the receiver the MDCT levels $x$ are reconstructed to $\bar{X}$ and traditional noise filling is employed up to the spectral bin $b_s$. Due to psychoacoustically motivated coarse quantization of $X$ in the encoder, $\bar{X}$ largely or fully consists of zero values within the HF bands and, since noise filling is not used in the IGF range, has spectral gaps. The envelope $E$ is recovered to

$$\bar{E}_k = \frac{1}{s} \cdot 2^{\frac{1}{4} e_k},\ 1 \leq k \leq n, \tag{6}$$

where the $e_k$ are the entropy decoded IGF scale-factors. With $\bar{X}$ the survived waveform-coded energy $S$, accounting for the loss due to zeroing of MDCT bins, is computed for all bands:

$$S_k = \sum_{i \in \omega_k} \bar{X}_i^2,\ 1 \leq k \leq n. \tag{7}$$

Again with $\bar{X}$ the IGF tile (or source) energy $\bar{T}$ is calculated:

$$\bar{T}_k = \sum_{i \in v_k} \bar{X}_{i-d_k}^2,\ d_k \geq \sum_{j=1}^{k} w_j,\ 1 \leq k \leq n, \tag{8}$$

where $v_k$ is the subset of indices $i$ in $\omega_k$ at which $\bar{X}_i = 0$ and $d_k$ is a tabulated minimum spectral distance for each band to arrive at LF non-zero MDCT data. As $\bar{E}$ reflects the original (or target) energy, the missing energy $M$ in $\omega$ can be derived:

$$M_k = w_k \cdot \bar{E}_k^2 - S_k,\ 1 \leq k \leq n. \tag{9}$$

Then, a tile gain $g_k$ can be computed for each band in grid $\omega$:

$$g_k = \begin{cases} \sqrt{\frac{M_k}{\bar{T}_k}} & \text{if } \bar{T}_k > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Finally, the content in the spectral gaps is reconstructed using

$$\bar{X}_i = \begin{cases} g_k \cdot \bar{X}_{i-d_k} & \text{if } \bar{X}_i = 0, \\ \bar{X}_i & \text{otherwise,} \end{cases} \tag{11}$$

for $i \in \omega_k,\ 1 \leq k \leq n$. This effectively fills zero-quantized bins in $v_k$ via tile copy-up, leaving non-zero bins unaffected. To obtain the output spectrum $X'$, $\bar{X}$ is TNS/TTS filtered.

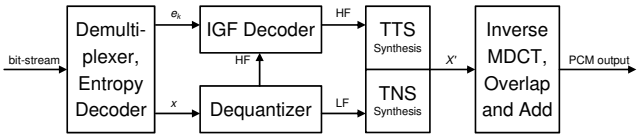


**Fig. 2**. *Audio decoder with IGF process before TTS decoding*

## 3. COMPLEX-VALUED ENVELOPE CALCULATION

The previous section demonstrated that IGF can preserve the tonality of a HF signal via tile transposition if both the HF and LF spectral regions exhibit roughly the same tonality, or by waveform coding due to the possibility of transmitting HF non-zero MDCT levels. Moreover, it was shown that the IGF envelope information can be computed using only real-valued MDCT samples. However, being a sub-sampled transform of $2N$ time inputs and $N$ frequency outputs, the MDCT does not satisfy Parseval's theorem [11] and as such contains undesirable time-domain aliasing components in its coefficients [12]. In case of spectrally flat, relatively coarse bands $\omega_k$ comprising a lot of bins, these disadvantages have only little impact, and the implicit averaging in the measurement of $e_k$ still leads to a good estimate of the true target energy. For narrow $\omega_k$ and tonal signals, particularly high-pitched ones where only one harmonic falls into each band, the effect is more obvious.

Similarly to illustration [11, Fig. 3], Figure 3 a shows how frame-wise MDCT energies of a stationary tone of constant level can vary with both tone frequency and phase. Depicted are MDCT energies measured on two time-domain oscillators of different frequency (and, in this example, fractional offset within the discrete MDCT bin-grid). It can be observed that these energies never reach unity, as they would for an energy preserving transform for which Parseval's theorem holds, and that they exhibit unequal temporal modulation patterns. If in each frame, the MDCT values of one tone – the source – are used for IGF of the zero-quantized MDCT bins of the second tone – the target – by means of (6 – 11), modulation artifacts also occur in the decoded output after inverse MDCT of $X'$ and overlap-add of the frames. This audible level fluctuation of up to a few dB can be measured on the reconstructed waveform and is visualized by the thin solid line in Figure 3 b.

Two changes to the real-valued IGF envelope calculation described in section 2.2 can serve to prevent modulation in an IGF processed decoding when the target signal is stationary. Firstly, a complex RMS value $E'_k$ can be determined instead of $E_k$, which uses only the real-valued $X$. This can be done via the modulated complex lapped transform (MCLT), which has the MDCT as its real part and whose imaginary part is the modified discrete sine transform (MDST) [12]. This approach delivers stable measurements (thick solid line in Fig. 3 a) and has been used by Fielder et al. in the parameter derivation for the spectral extension tool of Dolby Digital Plus [13]. In fact, when substituting $\sqrt{0.5} \cdot E'_k$ for $E_k$ in (5) – which, if ignoring quantization differences and assuming $S_k = 0$, is equivalent to the method described in [13] – the decoder-side modulation depth is reduced (dashed line in Fig. 3 b) but remains audible.

Secondly, a modification to (4) is therefore proposed that accounts for the time-domain aliasing inherent in the MDCT-only processing of section 2.3. Let $X$ be the real MDCT part, $Y$ the imaginary MDST part of a MCLT power spectrum $P$:

$$P := X^2 + Y^2. \tag{12}$$

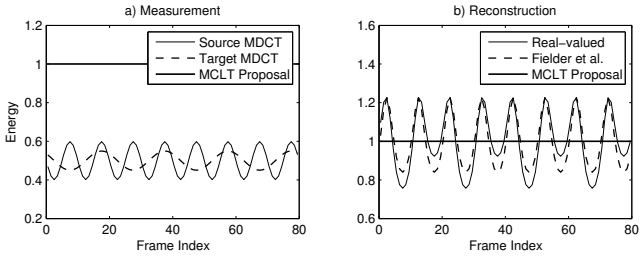


**Fig. 3**. *Effects of real- or complex-valued IGF calculations in the encoder (a) and real-valued adjustment in the decoder (b)*

### 3.1. Encoder – Transmitter Side

With the complex-valued $P$ the target envelope $E'_k$ is derived:

$$E'_k = \sqrt{\frac{1}{w_k} \sum_{i \in \omega_k} P_i},\ 1 \le k \le n. \tag{13}$$

This represents the complex RMS value noted earlier. Using the ratio between an estimate of the decoder-side energy $\bar{T}_k$,

$$T_k = \sum_{i \in \omega_k} X^2_{i-d_k},\ 1 \le k \le n \tag{14}$$

with $X$ instead of $\bar{X}$ as in (8), and the complex source energy,

$$T'_k = \sum_{i \in \omega_k} P_{i-d_k},\ 1 \le k \le n, \tag{15}$$

an energy-preserving modification of $E_k$ in (4) is derived, and similar to the real-valued calculation, quantization is applied:

$$e'_k = \lfloor 4 \cdot \log_2 \left( s \cdot \sqrt{\frac{T_k}{T'_k}} \cdot E'_k \right) \rceil,\ 1 \le k \le n, \tag{16}$$

using the same $s$ as in (5) for scaling to a convenient range for the quantization. By transmitting $e'_k$ instead of $e_k$, modulation can be avoided completely, as shown by the thick line in Fig. 3 b. At the same time, all decoder-side processing can remain unchanged, i. e. real-valued, which is advantageous in terms of computational complexity (no imaginary components need to be computed), an important issue on low-power devices.

Still, real-valued envelope calculation according to (4, 5) can be of benefit and was noted to be utilized whenever TNS and/or TTS filtering is applied. The motive for this decision is that, as depicted in Fig. 1, complex envelope measurement in a filter-residual domain would require filtering of the MDST in addition to the MDCT bins, adding unwanted complexity.

## 4. ADAPTIVE BAND COMBINING

The performance of parametric techniques often depends on an appropriate choice of parameter granularity. A low granularity emphasizes averaging effects, whereas a higher granularity allows for a detailed reconstruction. In case the original

spectral content is vastly different in terms of fine structure from the reconstructed content, one should prefer a coarse frequency band resolution and the implicit averaging that comes with it. The following example illustrates how a too high spectral band resolution can lead to these artefacts: when the content of the original spectral band comprises a tone and the reconstructed band contains just noise, the noise gets amplified up to the energy of the original tone. This is e.g. the case if all spectral lines that represent the tone are zeroed by hitting the dead-zone of the subsequent quantizer and the width $w_k$ being less than the distance between two adjacent harmonic lines in a tonal signal. The proposed adaptive band combining beneficially resorts to a lower spectral resolution in the above case: the width sequence $w_1, w_2, .., w_n$ starts with fine resolution $w_1, w_2, ...$ to basically enhance energy precision in lower frequency bands. If the signal is tonal, all bands $w_1, w_2, ...$ are combined pairwise to enlarge the width of bands. In the EVS coder, this is done at the decoder side if the whitening level is signaled to be "off". After combining bands, the transmitted energies $\bar{E}_k$ have to be added to reflect the energy level on the new combined bands. Now formulae (9) to (11) are applied with the new grid $\omega'$ and $S'_k$, $\bar{T}'_k$:

$$w'_k = w_{2k-1} + w_{2k},\ 1 \le k \le \frac{n}{2}. \tag{17}$$

Note that $n$ is even in EVS. Now a new grid $\omega'$ is defined with the sequence $w'_1, w'_2, ...$, see formula (1 – 3) for details.

$$\bar{E}'_k = \sqrt{\frac{1}{w'_k} \cdot \left(w_{2k-1}\bar{E}^2_{2k-1} + w_{2k}\bar{E}^2_{2k}\right)} \quad \text{for } 1 \le k \le \frac{n}{2}. \tag{18}$$

The survived energy and tile energy should also be combined:

$$S'_k = S_{2k-1} + S_{2k} \tag{19}$$

$$\bar{T}'_k = \bar{T}_{2k-1} + \bar{T}_{2k} \tag{20}$$

Figure 4 shows this effect on a coded tonal input signal. A blind listening test of the concepts in sections 2–4 will follow.

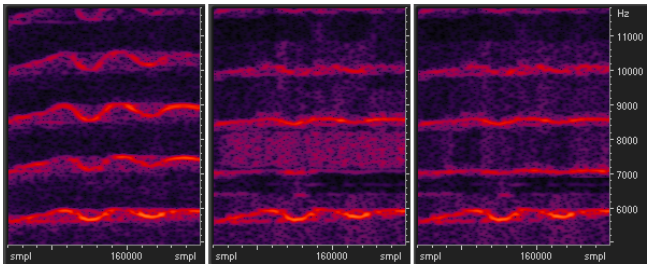


**Fig. 4**. *Effect of IGF band combining: original (left), decoded without band combining (mid), with band combining (right).*

## 5. ENTROPY CODING OF QUANTIZED ENVELOPE

When utilizing IGF, either independent or consecutive interdependent frames are generated. The quantized energy values $e_k$, with $1 \le k \le n$, are encoded depending on the type of frame and the value of $k$. The table below illustrates the possible context values $a$–$d$ available when encoding a value $e_k$.

| band index | previous frame | current frame |
|---|---|---|
| $k$ | $a$ | $e_k$ |
| $k-1$ | $c$ | $b$ |
| $k-2$ | | $d$ |

Using envelope contexts both across frequency and time is similar to image coding, where contexts across the x and y axis of an image are used, such as in [14], where a fixed linear predictor is used as a plane fitting and basic edge detection, and the prediction errors are coded. However, in our scheme the logarithmic domain linear prediction enables to accurately predict constant, fade-in, and fade-out IGF energy areas. The actual prediction error values are encoded with an arithmetic coding using probability distribution tables extracted from a representative training data set. An escape coding mechanism is used for the values outside of the distribution center.

The values $b-d$ and $a-c$ represent the expected amount of noisiness or variability across $k$ and the value $b-c$ represents the same across frames. For independent frames, when $k = 1$, $e_k$ is encoded using one probability table. When $k = 2$, $e_k - b$ is encoded using another probability table. When $k \ge 3$, the value $e_k - b$ is encoded using a probability table determined by $Q(b-d)$, with $Q(\cdot) = \text{sgn}(\cdot)\min(|\cdot|, 3)$. For dependent frames and $k = 1$, $e_k - a$ is encoded using one probability table. When $k \ge 2$, $e_k - a + b - c$ is encoded using a probability table determined by both $Q(a-c)$ and $Q(b-c)$.

## 6. CONCLUSION

This paper discussed the motivation and implementational details behind the parametric energy reconstruction utilized in the MDCT-based TCX core of the new EVS coding standard, and its relation to noise filling, gap filling (IGF), and tonality preservation. The IGF concept was examined using both real and complex-valued computation of the IGF energy representation. Moreover, the design of a context adaptive arithmetic coding scheme for bitrate-efficient transmission of the spectral envelopes was presented. It can be concluded that IGF, unlike previous QMF or MDCT-based bandwidth extension tools, can prevent both spectral gaps and tonality mismatch in a decoded signal with only little computational complexity and bitrate overhead, and no additional algorithmic delay.

## 7. ACKNOWLEDGMENT

The authors thank Christian Neukam for his help in the design as well as the implementation and evaluation of IGF and TTS.

## 8. REFERENCES


[1] 3GPP TS 26.445, "EVS codec; detailed algorithmic description; technical specification, release 12," Sep. 2014.

[2] C. R. Helmrich, G. Marković, and B. Edler, "Improved low-delay MDCT-based coding of both stationary and transient audio signals," in *Proc. of IEEE ICASSP '14*, May 2014, pp. 6954–6958.

[3] ISO/IEC MPEG, *N14747, Text of ISO/MPEG 23008-3 DIS on 3D Audio*, JTC1/SC29/WG11, Oct. 2014.

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

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

[6] 3GPP TS 26.290, "Audio codec processing functions; Extended Adaptive Multi-rate–Wideband (AMR-WB+) codec; Transcoding functions," Dec. 2004.

[7] M. Neuendorf, M. Multrus, N. Rettelbach, G. Fuchs, J. Robilliard, J. Lecomte, S. Wilde, S. Bayer, S. Disch, C. R. Helmrich, R. Lefebvre, P. Gournay, B. Bessette, J. Lapierre, K. Kjörling, H. Purnhagen, L. Villemoes, W. Oomen, E. Schuijers, K. Kikuiri, T. Chinen, T. Norimatsu, K. Chong, E. Oh, M. Kim, S. Quackenbush, and B. Grill, "The ISO/MPEG Unified Speech and Audio Coding Standard – Consistent High Quality for All Content Types and at All Bit Rates," *J. Audio Eng. Soc.*, vol. 61, no. 12, pp. 956–977, Dec. 2013.

[8] H. Zhong, L. Villemoes, P. Ekstrand, S. Disch, F. Nagel, S. Wilde, K.-S. Chong, and T. Norimatsu, "QMF based harmonic spectral band replication," in *Proc. of AES 131st Convention*, Oct. 2011, p. no. 8517.

[9] S. Disch, C. Neukam, and K. Schmidt, "Temporal tile shaping for spectral gap filling in audio transform coding in EVS," in *Proc. of IEEE ICASSP '15*, Apr. 2015.

[10] J.-M. Valin, T. B. Terriberry, C. Montgomery, and G. Maxwell, "A high-quality speech and audio codec with less than 10-ms delay," *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 18, no. 1, pp. 58–67, Jan. 2010.

[11] S. Zhang, W. Dou, and H. Yang, "MDCT sinusoidal analysis for audio signals analysis and processing," *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 21, no. 7, pp. 1403–1414, July 2013.

[12] H. Malvar, "A modulated complex lapped transform and its applications to audio processing," in *Proc. of IEEE ICASSP '99*, Mar. 1999, vol. 3, pp. 1421–1424.

[13] L. D. Fielder, R. L. Andersen, B. G. Crocket, G. A. Davidson, M. F. Davis, S. C. Turner, M. S. Vinton, and P. A. Williams, "Introduction to Dolby Digital Plus, an enhancement to the Dolby Digital coding system," in *Proc. of AES 117th Convention*, Oct. 2004, p. no. 6196.

[14] M. J. Weinberger and G. Seroussi, "The LOCO-I lossless image compression algorithm: Principles and standardization into JPEG-LS," 1999, Available online at http://www.hpl.hp.com/research/info_theory/loco/HPL-98-193R1.pdf.