LOW DELAY LPC AND MDCT-BASED AUDIO CODING IN THE EVS CODEC

Guillaume Fuchs[1], *Christian R. Helmrich*[2], *Goran Marković*[1], *Matthias Neusinger*[1],
Emmanuel Ravelli[1], *and Takehiro Moriya*[3]
[1] Fraunhofer Institut für Integrierte Schaltungen (IIS), Germany
[2] International Audio Laboratories Erlangen, Germany
[3] Nippon Telegraph and Telephone (NTT), Japan
amm-info@iis.fraunhofer.de

ABSTRACT

Speech coders operating in time domain can be extended with a frequency domain mode to improve encoding of music, even though this is challenging at low delay. In such a scenario, the short analysis window limits the benefit of the transform coder, while a delayless switch between the two coders constrains the system further. The paper presents an LPC and MDCT-based audio coder part of the new 3GPP codec for Enhanced Voice Services, which aims to solve the issues. Several advanced coding tools are introduced to alleviate the constraints: transient handling is improved, harmonic structures are better preserved, and the modeling of the zero-quantized frequencies is enhanced. Test results show that the obtained low-delay switched coder brings a clear improvement over a speech coder and is competitive even in comparison to audio coders with higher delay.

Index Terms— Audio coding, Low delay, LPC, MDCT, EVS

1. INTRODUCTION

3GPP Enhanced Voice Services (EVS) codec was recently finalized with the objective of providing enhanced speech and music quality for conversational services over LTE. It extends the traditional narrowband (NB) and wideband (WB) speech coders to superwideband (SWB) and fullband (FB) and also improves generic audio coding by supplementing the speech coding with a transform coding. The system switches between the two coding schemes depending on the content of the input signal. Speech-like signals are conveyed to an ACELP-based speech coder while music is handled by an MDCT-based generic audio coder.

Switched coders like AMR-WB+ [1] and USAC [2] have already shown their efficiency in handling both speech and music signals. They integrate elements from audio and speech coding into a single structure that is able to switch from one paradigm to the other one. However they rely on long windows for stationary music and their too high algorithmic delay prevents them to be utilized for low delay applications.

On the other hand, low delay transform audio coding has recently received special attention with the development of AAC-ELD [3] and OPUS [4]. However, such coders are not specifically designed for a switched system and don't scale efficiently towards low bitrates. Moreover they still suffer from typical low-delay artifacts [5].

The aim of the paper is to describe the LPC and MDCT-based audio coder employed in the EVS codec. It forms one of the two MDCT coding variants associated to ACELP. As it shares the LPC with ACELP, this eases the transitions between the two coding modes and is especially well suited for speech over background noise or music, where a frequent switching is needed. The audio coder is also particularly designed for handling transients and tonal music. The other MDCT coding variant, which is not in the scope of the paper, exploits a spectral noise shaping independent from LPC [6].

The audio coder is derived from the Transform Coded Excitation (TCX) mode of AMR-WB+ and USAC and was further amended for low-delay. It inherits the capacity to switch seamlessly and on demand to ACELP for voiced speech and strong transients. For other transients, or when ACELP quality saturates at high rates, a delayless switch to low-overlap windows or short blocks is possible. Moreover, the noise shaping is refined by an Adaptive Low Frequency Emphasis (ALFE) and by Temporal Noise Shaping (TNS). For tonal music and periodic signals, usually penalized by the poor frequency selectivity of low-delay windows, a harmonic model is added to the probability model of the entropy coder. A post-filter guided by side-information can further enhance the harmonic structure. The quantization is optimized and in case of coarse quantization, the spectral holes are coped by an enhanced noise and gap filling.

Subjective test results validate the relevance of a switched coding at low-delay. They show that the proposed MDCT-based TCX combined with ACELP can significantly improve the low-delay coding of mixed and music content. Moreover the obtained switched coder is a very competitive alternative to audio coders with higher delay for coding music at low bitrates.

The paper is organized as follows. It begins with a system overview section. The windowing and switching between different windows and with ACELP is presented in section 3. The following sections are dedicated to the main advances brought to the noise shaping, quantization, entropy coding, and post-processing. The evaluation is done in section 8 before concluding the paper in section 9.

2. SYSTEM OVERVIEW

The proposed MDCT-based TCX follows the traditional coding chain, i.e. signal transformation, noise shaping, quantization and entropy coding. However as it is illustrated in Figure 1, the coding scheme is extended by several additional predictions and analyses for a more efficient coding.

A transient detector controls the adaptive windowing. The window is also dependent whether the previous coding mode was ACELP.

The LP analysis is used by the Frequency Domain Noise Shaping (FDNS) [2]. It has the advantage to have a compact representation and to share with ACELP its perceptual model, which is generic enough for a large range of bitrates and audio materials. The LPC

978-1-4673-6997-8/15/$31.00 ©2015 IEEE

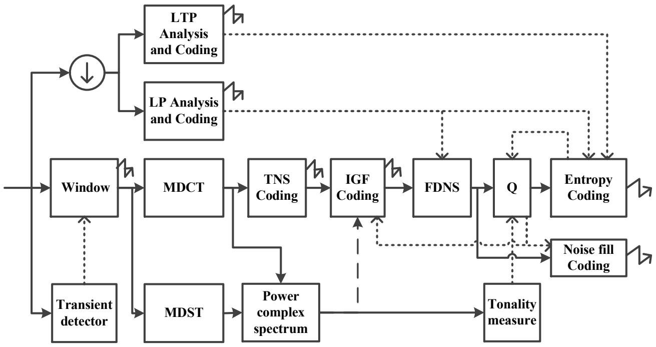

Fig. 1. Block diagram of the low-delay MDCT-based TCX encoder.

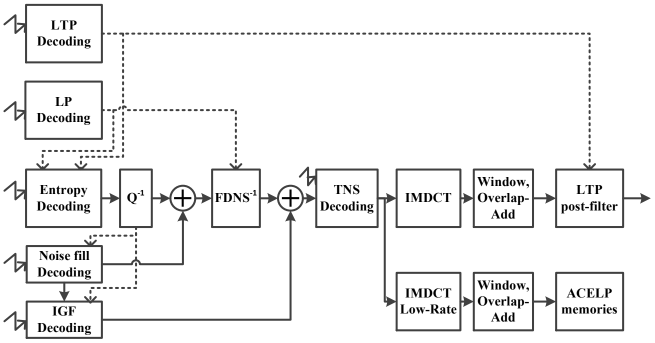

Fig. 2. Block diagram of the low-delay MDCT-based TCX decoder.

can be also exploited by the entropy coder. Besides, a frequency domain prediction is possible for TNS. The Long-Term Prediction is used by both the entropy coder and the decoder's LTP post-filter for enhancing periodic signals.

An MDST is adjoined to the MDCT in order to get a complex frequency representation of the signal [7]. It allows a more accurate computation of band energies and tonality measurements for the quantizer optimization and for the parametrization of the remaining unquantized MDCT lines. In low frequencies, the noise fill models the zeroed MDCT lines with a single energy level. For high frequencies, where more lines are set to zero, Intelligent Gap Filling (IGF) [8] complements the noise filling. IGF is an efficient and flexible parametric representation of spectral regions that upgrades beyond traditional bandwidth extension functionality and allows the quantizer to be relaxed in high frequencies by zeroing more MDCT lines.

The decoder is depicted in Figure 2 and consists in inverse processing of the encoder path. Specific to the decoder, the noise filling and gap filling are added to the quantized MDCT spectrum. It is also worth noting that a second Inverse MDCT is performed on the low band for getting a synthesis signal at ACELP sampling rate. The output is used for feeding ACELP memories, e.g. LP synthesis filter states and the adaptive codebook, in case of switching to ACELP in the next frame. Such a decoded signal is also available at the encoder side for synchronizing ACELP states at both sides. The decoding ends with an LTP post-filtering enhancing the harmonic structure of periodic signals.

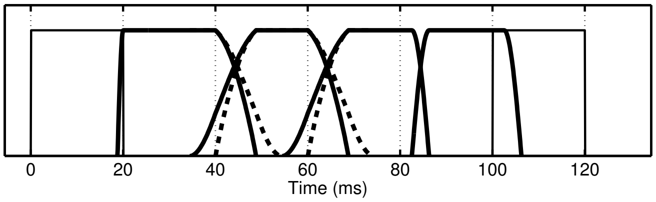

Fig. 3. Window switching between asymmetric and symmetric windows (thick) and with ACELP (thin). Analysis windows are in solid lines, synthesis windows in dashed lines for asymmetric windows.

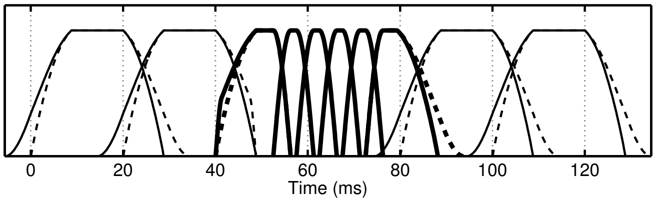

Fig. 4. Block switching between long (thin) and short windows (thick). Synthesis windows are in dashed lines for asymmetric windows.

3. WINDOWING AND SWITCHING

TCX in EVS works on a 20 ms framing using adaptively asymmetric and symmetric windows with a maximal overlap region of 8.75 ms. The asymmetric window was optimized for reducing the time modulation compared to AAC-ELD [3], while its long tail still brings a better frequency selectivity than symmetric windows. The symmetric windows are sine windows with half of the overlap, i.e. 4.375 ms or a minimum overlap of 1.25 ms. The symmetric windows produce no time modulation and limit time spreading of the quantization noise. They are more adapted for temporally structured signals and are selected in case of a transient being detected in the current frame or in the lookahead.

The window switching at bit-rates lower than 48 kbps is illustrated in Figure 3. ACELP represented by rectangular windows is selected for voiced speech and strong transients. The transitions between ACELP and TCX follow the same principle as in AMR-WB+. The first TCX window is extended by 25%, while the overlap of the last frame is discarded. The transitions are smoothed using the ringing of the LP synthesis filter of ACELP.

At higher bit-rates, TCX is no more accompanied by ACELP as the quality of the speech coder saturates. Switching to shorter windows is then needed for handling transients as illustrated in Figure 4. The delayless switching from asymmetric windows to short blocks is achieved by using a specific transition slope: the long tail of the asymmetric window is cut and smoothed on 1.25ms. It ensures a perfect Time Domain Aliasing Cancellation, at the price of a near-perfect reconstruction. Indeed such a window is not entirely complementary between the right and left side. However, the loss of energy is negligible and it avoids the need either to delay the switching or to use complicated perfect reconstruction transition windows or windows showing inefficient high time modulation [5].

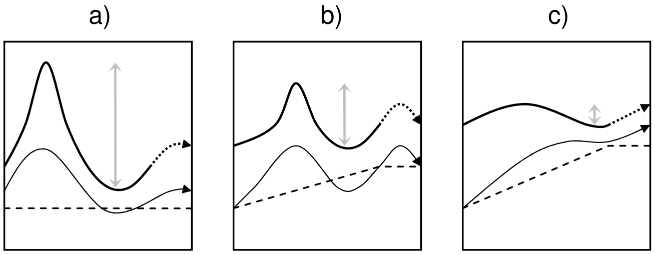

Fig. 5. Noise shaping effect of the new ALFE for three input signals based on low-frequency dynamic range (gray). Spectral envelopes (thick), quantization noise levels (thin) when employing LPC-based FDNS and the resulting low-frequency de-emphasis (dashed) in the decoder. ALFE is not used above a certain frequency (dotted).

4. NOISE SHAPING

4.1. LPC-Based FDNS

The spectral quantization noise shaping is derived from the weighted quantized LP filter coefficient and is applied in frequency domain by means of gain factors as in USAC [2]. Since LP analysis is common to ACELP and is performed on ACELP sampling rate, the remaining part of the high-frequency spectrum not covered by the LP analysis is processed using the last gain factor.

Quantization of LSFs is also common to ACELP except at lower bit-rates below 10 kbps where the LSF coding bits are reduced from 31 to 16 bits per frame. For such a bitrate, only the weighted LPC coefficients needed for the noise shaping are transmitted. The full predictive coefficients are estimated in case of transition to ACELP.

4.2. Temporal Noise Shaping (TNS)

TNS is used at rates above 16.4 kbps to control the temporal shape of the quantization noise [9, 10]. Up to 2 filters per window are used, one in the range from 600 to 4500 Hz and another one from 4500 Hz to the Nyquist frequency. TNS is used only if the prediction gain or the sum of the squared TNS reflection coefficients are high enough.

4.3. Adaptive Low-Frequency Emphasis (ALFE)

Due to the limited order of the LPC coefficients, artifacts may occur in the quantized decoded signal especially at low frequencies, where human hearing is most sensitive. To this end low-frequency emphasis (prior to quantization) and de-emphasis (upon decoding) were introduced in [1, 11] to increase the SNR and quality at low frequencies. In AMR-WB+, FDNS is not used, hence its (de-)emphasis can only be controlled asynchronously using the un-quantized input spectrum at the encoder and the quantized decoded spectrum at the decoder. The consequence is that the emphasis is not perfectly invertible, particularly when a spectrum is coarsely quantized.

A new, invertible ALFE method is therefore devised and controlled by the LPC shaping gains, available in both encoder and decoder, instead of the MDCT spectra. Moreover, ALFE is applied in a psychoacoustically optimized fashion such that only those low-frequency spectral regions which would cause quality degradation at coarse quantization are amplified (emphasis) and attenuated (de-emphasis), as depicted in Figure 5. Essentially, ALFE is maximized when the low frequency signal is spectrally flat, as in Figure 5 c.

5. OPTIMIZED QUANTIZATION

A Uniform Reconstruction and Unity Ratio Quantizer (URURQ) is applied after the noise shaping on the perceptually whitened MDCT coefficients. Such a quantization was shown to be quasi-optimal for entropy constrained scalar quantization of Laplacian-distributed sources [12]. URURQ employs the same simple decoding rule as a Uniform Quantizer (UQ) but outperforms it at low bit-rates and is asymptotically equivalent at high bit-rates. The quantization differs from UQ by its encoding rule and shows a step around 0 (deadzone) which is larger than the other steps. The encoding and decoding rules are summarized by the following formula:

$$\hat{X}(k) = \begin{cases} \lfloor X(k) + (1 - \frac{dz}{2}) \rfloor & \text{if } X(k) \geq 0, \\ \lceil X(k) - (1 - \frac{dz}{2}) \rceil & \text{if } X(k) < 0, \end{cases} \tag{1}$$

where the quantized values $\hat{X}(k)$ also represent the quantization indices, and $dz > 0$ is the deadzone size. A dz of 1.25 was found near-optimal for a large range of bit-rates and audio materials.

One advantage of URURQ is that the deadzone can be adaptively optimized in the entire or even in a part of the spectrum without sending any additional side-information, which is advantageous over a previous approach [13]. This is exploited for coarse quantization steps where some MDCT quantized magnitudes may toggle between 1 and 0 from frame to frame. This unpleasant musical noise is avoided by increasing dz to 2 in frequency regions where a line-wise tonality computed on the power spectrum, determined from the MDCT and MDST coefficients, is low.

Additionally, a perceptually driven mask applied in the high frequencies selects the tonal components to code and lets the IGF model the non-tonal components.

The quantization step is defined by a global gain and is adjusted per frame. The optimal global gain is iteratively searched for fulfilling at best the bit-budget while lowering the overall distortion. At each optimization step, the spectrum is normalized and quantized, and the entropy coding bit consumption estimated. For complexity reasons, the number of iterations is limited to 4. The optimal global gain is then quantized on a 7-bit logarithmic scale.

The quantization can even be further refined by a residual quantization which exploits the eventually unused bits after the rate-loop optimization. The quantized global gain is first refined on a maximum of 3 extra bits, followed by a refinement of the spectral lines, each with 1 extra bit, starting at the lowest-frequency line.

6. ENTROPY CODING

The quantized spectral coefficients are noiselessly coded by an arithmetic coding which can derive its probability models in different ways.

At bitrates above 10 kbps, a past context is used for deriving the probability model [14]. For robustness reasons in case of packet loss, the context is limited to only include lower frequency coded lines of the current spectrum and is independent from the past frames.

At low rates, the quantized spectrum is too sparse for exploiting efficiently a past context. The probability is then deduced from the formantic structure of the spectrum given by the LPC envelope [15].

In both cases, a harmonic model can refine the probability model by exploiting the expected residual correlation between harmonics of tonal signals.

7. POST-PROCESSING

7.1. Enhanced Noise and Gap Filling

Noise filling plays an important perceptual role as it conceals the spectral holes of a coarse quantization by injecting a random noise, replacing zero-quantized lines. As it was taught in [5], the injected noise is lowered for harmonic signals and a spectral tilt is added for matching better the true spectral envelope.

Moreover, Intelligent Gap Filling (IGF) enhances the noise filling for high frequencies. Instead of injecting only random noise, it exploits advantageously neighboring spectral portions for filling spectral gaps [8]. The filled gaps are frequency and time shaped by entropy-coded scale factors and a temporal shaping [16].

7.2. LTP Post-Filtering

The transmitted LTP parameters are exploited by an LTP post-filter for enhancing the harmonic structure of periodic signals. A coded fractional pitch lag and a quantized gain control an IIR comb-filter. By attenuating the inter-harmonic spectral components, the post-filter reduces significantly the coding distortion.

8. EVALUATION

To evaluate the new MDCT-based TCX, two listening tests were conducted for WB and SWB mixed and music content. The system was tested in combination with ACELP for an overall algorithmic delay of 32 ms, although the TCX delay is only 28.75 ms.

8.1. WB Mixed and Music

For WB, a subjective test based on the recommendation ITU-T P.800 DCR (Degradation Category Rating) [17] was conducted. 14 naive listeners took part in the test and assessed the tested conditions against the original. In total 24 items were tested, 12 items of mixed content and 12 music items of diverse genres. Stimuli were mono and played on binaural Sennheiser HD-280 Pro headphones.

As references, 3 MNRUs were used as well as 3GPP AMR-WB and ITU-T G.722.1 at 24 kbps. The first coder is an ACELP-based speech coder while the second one is an MDCT-based audio coder with 40 ms of delay, recommended for hands-free applications.

Figure 6 shows a clear advantage of the switched coder (LD-TCX+ACELP) over a standalone speech coder and even over the standalone transform coder. Results confirm that a switched system is a good choice even at low-delay: the MDCT-based TCX interacts seamlessly with ACELP while performing very well on music. On such a test, 16.4 kbps is already rated as good as the original uncoded reference.

8.2. SWB Mixed and Music

For SWB, the MUSHRA methodology [18] was followed. 8 expert listeners were asked to assess 5 mixed items (speech over background music, radio commercials and jingles) and 5 music items (indexed from 1 to 5 are singing pop, barrel organ, baroque music, dance and rock).

The tested coder was compared to the non low-delay switched coder AMR-WB+ and the low-delay coder OPUS. AMR-WB+ was used at 24 kbps with an algorithmic delay of 75 ms. OPUS version 1.1 was used with the default parameters, i.e. VBR coding and a 20 ms framing with a lookahead of 2.5 ms for the transform coder.

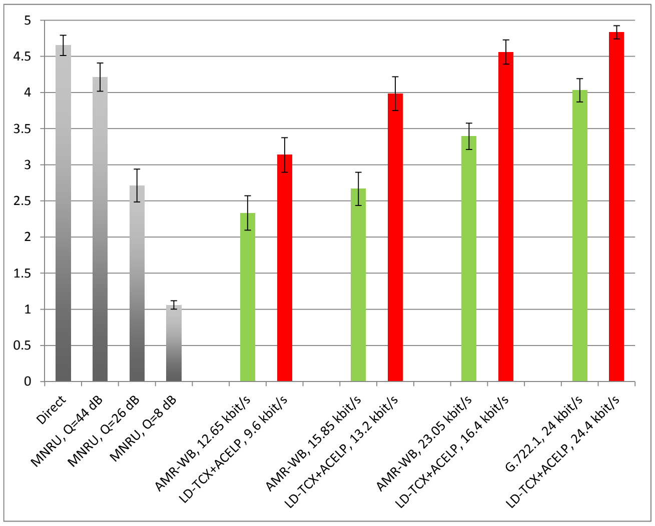

Fig. 6. P.800 DCR test results for WB Mixed and Music content. Average scores and 95% confidence intervals.

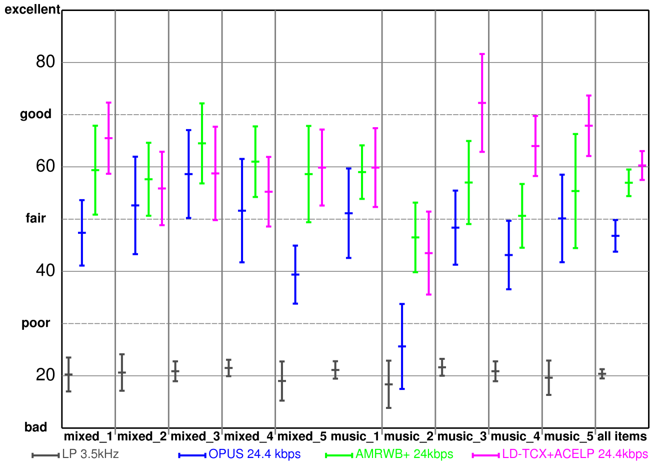

Fig. 7. MUSHRA test results for SWB Mixed and Music content. Average scores and 95% confidence intervals. The hidden reference scored always at 100 and is omitted.

Results are reported in Figure 7 and show that the new switched system is able to slightly (but significantly) exceed the quality of the AMR-WB+ codec at about the same bitrate. In contrast to OPUS, the new system can cope with the low delay even at low bitrates.

9. CONCLUSION

In the current paper, we presented an LPC and MDCT-based audio coder designed for a low-delay switched coding system. New advanced and optimized coding tools allow the transform coder to improve largely the quality of music for low-delay audio coding and to reach a quality comparable to the state-of-the-art non-low-delay audio coders at low-bitrates. The present work forms one of the main component of the new 3GPP EVS standard.

10. REFERENCES

[1] J. Makinen, B. Bessette, S. Bruhn, P. Ojala, R. Salami, and A Taleb, "AMR-WB+: a new audio coding standard for 3rd generation mobile audio services," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, March 2005, vol. 2, pp. ii/1109–ii/1112 Vol. 2.

[2] M. Neuendorf, M. Multrus, N. Rettelbach, G. Fuchs, J. Robilliard, J. Lecomte, S. Wilde, S. Bayer, S. Disch, C. Helmrich, R. Lefebvre, P. Gournay, B. Bessette, J. Lapierre, K. Kjrling, H. Purnhagen, L. Villemoes, W. Oomen, E. Schuijers, K. Kikuiri, T. Chinen, T. Norimatsu, C. K. Seng, E. Oh, M. Kim, S. Quackenbush, and B. Grill, "MPEG Unified Speech and Audio Coding - the ISO/MPEG standard for high-efficiency audio coding of all content types," in *Audio Engineering Society Convention 132*, Apr 2012.

[3] R. Geiger, J. Herre, M. Jander, M. Multrus, M. Schmidt, M. Schnell, and G. Schuller, "Enhanced MPEG-4 Low Delay AAC - low bitrate high quality communication," in *Audio Engineering Society Convention 122*, May 2007.

[4] J.-M. Valin, G. Maxwell, T. B. Terriberry, and K. Vos, "High-quality, low-delay music coding in the opus codec," in *Audio Engineering Society Convention 135*, Oct 2013.

[5] C. R. Helmrich, G. Marković, and B. Edler, "Improved low-delay MDCT-based coding of both stationary and transient audio signals," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 6954–6958.

[6] M. Dietz et al., "Overview of the EVS codec architecture," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2015.

[7] F. Küch and B. Edler, "Aliasing reduction for Modified Discrete Cosine Transform domain filtering and its application to speech enhancement," in *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*, Oct 2007, pp. 131–134.

[8] ISO/IEC JTC 1/SC 29 N, "Information technology - High efficiency coding and media delivery in heterogeneous environments - Part 3: 3D audio," April 2014.

[9] J. Herre and J. D. Johnston, "Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS)," in *Audio Engineering Society Convention 101*, Nov 1996.

[10] ISO/IEC International Standard 14496-3, "Coding of Audio-visual Objects, Part 3: Audio, Subpart 4 Time/Frequency Coding," 1999.

[11] 3GPP TS 26.290, "Extended AMR Wideband Codec - Transcoding," 2004.

[12] G.J. Sullivan, "Efficient scalar quantization of exponential and laplacian random variables," *Information Theory, IEEE Transactions on*, vol. 42, no. 5, pp. 1365–1374, Sep 1996.

[13] M. Oger, S. Ragot, and M. Antonini, "Model-based deadzone optimization for stack-run audio coding with uniform scalar quantization," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, March 2008, pp. 4761–4764.

[14] G. Fuchs, V. Subbaraman, and M. Multrus, "Efficient context adaptive entropy coding for real-time applications," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 493–496.

[15] T. Bäckström et al., "Arithmetic coding with probability prioris based on LPC shape," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2015.

[16] S. Disch, C. Neukam, and K. Schmidt, "Temporal tile shaping for spectral gap filling in audio transform coding in EVS," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2015.

[17] International Telecommunication Union, "ITU-T Recommendation P.800: Methods for subjective determination of transmission quality," August 1982.

[18] International Telecommunication Union, "ITU-R Recommendation BS.1534: Method for subjective assessment of intermediate quality levels of coding systems," June 2014.