OVERVIEW OF THE EVS CODEC ARCHITECTURE

Martin Dietz[1], *Markus Multrus*[2], *Vaclav Eksler*[3], *Vladimir Malenovsky*[3], *Erik Norvell*[4], *Harald Pobloth*[4],
Lei Miao[5], *Zhe Wang*[5], *Lasse Laaksonen*[6], *Adriana Vasilache*[6], *Yutaka Kamamoto*[7], *Kei Kikuiri*[8],
Stephane Ragot[9], *Julien Faure*[9], *Hiroyuki Ehara*[10], *Vivek Rajendran*[11], *Venkatraman Atti*[11],
Hosang Sung[12], *Eunmi Oh*[12], *Hao Yuan*[13], *Changbao Zhu*[13]

[1]Consultant for Fraunhofer IIS, [2]Fraunhofer IIS, [3]VoiceAge, [4]Ericsson AB, [5]Huawei Technologies Co. Ltd.,
[6]Nokia Technologies, [7]Nippon Telegraph and Telephone Corp., [8]NTT DOCOMO, INC., [9]Orange, [10]Panasonic,
[11]Qualcomm Technologies, Inc., [12]Samsung Electronics Co., Ltd., [13]ZTE Corporation

ABSTRACT

The new 3GPP codec for Enhanced Voice Services (EVS) offers important new features and improvements for low-delay real-time communication systems. Based on a novel, switched low-delay speech/audio codec, the EVS codec contains various advancements for better compression efficiency and higher quality for clean/noisy speech, mixed content and music, including support for wideband, super-wideband and full-band content. The EVS codec operates in a broad range of bitrates, is highly robust against packet loss and provides an AMR-WB compatible mode for compatibility with existing systems. This paper gives an overview of the underlying architecture as well as the novel technologies in the EVS codec and presents listening test results showing the excellent performance of the new codec in terms of compression and speech/audio quality.

Index Terms— *EVS, speech coding, audio coding*

1. INTRODUCTION

The codec for Enhanced Voice Services (EVS), standardized by 3GPP in September 2014, provides a wide range of new functionalities and improvements enabling unprecedented versatility and efficiency in mobile communication [1], [2]. It has been primarily designed for Voice over LTE (VoLTE) and fulfills all objectives defined by 3GPP in the EVS work item description [3], namely:

- Enhanced quality and coding efficiency for narrowband (NB) and wideband (WB) speech services.
- Enhanced quality by the introduction of super-wideband (SWB) speech.
- Enhanced quality for mixed content and music in conversational applications.
- Robustness to packet loss and delay jitter.
- Backward compatibility to the 3GPP AMR-WB codec.

The EVS codec builds upon earlier standards from the speech and audio coding world but adds important new functionalities and improvements, which are described in Sections 2 and 3, whereas section 4 focuses on test results confirming the performance of the codec.

2. NEW KEY FUNCTIONALITIES IN THE EVS CODEC

2.1. Switched Speech/Audio Coding at Low Delay

The EVS codec is the first codec to deploy content-driven on-the-fly switching between speech and audio compression at low algorithmic delay of 32 ms and bitrates down to 5.9 kbps (average) or 7.2 kbps (constant) as used in mobile communication. As a result, coding of generic content, such as natural background and music, is vastly improved compared to older systems. Figure 1 shows a high-level block diagram of the EVS encoder and decoder.

While the speech core is an improved variant of Algebraic Code-Excited Linear Prediction (ACELP) extended with specialized LP-based modes for different speech classes (Section 3.1), MDCT-based coding in different variants is used for audio coding. Major research has been done to increase the efficiency of MDCT based coding at low delay/low bitrates (Section 3.6), to obtain seamless transitions between speech and audio core and to achieve a robust and reliable decision which core to use (Section 3.7).

2.2. Super-wideband Coding and Beyond

The EVS codec is the first codec to offer super-wideband coding of speech up to 16 kHz bandwidth from bitrates as low as 9.6 kbps in combination with features such as support for discontinuous transmission (DTX) and advanced packet loss resiliency (Section 2.5). The EVS codec can also offer full-band (FB) coding up to 20 kHz bandwidth starting at 16.4 kbps.

In contrast to earlier speech/audio codecs, which use a core-independent bandwidth extension [4], the EVS codec uses different approaches depending on the core used. For the LP-based coding, the larger audio bandwidth is achieved by bandwidth extension technologies, namely a time-domain extension technology used during active speech [5]. For the MDCT cores, the coding of higher bandwidth is integrated within the respective algorithms. The result is higher efficiency across all types of content, but in particular for speech. Multi-bandwidth listening tests show a significant quality improvement for SWB compared to WB at all supported operation points (Section 4).

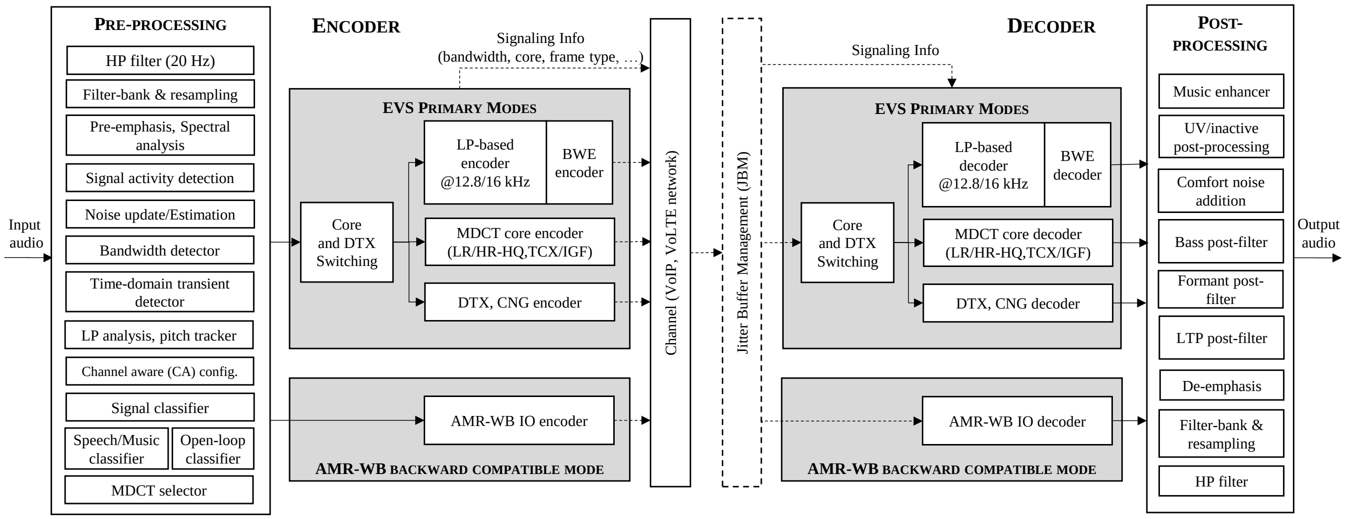

Figure 1. High-Level block diagram of the EVS codec.

2.3. AMR-WB Backward Compatibility

In addition to the EVS Primary modes (Section 2.4), the EVS codec enables backward compatibility with the AMR-WB bitrates through an interoperable (IO) mode, which may be used instead of legacy AMR-WB in terminals and gateways supporting the EVS codec. The AMR-WB-IO mode offers certain decoder improvements over legacy AMR-WB through improved post processing, especially notable for noisy channels and mixed content [7]. Better presence is achieved through bandwidth extension up to 7.8 kHz. Finally, dynamic scaling in the fixed-point implementation improves the performance for low-level input signals (e.g., -36 dBov). Terminals supporting the EVS codec can therefore provide improved quality even for calls restricted to AMR-WB coding. In addition, the integrated implementation allows for seamless switching between AMR-WB IO and EVS Primary modes.

2.4. Range and Switching of Operating Points

Compared to earlier 3GPP conversational codecs, the EVS codec offers a much wider range of operation points, stretching from highest compression to transparent coding. Namely, the EVS codec supports:

- Sampling rates of 8 kHz, 16 kHz, 32 kHz and 48 kHz
- Bitrates from 7.2 kbps to 24.4 kbps for NB
- Bitrates from 7.2 kbps to 128 kbps for WB
- Bitrates from 9.6 kbps to 128 kbps for SWB
- Bitrates from 16.4 kbps to 128 kbps for FB
- DTX and Comfort Noise Generation (CNG)

In addition, a source controlled variable bitrate (SC-VBR) mode at an average bitrate of 5.9 kbps is supported for NB and WB (see Section 3.2). SC-VBR coding is related to active speech segments with DTX/CNG always used for inactive speech coding. The EVS codec operates with a fixed frame length of 20 ms and an overall algorithmic delay of 32 ms. Internally, a set of low delay filters/filterbanks are used to resample the signal to an internal sampling rate of 12.8 kHz (for the common preprocessing as shown in Figure 1) as well as a potentially different sampling rate for coding (depending on bandwidth mode and bitrate). Finally, resampling is also used in the decoder.

The EVS codec may seamlessly switch between operation points at any frame boundary to adapt to the needs of the mobile transmission channel. To avoid inefficient coding for band-limited content, an integrated bandwidth detector will automatically switch to lower bandwidth coding modes for such content, regardless of the input sampling rate. As a result, the EVS codec is a highly flexible, dynamically reconfigurable codec spanning all quality ranges. EVS supports coding of stereo signals by means of coding two mono channels, additional tools for stereo coding are in planning for future Releases.

2.5. Advanced Error Resiliency

Multiple innovative measures have been taken in the EVS codec to provide a built-in, highly robust frame loss concealment to mitigate the impact of packet loss in mobile communications. Inter-frame dependencies in the core coding (e.g. in Linear Prediction (LP)-domain coding or entropy coding) have been minimized to arrest error propagation and thereby ensure fast recovery after lost packets, while various technologies are deployed for concealment of lost packets [8]. At higher bitrates, tools including efficiently coded assisting side information are used [8]. The “channel-aware” coding at 13.2 kbps offers even higher robustness on top of the concealment techniques in [8] through transmission of redundant information of previous frames [9].

Finally, the EVS decoder comes with a Jitter Buffer Management (JBM) solution to compensate for transmission delay jitter. Depending on the conditions in the transmission channel, the JBM uses time scaling methods and interacts with the decoder concealment to provide a well-balanced trade-off between delay and perceptual quality and thereby overall performance.

3. IMPROVEMENTS BROUGHT BY EVS

3.1. LP-based Coding

The speech core used in the EVS codec inherits coding principles of ACELP technology from the 3GPP AMR-WB standard [10], building blocks of which are part of the EVS codec to form the AMR-WB-compatible mode. For EVS Primary modes, the efficiency of the codec has been improved over AMR-WB through various advancements, such as:

- Classification of speech signals based on technologies introduced in the 3GPP2 VMR-WB standard [11] and further refined in the ITU-T G.718 standard [12]. Use of dedicated LP-based coding modes for different speech classes.
- Introduction of Generic Signal Coding (GSC), a LP-based time-frequency mode optimized for very low bitrate coding of music and generic audio [13].
- Support for 16 kHz internal sampling rate in addition to 12.8 kHz
- Use of bass post-filtering and formant enhancement
- Use of an adaptive lag-windowing for LP analysis
- Optimized open-loop pitch search, multi-stage multiple scale lattice and block-constrained trellis coded vector quantization and indexing of the LP coefficients [28].
- Use of a time domain bandwidth extension for active speech [5] for WB, SWB and FB; Use of a frequency domain bandwidth extension for inactive speech and mixed/music in conjunction with GSC [6]

As a further major improvement, the EVS codec detects not only voice activity, but also the level of background noise. If speech over background noise is detected, additional measures are taken, e.g.:

- Modified use of bass post-filtering and formant enhancement during active speech
- Use of dedicated cores for coding the background noise at bitrates of 24 kbps and below: Depending on the operation mode either a variant of GSC or the MDCT-based Transform Coded Excitation (TCX) core (Section 3.5)
- Use of comfort noise addition for a better rendering of the background noise at low bitrates and for masking coding distortions on active speech.

As a result, the EVS codec offers higher compression efficiency as well as speech quality levels unseen in mobile communication so far (Section 4).

3.2. Source-Controlled Variable Bitrate Coding

The EVS VBR mode includes source-controlled variable bitrate (SC-VBR) coding technologies based on the 3GPP2 EVRC-NW speech coding standard [14]. Depending on the input speech characteristics, SC-VBR coding uses an encoding bitrate from among 2.8, 7.2, or 8 kbps. Two new low bitrate (2.8 kbps) coding modes, namely, the prototype pitch period (PPP) and the noise-excited linear prediction (NELP) modes are introduced to encode stationary voiced and unvoiced frames, respectively. PPP encoding exploits the slow varying nature of pitch-cycle waveforms in voiced segments by coding a single representative PPP waveform in the frequency domain. At the decoder, the non-transmitted pitch-cycle waveforms are synthesized through PPP interpolation techniques [14]. In NELP coding, the prediction residual is modeled by shaping a randomly generated sparse excitation signal in both time and frequency domain.

Transient and generic frames that represent weakly correlated signals are encoded using the EVS native coding modes at 8 and 7.2 kbps, respectively. Using novel bitrate selection and bump-up techniques [1], the EVS VBR mode targets an average bitrate of 5.9 kbps by adjusting the proportion of 2.8 kbps and 7.2 kbps frames. SC-VBR coding offers the advantage of equal or better speech quality at a considerably lower average active speech bitrate compared to constant bitrate coding [2].

3.3. Improved Preprocessing and VAD

Signal preprocessing and Voice Activity Detection (VAD) in the EVS codec have been improved to cope with the enhanced requirements of the EVS codec. The VAD, in particular, needs to reliably distinguish between active speech, active music and inactive periods (recording noise, background noise) including a reliable estimate of the background noise level. This data is not only needed for the DTX mode operation (Section 3.4), if enabled, but is also essential for selection between LP-based or MDCT-based coding and the signal adaptive configuration of these cores. The VAD in the EVS codec combines an improved version of a VAD derived from G.718 that works on the spectral analysis of the 12.8 kHz sampled signal [15] with a VAD that operates on the sub-band filter that runs on the input sampling frequency to achieve highest reliability.

3.4. Improved Comfort Noise Generation

DTX operation is important for efficient use of spectrum and battery life in mobile communications. In DTX mode, transmission of background noise is replaced by CNG in the decoder. Apart from the improved VAD, the EVS codec implements two types of CNG to enhance the versatility of the DTX mode: an improved version of LP-based CNG [16] and a novel frequency-domain CNG algorithm [17]. Based on the characteristics of the background noise, the EVS encoder selects which type of CNG will be used. As a result, the EVS codec offers well performing DTX operation throughout all applicable modes up to the high quality level offered by 24.4 kbps EVS coding.

3.5. Improved Low Delay MDCT-based Coding

The delay constraints imposed by systems designed for real-time communication have so far prevented the use of MDCT-based coding for low bitrate mobile systems. In the EVS codec, the availability of efficient MDCT-based compression at low delay *and* low bitrate is, in combination with core switching (Section 3.6) the key enabler for efficient coding of mixed content and music with the EVS codec.

Given a frame length of 20 ms, a delay of 3.25 ms for resampling and other tools, and the design constraint of 32 ms overall delay, as little as 8.75 ms are available for overlap between consecutive frames, a low value compared to codecs

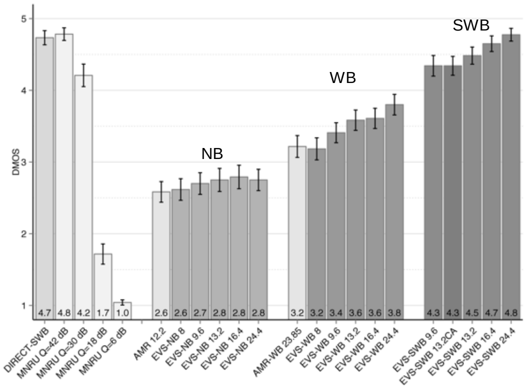

Figure 2. Clean speech multi-bandwidth test.

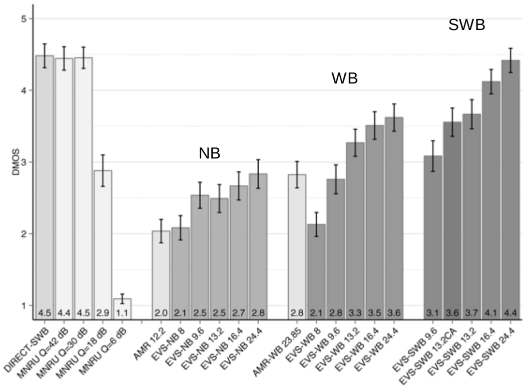

Figure 3. Mixed/music multi-bandwidth test.

for content distribution such as AAC [18]. To cope with this constraint, two improved variants of MDCT coding are implemented in the EVS codec: the Low-Rate/High-Rate High Quality-MDCT coding (LR/HR-HQ) [19][20], an advanced version of G.719 [21], and TCX, an enhanced low delay version of the homonymous core in the MPEG USAC standard [4]. Amongst several other tools, novel LTP post-filter and harmonic model have been added to the TCX algorithm to compensate the effects of the short overlap [29]. The HQ modes benefit from the introduction of improved techniques for, e.g., harmonic signals [32] and noise fill [31].

The EVS encoder selects the MDCT variant to be used depending on the operation mode and the characteristics of the input signal as analyzed in the preprocessing stage. Furthermore, at 7.2 kbps and 8.0 kbps (and rarely also at 13.2 kbps) the GSC mode is also used to code musical content.

3.6. Switching between Speech and MDCT Coding

Naturally, the decision whether to use the LP-based or the MDCT-based coding modes is essential to a switched codec. Embedded in the preprocessing stage, the EVS codec implements a speech/music classifier [22] as well as an SNR-based open loop classifier [23]. The latter is mainly used with the TCX MDCT core, as ACELP and TCX share the same LP-based coding algorithm, enabling selection of the core based on SNR rather than music classification.

Apart from the decision itself, significant efforts have been spent to ensure inaudible transitions between the two cores. While the actual transition happens in the time domain stage of the decoder, buffer updates are performed to enable seamless, signal-adaptive frame-by-frame switching between the cores. Consequently, the EVS codec offers unprecedented compression quality for mixed content and music at low delay and bitrate.

4. TEST RESULTS

Extensive testing has been performed by the contributing companies and within 3GPP to verify the performance of the EVS codec over a wide range of operation points and content types [24]-[26]. Figures 2 and 3 show the results of a multi-bandwidth test conducted by Fraunhofer according to the P.800 DCR test methodology [27] and give a high-level impression of the quality (in DMOS score) for clean speech (German) and mixed content and music (German):

- At 13.2 kbps, an operating point similar to popular bitrates in today’s mobile services, EVS-SWB and EVS-WB outperform AMR-WB 23.85 kbps significantly, in both normal and channel aware (CA) mode for improved error robustness
- EVS-SWB clean speech quality is already high for 9.6 kbps, outperforms AMR-WB 23.85 kbps significantly and increases further with bitrate towards transparency at 24.4 kbps
- EVS-SWB mixed content and music quality outperforms AMR-WB 23.85 kbps at any supported bitrate (9.6 kbps and higher). The quality benefit through increased bitrate is larger than for clean speech. 24.4 kbps is statistically not worse than the original (denoted “DIRECT”).
- For wideband services, EVS-WB is approximately twice as efficient as AMR-WB at 23.85 kbps and offers much higher quality for clean speech and music when using an equivalent bitrate (24.4 kbps)
- In case of NB input signals, the EVS codec performs significantly better than earlier standards especially for mixed content and music stimuli. This mode may be useful in case of inter-connections to NB fixed line networks.

It is well known that test results and their interpretation vary with language and material chosen. However, in the 3GPP Selection Phase, the EVS codec has been tested with 10 languages, 6 different background noises and various music material, showing excellent performance and improvement over earlier standards on a broad basis [25]. Furthermore, extensive performance characterization of the EVS codec will soon become available in the 3GPP Technical Report (TR) 26.952. Additional performance information is available in [26].

5. CONCLUSION

Various new features and improvements make the EVS codec, the latest 3GPP codec for enhanced voice services, the most efficient and versatile codec for high quality communication in any type of network, including the Internet and in particular mobile networks. The imminent introduction of the EVS codec in chipsets and gateways will allow mobile operators and their customers to greatly benefit from capabilities of the EVS codec in VoLTE services.

6. REFERENCES

[1] 3GPP TS 26.445, “EVS Codec Detailed Algorithmic Description; 3GPP Technical Specification (Release 12)”, 2014.

[2] S. Bruhn, et al., “Standardization of the new EVS Codec”, submitted to *IEEE ICASSP*, Brisbane, Australia, Apr. 2015.

[3] 3GPP Tdoc SP-100202, Work Item Description: Codec for Enhanced Voice Services.

[4] M. Neuendorf, et al., “The ISO/MPEG Unified Speech and Audio Coding Standard — Consistent High Quality for All Content Types and at All Bit Rates,” *Journal of the AES*, 61(12): 956—977, Dec. 2013.

[5] V. Atti, et al., "Super-wideband bandwidth extension for speech in 3GPP EVS codec", submitted to *IEEE ICASSP*, Brisbane, Australia, Apr. 2015.

[6] L. Miao, Z. Liu, et al., “A novel frequency domain BWE with adaptive delay and its associated BWE switching”, submitted to *IEEE ICASSP*, Brisbane, Australia, Apr. 2015.

[7] T. Vaillancourt, R. Salami, and M. Jelínek, “New Post-processing Techniques for Low Bit Rate CELP Codecs,” submitted *to IEEE ICASSP*, Brisbane, Australia, Apr. 2015.

[8] J. Lecomte, et al., “Packet Loss Concealment Technology Advances in EVS”, submitted to *IEEE ICASSP*, Brisbane, Australia, Apr. 2015.

[9] V. Atti, et al., "Improved error resilience for VOLTE and VOIP with 3GPP EVS channel aware coding", submitted *to IEEE ICASSP*, Brisbane, Australia, Apr. 2015.

[10] B. Bessette, et al., “The adaptive multi-rate wideband speech codec (AMR-WB),” *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 8, pp. 620-636, Nov. 2002.

[11] M. Jelínek and R. Salami, “Wideband Speech Coding Advances in VMR-WB standard,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1167-1179, May 2007.

[12] M. Jelínek, T. Vaillancourt, and Jon Gibbs, “G.718: A New Embedded Speech and Audio Coding Standard with High Resilience to Error-Prone Transmission Channels,” *IEEE Communications Magazine*, vol. 47, no. 10, pp. 117-123, Oct. 2009.

[13] T. Vaillancourt, et al., “Advances in Low Bitrate Time-Frequency Coding,” submitted to *IEEE ICASSP*, Brisbane, Australia, Apr. 2015.

[14] 3GPP2 C.S0014-D v3.0, "Enhanced Variable Rate Codec, Speech Service Options 3, 68, 70 & 73 for Wideband Spread Spectrum Digital Systems", Oct. 2010.

[15] Z. Wang, L. Miao, J. Gibbs, “Voice activity detector based on optimized sub-band SNR,” submitted to *IEEE ICASSP*, Brisbane, Australia, Apr. 2015.

[16] Z. Wang et al., “Linear Prediction Based Discontinuous Transmission System and Comfort Noise Generation“, submitted to *IEEE ICASSP*, Brisbane, Australia, Apr. 2015.

[17] A. Lombard, et al., “Frequency Domain Comfort Noise Generation for discontinuous transmission in EVS”, submitted to *IEEE ICASSP*, Brisbane, Australia, Apr. 2015.

[18] M. Bosi, et al., "ISO/IEC MPEG-2 advanced audio coding." Journal of the Audio engineering society 45.10 pp. 789-814, 1997.

[19] S. Nagisetty, et al., “Low bit rate high quality MDCT audio coding of the 3GPP EVS standard,” submitted to *IEEE ICASSP*, Brisbane, Australia, Apr. 2015.

[20] Volodya Grancharov, “MDCT Audio Coding with Pulse Vector Quantizers”, submitted to *IEEE ICASSP*, Brisbane, Australia, April, 2015.

[21] Recommendation ITU-T G.719, “Low complexity, full band audio coding for high quality, conversational applications”, International Telecommunication Union (ITU), Series G, Jun. 2008.

[22] V. Malenovsky, et al., “Two-Stage Speech/Music Classifier with Decision Smoothing and Sharpening in the EVS Codec,” submitted *to IEEE ICASSP*, Brisbane, Australia, Apr. 2015.

[23] E. Ravelli, et al., “Open loop switching decision based on evaluation of coding distortions for audio codecs, submitted to *IEEE ICASSP*, Brisbane, Australia, Apr. 2015.

[24] 3GPP Tdoc S4-130522, “EVS Permanent Document (EVS-3): EVS performance requirements”, Version 1.4.

[25] 3GPP, Tdoc S4-141065, "GAL report for EVS Selection Phase", Aug. 2014.

[26] A. Rämö and H. Toukomaa, "Subjective quality evaluation of the 3GPP EVS codec", submitted to *ICASSP 2015*, Brisbane, Australia, Apr. 2015.

[27] Recommendation ITU-T P.800, "Methods for Subjective Determination of Transmission Quality," International Telecommunication Union (ITU), Series P., Aug. 1996.

[28] A. Vasilache, et al., "Flexible spectrum coding in EVS speech and audio codec", submitted to *IEEE ICASSP*, Brisbane, Australia, Apr. 2015.

[29] G. Fuchs, et al., “Low delay LPC and MDCT-based Audio Coding in EVS,” submitted to *IEEE ICASSP*, Brisbane, Australia, Apr. 2015.

[30] T. Moriya, et al., “Harmonic model in combination with LPC envelope for MDCT coding,” submitted to *IEEE ICASSP*, Brisbane, Australia, Apr. 2015.

[31] J. Svedberg, et al., “MDCT Audio Coding with Pulse Vector Quantizers”, submitted to *IEEE ICASSP*, Brisbane, Australia, Apr. 2015

[32] V. Grancharov, et al., “Harmonic Vector Quantization”, submitted to *IEEE ICASSP*, Brisbane, Australia, Apr. 2015