SUPER-WIDEBAND BANDWIDTH EXTENSION FOR SPEECH IN THE 3GPP EVS CODEC

Venkatraman Atti, Venkatesh Krishnan*, Duminda Dewasurendra*, Venkata Chebiyyam*, Shaminda Subasingha*, Daniel J. Sinder*, Vivek Rajendran*, Imre Varga*, Jon Gibbs†, Lei Miao†, Volodya Grancharov‡, Harald Pobloth‡*

*Qualcomm Technologies, Inc., †Huawei Technologies Co. Ltd., ‡Ericsson AB

ABSTRACT

This paper describes the time-domain bandwidth extension (TBE) framework employed to code wideband and super-wideband speech in the newly standardized 3GPP EVS codec. The TBE algorithm uses a nonlinear harmonic modeling technique that incorporates principles of time-domain envelope-modulated noise mixing. At 13.2 kbps, the super-wideband coding of speech uses as low as 1.55 kbps for encoding the spectral content from 6.4-14.4 kHz. Subjective evaluation results from ITU-T P.800 Mean Opinion Score (MOS) tests are provided, showing significantly improved quality compared to the other standardized SWB codecs under both clean speech and speech with background noise.

Index Terms— 3GPP EVS codec, low bitrate bandwidth extension, super-wideband, harmonic nonlinear extension, temporal envelope modulation.

1. INTRODUCTION

Traditionally, speech signals carried by wired and wireless telecommunication systems were band-limited to 4 kHz. Such narrowband voice signals enabled high compression over band-limited channels while largely preserving intelligibility. Nevertheless, removal of higher frequency content in NB systems adversely impacts naturalness and sense of presence, and can also cause phonetic confusions for the listener, particularly for unvoiced sounds. The advent of networks with higher data rates and transcoder free operation, and also improvements in the electro-acoustics of handheld devices, has paved the way for the deployment of codecs that encode speech beyond narrowband. Today, wideband speech coders typically encode signal bandwidths from 50 Hz to nearly 7 kHz, and super-wideband and full-band vocoders extend the upper range of the coded bandwidth to 16 and 20 kHz, respectively [1]-[4].

Perceptual tolerance to spectral and temporal distortion introduced by codecs tends to be greater in higher frequency bands than it is at lower frequencies. Consequently, bandwidth extension methods [1]-[4] encode the spectral regions beyond narrowband, or even beyond wideband, with far greater spectral efficiency (*i.e.*, bits/Hz) than the signal content coding in the narrowband frequency range. One aspect for which much efficiency can be gained is in coding the pitch structure in the higher frequency bands. Since the structure of pitch harmonics in the higher bands of speech is closely related to the structure in the low band, explicit coding of the fine pitch structure of the high band can be avoided. Instead, the fine pitch structure can be derived from that of the low band by employing appropriate estimation models [1]-[3], [8], [9]. The correction factors that are needed to modify the estimated high band fine structure to match that of the input speech signal are then transmitted to the decoder to enable reconstruction of wideband (WB), super-wideband (SWB), or full-band (FB) speech. Coding these correction factors involves a trade-off between spectral efficiency and accurately reconstructing the high band signal in the presence of modelling approximations or errors.

2. TIME-DOMAIN BANDWIDTH EXTENSION

This paper describes the time-domain bandwidth extension (TBE) framework employed to code WB and SWB speech signals in the newly standardized 3GPP EVS codec [15]-[17]. As in other commonly used speech coders, time-domain coding in EVS is based on the Linear Predictive Coding (LPC) paradigm [5]-[7] in which the speech signal is generated by sending an excitation signal through an all-pole synthesis filter. The all-pole synthesis filter models the spectral envelope and shapes the fine pitch structure of the excitation signal when generating the output speech signal.

In the TBE framework, the input speech signal is first split into low frequency (LF) and a high frequency (HF) sub-band signals. The LF signal is coded using the LP-based algebraic code excited linear prediction (ACELP) algorithm [5]-[7]. The high-band signal is coded using a separate LPC based model in which the high-band excitation signal is derived from the low-band excitation. To generate a HF excitation signal that preserves the harmonic structure of the LF excitation signal, a nonlinear function can be used [10], [11]. This nonlinear function is applied to the LF excitation after sufficient over-sampling in order to minimize aliasing. Fixed or adaptive whitening can be applied to flatten the spectrum and reverse unwanted effects of the nonlinear function. As the lower frequencies of voiced speech signals generally exhibit a stronger harmonic structure than the higher frequencies, the output of the nonlinear function can lead to a HF excitation signal that is too harmonic, leading to objectionable, `buzzy'-sounding artifacts [1], [10]. To remedy this, a combination of a nonlinear function and noise modulation is used to produce a pleasant sounding high-band signal.

2.1. TBE Encoder Framework

Figure 1 shows a high level framework of the SWB TBE encoder. The upper path of Fig. 1 shows the steps to generate a high band target signal using a low complexity and low delay Quadrature Mirror Filter-bank (QMF) [15], [17]. The generated high band target signal is then used to estimate the spectral and temporal evolution. The lower path of the Fig. 1 shows the estimation of high band excitation from the low band ACELP core using a nonlinear harmonic extension combined with envelope-modulated noise mixing.

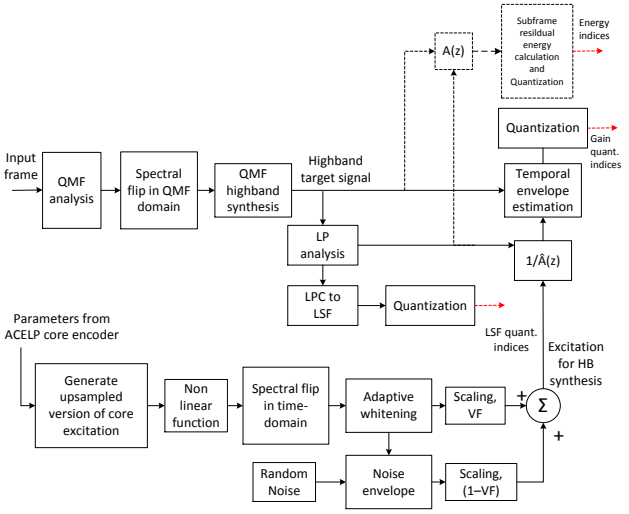

Figure 1. Super-wideband TBE encoder framework in EVS.

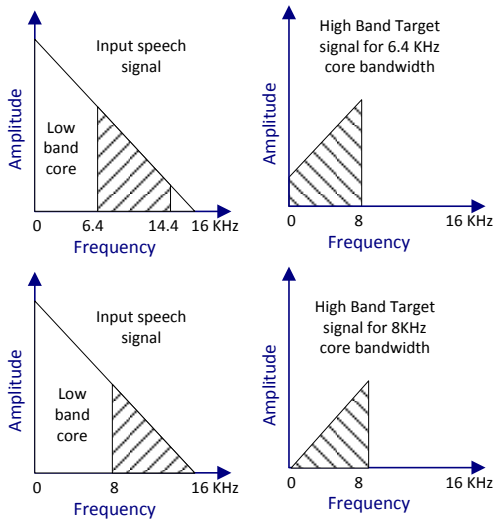

Figure 2. High band target signal generation in SWB-TBE for the ACELP core max bandwidths of 6.4 kHz (top) and 8 kHz (bottom).

2.1.1. *High-band target signal generation*

The process of deriving the SWB high band target signal is illustrated in Fig. 2. The input signal that is sampled at 32 kHz is segmented into frames of 20 ms and subsequently processed using a QMF analysis filter-bank to generate 40 subbands at a resolution of 800 Hz. Depending on the bandwidth up to which the low band ACELP core encodes, a spectral flip is performed in the QMF domain at two different cross-over frequencies [15]. In particular, as shown in Fig. 2, when the low band ACELP core is coded up to 6.4 kHz, the upper band from 6.4 to 14.4 kHz is flipped to generate the high band target signal. Similarly, when the low band ACELP core codes up to 8 kHz, the upper band from 8 to 16 kHz is flipped to generate the high band target signal. Table 1 elaborates on the low band core bandwidths coded by the ACELP for various bitrates.

2.1.2. *High-band LP analysis and quantization*

LP analysis is performed on a 33.75 ms high band signal (as shown in Fig. 3) that includes the current frame's 20 ms of the high band target signal along with 5 ms from the past frame and 8.75 ms of look-ahead. A Hanning window is applied to the high band LP analysis buffer followed by the estimation and pre-conditioning of autocorrelation coefficients [5]-[7]. A tenth-order LP analysis is then performed to estimate the LP coefficients [6]. The LP coefficients are quantized in the line spectral frequency (LSF) domain as follows. The first 5 LSFs are scalar-quantized using 4, 4, 3, 3, and 3 bits, respectively. The LSFs from 6 through 10 are re-estimated by mirroring/folding the first 5 LSFs to upper frequencies and aligning them to a pre-determined grid [15]. The grid parameters are quantized using 4 bits and transmitted along with the LSF indices as shown in Fig. 1. At certain bit rates (e.g., 9.6 kbps WB/SWB, 13.2 kbps WB) a single-stage vector quantizer is used to encode the LSFs with much fewer bits as shown in the bit allocation Table 1. LSF interpolation techniques [15] are used to improve the evolution of the LP synthesis filters from frame to frame. The quantized and interpolated LSFs are converted back to the LP domain to perform synthesis filtering, $1/\hat{A}(z)$.

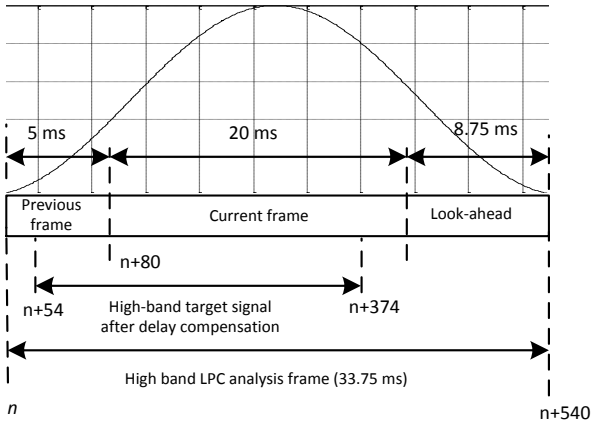

Figure 3. SWB-TBE LP analysis window and frame boundaries.

2.1.3. *Nonlinear harmonic extension of LB excitation*

The lower path of Fig. 1 shows the steps used to estimate the excitation for high band synthesis. As a first step, as shown in Fig. 4, the fixed codebook (FCB) contribution of the low band ACELP excitation is up-sampled by a factor of either $\beta = 5/2$ *or* 2 depending on whether the core sample rate is 12.8 or 16 kHz, respectively. The resulting up-sampled FCB is scaled by the FCB gain, g_c, and then mixed with a delayed (by $z^{-\beta T}$) contribution of the past up-sampled and scaled (g_p) LB-excitation, where T is the closed-loop pitch from the low band.

The up-sampled LB excitation is then processed using a nonlinear function, e.g., $\text{sign}(s(n))s^2(n)$, which extends the frequency harmonics from the low band to the upper band. The nonlinear excitation is then spectrally flipped in the time-domain (e.g., $s_{flipped}(n) = (-1)^n s(n)$) such that the high band portion of the excitation is modulated down to the low frequency region. The flipped excitation is then decimated by 2 to obtain the 16 kHz harmonic excitation.

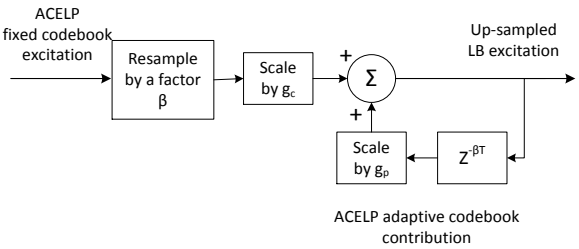

Figure 4: Generating the up-sampled low band excitation

TABLE I
BIT ALLOCATION FOR TIME-DOMAIN BANDWIDTH EXTENSION CODING

Bandwidth	WB		SWB		
Total bitrate (kbps)	9.6	13.2	9.6, 13.2†	13.2, 16.4	24.4, 32
ACELP core bandwidth (kHz)	6.4	6.4	6.4	6.4, 8.0	8.0
BWE bitrate (kbps)	0.3	1.0	0.9	1.55	2.75
Bit allocation across different parameters					
LSPs	2	8	8	21	21
Gain shapes, *GS*	0	5	5	5	5
Gain frame, *GF*	4	6	5	5	5
Subframe gains	-	-	-	-	15
Energy Threshold	-	-	-	-	6
Voicing	-	1	-	-	3

† 13.2 kbps primary frame SWB-TBE coding in channel aware mode [18].

2.1.4. *Envelope-modulated noise mixing*

Due to the nonlinear processing, the spectrum of the harmonic excitation may no longer be flat. An adaptive inverse filter (e.g., based on a fourth-order LP) is used in spectral whitening. The whitened harmonic excitation is further modified by adding a random noise whose amplitude is modulated according to the envelope of the whitened excitation (Fig. 1). The ratio at which the whitened excitation and the envelope-modulated noise are mixed is dependent on how strongly-voiced the speech segment is. In particular, given that the fine signal structure in the higher bands is closely related to that in the lower band, the mixing ratio may be estimated from the low band core ACELP parameters. For each subframe, i, the normalized correlation, α_i, from the low band is mapped to a voice factor parameter, VF_i

$$VF_i = 0.34 + 0.5\alpha_i + 0.16\alpha_i^2, \;\; i = 1,2, \ldots M \qquad (1)$$

where the number of subframes, $M = 4 \; or \; 5$, depending on whether the core sample rate is 12.8 or 16 kHz, respectively.

The voice factors undergo further smoothing to compensate for any sudden variations in the low band voicing within a frame [15]. Next the envelope-modulated noise is power normalized such that it is at the same level as that of the harmonic excitation. At each subframe, i, the harmonic excitation that is scaled by the factor, VF_i, and the normalized modulated noise that is scaled by the factor $(1 - VF_i)$ are mixed to generate the high band excitation as shown in Fig. 1. The high band excitation is then passed through the high band synthesis filter, $1/\hat{A}(z)$, to obtain the spectrally shaped excitation.

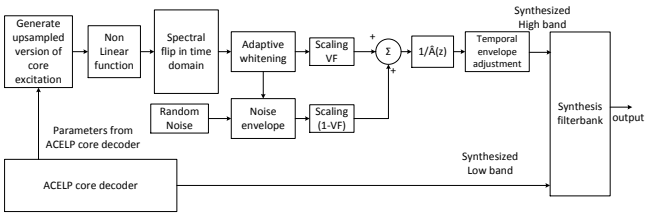

Figure 5. Super-wideband TBE decoder framework in EVS

The spectrally shaped excitation is further conditioned using a traditional post-processing filter [12] [13], e.g., $H(z) = \hat{A}(z/\gamma_1)/\hat{A}(z/\gamma_2)$, where $0 < \gamma_1 < \gamma_2 < 1$ and the factors $\gamma_1 \; and \; \gamma_2$ are adaptively estimated based on the high band spectral tilt [15]. The post-processing tends to improve the spectral matching of the high band excitation with that of the high band target signal.

2.1.5. *Gain shape and Gain frame estimation*

The high band target signal, $x(n)$, and the post-processed high band excitation, $e(n)$, are used to estimate the temporal gain shapes (over 5ms segments) and the overall gain per frame. The gain shape parameters are estimated using an overlap of 20 samples from the previous frame to avoid transition artifacts during the reconstruction at the decoder. The high band target signal is delayed by 27 samples to compensate for this overlap and the delay incurred during the high band excitation generation. This delay compensation also insures that the high band target signal, $x(n)$, and high band excitation, $e(n)$, are time-aligned for the gain shape, $GS(j)$, estimation,

$$GS(j) = \sqrt{\frac{\sum_{n=0}^{99} w(n)[x^2(n+j80-20)]}{\sum_{n=0}^{99} w(n)[e^2(n+j80-20)]}}, \;\; j = 0,1,2,3 \qquad (2)$$

where the window, $w(n)$, is a trapezoidal window,

$$w(n) = \begin{cases} n/20 & 0 \le n < 20 \\ 1 & 20 \le n < 79 \\ (1 - n/20) & 80 \le n < 99 \\ 0 & otherwise \end{cases} \qquad (3)$$

The four gain shapes $GS(j)$ are then normalized and subsequently smoothed for improved temporal evolution [15]. The smoothed gain shapes are vector quantized in the log domain using 5 bits.

In addition to the gain shape parameters, an overall gain parameter *GF* is calculated. First, the high band excitation, $e(n)$, is scaled using the quantized temporal gain shapes, $\widehat{GS}(j)$,

$$e_{gs}(n) = \sum_{j=0}^{3} w(n - j80)\widehat{GS}(j)e(n), \;\; n = 0, \ldots, 339 \qquad (4)$$

where the first 20 samples in the buffer $e(n)$ are from the previous frame. The overall frame gain GF is calculated as,

$$GF = \sqrt{\frac{\sum_{n=0}^{339} w\prime(n)[x^2(n)]}{\sum_{n=0}^{339} w\prime(n)[e_{gs}^2(n)]}} \qquad (5)$$

where the window $w'(n)$ is a trapezoidal window with tapering edges on the 20 sample overlap regions and $w'(n) = 1$ for $20 \le n < 319$. The gain frame parameter is quantized in the log domain using 5 bits for SWB and transmitted along with the gain shape parameters as shown in Fig. 1.

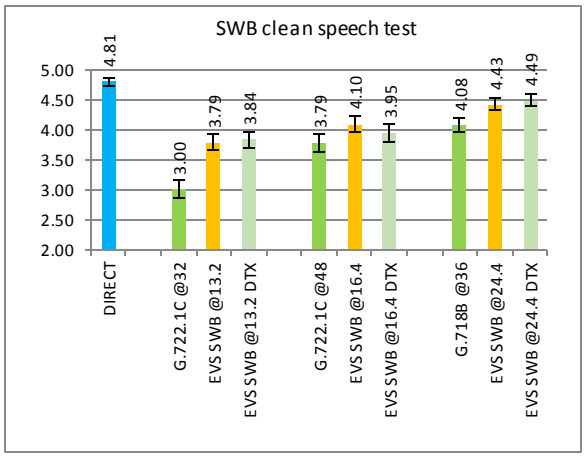

Figure. 6. EVS SWB clean speech ITU-T P.800 DCR MOS test results.

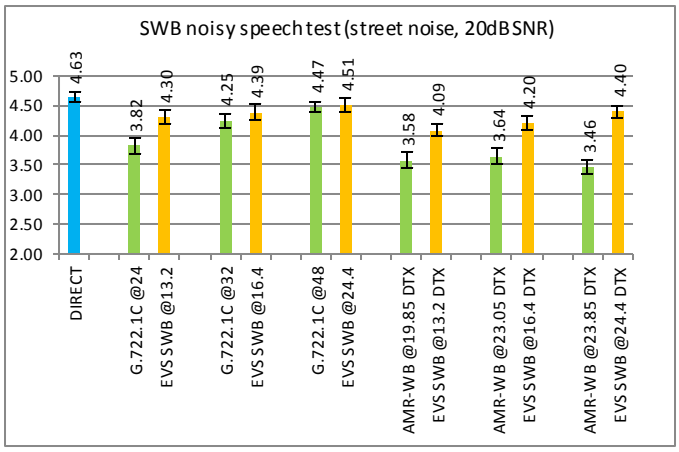

Figure 7. EVS SWB noisy speech ITU-T P.800 DCR MOS test results.

2.2. TBE Decoder Framework

Figure 5 shows a high level framework of the SWB-TBE decoder. The steps described in Section 2.1.3 and 2.1.4 are performed at the decoder to estimate the harmonic excitation and the envelope-modulated noise. The decoded LSPs are converted to the LP-domain to perform the high band spectral shaping and subsequently estimate the post-processed high band excitation signal, $e(n)$.

The post-processed excitation signal, $e(n)$, is then scaled using the decoded gain shapes as shown in Eq. (4). The scaled excitation is finally multiplied by the decoded gain frame to obtain the high band synthesized signal. For more details on various gain shape smoothing and gain frame attenuation techniques that are used to improve the temporal evolution of the synthesized high band, refer to [15].

The synthesized high band is up-sampled by 2 and flipped (reverse of Fig. 2, i.e., flip from low to high band) to generate a 32 kHz high band component associated with the final decoded speech. The low band is up-sampled to 32 kHz, and then delayed by 2.3125ms before mixing with the high band component to generate the SWB synthesis.

3. SUBJECTIVE QUALITY TESTS

The quality benefit achieved by the addition of the high band has been tested in an independent listening test laboratory through ITU-T P.800 Mean Opinion Score (MOS) tests using a Degradation Category Rating (DCR) methodology and 32 naïve listeners [21]. At each bit rate, EVS with TBE is compared against a standardized WB or SWB reference codec (or one of each). The references have been selected by 3GPP to serve as performance requirements during the development of the EVS codec. Similar subjective tests with these same reference conditions have been conducted by 3GPP during codec selection and characterization demonstrating comparable results [16], [17].

Figure 6 compares MOS scores for EVS SWB with TBE at 13.2, 16.4, and 24.4 kbps with the reference codecs for clean speech inputs. EVS conditions with and without discontinuous transmission (DTX) during non-active speech portions are shown. In all three cases, EVS with TBE, for both DTX on and off cases, shows a statistically significant quality improvement compared to the performance requirement condition, each of which has a significant bit rate advantage and does not use DTX, so the overall average bit rate in the EVS DTX case is quite substantial. Thus, these results also show the capacity benefits of EVS due to it substantially lower bit rate for the same, or better, quality than these other standard codec benchmarks.

MOS tests results using noisy speech inputs are shown in Figure 7. Specifically, street noise has been added to the speech input such that the combined input has a 20 dB signal-to-noise ratio. This leads to a considerably more challenging input for the codec, and does result in some unvoiced and inactive frames being encoded using the MDCT-based TCX core [17], [20] or frequency-domain bandwidth extension [17]. Despite the more challenging input, EVS SWB (without DTX) is statistically no worse than the reference codecs, which are again operating at higher bit rates. In the case of DTX, coding of noisy inputs is more susceptible to the introduction of artifacts. Therefore, the reference conditions for noisy inputs with DTX were chosen by 3GPP to be a codec which also supports DTX, namely the prior 3GPP standard AMR-WB. Here, due to a bandwidth difference and better coding quality, EVS SWB far outperforms AMR-WB, even at lower bit rates in each case.

4. CONCLUSIONS

This paper presents the highly efficient time-domain bandwidth extension technique which is an integral part of the EVS codec. In particular, the advanced modelling techniques used to recreate the WB and SWB frequencies with a fewer number of bits paved the way for EVS to become the most advanced feature rich conversational speech coder of its time. Extensive MOS testing has proven that the EVS codec with time-domain bandwidth extension outperforms all other standard codec references with significant margins, making it the ideal codec to be deployed in modern VoLTE networks and other VOIP networks such as VoWiFi.

5. REFERENCES

[1] D. J. Sinder, I. Varga, V. Krishnan, V. Rajendran and S. Villette, "Recent Speech Coding Technologies and Standards," in *Speech and Audio Processing for Coding, Enhancement and Recognition*, T. Ogunfunmi, R. Togneri, M. Narasimha, Eds., Springer, 2014.

[2] E Larson, R Aarts, *Audio Bandwidth Extension*, John Wiley & Sons, West Sussex, UK, 2005.

[3] B. Iser, G. Schmidt, W. Minker, *Bandwidth Extension of Speech Signals*, Springer, Lecture Notes in Electrical Engineering Series, 2008.

[4] B. Geiser, *et al.*, S, Bandwidth Extension for Hierarchical Speech and Audio Coding in ITU-T Rec. G.729.1 IEEE Transactions On Audio, Speech, And Language Processing, Vol. 15, No. 8, November 2007.

[5] A. Spanias, T. Painter, V. Atti, *Audio Signal Processing and Coding,* John Wiley and Songs, Feb 2007.

[6] W. C. Chu, *Speech Coding Algorithms*. New York:Wiley-Interscience, 2003.

[7] P. Kroon and W. B. Kleijn, "Linear prediction-based analysis-synthesis coding" in *Speech coding and synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., Elsevier, 1995.

[8] M. Dietz, *et al.*, "Spectral Band Replication, a novel approach in audio coding," *Audio Engineering Society Convention 112*. Audio Engineering Society, 2002.

[9] 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.

[10] J. Makhoul and M. Berouti, "High frequency regeneration in speech coding systems," *IEEE ICASSP*, 1979, pp. 428-431.

[11] V. Krishnan, V. Rajendran, A. Kandhadai, S. Manjunath., "EVRC-Wideband: The New 3GPP2 Wideband Vocoder Standard," *IEEE ICASSP*, Vol. 2, 2007.

[12] 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.

[13] 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.

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

[16] S. Bruhn, *et al.*, "Standardization of the new EVS Codec", submitted to *IEEE ICASSP*, Brisbane, Australia, April, 2015.

[17] M. Dietz, *et al.*, "Overview of the EVS codec architecture," submitted to *IEEE ICASSP*, Brisbane, Australia, April, 2015.

[18] 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.

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

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

[21] ITU-T P.800, Methods for Subjective Determination of Transmission Quality. International Telecommunication Union (ITU), Series P., August 1996.