HARMONIC MODEL FOR MDCT BASED AUDIO CODING WITH LPC ENVELOPE

Takehiro Moriya[1], Yutaka Kamamoto[1], Noboru Harada[1], Tom Bäckström[2,3], Christian Helmrich[2], and Guillaume Fuchs[3]

1 Nippon Telegraph and Telephone Corp. (NTT), Japan
2 International Audio Laboratories Erlangen, Friedrich-Alexander University (FAU), Germany
3 Fraunhofer Institute for Integrate Circuits (IIS), Germany

ABSTRACT

Conventional music coders, based on a modified discrete cosine transform (MDCT) suffer greatly when lowering their bit-rate and delay. In particular, tonal music signals are penalized by short analysis windows and the variable length coding of the quantized MDCT coefficients demands a significant amount of bits for coding the harmonic structure. For solving such an issue, the paper proposes a frequency-domain harmonic model aiming to amend the probability model of the variable length coding of the quantized MDCT coefficients. The new model was combined successfully with an envelope based arithmetic coding at rate lower than 10 kbps, and with a context based arithmetic coding at higher bit rates in the recent 3 GPP EVS (Enhanced Voice Services) codec standard. Objective and subjective quality tests indicate that the proposed harmonic model enhances the quality of music for low-delay audio coding.

Index Terms— MDCT, envelope, harmonic interval, arithmetic coding, EVS

1. INTRODUCTION

A low-bit-rate, low-delay speech and audio coding scheme is desired for future mobile communication systems such as VoLTE (Voice over Long Term Evolution). Efficient speech coding schemes have been developed, including ITU-T G.729 [1, 2] and 3GPP Adaptive Multi-Rate (AMR) [3], which are widely used in narrowband mobile and IP communications. AMR-WB [4] is a wideband speech coding scheme newly used in the VoLTE system. These schemes are all based on algebraic code-excited linear prediction (ACELP), which employs a source model optimized for speech, whereby it is inefficient for music.

In contrast, generic audio coding schemes have been developed such as MP3 [5], AAC (Advanced Audio Coding) [6] and HE (High Efficiency)-AAC [7], as well as integrated speech and audio coding schemes such as 3GPP Extended Adaptive Multi-Rate Wide-Band (AMR-WB+) [8] and MPEG-D Unified Speech and Audio Coding (USAC) [9-10]. However, these coding schemes cannot be applied for two-way communication because of their too long algorithmic delay (typically more than 100 ms) and high bit rates (typically more than 20 kbps).

The 3GPP EVS project has been launched in order to achieve low-bit-rate, low-delay, and super-wide-band speech and audio coding, and the specification was established in 2014 [11-13]. As a frequency domain core coder of EVS, a MDCT-based transform coded excitation (TCX) scheme was developed with a low-delay transform window, a LPC-based quantization noise shaping and frame-independent arithmetic coding of the quantized spectrum [14-16]. However, some quality degradation was encountered for tonal music signals due to inefficiency of the frame-independent arithmetic coding in a shorter overlap window. Therefore, we propose a harmonic model to enhance the efficiency of the arithmetic coding.

The aim of the harmonic model is to enhance the probability modeling in the arithmetic coding. Unlike Long Term Prediction (LTP) used in AAC [7] and OPUS [17], which exploits the irrelevance of tonal signals by shaping the quantization noise, the new method exploits the redundancy of the harmonic structure. Both concepts can be combined, which is actually the case in 3GPP EVS codec, where a LTP post-filter is also present.

2. FUNDAMENTAL STRUCTURE

3GPP EVS is a 32ms algorithmic delay multi-mode coder switching between a time domain ACELP and a frequency domain coder. ACELP is appropriate for speech, whereas frequency domain MDCT based TCX is appropriate for music. One of these will be activated as the core coder based on a decision made at the preprocessor. When MDCT based TCX is selected, the spectral coefficients are perceptually shaped by a smoothed frequency envelope derived from a LPC analysis [7, 9, 10, 14, 18], as illustrated in Fig. 1. LPC coefficients are efficiently quantized by line spectrum pair (LSP) parameters [19, 20]. The whitened MDCT coefficients are scalar quantized and further compressed by arithmetic coding.

For getting higher efficiency, two arithmetic coding schemes are adopted in MDCT based TCX of EVS. At high bit rates (over 10 kbps) a context based coding [15] exploits the information of the past decoded spectral lines for predicting the current lines to code. When the bit-rate becomes very low and the quantized spectrum is too sparse, the past context contains too few information for being efficiently exploited. In that case an envelope based

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

entropy coder [16] is used instead, assuming that the main quantized energy is located in the formants of the signal.

The proposed harmonic model is used to make the both types of arithmetic coding more efficient for tonal signals. Note that the perceptual noise shaping is carried out by a smoothed LPC envelope and does not take into account any harmonic model. On the other hand, the coding error is shaped by a harmonic post-filter at the end of the decoding process. The post-filter is controlled by parameters derived from a LTP analysis of the input signal.

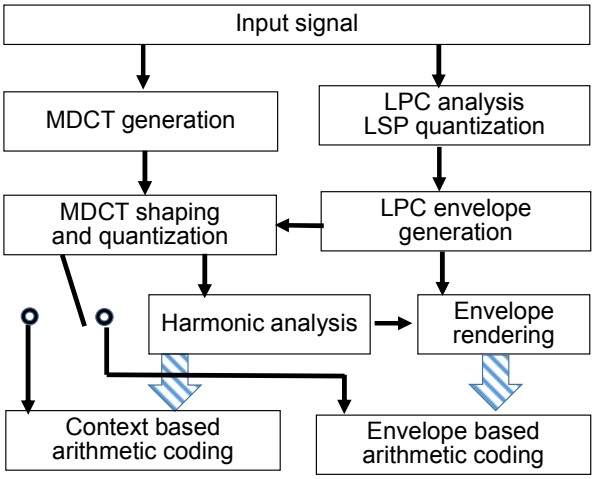

Fig. 1. Harmonic models (shaded arrows) either for context based or envelope based arithmetic coding.

3. HARMONIC MODEL

3.1. Activation in MDCT based TCX of EVS

Abrupt changes in the amplitude of MDCT coefficients consume quite a few bits with both context and envelope based arithmetic coders. To convey the position of abrupt peaks to the arithmetic coders, we need to capture the periodic peaks in the MDCT coefficients; these peaks are caused by a dominant periodic structure of the time domain signal. The frequency domain interval of harmonics is a key parameter when the harmonic model is enabled, and it is encoded for both types of arithmetic coders. The activation of the harmonic model is finally decided in a semi-closed loop, where the estimated bit consumption of arithmetic coding is checked, since there is a trade-off between saved bits for arithmetic coding and consumed additional bits (2-8 bits) for representing the periodicity.

3.2. Search for harmonic interval

In searching for the optimum harmonic interval, the encoder tries to find the index that maximizes $E_P()$, the weighted sum of the peak part of absolute MDCT coefficients $X_M(k)$. Let $E_{ABS}(k)$ be the sum of three neighboring samples of absolute values of MDCT coefficients as

$$E_{ABS}(k) = \sum_{j=0}^{2} \left| X_M(k+j-1) \right| , \qquad (1)$$

whereby

$$E_P(T_{MDCT}) = \left(\frac{1}{np}\right) \sum_{n=1}^{np} E_{ABS}\left(\lfloor n \cdot T_{MDCT} \rfloor\right) \left(\frac{3n-2}{255}\right)^{0.3} , \qquad (2)$$

where np is the maximum number n such that $\lfloor n \cdot T_{MDCT} \rfloor$ reaches the limit of samples in the frequency domain.

3.3. Encoding of harmonic interval

When the LTP post filter is activated, a pitch lag and the gain are encoded by the LTP analysis. The encoded lag values can be partly utilized for coding the interval of harmonics in the frequency domain. The process of searching harmonic intervals is also simplified because the integer multiple is predicted from the pitch lag. If an LTP is not used or if the pitch gain of the LTP is too small, a normal representation of the harmonic interval is applied as shown in Fig.2.

3.3.1. Encoding interval depending on pitch lag

If the integer part of the pitch lag in time domain d_{int}, which is used for the LTP, is less than the frame size of MDCT, L_{TCX}, the frequency domain interval unit (between harmonic peaks corresponding to the pitch lag) T_{UNIT} with 7-bit fractional accuracy is given by

$$T_{UNIT} = \frac{(2 \cdot L_{TCX} \cdot res\_max) \cdot 2^7}{(d_{\text{int}} \cdot res\_max + d_{fr})} , \qquad (3)$$

where d_{fr} denotes the fractional part of the pitch lag in the time domain and res_max denotes the maximum number of allowable fractional values whose values are either 4 or 6 depending on the conditions.

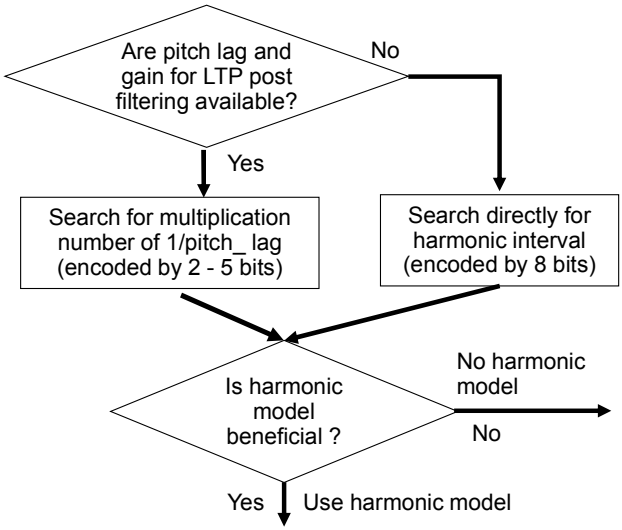

Fig. 2. Encoding parameters for harmonic model.

Since T_{UNIT} has a limited range, the actual interval between harmonic peaks in the frequency domain is coded by the index for the multiplication numbers relatively to T_{UNIT} using 2 to 5 bits depending on the value of T_{UNIT}. Among a preselected set of potential multiplication fac-

tors, the factors that give the most suitable harmonic interval of MDCT coefficients are selected. Using the pitch lag for LTP post-filtering enables the harmonic model to save several bits for the harmonic interval and to reduce the computational cost of the search.

3.3.2. Encoding interval independent from pitch lag

When an LTP is not used or the pitch gain is less than 0.46, we cannot use or rely on the time domain pitch lag. In these cases, normal encoding of the interval with unequal resolution is used. The unit interval of spectral peaks T_{UNIT} is coded as

$$T_{UNIT} = index + base \cdot 2^{Res} - bias \qquad (4)$$

and the actual interval T_{MDCT} is represented with a fractional resolution of Res as

$$T_{MDCT} = T_{UNIT} / 2^{Res} \quad . \qquad (5)$$

Each parameter is listed in Table 1, where "small size" refers to the case when a frame size is smaller than 256 or when the target bit budget is less than or equal to 150.

Table 1. Unequal resolution for coding harmonic intervals

	Res	*base*	*bias*
$index < 16$	3	6	0
$16 \le index < 80$	4	8	16
$80 \le index < 208$	3	12	80
"small size" or $208 \le index < 224$	1	28	208
$224 \le index < 256$	0	188	224

If the interval does not rely on the pitch lag of LTP, a hierarchical search is used to reduce the computational cost. If the index of the interval is less than 80, periodicity is checked by a coarse step of 4. After obtaining the optimum interval, a finer periodicity is searched around the optimum interval from -2 to +2.

4. CONTRIBUTIONS TO ARITHMETIC CODING OF QUANTIZED MDCT COEFFICIENTS

MDCT samples at harmonic peak regions are assumed to show larger amplitudes than samples in other regions. This remarkable characteristic can be exploited by the entropy coders for reducing the bit demand.

4.1. Context based variable length coding

By exploiting the past history of the decoded quantized values, the context based variable length coding can model more precisely the current statistics of the current samples to code. In the conventional context based entropy coder, the context points uniquely to the direct past neighborhood.

The harmonic model allows the context to adapt itself to the harmonic structure and to point to the previous harmonic peak in case the current sample to code belongs to one of the three samples considered to form a harmonic peak. In that way, the large amplitudes of the tones are better predicted. For regions other than harmonic peaks, the conventional context is kept. The principle is depicted in Fig. 3.

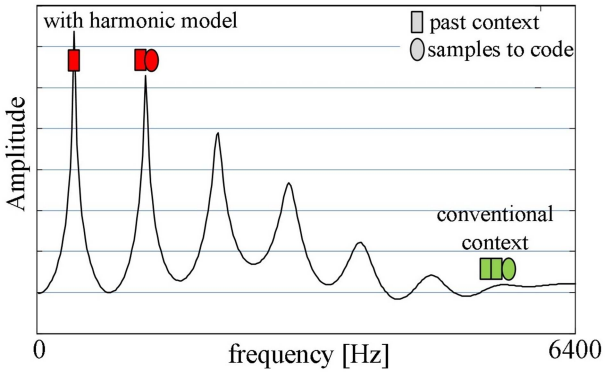

Fig. 3. Example of harmonic model exploited in a past context of the entropy coding.

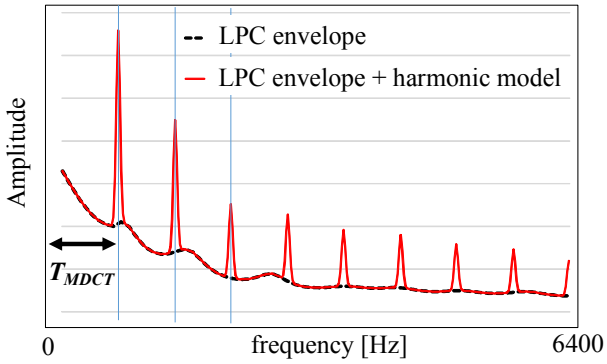

Fig. 4. Example of harmonic envelope combined with LPC envelope used in envelope based arithmetic coding.

4.2 Envelope based variable length coding

With envelope based arithmetic coding, the shape of the LPC envelope is used for arithmetic coding. The harmonic model is used in combination with the LPC envelope as shown in Fig. 4. The shape of the envelope is rendered according to the harmonic interval. We used a Gaussian envelope shape at the k-th sample as

$$Q(k) = h \cdot \left(-\frac{(k-\tau)^2}{2\sigma^2} \right), \quad (\tau - 4 \le k \le \tau + 4) \qquad (6)$$

and $Q(k) = 1.0$, when $(k < \tau - 4$ or $\tau + 4 < k)$, centered at τ, which is the U^{th} harmonic peak,

$$\tau = \lfloor U \cdot T_{MDCT} \rfloor . \qquad (7)$$

Variables, h and σ are the height and the width of each harmonics depending on the unit interval and defined as:

$$h = 2.8(1.125 - exp(-0.07 \cdot T_{MDCT})) \qquad (8)$$

$$\sigma = 0.5(2.6 - exp(-0.05 \cdot T_{MDCT})) \qquad (9)$$

These are experimentally decided according to the observation that larger height and width are preferable when harmonic interval T_{MDCT} gets larger. The original LPC

based spectral envelope $S(k)$ is modified by the harmonic shape $Q(k)$ at k as

$$S(k) = S(k) \cdot (1 + g_{harm} \cdot Q(k)) \,, \qquad (10)$$

where gain for the harmonic components g_{harm} is always set as 0.75 for the generic mode, and it is selected from {0.6, 1.4, 4.5, 10.0} with 2 bits for the voiced mode. The selection criteria is to minimize the 4th power of coefficients $X_M(k)$ normalized by a combined envelope $S(k)$ as

$$E_{norm} = \sum_{k=0}^{L_M - 1} (|X_M(k)| / S(k) / E_{ABSres})^4 \,, \qquad (11)$$

where L_{TCX} denotes frame length, and

$$E_{ABSres} = \sum_{k=0}^{L_{TCX} - 1} (|X_M(k)| / S(k)) \,. \qquad (12)$$

5. EVALUATION

5.1. Evaluated system

The proposed harmonic model can be used for both types of arithmetic coding. However, the present section focuses only on evaluating the combination with the envelope based coder for coding wideband signals at 9.6 kbps. For other conditions, extensive listening test results of the EVS codec can be found in the report on performance characterization [21].

5.2. Objective evaluation

The proposed harmonic model was tested within the EVS codec standard by comparing it to the baseline version of the codec, i.e. by deactivating the harmonic model tool. We compared the distortion of the harmonic model on top of the baseline by Perceptual Objective Listening Quality Assessment (POLQA) [22] for WB signals coded at 9.6 kbps. Four categories of input signals were under test: artificially mixed speech and music, a natural mix of speech and music, classical and modern music. Each category includes seven items of 8 s.

The absolute comparison of averaged scores in each category with 95 % confidence interval shown in Fig. 5 (left) indicates that there is a consistent tendency that the codec with the harmonic model enabled is scored better than without. But none of the differences are statistically significant due to large confidence intervals. In contrast, we can see statistically significant differences (95% double-sided student's t-test) in favor of the harmonic model tool if the scores are compared item by item as shown in Fig. 5 (right).

5.3. Subjective evaluation

A listening test with seven experienced listeners was carried out to compare the subjective quality. The Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) method [23] was used. Figure 6 shows comparisons of the EVS codec with and without harmonic model, both at 9.6 kbps, and the AMR-WB at 15.85 kbps. Input signals were selected from the mixture of speech and music, and from music sampled at 16 kHz. The test indicates a consistent tendency for TCX to perform better than AMR-WB for music signals, while AMR-WB performs better than TCX for mixed items. In addition, there is a consistent tendency for the harmonic model version to perform better compared to the baseline. None of these tendencies, however, are statistically significant.

In Fig. 7, the same MUSHRA scores are used, but the item-by-item difference scores with and without the harmonic model are averaged over all 7 listeners. This time, we can see statistically significant advantages (95% double-sided student's t-test) of the harmonic model for some music signals (two classics and one modern items) and on average. We observe that the proposed harmonic model doesn't show any merit for the 5th item (classic music) from the left. We suspect that this is due to its very complex harmonic structure.

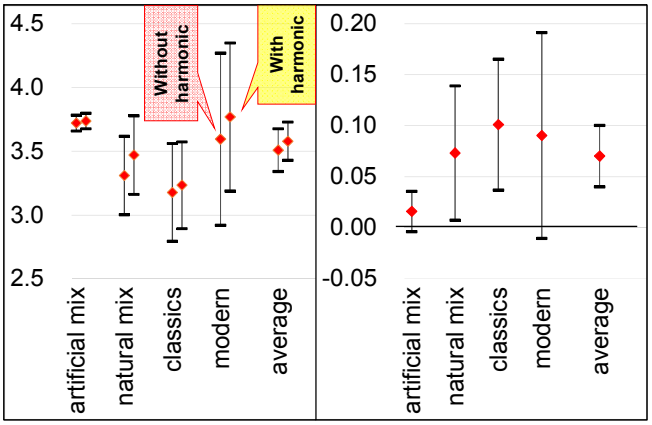

Fig. 5. Absolute POLQA scores of MDCT based TCX (left panel), and item by item difference scores (right panel, "with" minus "without" harmonic model)

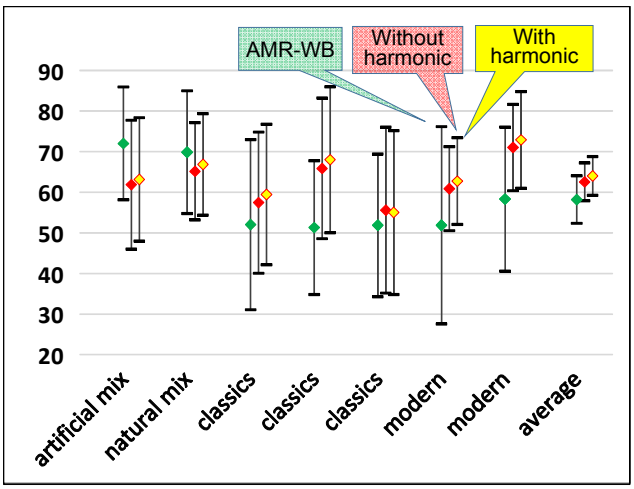

Fig. 6. Absolute MUSHRA scores comparing AMR-WB (15.85 kbps) and MDCT based TCX (9.6 kbps) "with" and "without" harmonic model. Number of listeners is 7.

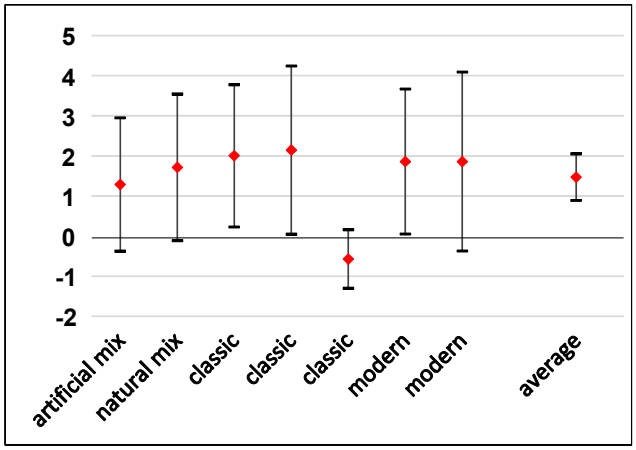

Fig. 7. MUSHRA difference scores ("with" minus "without" harmonic model) for MDCT based TCX.

6. CONCLUSION

A harmonic model has been devised for MDCT based TCX audio coding in order to achieve low-delay, low-bitrate speech and audio coding for VoLTE. This model effectively reduced the distortion by reducing the bit consumption of arithmetic coding even at the cost of encoding the harmonic interval as side information. This model can be used for both types (context based and envelope based) of arithmetic coding in MDCT based TCX of the EVS codec. Objective and subjective quality comparisons at 9.6 kbps indicated that the proposed harmonic model is effective in enhancing music quality. The same degree of improvements was also observed for higher rates with the context-based arithmetic coder. As a consequence, the proposed scheme has been adopted in the MDCT based TCX system of 3GPP EVS codec standard at bit rates from 9.6 to 128 kbps and for various bandwidths (narrow, wide, super-wide, and full-band).

7. ACKNOWLEDGMENTS

The authors are grateful to all those who contributed to the 3GPP EVS, in particular, Csaba Kos, who has made excellent software implementation.

REFERENCES

[1] ITU-T G.729, "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)," 2012.

[2] R. Salami, C. Laflamme, J.P. Adoul, A. Kataoka, S. Hayashi, T. Moriya, C. Lamblin, D, Massaloux, S. Proust, P. Kroon, and Y. Shoham, "Design and Description of CS-ACELP: A Toll Quality 8kb/s Speech Coder," IEEE Trans. Speech and Audio Processing, Vol. 6, No. 2, pp. 116 - 130, 1998.

[3] 3GPP TS 26.071 Mandatory speech CODEC speech processing functions; AMR speech CODEC; General description version 11.0.0 Release 11, 2012.

[4] 3GPP TS 26.190 Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; Transcoding functions, version 11.0.0 Release 11, 2012.

[5] ISO/IEC 11723-3, Information technology – "Generic coding of moving pictures and associated audio for digital storage media up to 1.5 Mbit/s", Part 3 Audio, 1993.

[6] ISO/IEC 13818-7, Information technology – "Generic coding of moving pictures and associated audio", Part 7 Advanced Audio Coding, 1997.

[7] ISO/IEC 14496-3, Information technology – "Generic coding of Audio and Visual Objects", Part 3 Audio.2009.

[8] 3GPP TS 26.290 Audio codec processing functions; Extended Adaptive Multi-Rate - Wideband (AMR-WB+) codec; Transcoding functions version 11.0.0 Release 11, 2012.

[9] S. Quackenbush, "MPEG Unified Speech and Audio Coding," MultiMedia, IEEE Computer Society, vol. 20, issue 2, pp. 72-78, 2013.

[10] M. Neuendorfet al,"MPEG Unified Speech and Audio Coding - The ISO/MPEG standard for high efficiency audio coding of all content types", in Proc. AES 132nd Convention Paper, #8654, Apr., 2012.

[11] 3GPP TS 26.441: Codec for Enhanced Voice Services (EVS); General Overview, version 12.1.0 Release 12, 2014.

[12] 3GPP TS 26.445: Codec for Enhanced Voice Services (EVS); Detailed algorithmic description, version 12.1.0 Release 12, 2014.

[13] M. Dietz, et al. "Overview of the EVS codec architecture," Proc. ICASSP 2015.

[14] G. Fuchs, M. Multrus, M. Neuendorf, and R. Geiger, "MDCT-Based Coder for Highly Adaptive Speech and Audio Coding," in Proc. EUSIPCO, pp. 1264-1268, 2009.

[15] G. Fuchs, C. Helmrich, G. Markovi´c, M.Neusinger,E. Ravelli1, and T. Moriya, "Low Delay LPC and MDCT-Based Audio Coding in EVS," proc. ICASSP 2015.

[16] T. B¨ackstr¨om and C. Helmrich, "Arithmetic Coding of Speech and Audio Spectra Using TCX Based on Linear Predictive Spectral Envelopes," ICASSP 2015.

[17] IETF RFC 6716. Definition of OPUS Audio Codec

[18] N. Iwakami, T. Moriya and S. Miki, "High quality Audi Coding at less than 64 kbit/s Using TwinVQ," Proc. ICASSP'95, pp. 937-940, 1995.

[19] F. Itakura, "Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals," J. Acoust. Soc. Am., 57, 533(A), 1975.

[20] F. Itakura, T. Kobayashi and M. Honda, "A Hardware implementation of a new narrow and medium band speech coding,", Proc. ICASSP 82, pp. 1964 – 1967, 1982.

[21] 3GPP TR 26.952: Codec for Enhanced Voice Services (EVS); Performance characterization, version 12.1.0 Release 12, 2014.

[22] ITU-T P.863. "Perceptual objective listening quality assessment."

[23] ITU-R BS.1534, "The Multiple Stimuli with Hidden Reference and Anchor (MUSHRA).