LOW BIT RATE HIGH-QUALITY MDCT AUDIO CODING OF THE 3GPP EVS STANDARD

Srikanth Nagisetty[1], Zongxian Liu[1], Takuya Kawashima[2],Hiroyuki Ehara[3], Xuan Zhou[4],Bin Wang[4],Zexin Liu[4],Lei Miao[4], Jon Gibbs[4], Lasse Laaksonen[5], Venkatraman Atti[6], Vivek Rajendran[6], Venkatesh Krishnan[6], Hosang Sung[7], Kihyun Choo[7]

[1]Panasonic R&D Center Singapore, [2]Panasonic System Networks R&D Lab, [3]Panasonic Corp.
[4]Huawei Technologies Co., Ltd, [5]Nokia Technologies, [6]Qualcomm Technologies, Inc., [7]Samsung Electronics Co., Ltd.

ABSTRACT

This paper presents a low bit-rate MDCT coder, which is adopted as a part of the recently standardized codec for Enhanced Voice Services. To maximize codec performance for NB to SWB input signals for low bit-rates (7.2 to 16.4 kbps), new adaptive bit-allocation and spectrum quantization schemes, which emphasize perceptually important spectrum while efficiently coding full spectrum, was introduced into the low bit-rate MDCT coder. Further, small symbol switched Huffman coding is exploited for reducing the bits consumption for quantizing band energies of the spectrum. Finally, the performance of the coder is illustrated with some listening test results.

Index Terms—*Speech coding, Audio coding, Adaptive bit allocation, spectrum gap filling, MDCT, low bit-rate*

1. INTRODUCTION

In communications, network resources are limited and speech and audio codecs are adapted to compress signals at low bitrates with acceptable quality. Accordingly, there is a continuing need to increase the compression efficiency over time when encoding speech and audio signals. To meet the demand, a new Enhanced Voice Services (EVS) codec [1] has been developed and standardized by 3GPP. Unlike previous 3GPP speech coding standards such as AMR [2], AMR-WB [3] and AMR-WB+ [4], EVS codec is capable of coding both speech and audio signals efficiently at a low algorithmic delay of 32 ms with significant better quality for both audio and speech. This is achieved at all the operating bitrates [5] and with high levels of network robustness. It consists of numerous coding schemes; each of which is tailored to a specific class of input signals over different bitrates. The low bit rate high quality (LR-HQ) MDCT coding mode is one such mode suitable for coding the full audio spectrum. This is applicable to narrowband (NB), wideband (WB) and super-wideband (SWB) and achieves high quality for bitrates as shown in Table 1.

The focus of the paper is to present the improvements brought by the LR-HQ MDCT coder in the transform core, alongside with parametric methods such as noise filling and/ or gap filling which is used for filling the un-coded (zero) spectral regions due to limitations of bit constraint. Details of which will be discussed later. In Section 2, the technical approach of LR-HQ MDCT coder is presented. Section 3 describes the spectral quantization technique used in LR-HQ MDCT coder. To prove the effectiveness of the technical advancements in LR-HQ MDCT coder, in section 4 we present listening results comparing the approach with various legacy codecs [2]-[4]. Section 5 concludes the successful adoption of LR-HQ in the EVS codec.

2. LR-HQ MDCT CODER

Figure 1 shows the overview of LR-HQ encoder and is designed to encode predominantly mixed and music signals. LR-HQ encoder operates at a fixed frame length of 20ms to satisfy the requirements of low delay and high quality at low bitrates. Each frame of input signal is encoded using one of the supported modes as shown in Table 1. Coding modes are decided based on input signal bandwidth and signal characteristics. The Transient detector in Figure 1 is the same as the one used in ITU-T G.719 [6], and non-transient signals are further classified into Harmonic (tonal) and Normal (non-tonal) for SWB input signals. For each selected mode of operation, MDCT coefficients (MDCTs) from one

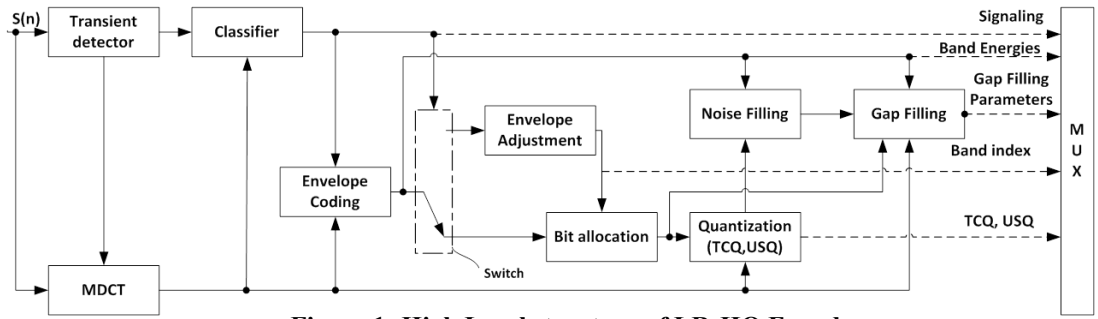

Figure 1: High Level structure of LR-HQ Encoder

frame of the input signal is split into limited number of bands with predefined width. Energy of each band is calculated and quantized using either a small symbol switched Huffman or large symbol Huffman coding method. This is followed by efficient bit allocation across bands based on coded band energies in order to achieve overall optimization for fine spectrum quantization using Trellis Coded Quantization (TCQ) and Uniform Scalar Quantization (USQ).

When coding the full spectrum at very low bit-rates, it is often not possible to allocate bits to all of the bands due to bits constraint. This would normally lead to audible artifacts. To alleviate this effect, un-coded bands (gaps) in the spectrum are filled with noise generated using quantized spectrum of bit-allocated bands. The approaches will be described in the next sections.

Table 1: LR-HQ core supported modes and bit-rates

Bit-rate [kbps]	Bandwidth	Supported modes
7.2, 8	NB	Non-Transient, Transient
13.2, 16.4	NB, WB	Non-Transient, Transient
13.2	SWB	Normal, Harmonic, Transient

2.1. Energy Envelope Coding

The band energies are scalar quantized, and then the indices are differentially encoded by adjusting the indices to constrain them to a specific range. The adjusted indices are then encoded in a more efficient manner by selecting either the Small symbol switched Huffman or the Large symbol Huffman coding method depending on a spanning range of indices and bit consumptions of coding methods. These techniques are described in the following sub-sections.

2.1.1. Small symbol switched Huffman coding

In the Small symbol switched Huffman coding method, differential indices are encoded by either a context based Huffman coding method or a re-sized Huffman coding method with the selection being based upon the number of bits consumed.

In the context based Huffman coding, a differential index in the previous band determines one out of two Huffman tables derived in a particular way for encoding the current band differential index.

In the resized Huffman coding method, a span of differential indices is narrowed down to a smaller range so that a Huffman table with fewer code words can be designed and used for encoding. The span of the current band indices is modified when the preceding band is out of predefined thresholds T and T' as demonstrated below, while able to perfectly reconstruct the original differential indices after quantizing the modified indices.

$$\Delta I'_M(b) = \begin{cases} \Delta I_M(b) + \min(\Delta I_M(b-1) - T, 3), \Delta I_M(b-1) > T \\ \Delta I_M(b) + \max(\Delta I_M(b-1) - T', -3), \Delta I_M(b-1) < T' \end{cases} \quad (1)$$

Where $b \in N$ b denotes the index of the band, $\Delta I_M(b)$ is the differential index for band b, $\Delta I_M(b-1)$ is the differential index for band b-1, $\Delta I'_M(b)$ is the new differential index for band b. The new differential indices are then quantized using the Huffman table derived in a particular way.

2.1.2. Large symbol coding

The Large symbol coding method uses eight symbol Huffman table [-4, 3] designed to deal the special cases where small symbol coding method fails to efficiently encode the differential indices when exceeding [-64, 63] range. It consists of pulse and scale modes, depending on the distribution of current differential indices either pulse or scale mode is selected for encoding the differential indices. When only one index exceeds [-4, 3] range and the others fit [-4,3] range, the pulse mode can encode them efficiently by indicating a position of the exceeding value and encoding it separately. In scale mode, to extend the capacity for encoding the differential indices beyond [-64, 63] range, differential indices are split into upper and lower bits. Upper bits are encoded using Huffman table while the lower bits are quantized directly.

2.2. Bit Allocation

Bits are allocated to the bands based upon the quantized band energies. In order to allocate bits to bands efficiently, two approaches are used. The first approach exploits the band energy relationships. The second approach adaptively groups the bands and exploits the relationship between the groups. The second approach is more suitable for tonal (Harmonic) like signals as the energy of the bands is mainly concentrated at discrete tones, while the first approach is more optimal for the other signals. Each approach is described in the following sub sections.

2.2.1. Energy envelope adjustments for NB and WB signals

For the Non-Transient mode, energy envelopes are adjusted prior to bit allocation for bit-rates less than 13.2 kbps, to address the tonal discontinuity between frames for perceptually important bands. In order to maintain tonal continuity between frames, bands with the most perceptual importance are identified based on the previous frame energy envelope, and the corresponding energy levels are adjusted to give high preference while allocating bits.

2.2.2. Adaptive bit allocation

The number of bits used to code the spectral components in bands is determined based on a dynamic bit allocation algorithm [1]. That is, if the number of bits, $m(b)$, used to encode the spectral components in a given band, b, is below a particular threshold, i.e., $m(b) < M_b$, then $m(b)$ is set to 0. Using a control logic, these unused bits in bands where $m(b) < M_b$ are redistributed among other bands. Sub-optimal coding is avoided using insufficient bits, but noise filling and gap filling techniques as described in Section 3.1 and 3.2 are used to encode these un-quantized bands. The noise

filling doesn't need any additional bits, while 1 or 2 bits are reserved for performing gap filling.

2.2.3. Bit allocation based on adaptive grouping

In this approach, bits are allocated in three stages: Firstly, dominant frequency bands (i.e. the frequency bands with the largest energies and those represent local maxima in the energy spectrum) are identified. The dominant frequency bands, including those regions adjacent to the energy peaks where the energy is decreasing, are grouped into a single group, referred to as the dominant group. The frequency bands between the dominant bands are grouped together and are known as the non-dominant group. Secondly, perceptual importance is determined according to both the energy and energy variance in the group. Based on the perceptual importance, bits are allocated to groups by emphasizing the groups with the largest energies and the largest energy variance. Finally, bits are allocated to the bands within each group in a way where more bits are allocated to frequency bands having larger energy.

3. TCQ AND USQ BASED SPECTRUM QUANTIZATION

Information on the spectral components of each frequency band is coded as the position, number, sign, and magnitude of the components. The magnitude information is quantized by joint USQ and TCQ with arithmetic coding, while the position, number, and sign are coded by arithmetic coding.

To support the Constant Bit-rate coding (CBR) required for the resource-allocation mechanism of 3GPP networks, the normalized signal is scaled by considering average bit-allocation for each spectral component in the band. If the average bit-allocation is larger than specified by the bit-allocation, then the signal is scaled more. With the scaled signal, a new buffer is constructed by the selected non-zero coefficients called Important Spectral Components (ISCs), as illustrated in Figure 2.

The joint USQ and TCQ has two coding methods: (1) TCQ and USQ with second bit allocation for NB and WB, and (2) Least Significant Bit (LSB) TCQ for USQ for SWB. In the TCQ quantization, a trellis is used for 8 states, 4-coset with 2 zero levels [9]. TCQ and USQ with second bit allocation quantize the ISCs for each band by using the selected encoding method, which is decided by the bit allocation depicted in Figure 2. LSB TCQ for USQ has the advantage of using both quantizer types, USQ and TCQ, in one scheme. LSB coding of quantized data is illustrated in Figure 3 where the sequence of LSBs can then be quantized by TCQ.

After the quantization, there will be several unused bits and two un-quantized bands previously allocated fewer bits, are identified to consume these unused bits. If any of the highest few bands are particularly harmonic or when one of the highest two bands has been quantized in the previous frame, the two bands are selected from these few highest

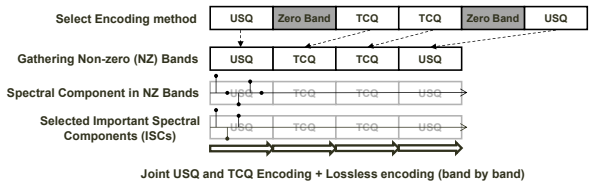

Figure 2: Concept of ISC selection and encoding

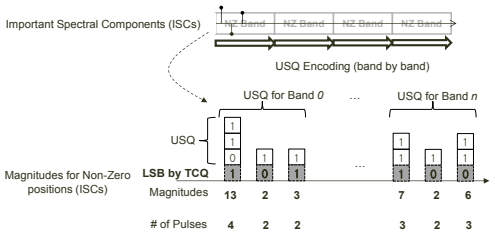

Figure 3: Concept of LSB TCQ for USQ coding

bands otherwise, the two bands to receive the unused bits will be selected from the lower bands.

3.1. Noise filling for un-saturated bands

The general noise filling method is only used to reconstruct bands where no bits have been assigned. However, it is also perceptually important to reconstruct zero-valued MDCTs when not all coefficients in a band are quantized. The quantized spectral bands is usually sparse due to bit restriction, to address the sparseness, noise is filled to the zero valued MDCTs for quantized bands. In this section, the concept of saturated and un-saturated bands is introduced to identify and fill zero bit MDCTs with comfort noise.

Saturated and un-saturated bands are classified according to the average number of bits allocated to the coefficients in a band. If the average number of bits allocated to a coefficient is 0.8 or greater, then the band is defined as saturated; otherwise, the band is defined as un-saturated. Noise is filled to the un-statured bands by adjusting the noise levels using the calculated noise gain.

3.2. Gap filling in high-frequency bands for SWB

Other than the noise filling described in the previous section, gap filling techniques for SWB are introduced when zero bits are allocated to high-frequency (HF) bands. Zero-bit-assigned bands cause spectral gaps, which lead to audible artifacts if left alone. To alleviate the effect, such bands in the HF region are identified and filled with spectra generated using coded spectra. The bandwidth extension (BWE) coding scheme of ITU-T G.718 Annex B [7], where the best match between the low-frequency (LF) band and the HF band is searched, is used for the gap filling methods. The methods have been refined for operation at low bit-rates, and several new features have been introduced. Following sub-sections will discuss the new features.

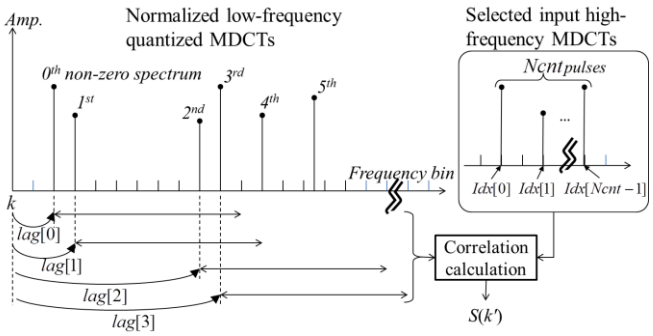

Figure 4: Sparse band search in Normal mode

3.2.1. Sparse band search in Normal mode

Once the normalized spectrum is obtained, The best match is identified between normalized spectrum and the input HF spectrum by finding the k' which maximizes the correlation measure, $S(k')$ according to.

$$S(k') = corr(k')^2 / Ene(k'), \quad k' = 0,..., Nlag-1 \qquad (2)$$

To reduce the computational load for $S(k')$, only the MDCTs selected with a dynamic thresholding process are used for similarity measure.

$$corr(k') = \sum_{k=0}^{Ncnt-1} X(Idx[k])\tilde{X}(k + lag[k'] + Idx[k]) \qquad (3)$$

$$Ene(k') = \sum_{k=0}^{Ncnt-1} \tilde{X}(k + lag[k'] + Idx[k])^2 \qquad (4)$$

where $Nlag$ is the number of lag candidates for a band, $X(i)$ is input HF MDCTs, $\tilde{X}(i)$ is normalized LF quantized MDCTs, and $Ncnt$ is the number of selected input high-frequency MDCTs. $Idx[k]$, $lag[k']$, and k are shown in Figure 4. This representation of lag candidates gives somewhat better coding performance when $Nlag$ is just two or four, i.e. used bits are one or two.

3.2.2. Gap filling for Harmonic signals using transposition

Similar to G.718 Annex B [7], best match is identified and based on best match information missing gaps in the spectrum are filled, however sometimes gap filling produces undesirable auditory artifacts such as perceivable roughness for tonal music items. The roughness is perceived when two tones lie in the range from 30 to 600 Hz and when the amplitudes of the two tones are rapidly changing. The intensity of the perceived roughness also depends on the spectral position of the tones [8]. This has been attributed to the situation when the missing portion in the coded spectrum is patched using the gap filling, especially when tonal peaks in reconstructed spectrum are in spectral vicinity to each other. To suppress the perceived roughness, the average spectral spacing is used to transpose the spectral positions of tonal components in the reconstructed spectrum. The average spectral spacing is the mean of the intervals between reconstructed spectral peaks.

4. PERFORMANCE EVALUATION

Performance evaluation for the LR-HQ mode has been conducted against a number of reference codecs [2]-[4]. The reference codecs have been selected by 3GPP to serve as performance requirements for the EVS codec [12]. The subjective evaluations for NB, WB, and SWB are performed under clean channel conditions for mixed and music content at the bit-rates given in Table 1. Figure 5 presents the NB test results using a P.800 Mean Opinion Score (MOS) test with an Absolute Category Rating (ACR) methodology [10], while Figures 6 and 7 present the results of WB and SWB experiments based on P.800 MOS tests with a Degradation Category Rating (DCR) [10]. The test results in Figure 5 illustrate the NB performance improvement at 7.2 kbps and above compared to reference codec AMR [2]. Similarly WB test results in Figure 6 show efficiency roughly twice that of AMR-WB [3] at 23.85 kbps. SWB performance at 13.2 kbps in Figure 7 shows the performance of LR-HQ which meets the EVS codec performance requirement and is not worse than AMR-WB+ [4] at 9.75kbps whose algorithmic delay is longer than twice of LR-HQ. Final EVS codec performance can be found in a report of 3GPP selection test results [11], which shows all the performance requirements for music and mixed contents are passed for the bit-rates given in Table 1.

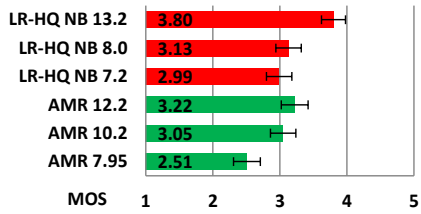

Figure 5: NB Mixed Music Performance Results

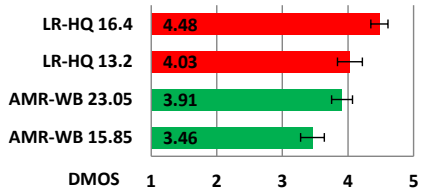

Figure 6: WB Mixed Music Performance Results

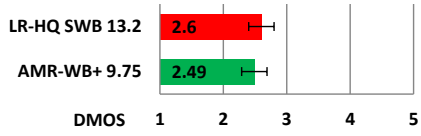

Figure 7: SWB Mixed Music Performance Results

5. CONCLUSIONS

LR-HQ MDCT coder is presented. It was adopted in the 3GPP EVS codec as one of the modes for coding mixed and music content due to its performance.

6. REFERENCES

[1] 3GPP TS 26.445: "Codec for Enhanced Voice Services (EVS); Detailed Algorithmic Description", 3GPP TS 26.445 (Release 12), Sep. 2014.

[2] "Mandatory speech codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; General description Speech", 3GPP TS 26.071.

[3] "Speech codec speech processing functions; Adaptive Multi-Rate-Wide-band (AMR-WB) speech codec; General description", 3GPP TS 26.171, Mar.2001.

[4] Audio codec processing functions; Extended Adaptive Multi-Rate Wide-band (AMR-WB+) codec", 3GPP TS 26.290.

[5] S. Bruhn, et al., "Standardization of the new EVS Codec", IEEE ICASSP, April 2015

[6] ITU-T G.719, "Low complexity, full band audio coding for high quality, conversational applications", ITU-T Recommendation G.719, June 2008.

[7] ITU-T G.718 Annex B, "Superwideband scalable extension for ITU-T G.718", ITU-T Recommendation G.718 Amendment 2, March, 2010.

[8] Fastl, H., Zwicker, E., Psychoacoustics: Facts and Models. Springer series in information sciences. Springer, 3rd edition, 2007.

[9] Thomas R. Fischer; Hosang Sung; Jie Zhan; Eunmi Oh, "High-quality audio transform coded excitation using trellis codes," ICASSP, 2008, pp.197-200.

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

[11] 3GPP, Tdoc S4-141065, Report of the Global Analysis Lab for the EVS Selection Phase, Aug. 2014.

[12] 3GPP, T-doc S4-130522, EVS performance requirements, April 2013.