Super-Wideband Bandwidth Extension for Wideband Audio Codecs Using Switched Spectral Replication and Pitch Synthesis

Bernd Geiser, Hauke Krüger, Peter Vary

Institute of Communication Systems and Data Processing (IND)
RWTH Aachen University, Germany
{geiser|krueger|vary}@ind.rwth-aachen.de

Abstract

This paper describes a new bandwidth extension algorithm which is targeted at high quality audio communication over IP networks. The algorithm is part of the Huawei/ETRI candidate for the ITU-T super-wideband (SWB) extensions of Rec. G.729.1 and G.718. In the SWB candidate codec, the 7-14 kHz frequency band of speech and audio signals is represented in terms of temporal and spectral envelopes. This description is encoded and transmitted to the decoder. In addition, the fine structure of the input signal is analyzed and compactly encoded. From this compact information, the decoder can regenerate the 7-14 kHz fine structure either by spectral replication or by pitch synthesis. Then, an adaptive envelope restoration procedure is employed. The algorithm operates in the MDCT domain to allow subsequent refinement coding by vector quantization of spectral coefficients. In the paper, relevant listening test results for the G.729.1-SWB candidate codec that have been obtained during the ITU-T standardization process are summarized. Good audio quality could be shown for both speech and music signals.

Codec Overview

The bandwidth extension (BWE) algorithm which is described in this paper has been implemented in the Huawei/ETRI candidate [1] for the super-wideband (SWB) extensions of ITU-T Rec. G.729.1 [2, 3] and G.718 [4, 5]. A brief overview of the SWB candidate codec is given in the following. The current implementation is based on ITU-T G.729.1 which has a total bit rate of 32 kbit/s and encodes wideband signals (16 kHz sampling rate).

The input signal, sampled at 32 kHz, is split into two critically sampled (16 kHz) subband signals $s_{\mathrm{wb}}(n)$ and $s_{\mathrm{swb}}(n)$ by means of the infinite impulse response quadrature mirror filter (IIR QMF) bank of [6]. The *wideband* signal $s_{\mathrm{wb}}(n)$ is then encoded by the G.729.1 core codec as described in [2]. The encoder and decoder for the 8–16 kHz signal $s_{\mathrm{swb}}(n)$ is depicted in Fig. 2. First, pre-processing by a 6 kHz lowpass filter and by an adaptive temporal envelope (ATE) normalization procedure [7] is conducted. The resulting normalized signal is transformed into the frequency domain with a modified discrete cosine transform (MDCT) using a window length of 40 ms, i.e., a frame shift of 20 ms. The MDCT domain signal is encoded in a hierarchical fashion: For low bit rates (e.g., 4 kbit/s on top of G.729.1), a parametric coding method is used. At higher bit rates (up to 32 kbit/s on top of G.729.1), vector quantization (VQ) of MDCT coefficients is employed, cf. [1]. In this paper, we focus on the parametric coding modes, i.e., the red processing blocks in the figure. For clarity, we omit the description of the extra processing that takes place for the 7–8 kHz frequency band.

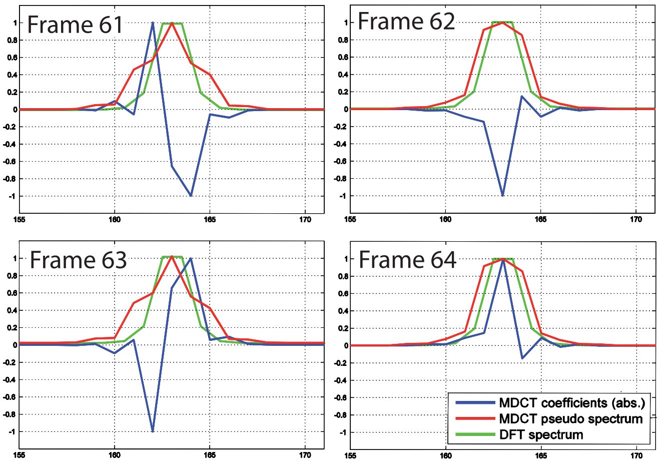

Figure 1: MDCT and pseudo spectrum representations of a stationary sinusoid over four successive signal frames. Comparison with DFT amplitude spectrum.

Bandwidth Extension Algorithm

As the first and most important part of the parameter set, the **BWE encoder** computes a *spectral envelope* which is formed of logarithmic subband gains for subbands of equal bandwidth. This envelope is quantized in a multi-stage approach: First, spherical VQ is applied to a 16-dim. gain vector. The VQ bit allocation is 4 bits for the vector mean and 34–37 bits for the radius and shape depending on the available bit budget. If necessary, the quality of the spectral envelope can be improved by scalar quantization of the VQ error and entropy coding of the quantizer indices (Huffman codes). This multi-stage quantization scheme efficiently captures outliers that stem from an insufficient VQ encoding. Note that the refined spectral envelope is reused for the high bit rate modes of the codec where VQ of MDCT coefficients is employed.

Afterwards, the *spectral fine structure* of the 8–14 kHz MDCT signal is analyzed. It is important to note, that the MDCT as such is not well suited for spectral analysis. This is easily illustrated with the MDCT representation of a stationary sinusoid signal as shown in Fig 1. The MDCT representation is *not* stationary which makes spectral analysis difficult. To avoid an additional (and costly) DFT, we conduct spectral analysis for each frequency bin with index k based on the MDCT "pseudo spectrum" [8]

$$S_k = \sqrt{y_k^2 + (y_{k-1} - y_{k+1})^2} \tag{1}$$

which approximates the DFT amplitude spectrum (green graph in Fig. 2). For practical reasons, an additional spectral whitening (tilt compensation) is applied in the codec. From the pseudo spectrum representation, first, a correlation based tonality value is computed and quantized with 3 bits. Secondly, a flag indicates whether the 8–14 kHz fine struc-

This work has been conducted in cooperation with Huawei Tech. Co., Ltd., Core Network Research Dept., Beijing, P.R. of China

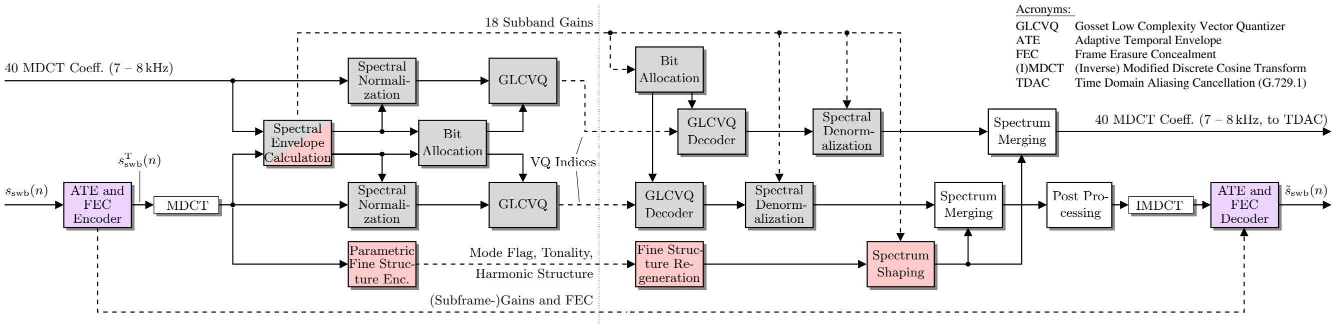
18 Subband Gains
40 MDCT Coeff. (7 – 8 kHz)
Spectral Normalization
GLCVQ
Spectral Envelope Calculation
Bit Allocation
$s^{\mathrm{T}}_{\mathrm{swb}}(n)$
$s_{\mathrm{swb}}(n)$
ATE and FEC Encoder
MDCT
Spectral Normalization
GLCVQ
VQ Indices
Parametric Fine Structure Enc.
Mode Flag, Tonality, Harmonic Structure
(Subframe-)Gains and FEC
Bit Allocation
GLCVQ Decoder
Spectral Denormalization
GLCVQ Decoder
Spectral Denormalization
Fine Structure Regeneration
Spectrum Shaping
Spectrum Merging
Spectrum Merging
40 MDCT Coeff. (7 – 8 kHz, to TDAC)
Post Processing
IMDCT
ATE and FEC Decoder
$\tilde{s}_{\mathrm{swb}}(n)$
Acronyms:
GLCVQ Gosset Low Complexity Vector Quantizer
ATE Adaptive Temporal Envelope
FEC Frame Erasure Concealment
(I)MDCT (Inverse) Modified Discrete Cosine Transform
TDAC Time Domain Aliasing Cancellation (G.729.1)

Figure 2: Block diagram of the super-wideband encoder and decoder [1].

ture is similar to the fine structure of the low frequency band (0–7 kHz). If this is the case, no additional information is transmitted. If dissimilarity is detected, the harmonic structure of the 8–14 kHz signal is explicitly transmitted. This explicit description consumes up to 14 bits (depending on the tonality) and consists of a fractional harmonic grid (resolution: 6.25 Hz) and an offset parameter. The additional offset is needed here because the pitch harmonics in the (downsampled) 8–14 kHz subband are equally spaced, but they are, in general, *not* placed at multiples of the first pitch peak position within this band.

The corresponding **BWE decoder** module is responsible for the regeneration of the *spectral fine structure* in the MDCT domain. If the transmitted mode flag indicates similarity between low and high band, the fine structure in the 8–14 kHz range is derived by spectral replication from the 1-7 kHz band. Additionally, the tonality of the replicated signal is adjusted according to the received tonality value by “peak sharpening” or “noise mixing.” If the mode flag indicates dissimilarity between low and high band, the fine structure in the 7–14 kHz range is generated as a mixture of pseudo random noise and synthetic harmonics which are placed according to the fractional harmonic grid and the offset parameters. The energy ratio of noise and harmonic components is controlled by the tonality value. Yet, as indicated by the instationarity of the transform coefficients for a stationary input signal (Fig. 1), signal synthesis in the MDCT domain is not straightforward. For example, placing single peaks in the MDCT domain (i.e., *stationary* MDCT coefficients) leads to an instationary time domain signal that is affected by annoying artifacts. Therefore, to obtain concise sinusoids, we synthesize the individual harmonic components by imitating the instationary MDCT domain behavior shown in Fig. 1.

Finally, the *spectral envelope* of the signal has to be restored. The generated MDCT coefficients are therefore spectrally shaped by multiplication with a gain correction factor which is the ratio of the desired subband gain and the measured gain of the artificially generated signal. However, in case of synthetic harmonic components, even a slight misplacement could move such a component into the neighboring (wrong) subband. Consequently, an incorrect subband gain would be applied, resulting in artifacts such as noise amplification. The described effect occurs in particular for high-pitched signals (e.g., violin). As a solution, we apply an interpolated gain factor if there are highly tonal subbands that contain synthetic pitch harmonics close to a subband boundary.

Test Results

The proposed BWE algorithm has been intensively tested during the ITU-T qualification phase for the G.729.1-SWB codec. In particular, the bit rates of 32+4 kbit/s and 32+8 kbit/s are relevant for the BWE performance. On the employed MOS scale with its upper limit of 5 (imperceptible impairment compared to the original signal), the following scores have been obtained in the ITU-T tests, cf. [1]:

- Clean speech at 36 kbit/s: 4.30,
- Clean speech at 40 kbit/s: 4.59,
- Music at 40 kbit/s: 4.37.

With these scores, the proposed coder is better suited for speech than the reference codec ITU-T G.722.1C at comparable bit rates of 32 and 48 kbit/s while it still performs reasonably well for music input. With little additional bit rate (32+16 kbit/s), the performance for music is statistically not worse than G.722.1C at the same bit rate (48 kbit/s).

Summary

We have outlined our proposal for a super-wideband bandwidth extension algorithm that operates entirely in the MDCT domain. Thereby, certain important characteristics of the MDCT have been considered. For spectral analysis, we have used the pseudo spectrum representation while, for signal synthesis, the specific behavior of the MDCT coefficients for stationary input has been taken into account, cf. Fig. 1. For speech signals, the BWE algorithm mostly selects the spectral replication mode, while pitch synthesis is used for music input if necessary. Together with the temporal envelope as described in [7] the total BWE bit rate is ≈ 4 kbit/s. This amount of information suffices to consistently synthesize additional frequency content for speech and music. In the ITU-T qualification tests, the proposed candidate codec passed all requirements for mono input signals.

References

[1] B. Geiser *et al.*, “Candidate proposal for ITU-T super-wideband speech and audio coding,” in *Proc. of IEEE ICASSP*, Taipei, Taiwan, Apr. 2009.

[2] ITU-T Rec. G.729.1, “G.729 based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729,” 2006.

[3] S. Ragot *et al.*, “ITU-T G.729.1: An 8-32 kbit/s scalable coder interoperable with G.729 for wideband telephony and Voice over IP,” in *Proc. of IEEE ICASSP*, Honolulu, Hawai’i, USA, Apr. 2007.

[4] ITU-T Rec. G.718, “Frame error robust narrowband and wideband embedded variable bit-rate coding of speech and audio from 8-32 kbit/s,” 2008.

[5] T. Vaillancourt *et al.*, “ITU-T EV-VBR: A robust 8-32 kbit/s scalable coder for error prone telecommunications channels,” in *Proc. of EUSIPCO*, Lausanne, Switzerland, Aug. 2008.

[6] H. W. Löllmann *et al.*, “IIR QMF-bank design for speech and audio subband coding,” in *Proc. of IEEE WASPAA*, New Paltz, NY, USA, Oct. 2009, pp. 269–272.

[7] B. Geiser and P. Vary, “Joint pre-echo control and frame erasure concealment for VoIP audio codecs,” in *Proc. of EUSIPCO*, Glasgow, Scotland, Aug. 2009.

[8] L. Daudet and M. Sandler, “MDCT analysis of sinusoids: exact results and applications to coding artifacts reduction,” *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 3, pp. 302–312, May 2004.