ENHANCED CODING OF HIGH-FREQUENCY TONAL COMPONENTS IN MPEG-D USAC THROUGH JOINT APPLICATION OF ESBR AND SINUSOIDAL MODELING

Tomasz Żernicki[*]

Telcordia Poland Sp. z o. o.
Applied Research Center
Umultowska 85
Poznań, Poland

Maciej Bartkowiak, Marek Domański

Poznań University of Technology
Chair of Multimedia Telecommunications
and Microelectronics
Polanka 3, Poznań, Poland

ABSTRACT

The new eSBR tool of MPEG-D Universal Speech and Audio Coding offers a great advantage in compression of high frequency content, however it produces audible artifacts for sounds whose pitch frequencies are strongly variable or exceeding the split frequency of eSBR. We propose an extension of the forthcoming standard by adding a high frequency sinusoidal tool. This tool introduces additional parametric information to the data bitstream in order to encode the challenging tonal components which are excluded from eSBR processing. Listening tests demonstrate benefits of the proposed approach for test items of strongly tonal character.

Index Terms— Bandwidth extension, audio coding, sinusoidal modeling

1. INTRODUCTION

The forthcoming MPEG-D Universal Speech and Audio Coding standard (USAC) [1] (currently at the stage of MPEG working draft [2]) features a number of novel compression tools allowing to preserve a good audio quality for mixed content (music and speech) at bit rates as low as 12 kb/s. First of all, a switched coding mode is introduced by adaptively selecting between a typical transform-based (frequency domain) core similar to AAC and a Linear Prediction based (time domain) core similar to ACELP with optional MDCT-based residual coding. A very important component allowing to achieve a great compression efficiency at very low bit rates is the enhanced Spectral Band Replication (eSBR) tool.

The SBR tool introduced within the MPEG-4 HE-AAC standard [3] allowed for significantly reducing the number of bits required for coding the high-frequency (HF) part of the spectrum. The HE-AAC encoder limits the bandwidth of the input signal handled by the transform core. The missing HF content is re-synthesized in the decoder by the SBR tool through frequency domain shift followed by appropriate energy scaling and temporal envelope shaping. This shift is simply realised by copying a part of the baseband signal time-frequency (TF) representation. The main disadvantage of SBR is that such crude generation of HF content results often in non-harmonic spectra since the translated harmonic partials of the low frequency signal are not placed at integer multiples of corresponding fundamental frequencies. Furthermore, frequency shift does not properly reconstruct the harmonic content with variable pitch frequency, because frequency deviations should increase their depths proportionally to the overtone number. Such spectra are significantly different than original full spectra, and sometimes result in unpleasant artifacts [4]. A technique based on adaptive complex modulation has been proposed to address the problem of inaccurate frequency shift [5], however it still does not produce appropriately scaled frequency deviations of partials.

As opposed to SBR, eSBR employs an optional method for generating the HF harmonic partials from the decoded baseband signal [6]. The Phase Vocoder (PV) technique is used for scaling in frequency a selected range of the LF content, which preserves its harmonic structure. The basic disadvantage of PV is that while scaling the spectrum with integer factors of 2, 3, etc., it creates sparse harmonic patterns with missing frequencies which are not multiples of these integer factors. This problem has been recently addressed by generating additional cross-products [7] that fill the gaps between scaled overtones through selectively applied modulation. Nevertheless, PV suffers from artifacts when operating on mixed spectra of many sound sources, where several physical partials exist within a single transform bin used for scaling. Moreover, spectra with rapidly varying pitch frequencies are not perfectly scaled by the PV due to the inappropriate phase evolution model. Last, but not least, eSBR (similarly to SBR) is not able to reconstruct harmonic sounds, whose fundamental frequencies are above the low operating limit of the eSBR tool. Such sounds are simply removed from the baseband signal by the lowpass filter applied before encoding. As a result, there is no respective content to replicate. In such situations, the SBR and eSBR tools employ a technique called "Sinusoidal Coding". It adds some components of fixed amplitude as well as fixed and crudely quantized frequency to the TF representation. While properly adjusting the amount of tonality in HF range, this addition also yields increased inharmonicity.

2. THE PRINCIPLE

The main idea behind the proposed technique is to augment the existing eSBR technique by an additional tool devoted exclusively to encoding selected HF sinusoidal partials which would otherwise be distorted. It shows an advantage in reconstructed audio quality in certain cases of music programmes including loud instruments playing solo in high registers, bells, and other percussive sounds which are typically very challenging for the eSBR tool.

The proposed technique, based on sinusoidal modeling and coding [8], operates as a pre-processor for the eSBR encoder (cf Fig. 1). It offers an efficient parametric representation of selected HF tonal components, leaving the remaining part of the signal to be processed by the eSBR tool. For this purpose, sinusoidal partials are detected in the original signal above f_{SBR} (the eSBR cut-off frequency) in suc-

[*]Part of this research has been done during this author's work at the Poznań University of Technology.

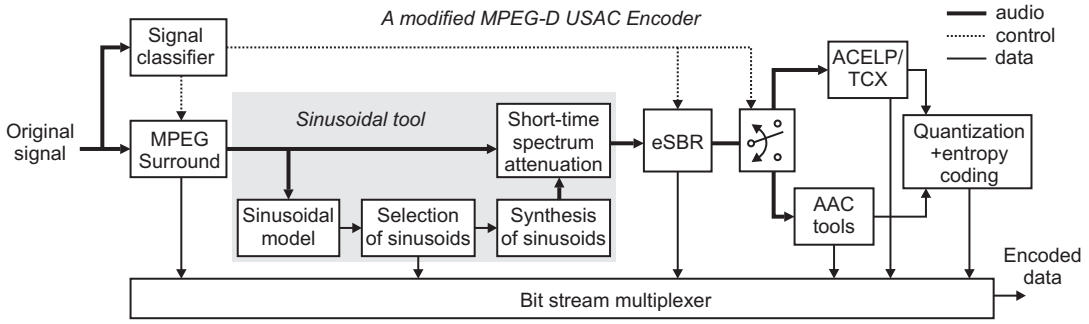

Fig. 1. The modified MPEG-D USAC encoder with additional blocks related to the proposed sinusoidal tool.

cessive data frames. These components are tracked by the sinusoidal model (SM) and some of them are subsequently removed from the signal. This small number of sinusoidal partials is selected based on psychoacoustic principles and subsequently encoded and embedded into the bit stream. The remaining signal is further handled by eSBR tool and the switched AAC/ACELP core.

At the decoder side the HF sinusoidal partials are synthesized from the data and mixed with the output of eSBR decoder thus providing a better quality bandwidth extension of the reconstructed bandlimited signal. An example encoded signal is shown in Fig. 3.

Sinusoidal modeling has been already proposed in the context of improving the SBR technique [9], however that approach acts solely at the decoder side as an enhancement of the existing patching algorithm and involves a fundamental frequency estimation which is only applicable to harmonic sounds and speech.

3. DESCRIPTION OF THE PROPOSED CODING PROCESS

3.1. Signal analysis

The input signal x is analysed by the FFT transform in consecutive frames of $N = 2048$ samples long, shifted by $H = 512$ samples. The spectral flatness measure (SFM) [10] is computed in each frame m of the spectrum $X_m(k)$. This measure is used to indicate a global level of tonality in each frame, so that if $SFM < 24dB$ holds for a sequences of frames, they are identified as containing mostly noise and are excluded from processing in this tool. Otherwise, prominent sinusoidal partials are detected and their parameters (amplitude, $A(m)$ and frequency, $f(m)$) are estimated. Partials of low frequencies are essentially excluded from this analysis, however the frequency limit is set significantly below the split frequency, f_{SBR}. The purpose of such margin is to prevent partials of varying frequencies near f_{SBR} from producing fragmented sinusoidal trajectories. The estimated partial parameters are subject to the tracking algorithm that is based on Linear Prediction (LP) [11].

3.2. Selection of tonal components for encoding

The established trajectories of sinusoidal parameters represent strong tonal components. Only a small subset of these trajectories is actually encoded in the proposed tool and removed from the signal prior to eSBR processing. The primary goal of such selection is to handle those partials which most likely would be audibly distorted by the standard SBR processing, and also to avoid encoding of remaining partials that would lead to suboptimal bit allocation. The selection is based on a couple of rules. First of all, trajectories with significant energy are considered. Secondly, the set is ordered according to their psychoacoustic relevance, taking into account the total energy represented by the sum of squared amplitudes and the sensitivity of human auditory system expressed by the A-weighting curve [12] denoted as $W_A(f)$. We use a ranking function

$$R_l = \sum_{m=1}^{M_l} A_l^2(m) \cdot W_A\left(f_l(m)\right), \tag{1}$$

where m is the frame number, l is the index of sinusoidal trajectory considered, and l-th trajectory is M_l frames long. Finally, a rate control loop rejects the weakest trajectories according to the lowest R_l, until the target bit rate is achieved.

3.3. Modification of eSBR input

For excluding of the encoded partials from redundant processing in eSBR, the HF tonal part in each frame is synthesized according to

$$s(n) = \sum_{l=1}^{M_l} \tilde{A}_l(n) \cos\left(\phi_l(n)\right), \tag{2}$$

$$\phi_l(n) = \phi_l(0) + 2\pi \sum_{i=1}^{n} \tilde{f}_l(n), \tag{3}$$

where sample index i restarts from 1 in each frame, $\phi_l(0)$ is the accumulated phase at the end of previous frame, and $\tilde{A}(n)$, $\tilde{f}(n)$ represent amplitudes and frequencies interpolated on a sample basis using a cubic spline.

This synthetic signal is used for masking the unwanted partials by the use of a modified Short-Time Spectral Attenuation (STSA) technique [13], fig. 2. The resulting signal y is obtained in consecutive STFT frames by $Y_m(k) = G_m(k) \cdot X_m(k)$, where $G_m(k)$ is a transfer function of attenuation filter defined as

$$G_m(k) = \left[\min\left(\epsilon \cdot |S_m(k)|, 1\right) * h(k)\right]^{-1}, \tag{4}$$

where ϵ is the attenuation factor, $h(k)$ is the impulse response of a short kernel operating in DFT domain (a moving average filter which smoothes the spectrum).

3.4. Encoding of selected partials

The parameter chains of sinusoidal trajectories selected for encoding are transmitted in the output bit stream by a sequence of quantized prediction residuals, separately for $A(m)$ and $f(m)$. No explicit phase information is transmitted. For each trajectory, a low order LP adaptive predictor is applied in a closed loop with a quantizer.

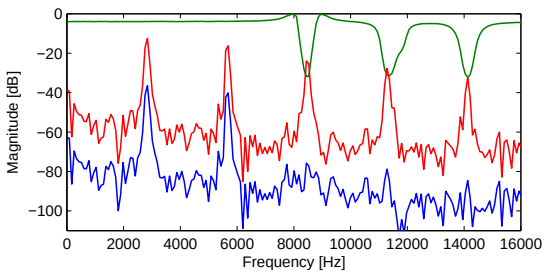

Fig. 2. STSA in a single frame: $X_m(k)$ (red) is attenuated by $G_m(k)$ (green). The resulting spectrum $Y_m(k)$ (blue) is shifted by -20 dB for clarity.

The quantization steps are uniform in a perceptual scale and correspond to 3 dB and 40 cents, which were determined experimentally as not causing significant degradation. The first element in each trajectory is encoded relatively to its closest already transmitted element, which is implicitly signalled by appropriate ordering of data. If there is no other trajectory, the absolute value is sent. The symbols are encoded using adaptive Huffman codes.

3.5. Bit rate control

The maximum bit-rate of the sinusoidal model parameters (B_{sin}) is limited to 3 kb/s. It means that the bit-rate of MPEG-D USAC must be reduced to obtain the total target bit-rate (B_T) , $B_{USAC} = B_T - B_{sin}$. The rate control loop attempts to fit in the budget of B_{sin} by iteratively encoding a reduced set of trajectories and calculating the length of the bitstream. This is repeated until the target is met within a margin of 50 b/s.

3.6. Decoding

The decoding process is straightforward. The reconstructed sinusoidal parameters are used for synthesizing the respective HF tonal components with continuously varying frequencies and amplitudes. The HF signal may be synthesized with the same sampling frequency as used by the USAC codec, or it may be synthesized with a higher sampling frequency thus allowing to reconstruct components in the whole audible range, exceeding the operating range of USAC at given bit rate.

Due to computational complexity constraints of the decoder, an IFFT–based chirp synthesis technique is employed [14]. The computational cost of such synthesis is dominated by the IFFT calculation which is estimated at 0.79 MOPS at sampling rate of 32 kHz.

4. LISTENING TEST RESULTS

A formal blind listening test was arranged to check the proposed technique against the USAC WD7 codec according to MUSHRA methodology [15]. The test was performed using professional grade studio monitoring speakers in an acoustically treated room. There was one test session to assess the audio quality at 16 and 20 kb/s. 10 trained and experienced listeners had to assess the quality of 12 standard test items selected by the MPEG committee and 5 new. For each test item, four conditions were listened to with a randomized and hidden order. The statistical analysis of the test results had been performed based on 8 subjects, excluding 2 which did not pass the post-screening. 95% confidence intervals of the mean results had been calculated according to the requirements of MPEG Audio CE methodology.

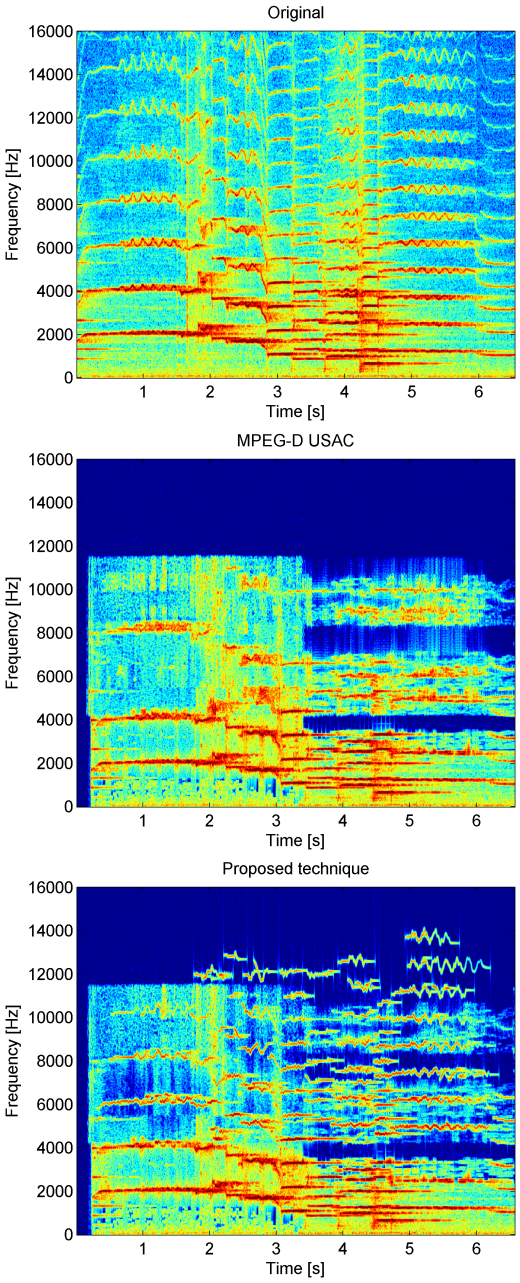

Fig. 3. Example spectrograms for signal violin2 and target bit-rate of 16 kb/s, from the top: input signal, full reconstructed signal in the decoder (MPEG-D USAC), full reconstructed signal in the decoder (proposed technique).

The absolute scores per system for the individual items in the listening test did not reveal a statistically relevant difference except one item (altosax) which showed an improvement of 20 points in MUSHRA scale at 16 kb/s. Figure 4 shows the results of differential scores for individual items relative to USAC WD7 calculated for all subjects and as mean over all items. It can be observed that the proposed technique (WD7-CE) significantly improves the performance

of USAC codec for 3 test items: (altosax, accordion-ebu and violin2) at 16 kb/s and for 4 items (altosax, accordion-ebu, te1_mg54_speech and violin2) at 20 kb/s. The proposed technique did not exhibit a statistically relevant lower result in any of the tests, neither in absolute nor differential scores.

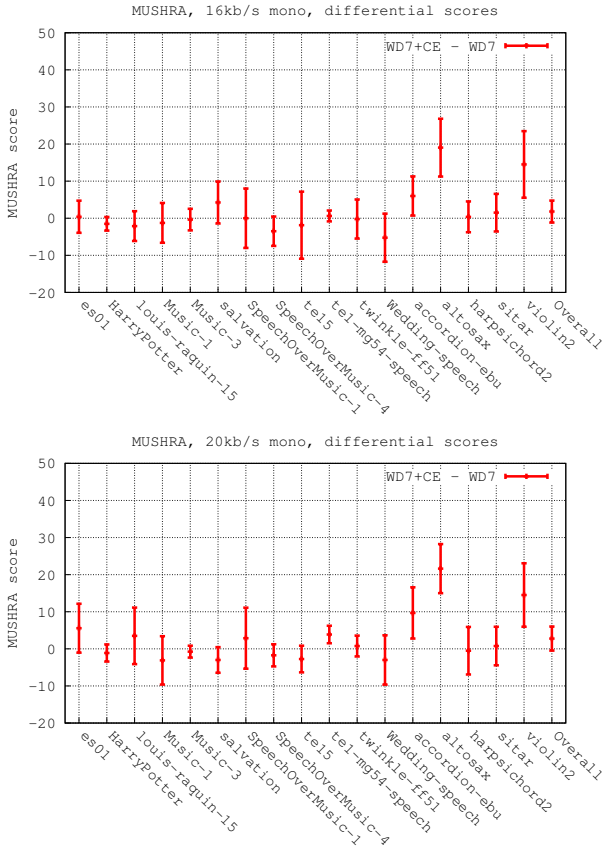

Fig. 4. Subjective test results at 16 kb/s (top) and 20 kb/s (bottom) - differential scores (WD7-CE vs WD7). The 95 % confidence intervals (8 listeners) are plotted.

5. CONCLUSIONS

Formal listening tests confirmed that the proposed technique offers a significant increase of quality compared to the MPEG-D USAC WD7 codec at 16 kb/s and 20 kb/s. Moreover, presented results confirmed the results previously obtained for MPEG-4 AAC HE [16, 17]. The sinusoidal coding technique shows particular advantages for signals with strong tonal HF components, where SBR and eSBR patching algorithm still introduce audible artifacts. Therefore, we propose to use this tool as an extension for audio coding techniques utilizing SBR and eSBR [18].

6. REFERENCES

[1] M. Neuendorf at al., "Unified speech and audio coding scheme for high quality at low bitrates," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, 2009, pp. 1–4.

[2] ISO/IEC JTC1/SC29/WG11, "Working Draft 7 of Unified Speech and Audio Coding," Doc. N11299, Dresden, Germany, Apr. 2010.

[3] ISO/IEC 14496-3:2001, "Coding of Audio-Visual Objects: Audio," Int. standard, 2001.

[4] Chi-Min Liu, Han-Wen Hsu, and Wen-Chieh Lee, "Compression Artifacts in Perceptual Audio Coding," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 681–695, 2008.

[5] F. Nagel, S. Disch, and S. Wilde, "A continuous modulated single sideband bandwidth extension (accepted)," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, Texas, USA, 14-19Mar. 2010.

[6] F. Nagel and S. Disch, "A harmonic bandwidth extension method for audio codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, 19-24 Apr. 2009, pp. 145–148.

[7] ISO/IEC JTC1/SC29/WG11, "Finalization of CE on improved harmonic transposer in USAC," Doc. N11299, Guangzhou, China, Oct. 2010.

[8] F.P. Myburg, *Design of a scalable parametric audio coder*, Ph.D. thesis, Eindhoven University of Technology, 2004.

[9] S.-U. Ryu, K. Rose, and J.-H. Chang, "Effective High Frequency Regeneration Based On Sinusoidal Modeling For MPEG-4 HE-AAC," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 16 –19 Oct. 2005.

[10] J.D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314 –323, Feb. 1988.

[11] M. Lagrange, S. Marchand, and J.-B. Rault, "Enhancing the tracking of partials for the sinusoidal modeling of polyphonic sounds," *IEEE Transactions on Audio Speech and Language Processing*, vol. 15, no. 5, pp. 1625–1634, July 2007.

[12] IEC, "IEC 61672:2003: Electroacoustics sound level meters," Tech. Rep., IEC, Geneva, 2003.

[13] O. Cappe and J. Laroche, "Evaluation of short-time spectral attenuation techniques for the restoration of musical recordings," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 84–93, 1995.

[14] X. Rodet and P. Depalle, "Spectral envelopes and inverse FFT synthesis," in *93rd Convention of the Audio Engineering Society*, Oct. 1992, AES Preprint 3393.

[15] ITU-R Recommendation BS.1534-1, "Method for the subjective assessment of intermediate quality levels of coding systems," Tech. Rep., International Telecommunication Union, Geneva, Switzerland, Jan. 2003.

[16] T. Żernicki and M. Domański, "Improved coding of tonal components in MPEG-4 AAC with SBR," in *Proc. of IX European Signal Processing Conference, EUSIPCO'08*, Lausanne, Switzerland, Aug. 25-29 2008.

[17] ISO/IEC JTC1/SC29/WG11, "Improved coding of tonal components in audio techniques utilizing the SBR tool," Doc. M17914, Geneva, Switzerland, July 2010.

[18] ISO/IEC JTC1/SC29/WG11, "Telcordia and PUT listening test results for CE on improved tonal component coding in eSBR (USAC)," Doc. M18532, Guangzhou, China, Oct. 2010.