Adaptive TCX Windowing Technology for Unified Structure MPEG-D USAC

Taejin Lee, Seungkwon Beack, Kyeongok Kang, and Whanwoo Kim

The MPEG-D unified speech and audio coding (USAC) standardization process was initiated by MPEG to develop an audio codec that is able to provide consistent quality for mixed speech and music contents. The current USAC reference model structure consists of frequency domain (FD) and linear prediction domain (LPD) core modules and is controlled using a signal classifier tool. In this letter, we propose an LPD single-mode USAC structure using an adaptive widowing-based transform-coded excitation module. We tested our system using official test items for all mono-evaluation modes. The results of the experiment show that the objective and subjective performances of the proposed single-mode USAC system are better than those of the FD/LPD dual-mode USAC system.

Keywords: MPEG-D USAC, TCX, AAC.

I. Introduction

Traditionally, voice production model-based speech coders and human auditory model-based audio coders have progressed independently due to the unique characteristics of each input signal and their different application areas. Due to their different design concepts, no state-of-the-art speech or audio coders can provide consistent sound quality for both speech signals and music signals. However, with the increasing number of mobile devices, there is a strong demand for a unified codec that is able to provide consistent quality for mixed contents. To satisfy this market need, a process has been initiated by MPEG that aims to standardize a new codec with consistent high quality for speech, music, and mixed contents over a broad range of bitrates [1]. ISO/IEC 23003-3 unified speech and audio coding (USAC) is a new audio coding standard that allows for the coding of speech and music at any mixture with consistent audio quality within a wide range of bitrates. It also supports single- and multi-channel coding at high bitrates providing perceptually transparent quality [2].

The USAC system is a hybrid audio coder that combines efficient MPEG audio coding technologies, such as advanced audio coding [3], spectral band replication (SBR) [4], and

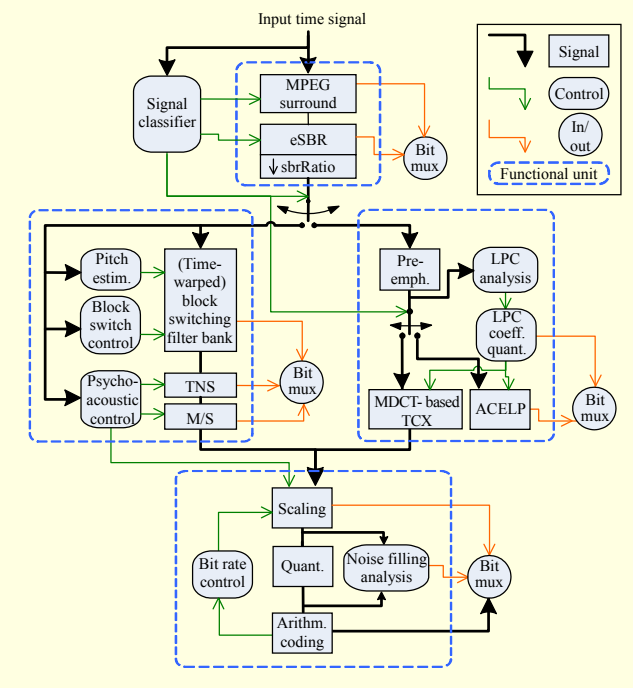

Fig. 1. MPEG-D USAC encoder structure.

Manuscript received Sept. 21, 2011; revised Nov. 1, 2011, accepted Nov. 14, 2011.

This research was supported by the KCC (Korea Communications Commission), Korea, under the ETRI R&D support program supervised by the KCA (Korea Communications Agency) (KCA-2011-11921-02001).

Taejin Lee (+82 42 860 5713, tjlee@etri.re.kr), Seungkwon Beack (skbeack@etri.re.kr), and Kyeongok Kang (kokang@etri.re.kr) are with the Broadcasting & Telecommunications Convergence Research Laboratory, ETRI, Daejeon, Rep. of Korea.

Whanwoo Kim (wwkim@cnu.ac.kr) is with the Department of Electronics Engineering, Chungnam National University, Daejeon, Rep. of Korea.

http://dx.doi.org/10.4218/etrij.12.0211.0404

MPEG Surround [5], with those of efficient speech coder technologies, such as algebraic code excited linear prediction and TCX [6].

Figure 1 shows a block diagram of the USAC encoder. The input signal, assumed to be stereo, is first processed by an MPEG Surround module, which produces parametric stereo information for transmission, as well as a downmixed mono signal. The mono signal forms an input for an enhanced SBR (eSBR) module. The eSBR module outputs parametric information for high band regeneration at the decoder (SBR info), as well as a lower band signal. The low band mono signal is then encoded by two specific frequency domain (FD) and linear prediction domain (LPD) core coders based on the input signal characteristics. The FD and LPD core modules process music- and speech-like input signals, respectively. The FD/LPD core modules are controlled by a signal classifier, and thus the performance of the USAC system depends heavily on the performance of the signal classifier tool [2], [7]. In this letter, we propose an LPD single-mode USAC system that does not require a signal classifier.

II. LPD Single-Mode USAC Using Enhanced TCX

1. Enhanced TCX Module Based on Adaptive Windowing

The current USAC system uses a signal classifier tool for the control of different coding modes. However, we cannot be certain that the signal classifier tool always selects the optimal coding scheme for a number of different items. Therefore, we propose an LPD single mode USAC system based on adaptive TCX windowing technology. The adaptive TCX windowing technology comes from two fundamental concepts: i) a revised analysis window that is designed to follow a sine-shaped form by removing the rectangular part from the original form as much as possible and ii) the last 50% overlap process for utilizing modified discrete cosine transform (MDCT) time domain aliasing cancellation (TDAC).

The first concept comes from common audio coding knowledge that a symmetrical sine-shaped analysis window for quantization has lower spectral leakage and better frequency selectivity than a rectangular window and hence could reduce the unfortunate distortion at the block boundaries.

The major factor that comes into play in the design of filter banks for audio coding is maximizing the ability to separate the frequency components of the signal while minimizing the audibility of blocking artifacts. Given the tonal input, an inadequate frequency resolution can produce unreasonably high signal-to-noise ratio requirements within individual subbands, resulting in high bitrates. Thus, for a tonal signal, good frequency selectivity is essential for low bitrates. For a transient signal, a good time resolution is essential for low bitrates. Unfortunately, most audio source material is highly non-stationary and contains significant tonal and atonal energy, as well as both steady-state and transient intervals. Therefore, no single-shape window is optimal for all signals. Based on the signal characteristics, one should dynamically select the window shape while satisfying the perfect reconstruction conditions of the window.

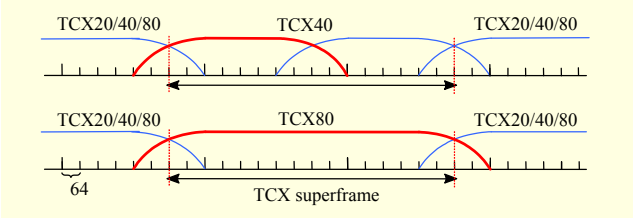

Fig. 2. Current USAC TCX40/80 window shapes.

The second concept is related to the MDCT process. To reduce blocking artifacts from windowing, we have to set up large overlap regions. When the perfect sine-shaped window is applied to the analysis frame before transformation, a 50% overlap process is requested as a corresponding decoding process to remove the window effect from the decoded signals. The 50% overlap results in a doubling of the data rate. However, the TDAC processing does not increase the data rate (that is, a critically sampled rate), despite perfectly removing the effect of the sine-shaped window from the decoded signals. Therefore, it is imperative that this concept be combined with the first concept to avoid a sacrifice in data rate. The above two concepts based on TDAC processing can provide more reliable frequency data to a spectral quantizer due to the window shape, and the windowing effect can be removed using a synthesis process with an additional 50% overlap, preserving the critical data rate.

The USAC system uses a 2048-sample MDCT as a TCX80, a 1024/512-sample MDCT as a TCX40/20, and a fixed 256-sample overlap for the TCX80/40/20 switching as shown in Fig. 2. Because of the fixed 256-sample overlap, the current USAC window shapes of TCX80 and TCX40 partly contain a rectangular shape in the middle part of the window, which might be a reason to make the original frequency response dull. In this letter, we propose removing the rectangular part from the current USAC TCX window as much as possible and adaptively changing the overlap region based on the input signal characteristics.

The window shape can be alternatively selected according to the temporal transition in the junction between the superframes. If the transition is detected, the windowing normally follows the current USAC TCX windowing. Otherwise, a long overlap process is applied between the TCX-based superframes. The

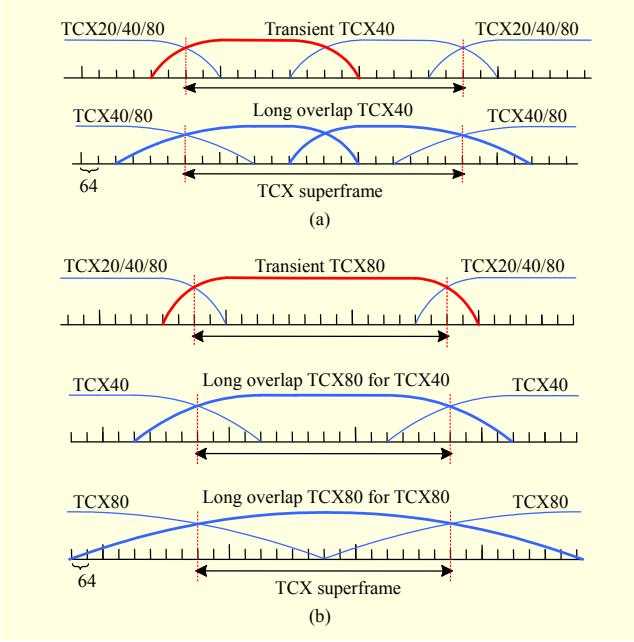

Fig. 3. Proposed USAC (a) TCX40 and (b) TCX80 window shapes.

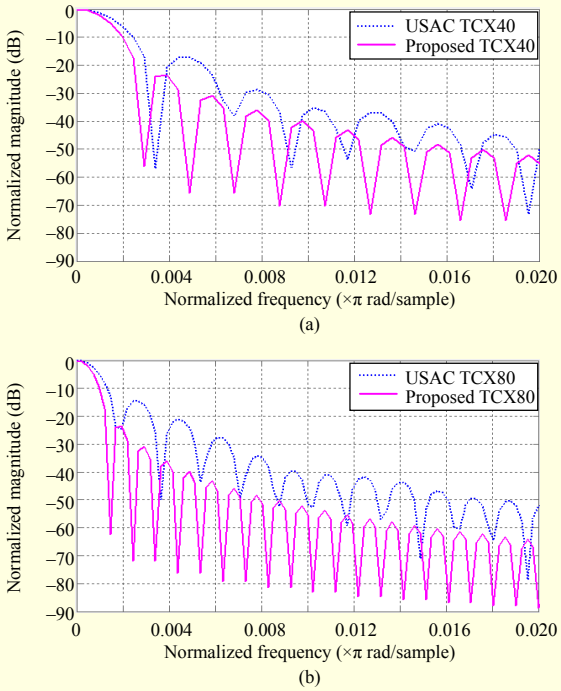

Fig. 4. Comparison of (a) TCX40 and (b) TCX80 window properties.

50% overlap process is used to enhance the coding gain when a stationary interval is guaranteed in the overlap region with a more sine-shaped window. Otherwise, the shorter overlap (that is, the current USAC overlap) is applied to code the temporal transition and results in a prevention of the spread of quantization noises.

As shown in Fig. 3(a), the proposed TCX40 window uses a long overlap between superframes; however, if there is a transient between superframes, we follow the same TCX40 window shapes. For the TCX80 case, as illustrated in Fig. 3(b), a long overlap window is used for TCX40 and TCX80 switching; however, if there is a transient between superframes, we follow the same TCX80 window shapes.

The window shape is faithfully designed from a perfectly shaped sine window between the overlap regions of the superframes as much as possible. This can help the quantizer estimate more precisely the peak of the spectrum and yield a reduction of the quantization noise.

The flat part of the middle of the current USAC window has a rectangular window-like property that causes it to emphasize the interference part of the frequency peaks from the desired original response due to relatively better attenuation of the sidelobe response of the window. On the other hand, the proposed sine-shaped window gives relatively lower sidelobe attenuation with the narrow width of the mainlobe and thus provides a faithful representation of the original peaks and is less affected by the interference parts through the windowed short frame analysis in the frequency domain.

Figure 4 and Table 1 show the frequency responses and parameters related to the properties of the current USAC TCX window and the proposed window shapes. These parameters show that the proposed sine-shaped window has superior properties with a lower spectral leakage ratio. The USAC TCX window is in no way better than the proposed window in terms of analysis tools for the frequency response.

Table 1. Window characteristics of proposed and current USAC.

	USAC			Proposed		
	Mainlobe width (– 3 dB)	Sidelobe attenuation (dB)	Leakage factor (%)	Mainlobe width (–3 dB)	Sidelobe attenuation (dB)	Leakage factor (%)
TCX40	0.0027	–17.0	2.13	0.0022	–23.1	0.48
TCX80	0.0016	–14.4	4.71	0.0011	–23.0	0.48

2. Delay and Complexity of Proposed System

The delay of the proposed system is identical to that of the USAC system on the decoder side. The proposed output buffering from the TCX frame is not changed if a long overlap is applied in the TCX frames. This means that the delay of the proposed USAC system occurring by an MDCT 50% overlap does not exceed the buffer delay of the current USAC system.

The increased complexity from applying an enhanced TCX module to the USAC system is negligible. A simple

windowing operation and an overlap operation are additionally necessary when a long overlap window is applied.

III. Evaluation

For an evaluation of the proposed new codec architecture, we carried out a formal listening test with twelve expert listeners for all official USAC mono-operating bitrates (12 kbps, 16 kbps, 20 kbps, and 24 kbps). The test items covered three categories: speech signal, music signal, and mixed signal. All twelve official USAC test items were used, with four items in each category [8]. The tests were carried out according to the MUSHRA methodology (ITU-R BS.1543-1).

Figure 5 shows the average differential scores between the proposed LPD single-mode USAC system and the FD/LPD dual-mode USAC system with a 95% confidence interval. The plus value means that the proposed system is statistically better than the current USAC system. As we can see in the figure, the proposed system shows significantly better sound quality for all test bitrates: five items (three music, two mixed) for 12 kbps, four items (three music, one mixed) for 16 kbps, four items (one speech, two music, one mixed) for 20 kbps, and five items (one speech, two music, two mixed) for 24 kbps. Additionally, there are many cases where our system performs significantly better than the USAC system for music items in all test bitrates.

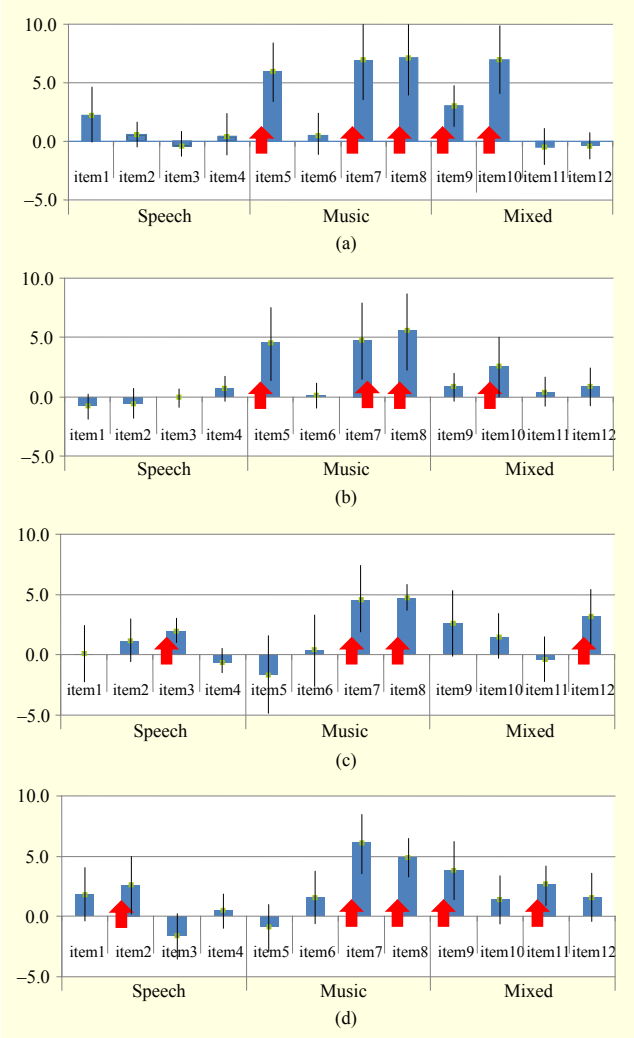

Fig. 5. Average differential scores for items at (a) 12 kbps, (b) 16 kbps, (c) 20 kbps, and (d) 24 kbps.

IV. Conclusion

The current USAC system uses a signal classifier tool for the control of two different FD/LPD coding modes. However, we cannot be sure that the signal classifier always selects the optimal coding scheme for various kinds of items. Thus, the sound quality of the current USAC system varies based on the performance of the signal classifier.

In this letter, we proposed an LPD single-mode USAC system using an enhanced TCX module, which does not cause an increase in delay or complexity. Through an objective window characteristic evaluation and subjective listening test of the LPD single-mode USAC system, it has been confirmed that the enhanced TCX module works successfully, regardless of the characteristics of the input signals (for example, without signal classification). Our proposed system also shows better quality compared with the optimal mode of the USAC system based on two different core structures.

References

[1] ISO/IEC SC29 WG11 N9519, "Call for Proposals on Unified Speech and Audio Coding," MPEG, Oct. 2007.

[2] ISO/IEC SC29 WG11 N12013, "Study on ISO/IEC 23003-3:201x/DIS of Unified Speech and Audio Coding," MPEG, Mar. 2011.

[3] ISO/IEC Std. 2003, "Information Technology—Coding of Audio-Visual Objects—Part 3: Audio," ISO/IEC 14496-3.

[4] ISO/IEC Std. 2003, "Bandwidth Extension," ISO/IEC 14496-3, AMD. 1.

[5] Y. Lee et al, "Design and Development of T-DMB Multichannel Audio Service System Based on Spatial Audio Coding," *ETRI J.*, vol. 31, no. 4, Aug. 2009, pp. 365-375.

[6] 3GPP TS 26.290 V6.3.0, "Extended Adaptive Multi-rate–Wideband (AMR-WB+) Codec," 2007.

[7] M. Neuendorf et al., "Unified Speech and Audio Coding Scheme for High Quality at Low Bitrates," *ICASSP*, 2009.

[8] ISO/IEC SC29 WG11 N9638, "Evaluation Guidelines for Unified Speech and Audio Proposals," MPEG, Jan. 2008.