Fixed-Point Implementation of MPEG-D Unified Speech and Audio Coding Decoder

Eunwoo Song, and Hong-Goo Kang
Digital Signal Processing Lab.
Yonsei University
Seoul, Korea
sewplay@dsp.yonsei.ac.kr

Joonil Lee
Digital TV Lab.
LG Electronics Inc.
Seoul, Korea
joonil.lee@lge.com

***Abstract*—This paper describes a fixed-point implementation method of the unified speech and audio coding (USAC) decoder that has been recently standardized by moving picture experts group (MPEG). Since the structure of USAC is too complicated to support both speech and audio signals, the quality and complexity issues must be carefully reviewed while performing fixed-point implementation. By analyzing the structure of the USAC decoder, this paper describes key ideas to successfully realize the fixed-point system. Subjective and objective test results verify that the implemented fixed-point decoder shows equivalent quality to the floating-point decoder. The average and worst cases of complexity depending on the type of encoding modes are also given in detail.**

***Index Terms*—Unified speech and audio coding (USAC); speech coding, audio coding; fixed-point implementation;**

I. Introduction

As mobile and multimedia devices advanced, the need to support both audio and speech contents increased significantly. Note that traditional audio coders have relatively poor quality for speech contents, and speech coders also have poor quality for music contents at low bitrates because each coding scheme highly depends on the characteristics of input signals.

In early 2012, the MPEG-D audio subgroup standardized unified speech and audio coding (USAC) that provides high quality for both speech and music contents even in low bitrates [1]. The core idea of USAC is switching between two state-of-the-art speech and audio coders depending on the characteristics of input signal: AMR-WB+ operates for speech-like signals, and HE-AAC v2 does for music-like signals [2][3].

Unlike voice communication oriented standard codecs, the standard USAC decoder software is released as a floating-point version. In commercialized products, however, the algorithm is typically implemented by a fixed-point processor due to the constraints on chip cost. Since the fixed-point system has limited arithmetic accuracy, how well the fixed-point algorithm is designed is very important to maintain high quality.

Our aim here is to design an efficient fixed-point algorithm for the USAC decoder by considering complexity and quality issues. A good starting point for accomplishing the objective is utilizing fixed-point software that has been implemented before such as AMR-WB+ and HE-AAC v2 [4][5]. Since many modules have been modified or newly introduced while making USAC standard, they cannot be used directly. For instance, a transform coded excitation (TCX) embedded in the AMR-WB+ codec is modified by a weighted linear prediction transform (wLPT), and the Huffman coder used for HE-AAC v2 codec is substituted by a context adaptive arithmetic coder. In addition, a transition windowing with a forward aliasing cancellation (FAC) module is introduced to solve the switching problem between two core codecs [1].

By investigating the functional modules of the USAC decoder in detail, this paper suggests various methods to efficiently implement it into a fixed-point system. It includes key ideas on how to handle the parameters that are sensitive to quality, how to efficiently transform the frequency domain signal into the time domain signal, and how to represent the complicated functions accurately. Experimental results verify that the quality of the implemented fixed-point version is equivalent to that of the floating-point version while its complexity is reasonably low. All of these techniques and system requirements provide a guideline of DSP instructions for the USAC audio services.

II. Overview of USAC decoder

Fig. 1 depicts a block diagram of the USAC decoder that consists of the core decoder and a bandwidth extension module. In the core decoding module, either the linear prediction domain (LPD) or the frequency domain (FD) decoder is selected depending on the encoding mode bit. The baseline schemes of the LPD and FD decoders are similar to the AMR-WB+ and HE-AAC v2 standards, respectively. The LPD decoding mode that is typically used for speech signals generates synthesized signals by passing excitation parameters through dequantized spectral coefficients. The excitation signal is decoded by one of the two types of decoding schemes; algebraic code-excited linear prediction (ACELP) or wLPT. In the ACELP-based decoding, the time domain excitation signal is reconstructed by combining adaptive and innovation codebook vectors with their gains [2]. On the other hand, the wLPT scheme reconstructs the excitation signal in the frequency domain. It requires an inverse modified cosine transform (IMDCT) to obtain time domain excitation signal [1]. When the decoding mode is FD, the IMDCT-based transform is also performed. Firstly, the spectral coefficients are obtained

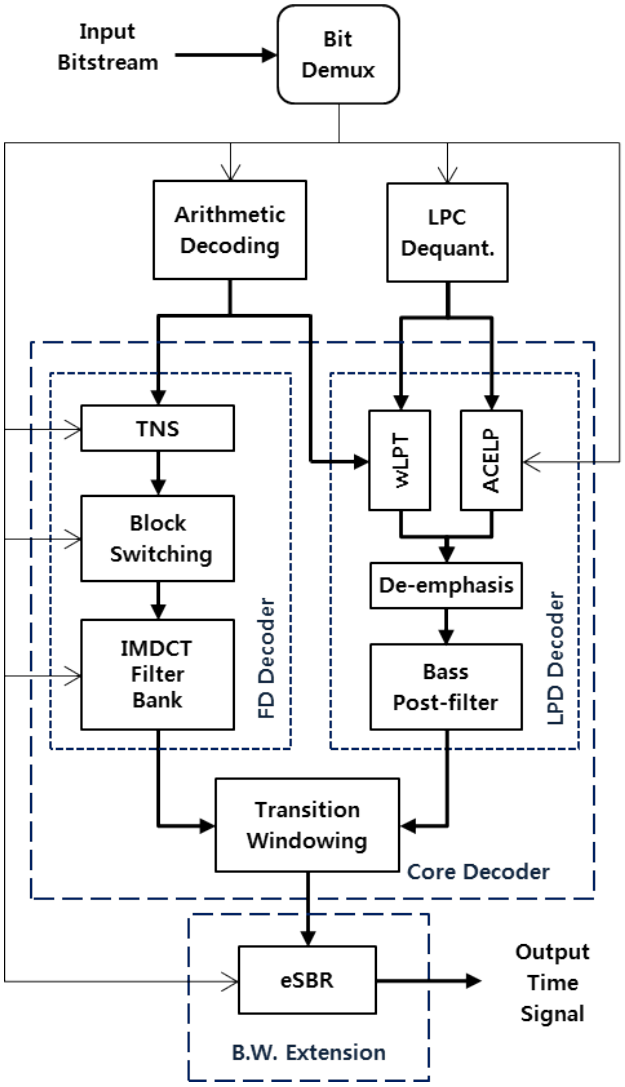

Fig. 1. USAC decoder block diagram

by taking an arithmetic decoding to the bitstream. Then, the decoded spectral coefficients are transformed into the time domain signal using a block switching and IMDCT tool [3].

Three decoding schemes (ACELP, wLPT, FD) mentioned above can be recategorized into two depending on the type of window. The IMDCT-based decoding type (wLPT and FD) uses an overlapping window, and the ACELP uses a rectangular non-overlapping window. Therefore, an additional transition windowing tool is needed to smoothly decode the transition region to avoid aliasing [1].

To decode high frequency region, the USAC codec introduces a spectral band replication (SBR) module [1]. The basic concept of SBR is a replication of the harmonic sequences from low frequency band signals (i.e. signals decoded by the core decoder) to high frequency band signals. To well represent the spectral characteristics of the high frequency bands, the parametric features such as spectral energy, noise level, and sinusoidal level are also utilized in the quadrature mirror filterbank (QMF) domain. The spectral energy adjusts the spectral envelope, and the other parameters adjust the level of harmonic strength. Compared with the conventional SBR, the enhanced SBR (eSBR) embedded in USAC significantly improves quality because it includes flexible crossover frequency control, adaptive noise floor control, higher temporal resolution, and phase-vocoder-based harmonic transposer [1]. Therefore, its complexity is very high.

III. Fixed-point Implementation

This section describes the detailed idea for efficiently implementing fixed-point algorithm, and measures the computational complexity of each module. The fixed-point library is selected from the ITU-T standard software tools defined in [6]. It includes 16 and 32 bit operations and provides basic operators such as addition, subtraction, multiplication, square root, etc. Each operator has a weight that reflects the number of DSP cycles which results in weighted million operations per second (WMOPS). The floating-point version of target decoder is selected from the USAC standard reference software given in [7]. Only the modules that are related to mono signal decoding are described in this paper.

A. Implementation issues

1) Core decoder: The ACELP module contains two autoregressive (AR) filters. One is a long-term prediction (LTP) filter that reconstructs adaptive codebook vectors from past excitation samples. The other is a short-term prediction (STP) filter that synthesizes the speech signal from past synthesized speech samples. Since past samples are recursively used for the filtering process, the resolution of excitation and synthesized speech signals should be very high to avoid the error propagation effect that results in quality degradation. The most important thing in this part is how to determine the appropriate Q-formats because the dynamic range of the input signals is not consistent. To represent the maximum resolution of excitation and synthesized speech signals, a variable Q-format is introduced to these variables. Although additional memory to save the Q-format of each variable and a scaling process to compensate the Q-format mismatch between previous and current frame are needed, it is more profitable to use a variable Q-format approach to maintain high quality.

To implement the IMDCT module to the wLPT and FD decoder, it is very important to find an appropriate trade-off point in view of complexity and quality. Typically, the inverse fast Fourier transform (IFFT) is leveraged with pre-twiddle and post-twiddle process rather than implementing IMDCT directly [8]. In USAC, an IFFT algorithm that supports unspecialized IMDCT length of 120, 240, 480, and 960 point window is embedded to utilize time-warped (TW) IMDCT. Since this optional module is not included in the proposed system, however, we do not need to use such a complicated algorithm. To design a low cost transform, the complex IFFT in USAC is replaced by the radix-2 IFFT algorithm. Furthermore, the twiddle operations are replaced

by the integer look-up table. As a result, it has much lower complexity than the original IMDCT while fully supporting the length of 128, 256, 512, and 1024 point window frames.

If there is a mode transition between two types of decoders, the transition windowing process is needed. Since the ACELP decoder does not have any overlap, the zero input response of LPC filter must be applied beforehand if the mode of following one is determined as the IMDCT-based decoder. If the Q-format of the current frame differs from adjacent frames, the scaling process is also needed to avoid quality degradation.

2) Bandwidth extension module: The eSBR module is complicated and requires many special functions such as square root, exponential, logarithm, division, etc. It is not easy to implement them because improper approaches result in very high complexity. This paper presents an example of solution to handle these complex functions.

To provide stable and predictable high frequency components, the eSBR introduces an additional gain adjuster that controls the low frequency band signals [9]. The gain is adjusted by

$$gain \;=\; 10^{\frac{1}{20}\left\{\left(\frac{1}{B}\sum_{k=0}^{B-1}10\log_{10}(E[k])\right)-\widetilde{E}\right\}}, \qquad (1)$$

where B is the number of sub-bands, k is a sub-band index, $E[k]$ is the average sub-band energy of QMF-domain signals, and $\widetilde{E}$ is the envelope of $E[k]$. By considering the fixed-point operators given in [6], an alternative form of the equation can be obtained by changing the factor of ten to two and rearranging the constants as follows :

$$gain \;=\; 2^{\left\{\left(\frac{\alpha}{B}\sum_{k=0}^{B-1}\log_{2}(E[k])\right)-\beta\cdot\widetilde{E}\right\}}, \qquad (2)$$

where the constant value α and β is 0.5 and $\log_2 10 \cdot 0.05 \approx 0.1661$, respectively. It is also not recommended to perform a division operation due to its high complexity, thus $1/N$ operation is replaced by a table look-up approach.

B. Complexity analysis

To estimate the precise complexity, the WMOPS in each encoding mode is separately measured using the counter operation of fixed-point library. In case of FD, ACELP, and wLPT, the encoding mode is forced to operate each mode only. On the other hand, in case of switching mode (normal encoding), the encoding mode selectively operates FD, ACELP, and wLPT depending on the input signal characteristics. The average and worst cases of WMOPS in each encoding mode are summarized in Table I. Note that the index 256 means the sub-frame length of wLPT.

In core decoding, the transform-based coders such as FD and wLPT have relatively higher complexity than the ACELP module. The difference between the worst case WMOPS of the switching mode and that of the wLPT256 can be considered as the influence of the window transition module. It is straightforward to conclude that the complexity of eSBR module is higher than that of core decoding module. The main reason can be found from the analysis and synthesis processing of QMF modules that require large number of iterations per frame. Note that the average and worst cases of QMF modules at 24 kbps are 15.804 and 15.813, respectively. A simplified algorithm needs to be considered to further reduce the complexity in the future.

TABLE I
WMOPS ESTIMATES OF USAC DECODER

Bitrate	Decoding module		Average	Worst
12 kbps	Core	FD	6.368	6.816
		ACELP	5.382	7.425
		wLPT256	9.620	12.650
		Switching	7.935	13.456
	eSBR		19.311	19.390
16 kbps	Core	FD	7.119	8.710
		ACELP	6.094	8.924
		wLPT256	10.809	14.519
		Switching	8.928	15.461
	eSBR		21.636	21.636
20 kbps	Core	FD	7.921	10.080
		ACELP	6.793	9.275
		wLPT256	12.098	16.009
		Switching	9.980	17.419
	eSBR		24.736	24.829
24 kbps	Core	FD	8.621	9.718
		ACELP	7.397	10.623
		wLPT256	12.845	17.269
		Switching	10.666	18.280
	eSBR		25.890	25.990

IV. PERFORMANCE EVALUATION

This section discusses the performance of the fixed-point USAC decoder. Twelve mono items such as four speech, four music, and four mixed signals used for the USAC evaluation tests are chosen [10]. The encoded bitstream is obtained by the USAC common encoder "JAME" that is developed and released by MPEG [11]. Each item is encoded at the bitrate of 12, 16, 20, and 24 kbps. A bitstream from the reference quality encoder (RQE) at 20 kbps is also included [12]. The objective and subjective performance is measured by three methods: a floating-point to fixed-point error ratio (FfER), a log-spectral distance (LSD), and a multiple stimuli with hidden reference and anchor (MUSHRA) test [13].

A. Objective quality: FfER and LSD

To evaluate the performance quantitatively, the FfER and LSD are measured by the following equations:

$$FfER\;[dB] = \frac{1}{T}\sum_{t=1}^{T}10\log_{10}\frac{\sum_{n=1}^{N_{fr}}|s_1[t,n]|^2}{\sum_{n=1}^{N_{fr}}|s_1[t,n]-s_2[t,n]|^2}, \quad (3)$$

$$LSD\;[dB] = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{N_{FFT}}\sum_{k=1}^{N_{FFT}}\left|10\log_{10}\frac{P_1[t,k]}{P_2[t,k]}\right|^2}, \quad (4)$$

TABLE II
FFER FROM JAME BITSTREAM [DB]

Bitrate	Speech	Music	Mixed
12 kbps	31.368	39.589	35.916
16 kbps	35.555	42.464	40.898
20 kbps	35.420	42.201	40.650
24 kbps	31.194	39.511	35.657

TABLE III
FFER FROM RQE BITSTREAM [DB]

Bitrate	Speech	Music	Mixed
20 kbps	33.752	45.267	42.331

TABLE IV
LSD FROM JAME BITSTREAM [DB]

Bitrate	Speech	Music	Mixed
12 kbps	0.871	0.661	0.843
16 kbps	0.499	0.463	0.463
20 kbps	0.548	0.480	0.499
24 kbps	0.901	0.667	0.876

TABLE V
LSD FROM RQE BITSTREAM [DB]

Bitrate	Speech	Music	Mixed
20 kbps	0.519	0.399	0.426

where T is the number of the frame, N_{fr} is the length of the frame, and N_{FFT} is the number of frequency bins in each frame. $s_1[t,n]$ and $s_2[t,n]$ denote the decoded signal from the floating-point and fixed-point USAC decoder, respectively. $P_1[t,k]$ and $P_2[t,k]$ represent the power spectrum of decoded signal from the floating-point and fixed-point USAC decoder, respectively. The FfER and LSD of 12 test items at each bitrate from JAME encoder and at 20 kbps from RQE encoder are measured. The average of the FfER result is listed in Table II and Table III, respectively. Despite lower FfER of speech contents due to the error propagation in AR filters, it shows that the FfER of all contents is higher than 30dB. In the same vein, average LSD listed in Table IV and Table V also shows reasonable result that the LSD of all contents is lower than 1dB.

B. Subjective quality: MUSHRA test

The perceptual quality of fixed-point USAC decoder is evaluated by listening tests using the MUSHRA methodology. The bitstream from RQE encoder at 20 kbps is utilized for the comparison. In the test, eight experienced listeners are asked to make a quality judgement in an acoustically isolated room with Sennheiser HD650 headphone. The mean scores of MUSHRA test with its 95% confidence intervals are plotted in Fig. 2. The result verifies that the perceived quality of the proposed fixed-point USAC decoder is equivalent to that of the floating-point one.

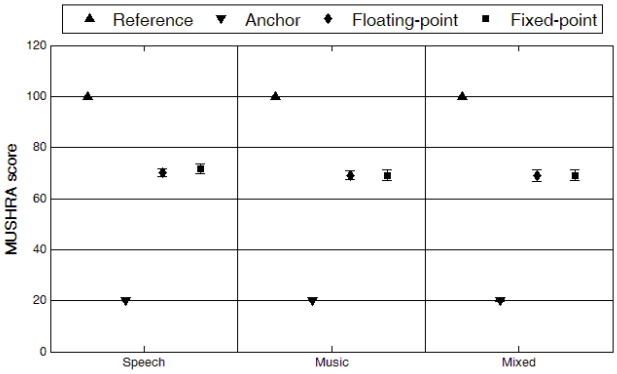

Fig. 2. MUSHRA test results

V. CONCLUSION

The fixed-point version of the MPEG-D USAC decoder has been implemented. By analyzing the structure of the USAC decoder first, this paper described several issues on fixed-point implementation such as complexity, design method, and quality. Objective and subjective evaluation tests confirmed that the performance of fixed-point decoder was equivalent to that of floating-point decoder.

REFERENCES

[1] M. Neuendorf, et al., "MPEG unified speech and audio coding - The ISO/MPEG standard for high-efficiency audio coding of all content types," in *Proc. of the 132nd AES Convention*, April 2012.
[2] 3GPP, "Audio codec processing functions; Extended Adaptive Multi-Rate - Wideband(AMR-WB+) codec; Transcoding functions," 3GPP TS 26.290, 2004.
[3] ISO/IEC 14496-3:2009, "Coding of Audio Visual Objects, Part 3: Audio," 2009.
[4] 3GPP, "ANSI-C code for the fixed-point Extended Adaptive Multi-Rate - Wideband (AMR-WB+) speech codec," 3GPP TS 26.273, 2004.
[5] 3GPP, "General audio codec audio processing functions; Enhanced AAC-Plus general audio codec; Fixed-point ANSI-C code," 3GPP TS 26.411, 2004.
[6] International Telecommunication Union, "Software Tools for Speech and Audio Coding Standards," ITU-T, Recommendation G.191, Geneva, Switzerland, 1994.
[7] ISO/IEC 23003-3:2012, "MPEG-D (MPEG audio technologies), Part3: Unified speech and audio coding," 2012.
[8] Y. Hou and S. You, "Implementation of IMDCT for MPEG2/4 AAC on 16 bit fixed-point digital signal processors," in *Proc. of the 2004 IEEE Asia-Pacific Conference on Circuits and Systems*, IEEE, December 2004.
[9] ISO/IEC JTC1/SC29/WG11, "Finalization of CE on improved SBR," Guangzhou, China, October 2010, MPEG2010/N18378.
[10] ISO/IEC JTC1/SC29/WG11, "Evaluation Guidelines for Unified Speech and Audio Proposals," Antalya, Turkey, January 2008, MPEG2008/N9638.
[11] J. Song, H. Oh, and H. Kong, "Enhanced long-term predictor for unified speech and audio coding," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2011.
[12] ISO/IEC JTC1/SC29/WG11, "DIS of Unified Speech and Audio Coding," Daegu, Korea, January 2011, MPEG2011/N11863.
[13] International Telecommunication Union, "Method for the subjective assessment of intermediate sound quality (MUSHRA)," ITU-R, Recommendation BS. 1543-1, Geneva, Switzerland, 2001.