ISO/IEC MPEG | USAC | Unified Speech and Audio Coding

Technology Reports

MPEG Unified Speech and Audio Coding Enabling Efficient Coding of both Speech and Music

The standardization of MPEG USAC in ISO/IEC is now in its final phase. USAC has the advantage of efficient compression performance both for speech and music signals. This breakthrough has been realized by resolving the issue of the relatively poor compression performance for speech signals compared to music signals, which existing audio codecs have. NTT DOCOMO has proposed the inter-TES technology for USAC which has been adopted as a part of the specification. This technology focuses on improving the subjective quality for transient audio components such as applause sounds and percussive sounds that conventional bandwidth extension schemes do not cope with efficiently. With the advent of USAC which promises better audio quality, improvement in mobile audio services can be expected in the future.

Research Laboratories

Kei Kikuiri

Nobuhiko Naka

1. Introduction

Traditionally, speech and audio signals have been encoded using different coding schemes because of their different applications. Regarding speech, speech coding schemes that encode narrow band signals (up to 3.5 kHz) at lower bitrates have been used for telephony applications. In the case of music, audio coding schemes that encode broad-band signals (up to 20 kHz) at higher bitrates have been used mainly for broadcasting and storage purposes. Therefore, in order to encode mixed speech and music content at low to medium bitrate suitable for mobile applications, there has been the issue that either the speech quality or the music quality deteriorates regardless of whether a speech coding scheme or an audio coding scheme is used. Under such circumstances, the International Organization for Standardization (ISO)[*1] /International Electrotechnical Commission (IEC)[*2] started to standardize Moving Picture Experts Group (MPEG)[*3] Unified Speech and Audio Coding (USAC) so that a single coding scheme can encode both speech and music signals without any disadvantage of degrading the quality of either.

This article describes an overview

©2011 NTT DOCOMO, INC.
Copies of articles may be reproduced only for personal, noncommercial use, provided that the name NTT DOCOMO Technical Journal, the name(s) of the author(s), the title and date of the article appear in the copies.

*1 **ISO**: An organization for standardization in the information technology. Sets international standards for all industrial fields except electrical and telecommunication fields.

*2 **IEC**: An organization for standardization in the information technology. Sets standards in the electrical and telecommunication fields.

of MPEG USAC as well as the inter-subband-sample Temporal Envelope Shaping (inter-TES) which is a technology for improving quality of bandwidth extension schemes. This technology was proposed by NTT DOCOMO and has been adopted as a part of USAC specification.

2. USAC

2.1 Requirements and Envisioned Use Cases for USAC

A call for proposals for USAC, which was issued in October 2007, included the requirement to achieve the comparable quality of the best existing speech coding and audio coding schemes regardless of the contents of the input signals. In July 2008, a Reference Model was selected out of the seven proposed technologies and was used as the basic algorithm for USAC. Following several core experiments performed to improve the performance of the Reference Model, a Final Draft International Standard was issued in September 2011 and is expected to be adopted before the end of the year. Use cases envisioned for USAC include:

- Multi-media download to mobile devices
- User Generated-Content (UGC) services
- Digital radio
- Mobile TV
- Audio books

All these applications deal with a mixed content of speech and music signals and it is expected that USAC will further improve the quality of such signals in these applications.

2.2 Technological Features of USAC

The basic algorithm of USAC is comprised of

- Core coding which is a combination of time-frequency transform coding[*4] and linear prediction coding[*5]
- Bandwidth extension by enhanced Spectral Band Replication (eSBR)[*6]
- Multi-channelization by MPEG Surround (MPS)[*7].

Figure 1 shows the basic configuration of the USAC decoder.

1) Core Coding

The core decoder performs decoding of signals in the low frequency band. The coding scheme used for core coding is switched according to input signal characteristics. Generally, time-frequency transform coding is used for signals of a steady nature such as music, and linear prediction coding for signals of a transient nature such as speech.

Modified Advanced Audio Coding

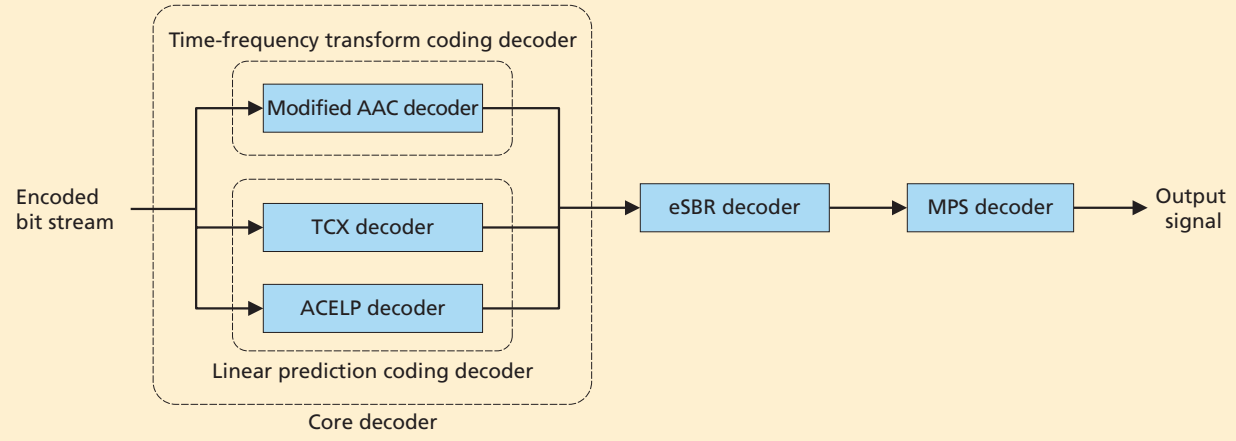

Figure 1 USAC decoder configuration

*3 **MPEG**: Technical standards for coding and transmission of digital audio and video. Standards developed by a working group under a Joint Technical Committee of ISO and IEC. MPEG-2 is used for digital TV and DVD while MPEG-4 is a coding scheme with extended application areas including mobile terminals operating at low bitrates.

*4 **time-frequency transform coding**: A type of coding schemes where a temporal signal is converted into the frequency domain using orthogonal transforms represented by Discrete Cosine Transform (DCT) and MDCT (see *16), and then compressed in the frequency domain.

*5 **linear prediction coding**: Coding schemes that compress redundancy by removing predictable components taking advantage of linear prediction analysis.

*6 **eSBR**: Technology developed through the extension of SBR - the bandwidth extension scheme standardized by ISO/IEC.

*7 **MPS**: One of the multi-channel audio coding schemes standardized by ISO/IEC which realizes high-quality coding at low bitrates by expressing multi-channel signals by audio signals of a smaller number of channels than the actual number and parameters.

(AAC) is used for time-frequency transform coding. Modified AAC is almost the same as MPEG-4 AAC[*8] [1] except for spectral noiseless coding[*9] which adopts arithmetic coding[*10] instead of Huffman coding[*11], for improving its coding efficiency. Linear prediction coding is composed of a combination of Transform Coded Excitation (TCX)[*12] and Algebraic Code Excited Linear Prediction (ACELP)[*13]. Both TCX and ACELP use linear prediction. However, TCX encodes the prediction residual signal[*14] in the frequency domain, whereas ACELP encodes it in the time domain. This configuration is basically the same as that used in the 3GPP AMR-WB+[*15] [2]. The differences are that TCX in USAC employs Modified Discrete Cosine Transform (MDCT)[*16], and that transform coefficients are encoded by arithmetic coding as in the case of modified AAC. In general, ACELP will be used for speech and for signals having large temporal fluctuations and TCX will be used for signals with relatively small temporal fluctuations.

2) Bandwidth Extension

eSBR is developed from MPEG-4 Spectral Band Replication (SBR) [1] which is a bandwidth extension scheme that generates signals in the higher frequency band from signals in the low frequency band. In addition to the functions in MPEG-4 SBR, eSBR has the following new functions: (a) Harmonics Transposer that enables the regular replication of harmonics[*17] in the low frequency band to the high frequency band, (b) Predictive Vector Coding (PVC) that encodes the high frequency band spectral envelope[*18] using the spectral envelope of the low frequency band, (c) inter-TES that shapes the high frequency band temporal envelope using the low frequency band temporal envelope, and (d) functions to enable the setting of ratios between the sampling frequencies for the low and high frequency bands to 3:8 or 1:4 in addition to the conventional 1:2. These technologies have resulted in improvements in qualities at low bitrates compared to those of the conventional SBR.

3) Multi-channelization

MPS can efficiently encode multi-channel signals using signals with a smaller number of channels and spatial parameters[3]. In USAC, the extension of monaural signals to stereo signals using MPS has been specified, and transmission of Inter-channel Phase Difference (IPD)[*19] as a parameter has become possible. In addition, Transient Steering Decorrelator[*20] (TSD) has been added for transient signals.

3. inter-TES

The inter-TES shapes the temporal envelope of the high frequency band generated by SBR using the temporal envelope of the low frequency band signal decoded at the core decoder. For this shaping the similarity between temporal envelopes of audio signals in the low frequency band and the high frequency band is employed. The configuration of SBR is shown in **Figure 2**. In SBR, frames are divided into time segments called SBR envelopes. These SBR envelopes are further divided into a number of sub-bands in the analysis Quadrature Mirror Filter (QMF) bank[*21], where power envelope is shaped for each of these sub-bands. This means that the temporal resolution for shaping the SBR power envelope is equal to that of SBR envelopes. When an input signal has a large temporal fluctuation, a distortion called pre-echo[*22]/post-echo[*23] is generated in front of and after the section with the large temporal fluctuation if the SBR envelope is not short enough. In contrast, if we make the SBR envelope short, then more data is needed to shape the temporal power envelopes which will lead to more bits being required for encoding.

In order to solve this issue, it is necessary to control the power envelopes with as small a number of bits as possible and as high a temporal resolution as possible. inter-TES proposed by NTT DOCOMO is a solution to the issue.

Figure 3 shows the configuration of inter-TES. First, the temporal enve-

*8 **AAC**: One of audio coding schemes specified by ISO/IEC which encodes signals that are time-frequency transformed by MDCT (see *16).

*9 **spectral noiseless coding**: A scheme to compress quantized MDCT (see *16) coefficients in AAC using Huffman coding (see *11).

*10 **arithmetic coding**: A kind of entropy coding in which the code assignment is determined on the basis of symbol occurrence probability; It is usually considered to have better compression efficiency than Huffman coding (see *11).

*11 **Huffman coding**: A kind of entropy coding in which the code assignment is determined on the basis of symbol occurrence probability.

*12 **TCX**: A coding scheme in which a synthesis filter, derived through linear prediction, is driven by excitation signals (input to synthesis filter) encoded in the frequency domain.

*13 **ACELP**: A coding scheme in which a synthesis filter, derived through linear prediction, is driven by an adaptive code book comprising past excitation signals and an algebraic code book comprising multiple pulse sequences.

*14 **prediction residual signal**: A signal component which is the difference between a reference signal and the predicted signal.

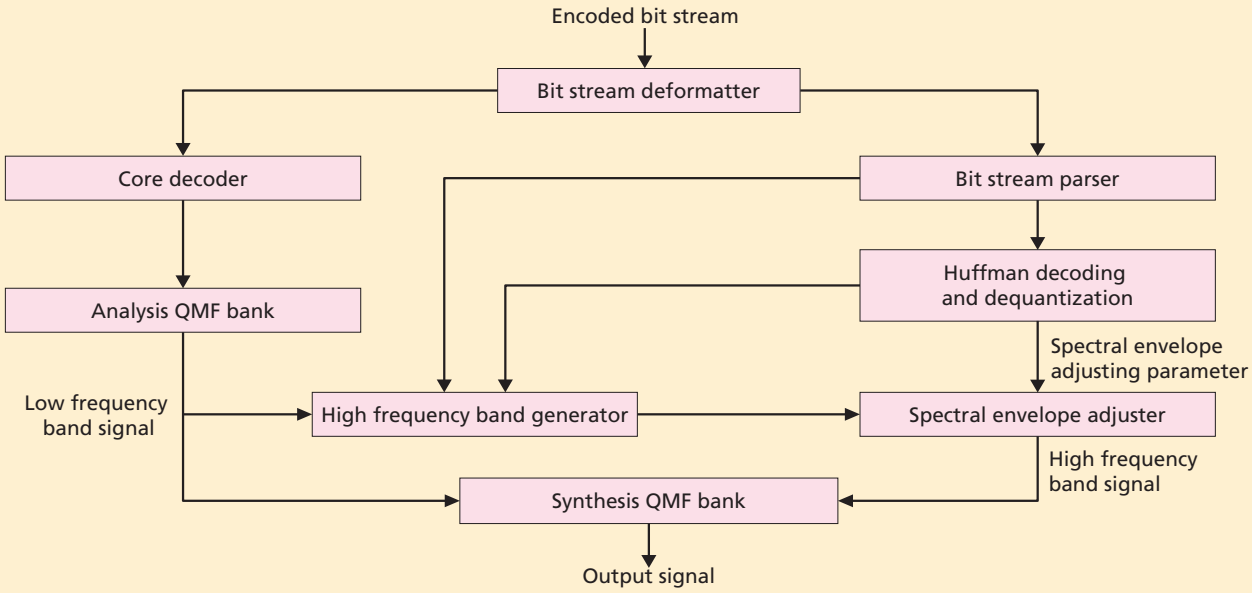

Figure 2 SBR configuration

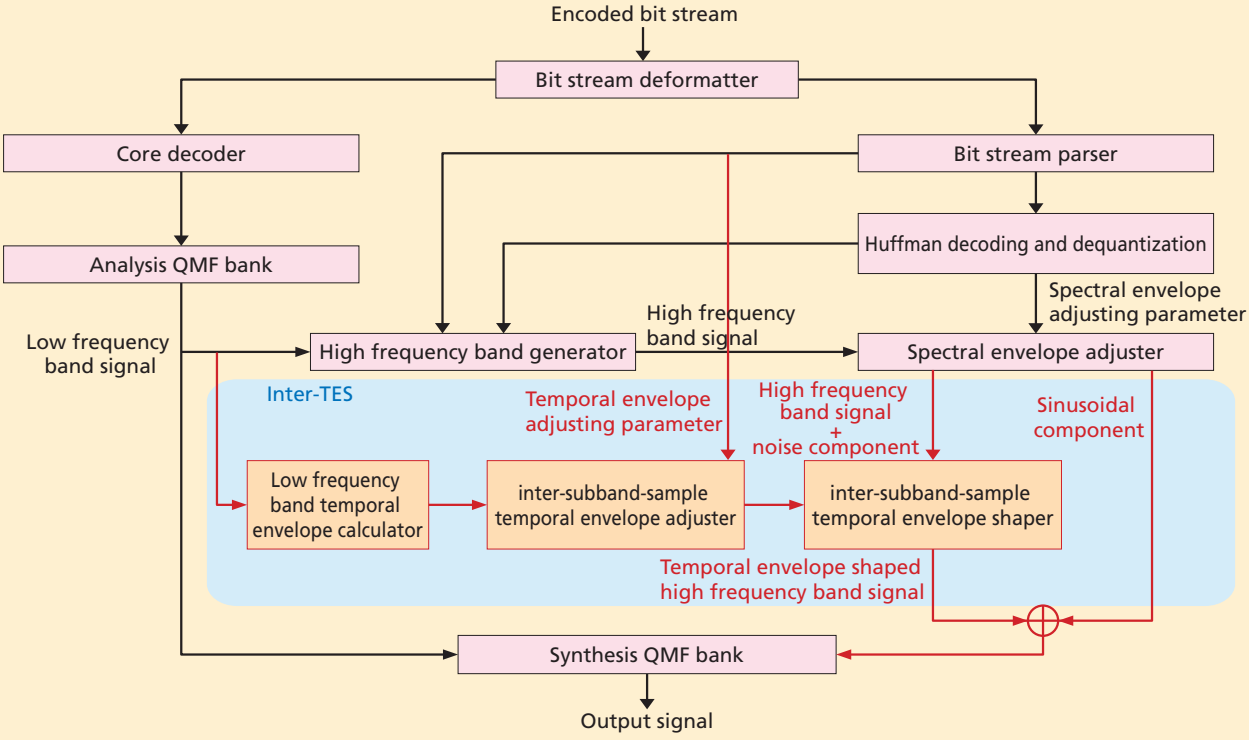

Figure 3 inter-TES configuration

*15 **AMR-WB+**: An extended coding scheme of AMR-WR, the speech coding scheme standardized by 3GPP, which enables it to be used for general audio signals such as music.

*16 **MDCT**: A method for converting a time–series signal to its frequency components. It is able to avoid distortion at block boundaries without losing information by windowing and overlapping transform with the preceding and following blocks, so it is widely used for audio encoding.

*17 **harmonics**: Signal components comprising a base frequency and its multiples.

*18 **frequency band spectral envelope**: The contour of the frequency spectrum.

*19 **IPD**: The difference between the phases of each channel in a multi-channel signal.

*20 **decorrelator**: The function to generate signals that have less correlation with the main signal component.

*21 **QMF bank**: A kind of filter bank that divides an input signal into more than one frequency components.

*22 **pre-echo**: A phenomenon in which a frequency domain quantization error just prior to an onset of attack in audio signal is perceived as an echo-like distortion.

*23 **post-echo**: A phenomenon in which a frequency domain quantization error just after an offset of attack in audio signal is perceived as an echo-like distortion.

lope of the low frequency band signal decoded at the core decoder is calculated. Then, the temporal envelope of the low frequency band signal is adjusted using the temporal envelope adjusting parameter contained in the encoded bit stream[*24], which is followed by a calculation of the gain for shaping the temporal envelope of the high frequency band signal. By applying the gain to the high frequency band signal, the high frequency band signal with a shaped envelope is obtained.

Figure 4 shows the temporal waveforms of a high frequency band signal before and after its temporal envelope has been shaped by inter-TES. The horizontal and vertical axes show time and signal amplitude, respectively. While the input high frequency band signal (a) has steep onset and offset for an attack like an impulse signal, the high frequency band signal generated by the SBR (b) has distortion at the offset. In the signal whose temporal envelope has been shaped by inter-TES (c), the distortion that existed at the offset is now suppressed and it is observed that the temporal envelope is very similar to that of the input signal.

The authors have evaluated the performance of inter-TES by subjective listening tests. The Multi Stimuli with Hidden Reference and Anchors (MUSHRA)[*25] method [4] was used for the tests with eight participants. The tested bitrates were 16 kbit/s and 24 kbit/s (monaural), and the evaluated sound sources were two types of applause sounds. **Figure 5** shows the test results for the signals whose temporal envelope has been shaped by inter-

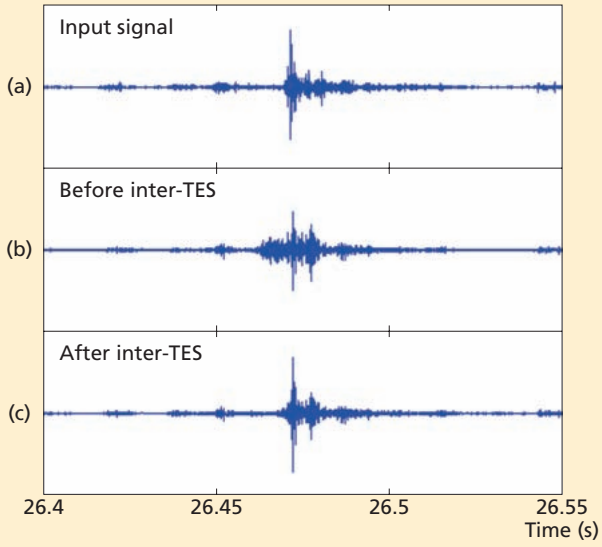

Figure 4 Temporal waveforms of high frequency band signal

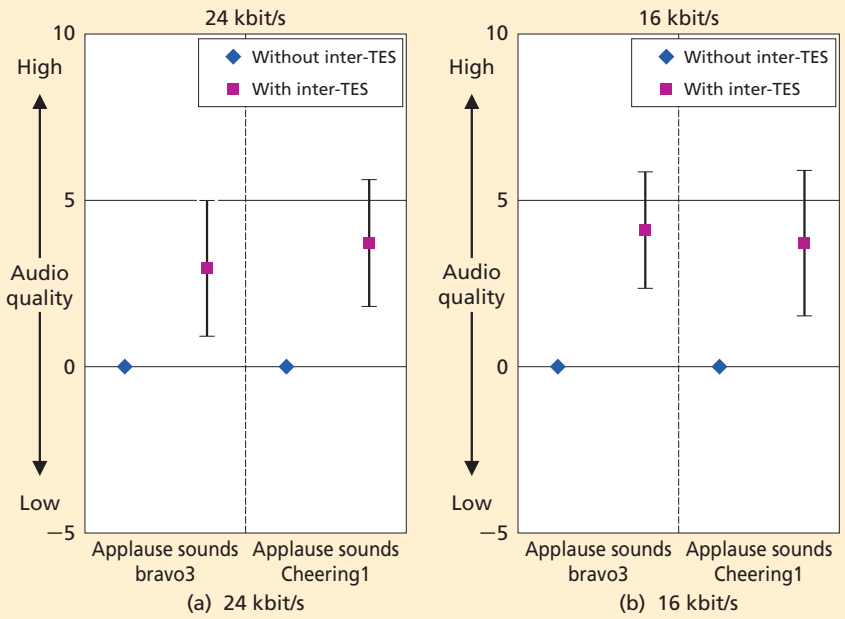

Figure 5 Subjective evaluation test results for inter-TES

*24 **bit stream**: Sequence of encoded bits.
*25 **MUSHRA**: One of the subjective evaluation test methods to evaluate performance of speech and audio coding schemes. The signals for evaluation are evaluated on a 0-100 scale relative to their original signals The signals for evaluation include the original signal and its band-limited signals, in addition to the decoded signal of the coding scheme under the test.

TES. Here, the differential scores for tests with and without inter-TES are plotted. The error bars indicate 95% confidence intervals. These test results confirmed the subjective quality improvements achieved by the introduction of inter-TES for 16 kbit/s and 24 kbit/s.

4. Conclusion

In the course of the USAC standardization work by MPEG, performance verification tests were conducted in June 2011 and the test results obtained at the sites that participated in the tests were reported in July 2011. A final report was issued in September 2011. The results at the test sites confirmed that USAC has a superior performance against the state-of-the-art speech and audio coding schemes, thereby fulfilling the USAC requirements. USAC is expected to succeed High-Efficiency (HE) -AAC[*26] v.2 [1]. HE-AAC v.2 is currently used in music delivery services such as "Chaku-Uta®[*27]", video delivery services such as BeeTV™[*28], and mobile broadcasting services such as One Seg. USAC can also be applied to other similar services. We thereby are sure that USAC will contribute to further improving the audio qualities of mobile audio services.

REFERENCES

[1] ISO/IEC 14496-3:2009: "Information technology - Coding of audio-visual objects - Part 3: Audio," Aug. 2009.
[2] 3GPP TS.26.290 V. 6.3.0: "Audio codec processing functions; Extended Adaptive Multi-Rate - Wideband (AMR-WB+) codec; Transcoding functions," Jun. 2005.
[3] ISO/IEC 23003-1:2007: "Information technology - MPEG audio technologies - Part 1: MPEG Surround," Jan. 2007.
[4] ITU-R Recommendation BS.1534-1: "Method for the subjective assessment of intermediate quality level of coding systems," Jan. 2003.

*26 **HE-AAC**: A speech compression coding scheme that achieves the same quality as MPEG-4 AAC with half its bitrate which is an enhanced specification of MPEG-4 AAC.
*27 **ChakuUta®**: A registered trademark of Sony Music Entertainment Inc.
*28 **BeeTV™**: A trademark of Avex Entertainment Inc.