Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

Unification of Speech and Audio Coding Technologies for Mobile Devices

Taejin Lee (1), Kyeongok Kang (1) and Whan-Woo Kim (2)

(1) Electronics and Telecommunications Research Institute, Daejeon, KOREA
(2) Chungnam National University, Daejeon, KOREA

PACS: 43.60.-c Acoustic signal processing

ABSTRACT

As mobile devices become multi-functional, and multiple devices converge into a single device, there is a strong market need for an audio codec that is able to provide consistent quality for mixed speech and music content. In this paper, we propose unified speech and audio coding technology which could provide best quality for both speech and music contents. For the evaluation of the new codec architecture, we followed MPEG audio listening test procedures. Three categorized (music, speech, mixed) items are used and audio experts are joined for the evaluation. The listening test results show that the performance of new codec is statistically better than that state of the art speech and audio codec for each categorized item. The new codec architecture can be used for digital radio, mobile TV, audio books and so on, which need consistent quality for both speech and music signals.

INTRODUCTION

As mobile devices become multi-functional, and multiple devices converge into a single device, it is becoming prevalent for various types of content, including content that is a mix of speech and music, to be played on or streamed to mobile devices. Hence, there is a strong market need for a codec that is able to provide consistent quality for mixed speech and music contents and to do so with a quality that is better than that of codecs that are optimized for either speech content or music content [1].

To provide consistent quality for mixed speech and music contents, we designed new codec architecture based on two strategies.

Firstly, we reused tools from existing audio and speech codec and found the best combination. Instead of designing new coding tools for enhancing the efficiency, we investigate the performance of existing tools for reusing. Each tool is individually evaluated for its performance by appropriate evaluation procedure and found the best combination.

Secondly we revised each tools and developed new tools for harmonization of each tool. To enhance the performance, each tool included in the selected combination is revised. The selected combination should be suitably changeable according to the characteristics of input signals for the consistent harmonizing performance. To do this, signal analysis tool and harmonization tool are newly introduced.

In this paper, we will explain detailed information of new speech and audio codec architecture and subjective assessment results.

STATE OF THE ART AUDIO AND SPEECH CODEC

The HE-AAC (High-Efficiency Advanced Audio Coding) is a lossy audio coding technology for digital audio defined as a MPEG-4 Audio profile [2]. It is an extension of AAC (Advanced Audio Coding) optimized for low-bitrate applications. HE-AAC version 1 profile (HE-AAC v1) uses SBR (Spectral Band Replication) to enhance the compression efficiency in the frequency domain for higher frequency band [3]. HE-AAC version 2 profile (HE-AAC v2) couples SBR with PS (Parametric Stereo) to enhance the compression efficiency of stereo signals [4]. Scientific testing by the European Broadcasting Union has indicated that HE-AAC at 48kbit/s was ranked as "Excellent" quality using the MUSHRA scale. The MP3 in the same testing received a score less than half that of HE-AAC and was ranked "Poor" using the MUSHRA scale [5].

The state of the art speech codec, AMR-WB+ (Adaptive Multirate – Wide Band) is extention of AMR-WB to support higher sampling rates and stereo signals [6]. File storage of AMR-WB+ encoded audio is specified within the 3GPP container format, 3GPP-defined ISO-based multimedia file format. The AMR-WB+ codec has a wide bit-rate range, from 5.2–48 kbit/s. Mono rates are scalable from 5.2–36 kbit/s, and stereo rates are scalable from 6.2–48 kbit/s, reproducing bandwidth up to 20 kHz (approaching CD quality). Moreover, it provides backward compatibility with AMR wideband.

The figure 1 shows the sound quality of HE-AAC V2 and AMR-WB+ for speech and music stereo signal at 18kbps [7]. As we could see in this figure, even thought, HE-AAC V2 support high quality for music signal, there is a significant degradation of sound quality for speech signal. The AMR-

WB+ provide high quality for speech signal, but there is a degration of sound quality for music signals. Hence we designed new speech and audio codec architecture for consist sound quality both speech and music signals.

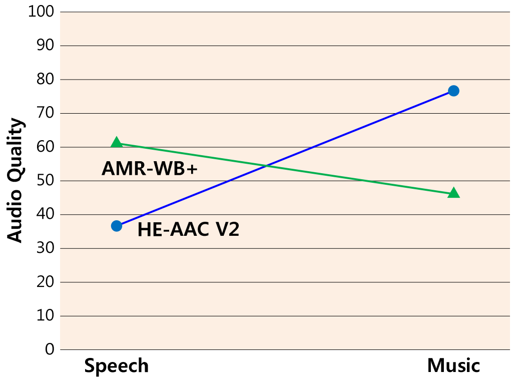

Figure 1. The sound quality of HE-AAC V2 and AMR-WB+ for speech and music signals [7].

THE NEW SPEECH AND AUDIO CODEC ARCHITECTURE

The new speech and audio codec architecture is a kind of hybrid coder, which combines MPEG technologies and time-domain speech coding technologies.

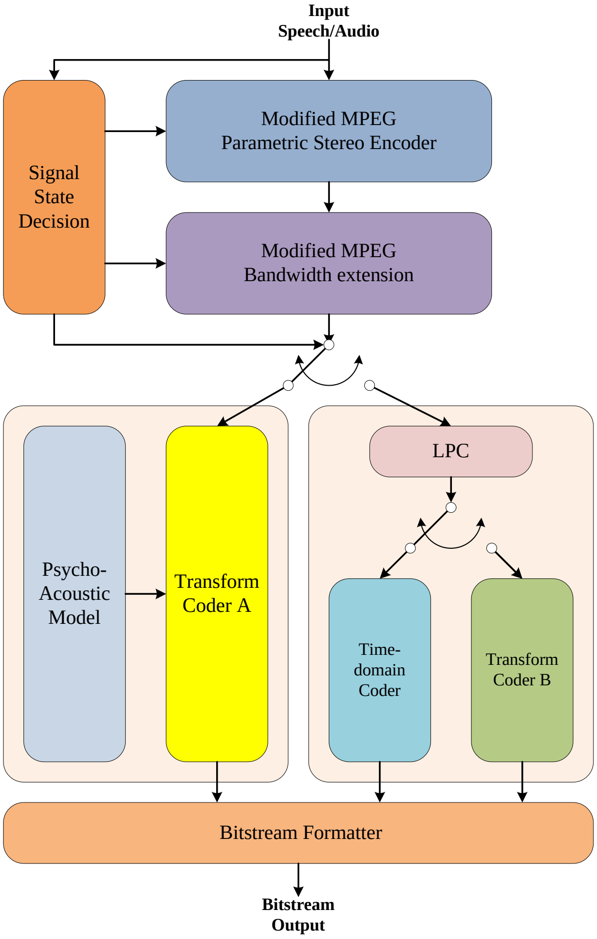

Figure 2. The new speech and audio codec encoder structure.

For the development of new speech and audio codec, we first evaluated state of the art audio codec HE-AAC V2 and state of the art speech codec AMR-WB+ in search of best combination. After that, tools such as MPEG PS, SBR, AAC and LPC-based residual coder (ACELP) were selected and combined in new codec. However, in prior to combination, each of selected tools had been modified for better performance.

Figure 2 shows the new speech and audio codec encoder structure based on combination of HE-AAC V2 and AMR-WB+, to provide consist quality both speech and music input signals.

Firstly, input frame is analyzed using Signal State Decision (SSD) module which is newly introduced in order to utilize the characteristics of input signals and categorizing input signal whether steady state signal or complex state signal. The SSD module analysis input signal and classify into Steady State Harmonic/Noise, Complex State Harmonic/Noise and Silence. After analysing previous and current state, SSD module determines core codec modes.

The second module, parametric stereo coding tool is reused after stereo image parameters modification to enhance the performance. The filterbank used in this tool is identical to the MPEG PS (HE-AAC V2).

The third module, bandwidth extension tool is adopted after modification for enhancing performance. For speech-like input signal, the noise floor is adaptively changed according to the characteristics of input signal. The figure 3 shows the performance of modified bandwidth extension tool compare with current standard (HE-AAC V2). As we could see in this figure, new bandwidth extension tool could provide better harmonic representation for speech-like signals.

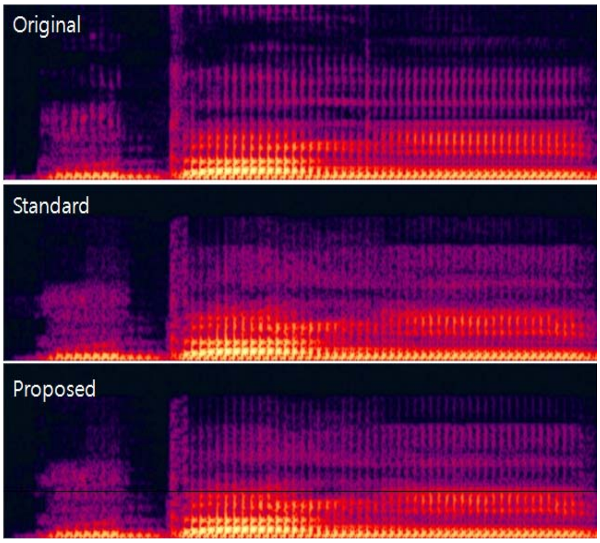

Figure 3. The performance comparison of modified bandwidth extension tool.

All modules are controlled by the output of SSD module. For instance, if input signals have speech-like intervals, the SSD module indicates this information to the other modules. In that case, the LPC-based core coding module is activated and encodes the input speech-like signals efficiently. In other case, e.g., music-like intervals, the Transform-based core coding module is activated and encodes the input music-like signals.

The coding scheme of LPC-based core coding module is based on the ‘LPC-residual coding’. It means that LPC prefiltering process is firstly applied into the input signals, and then the residual signal is efficiently coded by the ACELP based method or transform-based method. For the determination of coding mode in LPC-based core coding mudule, SNR based closed-loop mode decision is used. The input frame is devided into quarter sub-frames and each sub-frame is coded ACELP or Transform Coder, then each sub-frame SNR is compared and finally coding mode is decided.

Another core coding module for encoding the complex state signal, the transform based coding scheme can be adopted. In our proposed system, we adopted 'AAC' as a core band codec for music-like signals.

In our new speech and audio codec architecture, two different core codec is switched frame-by-frame based on input signal characteristics, so, it is important to harmonizing each different core codec. So for compensate the problem of block discontinuity or artifact between transform-based core coding module and LPC-based core coding module, we need windowing technique. We adopted adaptive windowing technology for smoothly change in LPC-based core coding module and used the side-information for transition between LPC-based core coding module and transform-based core coding modue. We used the TDAC (Time-Domain Aliasing Cancellation) parameters for side information.

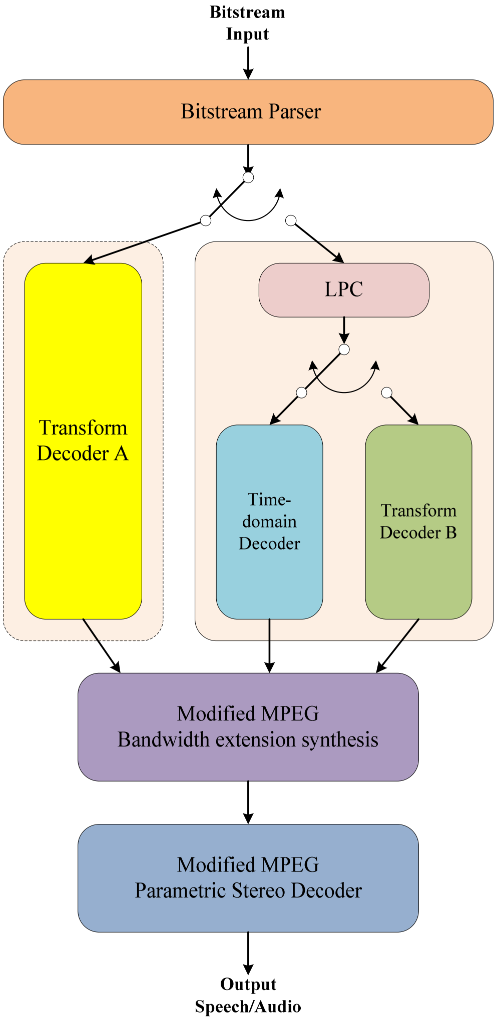

Figure 4. The new speech and audio codec decoder structure.

In decoder side, the procedure of decoding flow is reverse of encoding procedure depicted as in the figure 4. The Bitstream Parser module parses the input bitstream and sends demuxed bitstream into appropriate module for decoding. The Transform Decoder A decoding the transform coded bitstream at encoder for music-like input signals and LPC-based decoder decoding the LPC-based coded bitstream at encoder for speech-like input signals. The modified bandwidth extention module generate higher frequency region using side information and core band restored signals. If input signals were stereo, modified MPEG parametric stereo decoder module generate stereo information using side information and mono downmixed signals.

To evaluate the complexity of proposed system, we calculated the execution time of decoder. The description of the platform used in complexity evaluation is summarized in table 1.

Table 1. Description of the platform for complexity evaluation.

CPU	Intel Core 2 Duo CPU @ 2.4 GHz
RAM	2.0 GB
OS	Windows XP Service Pack 3

Table 2 shows the decoding time of 196 seconds input signal and CPU load. CPU load is calculated by (decoder execution time)/(duration of decoded signal) in %.

Table 2. Decoding time for each test.

Test Number	Elapsed time	CPU load
64kbps Stereo	9.02	4.60%
24kbps Stereo	8.69	4.43%
24kbps Mono	4.86	2.48%
20kbps Mono	4.76	2.43%

EVALUATION OF NEW SPEECH AND AUDIO CODEC

The basic motivation of our new speech and audio codec is to make bridge codec between the state of the art speech codec and audio codec, especially bellow 24 kbps. Moreover, our proposed architecture simply accommodate structure of the HE-AAC V2 at the high bitrate coding, and accommodate the structure of ACELP coder at the low bitrate coding, because SSD module and harmonization technique make it possible to flexible change the coding scheme without any artifacts.

For the evaluation of proposed new codec architecture, we adopted MUSHRA (Multiple Stimulus Hidden Reference and Anchor) test [8]. MUSHRA is a methodology for subjective evaluation of audio quality, to evaluate the perceived quality of the output from lossy audio compression algorithms. It is defined by ITU-R recommendation BS.1534-1[8]. The MUSHRA methodology is recommended for assessing "intermediate audio quality". 9 expert listeners are participated from ETRI and 15 items (5 music, speech, and mixed items) are used for evaluation.

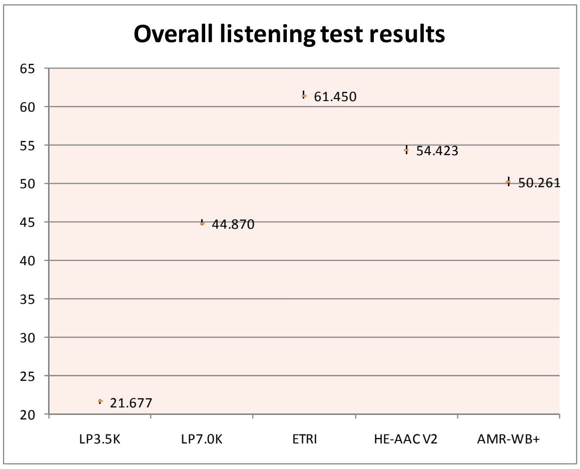

Figure 5. The listening test results: overall score.

The figure 5-8 show the performance of proposed system compared with HE-AAC V2 and AMR-WB+. The figure 5 shows the overall score for each system. In overall score comparison, our proposed system shows the statistically better performance compare with HE-AAC V2 and AMR-WB+.

The figure 6 shows the speech category listening test results for each system. In speech category lintening test, our proposed codec architecture shows the statistically better performance compare with other systems and AMR-WB+ shows the better performance compare with HE-AAC V2 .

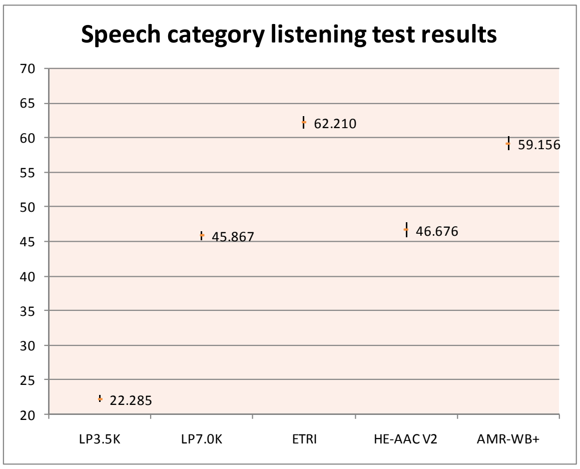

Figure 6. The listening test results: speech category.

The figure 7 shows the speech and music mixed category listening test results for each system. In mixed category lintening test, our proposed codec architecture shows the statistically better performance compare with other systems and HE-AAC V2 shows the better performance compare with AMR-WB+.

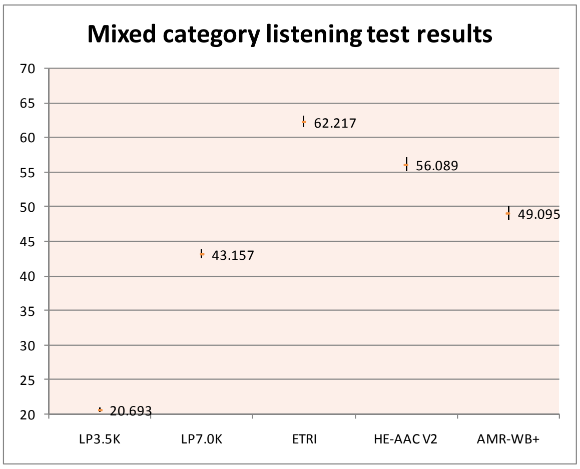

Figure 7. The listening test results: mixed category.

The figure 8 shows the music category listening test results for each system. In music category lintening test, our proposed codec architecture shows the statistically same performance compare with HE-AAC V2 and statiscically better performance compare with AMR-WB+. In music category listening test, HE-AAC V2 shows the better performance compare with AMR-WB+.

As we can see in these figures, proposed system shows statistically better performance for overall result, speech and mixed category results and shows statistically same performance for music category results.

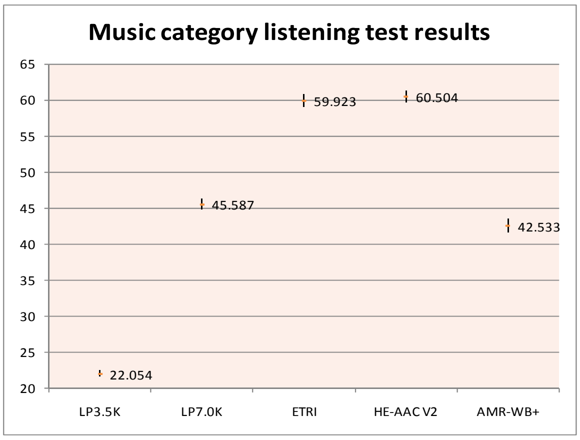

Figure 8. The listening test results: music category.

CONCLUSIONS

In this paper, we propose new codec architecture which can provide consist quality both speech and audio signal. We developed new speech and audio codec architecture based on two strategies. Firstly, we reuse of tools from existing audio codec and finding the best combination. Secondly we revised each tool and harmonizing the performance.

Based on above strategies, we first evaluated state of the art music codec HE-AAC V2 and state of the art speech codec AMR-WB+ in search of best combination. After that, tools such as MPEG PS, SBR, AAC and LPC-based residual coder (ACELP) were selected and combined in new codec architecture. However, in prior to the combination, each of selected tools had been modified for better performance.

The listening test results show that performance of new codec archtecture is statistically better for overall score, speech and mixed category and statistically the same for music category.

The new speech and audio codec architecture can be used for digital radio, mobile TV, audio books and so on, which need consistent quality for both speech and music signals.

ACKNOWLEDGEMENT

This work was supported by the IT R&D program of MKE/KCC/KEIT. [KI001932, Development of Next Generation DTV Core Technology]

REFERENCES

1 N9519, “Call for Proposals on Unified Speech and Audio Coding” *ISO/IEC* (2007)
2 ISO/IEC 14496-3, “Information technology – Coding of audio-visual objects – Part3: Audio, *ISO/IEC*(2005)
3 ISO/IEC 14496-3, “AMD. 1, Bandwidth Extension”, *ISO/IEC*(2003)
4 ISO/IEC 14496-3, “AMD. 2, Parametric Coding of High Quality Audio”, *ISO/IEC*(2004)
5 http://www.ebu.ch/CMSimages/en/tec_doc_t3296_tcm6-10497.pdf
6 3GPP TS 26.290, “Audio codec processing functions; Extended Adaptive Multi-Rate – Wideband (AMR-WB+) speech codec; General description,” *3GPP*(2004)
7 3GPP S4-040099, “Revised Global Analysis Laboratory Report on 3GPP Low-Rate Audio Codec Exercises”, *3GPP*(2004)
8 ITU-R recommendation BS.1534, “Method for the subjective assessment of Intermediate quality level of coding systems,” *ITU-R*(2003)