UNIFIED SPEECH AND AUDIO CODING SCHEME FOR HIGH QUALITY AT LOW BITRATES

M. Neuendorf[1], P. Gournay[2], M. Multrus[1], J. Lecomte[1], B. Bessette[2], R. Geiger[1], S. Bayer[1], G. Fuchs[1], J. Hilpert[1], N. Rettelbach[1], R. Salami[3], G. Schuller[4], R. Lefebvre[2], B. Grill[1]

[1]Fraunhofer IIS, Erlangen, Germany, [2]University of Sherbrooke, Sherbrooke, Canada, [3]VoiceAge Corp., Montréal, Canada, [4]Fraunhofer IDMT, Ilmenau, Germany

ABSTRACT

Traditionally, speech coding and audio coding were separate worlds. Based on different technical approaches and different assumptions about the source signal, neither of the two coding schemes could efficiently represent both speech and music at low bitrates. This paper presents a unified speech and audio codec, which efficiently combines techniques from both worlds. This results in a codec that exhibits consistently high quality for speech, music and mixed audio content. The paper gives an overview of the codec architecture and presents results of formal listening tests comparing this new codec with HE-AAC(v2) and AMR-WB+. This new codec forms the basis of the reference model in the ongoing MPEG standardization activity for Unified Speech and Audio Coding.

Index Terms— Audio coding, speech coding

1. INTRODUCTION

With the increasing number of portable and wireless devices, there is a growing demand for low bitrate audio codecs. In several applications, for example broadcasting, audiobooks and audio/video playback, the content can be varied and is not limited to speech or music only. Hence, a unified audio codec that can deal equally well with all types of audio content is highly desired.

Audio coding schemes, such as MPEG-4 High Efficiency AAC (HE-AAC) [1, 2], are advantageous in that they show a high perceived quality at low bitrates for music signals. However, the subband and transform-based models used in such audio coding schemes do not perform well on speech signals, i.e. they can not use a small bit budget as efficiently as linear predictive (LP) coders when encoding speech.

LP coding (or LPC), and in particular CELP coding, is well suited for representing speech at low bitrates. The excitation-filter paradigm in LP coders closely follows the speech production process. State-of-the-art speech coders include the 3GPP AMR-WB standard [3, 4], which can produce high quality wideband speech at less than 1 bit per sample. In general, speech coding schemes show a high quality for speech even at low bitrates, but show a poor quality for music.

Attempts to unify speech and audio coding were made by the 3GPP AMR-WB+ standard [5, 6]: The AMR-WB speech-coder was extended by selectable frequency domain coding and a stereo mode. In this way, the capability of coding music was significantly improved. But still, the AMR-WB+ audio coding model is not as optimal as HE-AAC(v2) for music signals.

In [7] a unified speech and audio codec was built by combining AMR-WB and HE-AAC. However, for speech signals the performance of AMR-WB was not preserved.

In this paper, a new coding model is presented, which retains all the advantages of state-of-the-art speech and audio codecs. Techniques from both HE-AAC and AMR-WB+ are combined in order to allow seamless switching between a more general music coding mode, and a speech-specific coding mode. Formal listening tests show that the resulting codec is for each signal category at least as good as the better of HE-AAC(v2) and AMR-WB+, and thus the goal of a unified speech and audio codec, as stated in [8], is reached.

2. STATE OF THE ART

2.1. HE-AAC(v2) and MPEG Surround

Frequency domain coding schemes such as AAC [1] are based on three main steps: (1) a time/frequency conversion; (2) a subsequent quantization stage, in which the quantization error is controlled using information from a psychoacoustic model; and (3) an encoding stage, in which the quantized spectral coefficients and corresponding side information are entropy-encoded using code tables. This results in a source-controlled, variable-rate codec which adapts to the input signal statistics as well as to the characteristics of human perception.

To further reduce the bitrate, HE-AAC combines an AAC core in the low frequency band with a parametric coding approach for the high frequency band (SBR) [2]. The high frequency band is reconstructed from replicated low frequency signal portions, controlled by parameter sets containing level, noise and tonality parameters.

Although HE-AAC has generic multi-channel capabilities, it can also be combined with a joint stereo or a multi-channel coding tool to further reduce the bitrate. The combination of "Parametric Stereo" [1, 9] and HE-AAC is known as HE-AACv2 and is capable of representing stereo signals by a mono downmix and corresponding sets of inter-channel level, phase and correlation parameters. By usage of "MPEG Surround" [9, 10] this principle is extended to transmit N audio input channels via M transmission channels (where $N \geq M$) and corresponding parameter sets.

2.2. AMR-WB and AMR-WB+

Efficient speech coding schemes, such as AMR-WB, typically have three major components: (1) a short-term linear prediction (LP) filter, which models the vocal tract; (2) a long-term prediction (LTP) filter, which models the periodicity in the excitation signal from the vocal chords; and (3) an innovation codebook, which essentially encodes the non-predictive part of the speech signal. In AMR-WB, the innovative codebook uses the ACELP model. In ACELP, a short block of excitation signal is encoded as a sparse set of pulses and associated gain for the block. The gain, signs and positions of the pulses are found in a closed-loop search (analysis-by-synthesis). The

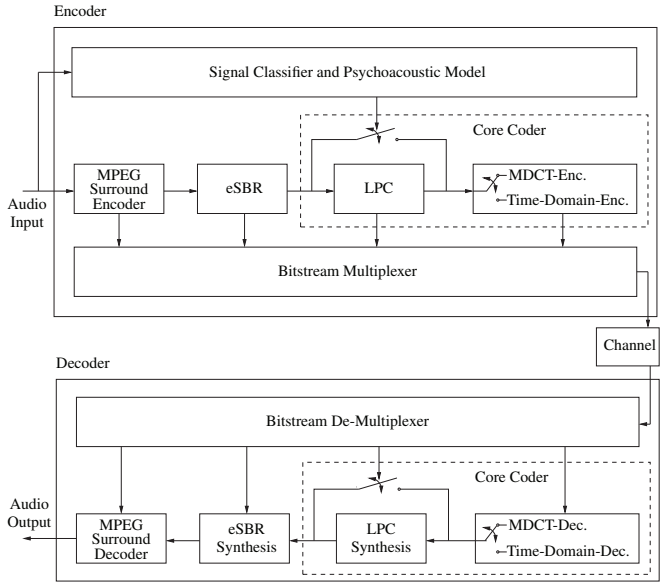


Fig. 1. Encoder and decoder of unified speech and audio codec

pulse codebook is not stored, but represented in lattice or algebraic form. The encoded parameters in a speech coder are thus: the LP filter, the LTP lag and gain, and the innovative excitation shape.

To properly encode music signals, in AMR-WB+ the time-domain speech coding modes were extended by a transform coding mode for the innovative excitation (TCX). The AMR-WB+ standard also has a low rate parametric high frequency extension as well as parametric stereo capabilities.

3. TECHNICAL APPROACH

3.1. System Overview

The proposed system consists of a hybrid audio coder, which combines the strengths of efficient MPEG audio coding technology, such as AAC, SBR and MPEG Surround with efficient speech coder technology. Figure 1 shows a block diagram of the encoder and decoder of the proposed system. The input audio, assumed to be stereo, is first processed by the MPEG Surround encoder, which produces the parametric stereo information to be transmitted, as well as a downmixed (mono) signal. The mono signal forms the input to an enhanced SBR module (eSBR). The eSBR module outputs the parametric information for high band regeneration at the decoder (SBR info), as well as the lower band signal. The crossover frequency between the low band (encoded with the core) and the high band (encoded with eSBR) depends on the desired average bitrate. The low band mono signal is then encoded in the remaining blocks. The use of these blocks is controlled by switches. Specifically, an LPC processing block can be activated or not. When this block is activated, the signal is encoded in a framework similar to that of AMR-WB+ and both MDCT encoding and time-domain encoding are allowed in the last block. Alternatively, when the LPC block is not activated, the signal is encoded in a framework similar to AAC and only the MDCT encoding functionality of the last block is allowed. Switching between these different modes is controlled by a signal classifier and psychoacoustic model which analyzes the input signal.

In the next subsections, the different modules of the proposed codec will be described.

3.2. Stereo Processing using MPEG Surround

In the proposed codec, stereo signals are processed using MPEG Surround technology with some modifications. When the audio input is stereo, a high quality mono downmix from the stereo input signal is first produced. Then, a set of spatial parameters is extracted from the two stereo channels. On the decoder-side, a stereo output signal is generated using the decoded mono downmix in combination with the extracted and transmitted spatial parameters. A low bitrate 2-1-2 mode has been added to the existing 5-x-5 or 7-x-7 operating points in MPEG Surround, using a simple tree structure that consists of a single OTT (one-to-two) box in the MPEG Surround upmix. Some of the components have received modifications to better adapt to the speech reproduction. For higher bitrates, such as 64 kbps stereo, the core coder is using discrete stereo coding (Mid/Side or L/R) and also allows discrete multichannel coding. Hence, MPEG Surround is only used at the lower bitrates.

3.3. Bandwidth Extension using eSBR

The bandwidth extension is based on MPEG SBR technology. The filter bank to separate the lower and higher bands is identical to the QMF filterbank in MPEG Surround and SBR. This allows sharing of QMF domain samples between MPEG Surround and SBR without additional analysis/synthesis steps. Compared to the standardized SBR tool, eSBR features a more flexible crossover frequency control, a signal-type-adaptive noise floor control, and a higher temporal resolution by using an increased number of envelopes per frame. A new phase-vocoder based harmonic frequency reconstruction module is included, which is better suited for very low bitrates and low crossover frequencies [11].

As known from the combination of SBR and AAC, this feature can be de-activated globally, leaving coding of the whole frequency range to the core coder.

3.4. LPC processing and Core Coder

As seen in Figure 1, the core coder part of the system can be seen as the combination of an optional LPC filter and a switchable frequency-domain/time-domain core coder. The LPC filter provides the basis for a source model for speech signals. The LPC processing can be enabled or disabled (bypassed) globally or on a frame-by-frame basis. It is enabled by the signal classifier and psychoacoustic model when the input signal exhibits speech characteristics.

The LPC filter residual is encoded (in the perceptual domain) using either a time-domain or transform-based frequency-domain approach, similar to AMR-WB+. The transform-based approach is similar to the TCX mode in AMR-WB+, with the following modifications: (1) non-critical sampling windows and FFT are replaced by critical sampling windows and MDCT; (2) lattice quantization of the spectral coefficients is replaced with scalar quantization followed by arithmetic coding of the quantized coefficients. The time-domain approach is based on the ACELP-technology of AMR-WB [4].

The LPC quantization uses a new scheme which takes advantage of the bit reservoir of the proposed codec. The 16 LPC coefficients can be quantized relative to adjacent LPC filters (using 0, 8 or 24 bits to quantize the interpolation difference). Alternatively, the 16 LPC coefficients can be quantized using a trained 3-stage absolute quantizer using 46 bits.

3.5. Mode Transitions and Windowing

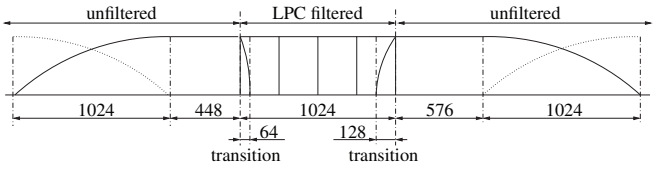
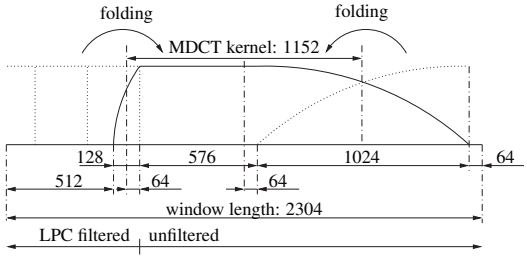


Fig. 2. Transition scheme between core coder modes

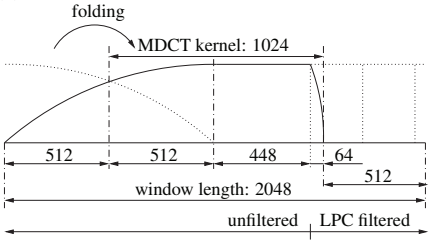
The main challenges for a unified speech and audio coder, which is based on a switched core codec, are firstly a fast adaptation to either speech, music or mixed content and secondly a smooth transition between the signal types. Transition regions can be prone to loss of coding efficiency and a cause for artifacts. Switching between speech, music and mixed signals is done by usage of LPC processing on one hand, or bypassing this filter on the other hand on an AAC frame basis (1024 samples). When switching between these two coding modes, several issues have to be addressed: (1) transition between regions with time-domain-aliasing (TDA)[12], i.e. MDCT domain (AAC-like), and time-domain, i.e. LPC filtered domain (AMR-WB+ like), (2) no blocking artifacts, (3) minimum transition overhead and (4) constant framing. To solve these issues, new transition windows for the AAC-like coding mode were designed, since the AAC-like coding mode is more flexible in terms of bit allocation and number of coefficients to transmit than the AMR-WB+ like coding mode. Figure 2 presents the scheme for transitions between unfiltered and LPC filtered domain within 3 successive frames. These windows are similar to regular AAC transition windows, with either Kaiser-Bessel-Derived (KBD) or sine window slope on the side facing the unfiltered signal portion. Further on, these windows consist of a flat top region of 576 or 448 samples resp., followed by a sine window slope of size 128 or 64 samples resp. and a consecutive number of zero samples, completing the windows in Figure 3 (a) or (b) resp. For the transition from LPC to unfiltered domain, the right side of the window (Figure 3 (a)) is completed with 64 zero samples. Here, four subframes of size 256 samples form one LPC processed block. This corresponds to four ACELP frames.

Figure 3 (a) shows the transition from LPC filtered to unfiltered domain. To fulfill the requirement on constant framing, the window size had to be enlarged from 2048 to 2304 samples. In the overlap region of size 128, TDA is introduced by folding the window signal before the MDCT kernel. In this way, only 64 samples overhead are introduced. In order to guarantee a correct TDA cancellation in this overlap region, appropriate aliasing components have to be imposed on the decoded LPC processed signal first. Due to the increased window size, an MDCT of kernel length 1152 is used.

In contrast to that, for the transition between unfiltered and LPC filtered domain, a crossfade of length 64 samples without TDA is used, as show in Figure 3 (b). Due to the start-up of the internal filters in the LPC, the first reconstructed signal samples are in general inaccurate. Folding these samples as in Figure 3 (a) would propagate this error.



(a) transition LPC filtered to unfiltered domain



(b) transition unfiltered to LPC filtered domain

Fig. 3. Transition windows for the transitions (a) LPC filtered to unfiltered domain and (b) unfiltered to LPC filtered domain

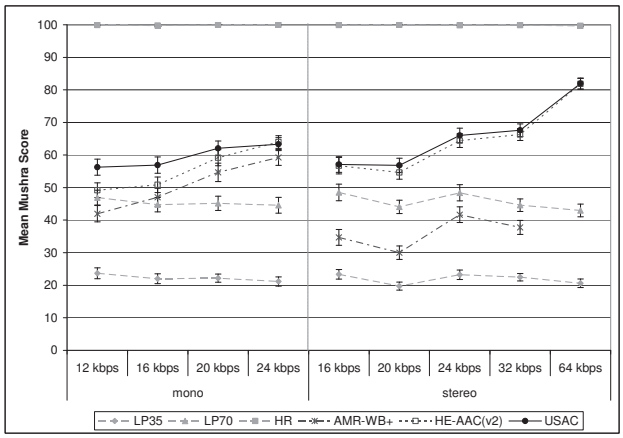
4. RESULTS

To assess the performance of the proposed codec, formal listening tests for various operating modes were carried out. Audio items covered three categories: speech, music and mixed speech/music signals. Twelve audio items were used, with four items in each of the three categories. The items were encoded using AMR-WB+, HE-AAC(v2) and the proposed codec (USAC). Additionally, a hidden reference (HR) and two lowpass anchors with a bandwidth of 3.5 kHz (LP35) and 7.0 kHz (LP70) were included in the tests. For a bitrate of 64 kbps, AMR-WB+ was not included in the test, since this codec does not support this operating mode. The tests were carried out following the MUSHRA methodology (ITU-R BS.1543-1). Per test, a minimum number of 39 experienced listeners participated.

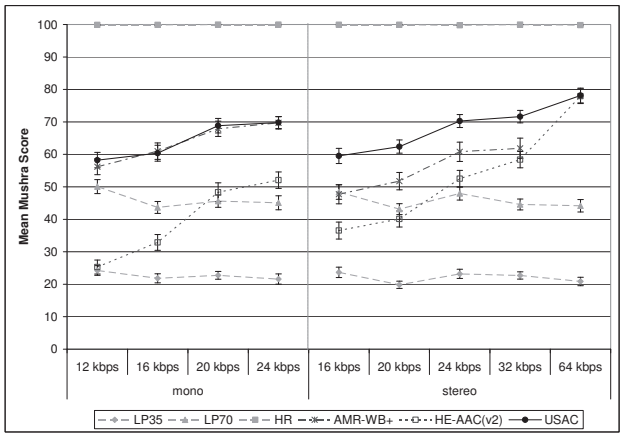
In total 9 separate listening tests were carried out: four tests in mono at 12 to 24 kbps, and five tests in stereo at 16 to 64 kbps. The test results are summarized in Figures 4 (a) - (c).

The first observation is that the reference codecs, namely AMR-WB+ and HE-AAC(v2), exhibit uneven performance depending on the operating mode and the audio content type. For music signals, HE-AAC(v2) performs significantly better than AMR-WB+ for 7 of 8 operating modes, with an enlarging gap when comparing stereo results. Conversely, for speech content AMR-WB+ performs significantly better than HE-AAC(v2) for 7 of 8 operating modes. For the mixed category, AMR-WB+ shows a better performance for lower mono bitrates, while HE-AAC(v2) is performing significantly better for all stereo modes. Hence, neither AMR-WB+ nor HE-AAC(v2) exhibit the consistent performance that would be required from a unified speech/audio codec.

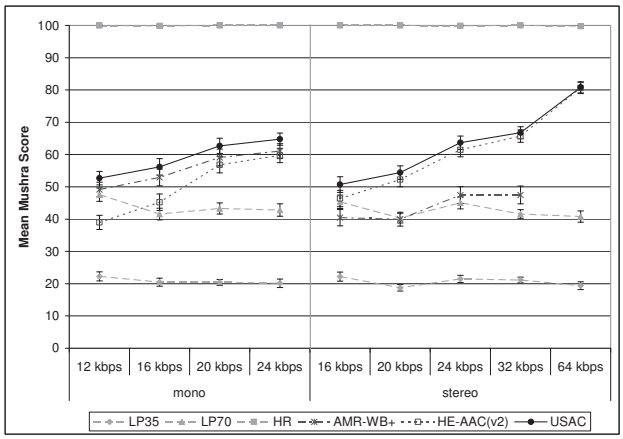
Comparing these results with the performance of the proposed unified codec architecture, it can be seen from Figure 4 that for each operating mode, and for each content type, the proposed architecture performs at least as good as the better of AMR-WB+ and HE-AAC(v2). Additionally, there are many cases where it performs significantly better than AMR-WB+ and HE-AAC(v2). This is the case for music at 12 and 16 kbps mono, and for speech at 16 - 32 kbps stereo. Scaling towards higher datarates (above 64 kbps), the performance of HE-AAC(v2) is expected to be at least matched, since the codec is able to fall back in a mode very similar to HE-AAC(v2).



(a) music content



(b) speech content



(c) mixed content

Fig. 4. Listening tests results for (a) music, (b) speech and (c) mixed content, at 12 to 24 kbps mono and 16 to 64 kbps stereo; mean values and according confidence intervals (95% level of significance).

5. CONCLUSION

In this paper, a new approach for unified speech and audio coding was presented. The new codec is based upon state-of-the-art speech and audio coding technology, such as AMR-WB+, HE-AAC and MPEG Surround and contains several enhancements of these components. This results in a codec which is capable of delivering state-of-the-art quality for speech, music and mixed content. For a bitrate range of 12 kbps to 64 kbps listening tests have shown that the codec performs at least as good as the better of AMR-WB+ and HE-AAC(v2). Operated at higher rates, the codec converges to AAC and reaches transparent quality.

6. ACKNOWLEDGMENT

The authors would like to thank the following people for their valuable contribution to this project: S. Disch, B. Edler, J. Herre, J. Hirschfeld, U. Krämer, J. Lapierre, F. Nagel, M. Neusinger, J. Robillard, C. Spenger, M. L. Valéro, S. Wabnik, Y. Yokotani.

7. REFERENCES

[1] ISO/IEC 14496-3:2005, "Coding of Audio-Visual Objects, Part 3: Audio," 2005.

[2] M. Wolters, K. Kjörling, D. Homm, and H. Purnhagen, "A closer look into MPEG-4 High Efficiency AAC," in *115th AES Convention*, New York, NY, USA, Oct. 2003, preprint 5871.

[3] 3GPP, "Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General description," 2002, 3GPP TS 26.171.

[4] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Jarvinen, "The adaptive multirate wideband speech codec (AMR-WB)," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 8, pp. 620–636, Nov. 2002.

[5] 3GPP, "Audio codec processing functions; Extended Adaptive Multi-Rate - Wideband (AMR-WB+) codec; Transcoding functions," 2004, 3GPP TS 26.290.

[6] J. Makinen, B. Bessette, S. Bruhn, P. Ojala, R. Salami, and A. Taleb, "AMR-WB+: a new audio coding standard for 3rd generation mobile audio services," in *Proc. IEEE ICASSP'05*, March 2005, vol. 2, pp. 1109–1112.

[7] Sang-Wook Shin, Chang-Heon Lee, Hyen-O Oh, and Hong-Goo Kang, "Designing a unified speech/audio codec by adopting a single channel harmonic source separation module," in *Proc. IEEE ICASSP'08*, Las Vegas, USA, 2008.

[8] ISO/IEC JTC1/SC29/WG11, "Call for Proposals on Unified Speech and Audio Coding," Shenzhen, China, Oct. 2007, MPEG2007/N9519.

[9] J. Breebaart and C. Faller, *Spatial Audio Processing: MPEG Surround and Other Applications*, John Wiley & Sons Ltd, West Sussex, England, 2007.

[10] ISO/IEC FCD 23003-1, "MPEG-D (MPEG audio technologies), Part 1: MPEG Surround," 2006.

[11] Frederik Nagel and Sascha Disch, "A Harmonic Bandwidth Extension Method for Audio Codecs," ICASSP'09, Taipei, Taiwan, 2009.

[12] John P. Princen and Alan B. Bradley, "Analysis/Synthesis Filter Bank Design Based on Time Domain Aliasing Cancellation," *IEEE Trans. ASSP*, vol. 34, no. 5, pp. 1153–1161, 1986.