Audio Engineering Society

Convention Paper 7713

Presented at the 126th Convention
2009 May 7–10 Munich, Germany

A Novel Scheme for Low Bitrate Unified Speech and Audio Coding - MPEG RM0

Max Neuendorf[1], Philippe Gournay[2], Markus Multrus[1], Jérémie Lecomte[1], Bruno Bessette[2], Ralf Geiger[1], Stefan Bayer[1], Guillaume Fuchs[1], Johannes Hilpert[1], Nikolaus Rettelbach[1], Frederik Nagel[1], Julien Robilliard[1], Redwan Salami[3], Gerald Schuller[4], Roch Lefebvre[2], Bernhard Grill[1]

[1] *Fraunhofer IIS, Erlangen, Germany*

[2] *Université de Sherbooke, Sherbrooke, Canada*

[3] *VoiceAge Corporation, Montreal, Canada*

[4] *Fraunhofer IDMT, Ilmenau, Germany*

Correspondence should be addressed to Max Neuendorf (amm-info@iis.fraunhofer.de)

ABSTRACT

Coding of speech signals at low bitrates, such as 16 kbps, has to rely on an efficient speech reproduction model to achieve reasonable speech quality. However, for audio signals not fitting to the model this approach generally fails. On the other hand, generic audio codecs, designed to handle any kind of audio signal, tend to show unsatisfactory results for speech signals, especially at low bitrates. To overcome this, a process was initiated by ISO/MPEG, aiming to standardize a new codec with consistent high quality for speech, music and mixed content over a broad range of bitrates. After a formal listening test evaluating several proposals MPEG has selected the best performing codec as the reference model for the standardization process. This paper describes this codec in detail and shows that the new reference model reaches the goal of consistent high quality for all signal types.

1. INTRODUCTION

Nowadays, a unified audio coding scheme dealing equally well with all content types at low bitrates is highly desired. Over the last years the market of wireless and portable devices has grown significantly. Since the capacity of the transmission channels is still limited and the number of applications ever increasing, low bitrate audio coding faces a growing

demand. Furthermore, in many application areas, such as broadcasting, audio books, and audio/video playback, the content is highly variable and is not restricted to speech or music only.

Traditional audio coding schemes were not designed to code both speech and music at low bitrates equally well. On the one hand, general audio coding schemes such as MPEG-4 HE-AAC(v2) [1] show a high perceived quality for music signals at low bitrates. Such schemes are usually built based on an information sink model, primarily exploiting the characteristics of the human hearing in the frequency domain. Typical implementations use a transform or subband-based approach. However, for speech signals this approach cannot make use of a given bit budget as efficiently as dedicated speech coders. On the other hand, codecs using Linear Predictive Coding (LPC), and in particular a CELP coding scheme, are built based on an information source model. The excitation filter paradigm is based on a physical model of the speech production process of the human glottis and vocal tract, and closely fits the characteristics of a speech signal. State of the art speech coders, such as 3GPP AMR-WB [2], can thus reproduce human speech signals very efficiently at low bitrates, but fail on general audio signals.

Past attempts to unify both coding schemes did not manage to meet the market demand in every aspect. In the 3GPP AMR-WB+ standard [3], an AMR-WB speech coder was extended by selectable frequency domain coding, parametric bandwidth extension, and parametric stereo coding. In this way, the capability of coding music was improved significantly. Nevertheless, the perceived quality for music signals was still inferior to HE-AAC(v2).

An MPEG standardization process with the working title of "MPEG-D Unified Speech and Audio Coding" (USAC) has been started in October 2007 [4], resulting in a first reference model architecture (RM0) in October 2008, which already combines all advantages of state of the art speech and general audio coders.

The new codec architecture combines techniques from both HE-AAC and AMR-WB+ by means of switching between the core coders of the two standards. The high performance of this switched coding scheme arises from the intelligent interaction between the two coding paradigms, controlled by a signal classification module. Special attention was paid to the transitions between the two core coders. Efficient transition windows, reset and start-up procedures were introduced. Moreover, the two coding schemes share common tools, namely a parametric stereo coding scheme based on MPEG Surround and an enhanced spectral band replication, which harmonize the final rendering of the synthesis and avoid typical switching artefacts.

This paper will describe in detail the switched core coder and all innovative technology, such as a harmonic bandwidth extension scheme, an improved LPC filter quantization scheme, a weighted Linear Predictive Transform coding (wLPT), a time-warping functionality, an enhanced entropy coding for spectral coefficients, and an additional noise-filling tool. Additionally, subjective tests including the selected reference model, three anchors (hidden reference and two low-pass filtered references), and two reference systems (HE-AAC(v2) and AMR-WB+) are presented.

2. STATE OF THE ART

2.1. HE-AAC(v2) and MPEG Surround

Modern general audio coding schemes such as AAC [1] are based on the following generic structure: (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 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 bandwidth extension scheme: Spectral Band Replication (SBR) [5] reconstructs the high frequency content by replicating the 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, 6]

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" [6, 7] this principle is extended to transmit N audio input channels via M transmission channels (where $N \geq M$) and corresponding parameter sets as side information.

2.2. AMR-WB and AMR-WB+

Efficient speech coding schemes, such as AMR-WB, typically feature the following major components: (1) a short-term linear prediction filter (LPC filter), which models the spectral envelope; (2) a long-term prediction (LTP) filter, which models the periodicity in the excitation signal; and (3) an innovation codebook, which essentially encodes the non-predictive part of the speech signal. In AMR-WB, the ACELP algorithm is used to model the innovative codebook. 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 pulse codebook is not stored, but represented in algebraic form. The encoded parameters in a speech coder are thus: the LPC 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.

2.3. Switched Coding

Switching between two or more coding schemes is the most obvious and straightforward way to deal robustly with a diverse range of audio types. In the past such methods were already adopted in different attempts to unify speech and audio coding [8, 9, 10, 11, 12]. The switched coding paradigm consists of appropriately selecting the coding mode, which is designed to perform best for a certain category of signal. However, the switching approach faces several problems. First, dissimilarities between the coding modes may compromise the quality in the transitions from one mode to another. Changeovers are prone to produce overhead information and may lead to a loss of efficiency and be a cause for blocking effects. In addition, juxtaposition of different types of distortion can emphasize switching artifacts. The second problem arises from the switching decision which can be taken by a closed-loop or an open-loop approach. The closed-loop approach runs an exhaustive search by encoding and decoding the signal for each mode and comparing their performance. It is especially appropriate if the coding modes share certain properties, like the same objective function to optimize. For instance in AMR-WB+, both TCX and ACELP minimize the quantized error in the same weighted domain. The same objective is then considered by the closed-loop decision. The open-loop approach is usually less complex. Moreover, it may be the only practical solution in cases when dealing with too complex memory management or when generating too much delay to produce a local synthesis. The open-loop decision can be based on internal states of the codec [10] or make use of a previous signal analysis [9, 12]. The input is then classified as speech, music or other categories. The coding quality relies greatly on the classification and may be very sensitive to misclassification.

3. TECHNICAL APPROACH

3.1. General Overview

Figures 1 and 2 give an overview over the most important parts of the system. For clarity the most essential processing blocks have been grouped by shaded areas.

The coder consists of a common preprocessing block which precedes the following mutually exclusive core coder blocks. These are the frequency domain (FD) coder on one hand and the linear prediction domain (LPD) coder on the other hand. Both core coder blocks share a common quantization and coding of MDCT-spectral lines, including a noise filling tool used to compensate excessive spectral holes in the quantized spectra. The choice of core coder depends on the output of a signal classifier, discriminating signals based on statistical properties of extracted features. Non-speech material is directed to the FD coder, while speech-like signals are handled by the LPD coder.

The common preprocessing stage consists of a parametric stereo and an artificial bandwidth extension.

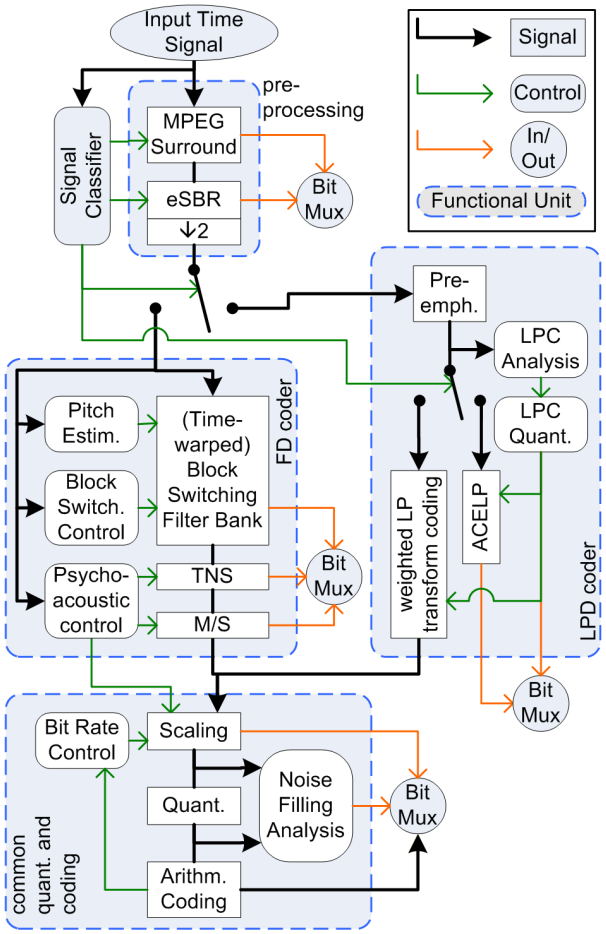

Fig. 1: Overview over the RM0 encoder

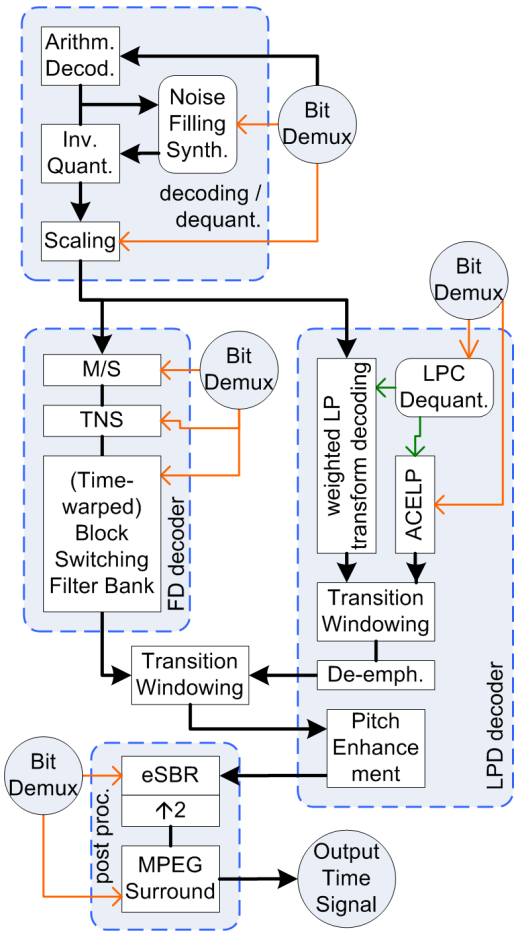

Fig. 2: Overview over the RM0 decoder

The parametric stereo uses a specifically designed 2-1-2 mode of the standard MPEG Surround and outputs a downmixed mono signal along with a set of coded parameters. The bandwidth extension is an enhanced version of the well-known spectral band replication (SBR) from e. g. HE-AAC. The original SBR was extended in order to work within the switched coding structure. Additionally, improvements were brought for getting better rendering when dealing with speech and when scaling the coding towards low bitrates.

Based on the output of the signal classifier, a first switch toggles between the two core coders, i. e. FD and LPD coder. This switching is done at granularity of one frame, i. e. 1024 samples.

For non-speech signals, like music, the FD coder is selected where coding techniques derived from the legacy AAC coder are adopted. Indeed, several tools like block switching control, temporal noise shaping (TNS) and M/S stereo coding are directly inherited from AAC. However, the filter bank is designed to be more flexible. The modified discrete cosine transform (MDCT) can now be warped in time domain. By means of a continuous pitch estimation, the warped MDCT can adapt its spectral representation for a better energy compaction. Moreover, the range of allowed windows has been amended especially for efficient handling of transitions between the FD coder and the LPD coder. Compared to AAC, the FD coder includes a new noise filling tool and uses a more efficient entropy coding namely a

context adaptive arithmetic coder.

For speech-like signals, the core coding mode is switched to the LPD coder which is inspired by AMR-WB+. The signal is coded in the linear prediction domain after an LPC analysis. A second switch permits to separate speech segments from non-speech segments more accurately because it operates with a higher granularity of 256 samples. While staying in the LPD, this switching, which resembles AMR-WB+ behavior, is highly signal adaptive and flexible. It can very promptly change back and forth from a time domain coding to a frequency domain coding, which is very convenient for mixed content. Segments fitting the speech production model are coded in the time domain by the algebraic code excitation linear prediction (ACELP) coding. Alternatively the frequency domain coding is used for remaining segments by processing the weighted LPC residual as it is also done by the transform coded excitation (TCX) in AMR-WB+. Furthermore, this weighted Linear Predictive Transform coding (wLPT) can adapt itself to the signal by selecting amongst three different transform lengths of 256, 512 and 1024 samples. As the FD coding branch, wLPT uses an MDCT and the subsequent quantization and coding chain.

3.2. Stereo Coding with MPEG Surround 2-1-2 Mode

MPEG Surround captures the spatial image of a multi-channel audio signal into a compact set of time- and frequency-variant parameters that can be used to synthesize a high quality, multi-channel representation from the transmitted downmix signal. The parametric representation of human's auditory cues for spatial perception results in a significant compression gain over conventional multi-channel audio codecs [6]. The MPEG Surround 2-1-2 mode is a natural subset of the standardized multi-channel MPEG Surround technology. This mode operates on a stereo input signal (encoder side) and generates a high quality mono downmix, together with a set of corresponding parameters such as inter-channel level differences and inter-channel correlation. On decoder side, a stereo output signal is generated using the processed mono downmix in combination with the transmitted parameters. A few modifications were applied to the system, compared to the well known 5-x-5 or 7-x-7 operating points [7]. The bitstream syntax was first modified in order to match exactly the requirements of low bitrate stereo coding. An optimized set of filter coefficients for the decorrelators was then specifically introduced for the 2-1-2 mode. It is meant to produce optimal results for stereo reproduction, also via headphones.

3.3. SBR Enhancements

The standard SBR as defined in [1] was enhanced and extended. Three major changes were carried out regarding firstly an adaptive crossover frequency, secondly an increased adaptive time resolution, and thirdly a modified patching algorithm.

The former two modifications aim at increasing the flexibility of SBR. The range over which the crossover frequency may be varied was increased, now allowing a variation over almost the entire SBR frequency range. A better time resolution was achieved by doubling the number of envelopes per SBR frame compared to what was previously allowed in the HE-AAC standard, thus facilitating a tighter adaptation to energy fluctuations in the replicated frequency bands.

The third modification affects the patching inside the quadrature mirror filterbanks (QMF). The standard patching algorithm in SBR copies low frequency bands to high frequencies within a QMF representation [5, 13, 14]. For certain signals this is replaced by a phase vocoder driven patching algorithm [15]. This algorithm intrinsically preserves the harmonic structure of the bandwidth extended signal.

Some items coded with SBR occasionally suffer from unwanted auditory artifacts such as unpleasant timbre or roughness: First, the harmonic structure of the signal is neglected and thus tonal peaks can be placed in the close vicinity of each other by simple patching. This can lead to amplitude modulation which is perceived as roughness [16]. Moreover, densely spaced tonal peaks that lie in separate critical bands in the low frequency part are transposed into one single critical band in the high frequency part. This is due to the fact that the width of critical bands increases with frequency. Since the phase vocoder stretches the spectrum, this problem is also counteracted.

As the new patching preserves harmonic relation, it was named harmonic bandwidth extension (HBE).

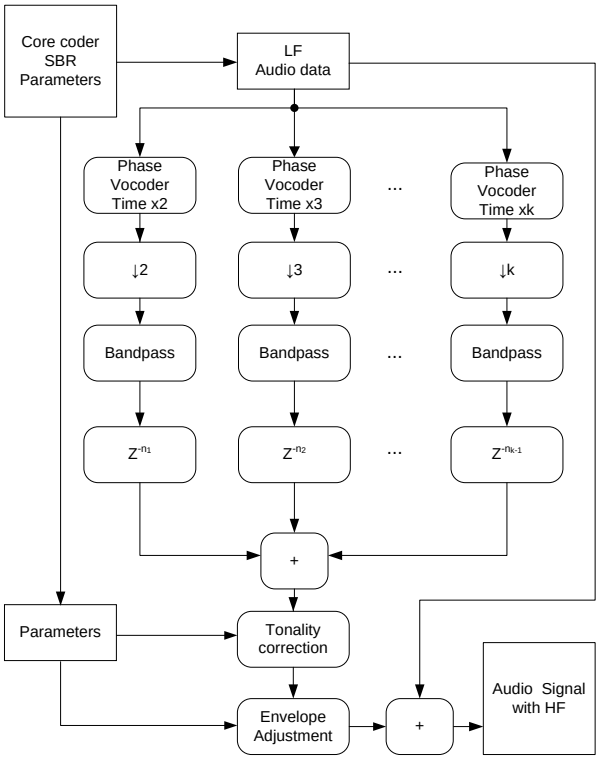

Fig. 3: Overview over the HBE bandwidth extension.

Its structure is illustrated in Figure 3.

The advantage of HBE over standard SBR was shown in a listening test in which only the patching algorithm of SBR was replaced by HBE [17]. All other SBR tools remained unchanged in that experiment.

3.4. Noise Filling

Modern audio coding schemes increase coding efficiency by replacing noise like signal parts with a parametric representation. In MPEG-4 AAC, the Perceptual Noise Substitution Tool (PNS) [18] implements this functionality. Instead of transmitting the quantized spectral lines of a noise like signal part, only its energy is transmitted. In the decoder the missing information is substituted by random values scaled to match the transmitted energy level.

The newly introduced Noise Filling Tool maintains this functionality, but at the same time addresses another problem of low bitrate coding: the coarse quantization of wide scale factor bands might generate artificial tonality, if only few spectral lines in a scale factor band remain after quantization. This increases the perceived tonality of the decoded signal.

While PNS can be applied to complete scale factor bands only, the new scheme is not limited to the entire band and enables an in-band "mix" of transmitted spectral lines and artificially generated noise. In the decoder, zero lines are replaced by a value representing the mean quantization error in the quantized domain. For the whole spectrum, only one quantization error value is calculated in the encoder and transmitted to the decoder. The scale factors are used to shape the energy of the artificially generated noise. In contrast to PNS, only a small amount of extra side information needs to be transmitted.

3.5. LPC Filter Quantization

When the codec operates in its LPD core coding mode, four LPC filters (LPC1 to LPC4) are estimated for each frame. At transitions from FD coding to LPD coding, an additional LPC filter LPC0, which corresponds to an LPC analysis centered at the end of the previous frame, is also estimated. This set of LPC filters is converted into the immitance spectral frequency (ISF) domain and then quantized.

Each LPC filter is quantized using either an absolute (48 bits) or one of three differential (0 bits, 8 bits or 24 bits) quantization modes, all based on vector quantization with trained codebooks. LPC4, which is always transmitted regardless of the ACELP/wLPT mode selection, is only quantized using the absolute mode. The other LPC filters are quantized using either the absolute or the differential modes. The reference used for differential quantization depends on the position of the LPC filter within the frame, as shown in Figure 4: LPC2 uses the quantized LPC4 as a reference, LPC1 uses the quantized LPC2 as a reference, LPC3 uses the average between the quantized LPC4 and the quantized LPC2 as a reference, and the optional LPC0 uses the quantized LPC4 as a reference. For each LPC filter, the decision between absolute and differential quantization is a tradeoff between spectral distortion and bitrate.

Compared to the fixed rate (46 bits per LPC filter) 1st order moving-average LPC filter quantizer

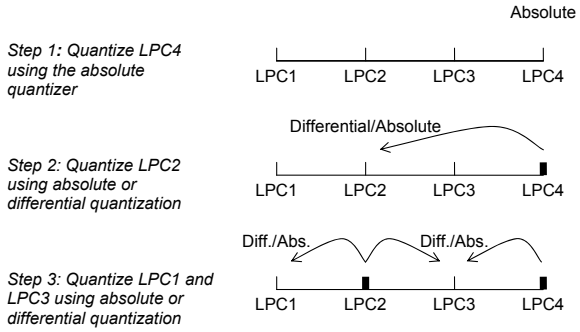

Fig. 4: Principle of the absolute/differential LPC filter quantization approach.

used by AMR-WB+, this scheme presents several advantages. Firstly, since LPC4 is always quantized using the absolute mode, the LPC filter quantization here does not introduce any inter-frame dependency. This makes transitions from FD coding to LPD coding easier to handle, and is also potentially more robust to transmission errors. Secondly, the intra-frame dependencies introduced by differential quantization are such that the entire set of filter parameters can be quantized before performing the ACELP/wLPT mode selection procedure. Once this selection is finalized, the quantization indices corresponding to the unnecessary LPC filters are simply skipped from the transmission with no effect on the way the other filters are transmitted or decoded at the receiver. This is unlike in AMR-WB+, where some LPC filters had to be encoded several times until the entire frame was encoded.

3.6. Weighted Linear Predictive Transform Coding

The weighted Linear Predictive Transform coding (wLPT) is an enhanced version of transform-based coding (TCX) used in AMR-WB+. The original TCX is built around a Discrete Fourier Transform (DFT) and suffers from some limitations. In particular during the transitions, TCX produces significant overhead information and rather abrupt changeovers. For allowing high granularity switching between the coding modes, low overlapping windows are applied. In fact, three different transformation sizes can be chosen from 256, 512 and 1024 samples. An additional overlap with the next block of 32, 64 or 128 samples is appended respectively. Because of the non-critical sampling of the DFT, the produced overhead information represents 1/8th of the total bitrate allocated to TCX.

Replacing the DFT by an MDCT permits to go towards critical sampling and furthermore, to improve the frequency response and get smoother transitions. The wLPT is critically sampled when staying in wLPT mode. As a result, longer and smoother overlaps of 128 samples are adopted without compromising the coding efficiency. The only remaining non-critical sampling occurs during transitions from wLPT to ACELP or AAC, where the 64 time-domain aliased samples are simply discarded.

Furthermore, the original Lattice Vector Quantization (LVQ) was replaced advantageously by an entropy-based scalar quantization. The transformed and quantized coefficients are processed by an efficient context-adaptive arithmetic coder described in the next section.

3.7. Context-Adaptive Arithmetic Coder

The quantized spectral values from the FD coding branch and from the wLPT part of the LPD coding branch are both fed to the context-adaptive arithmetic coder. The entropy coder works on 4-tuples $q(n, m)$, i. e. quantized spectral coefficients which are neighbored in frequency and coming from the same transformation block of time-index n. The 4-tuples are further decomposed into bit planes. The two most significant signed bit planes form a symbol which is encoded using probabilities derived from the surrounding context as proposed in [19]. The remaining least significant planes are encoded using a uniform probability distribution assumption. The symbols coming from the two most significant signed bit planes and the remaining planes are fed into an arithmetic coder with their respective probabilities. A binary code is generated by mapping the probability interval, where the set of symbols lies, to a variable length codeword [20].

The context is calculated synchronously at both encoder and decoder sides considering already coded 4-tuples previous in time and/or in frequency. As illustrated in Figure 5, four 4-tuples lying in the direct neighborhood of the current 4-tuple determine the context. The obtained context is then mapped

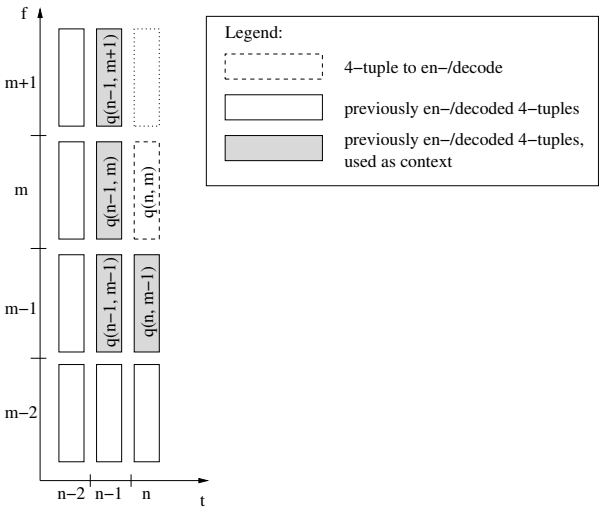

Fig. 5: Context-adaptive Arithmetic Coder: Neighbored 4-tuples are taken into account for the context.

to one of the probability models generated during a training phase.

Because of the transformation size adaptation occurring both in the FD part and during the wLPT coding, the context can gather 4-tuples of heterogeneous time/frequency resolutions. Therefore, a resolution conversion is necessary for keeping the advantage of the context adaptation. In required cases, the 4-tuples previous in time are resampled before using them in the context calculation. The resampling procedure is defined as follows:

$$q_r(n-1,m) = q(n-1, \lfloor m/r_f \rfloor) \quad \forall m \in \{0 \ldots \frac{N(n)}{4} - 1\}$$

where the resampling factor r_f is defined as the ratio between the current transform length $N(n)$ over the last transform length $N(n-1)$:

$$r_f = \frac{N(n)}{N(n-1)}$$

3.8. Time-Warped MDCT

For the transform coding branch, a new time-warped MDCT filterbank [21] (TW-MDCT) is introduced. One important property of the MDCT is the energy compaction for tonal signals. This means that ideally for strongly tonal or harmonic signals only a few peaks representing the partial tones of the signal would remain in the MDCT spectrum. Unfortunately this does not hold for tonal signals with a rapidly varying fundamental frequency. The varying frequency leads to a smearing of the partial tones within the MDCT spectra and thus to a loss of coding efficiency.

To overcome this problem, a time varying resampling corresponding to the frequency variation of the signal is applied locally within every block. The resampling ideally leads to a constant pitch within the block and therefore removes the smearing of the partial tones. Because the sampling rate in the warped time domain can differ for the overlapping parts of two consecutive blocks, the window shapes have to be adapted accordingly to retain the perfect reconstruction (PR) property of the MDCT. The new TW-MDCT also ensures a constant framing in the linear time domain, which e. g. could not be guaranteed by applying a continuous varying resampling to the time signal before a conventional MDCT filterbank.

The pitch contour needed both in the encoder and the decoder is coded very efficiently, resulting in a low average side info rate well below 1 kbps per channel. Figures 6 and 7 show the block diagrams of the processing chains for encoder and decoder.

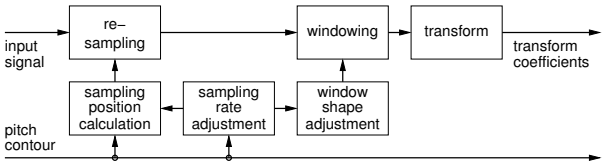

Fig. 6: Processing steps of the time-warped MDCT.

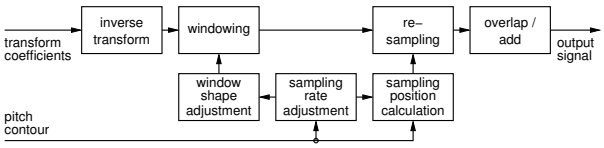

Fig. 7: Processing steps for inverting the time-warped MDCT.

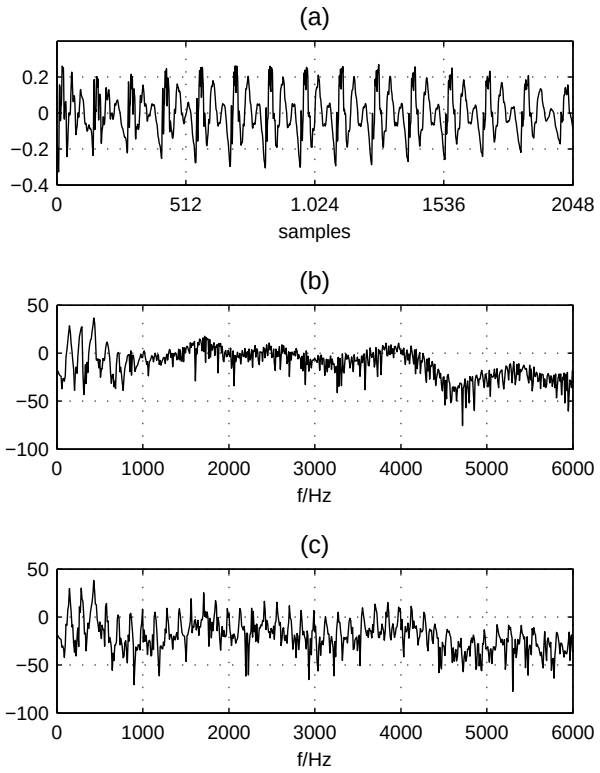

Fig. 8: Example of the energy compaction property of the TW-MDCT. (a) Time signal of a block containing a tonal signal at 20 kHz sampling rate. (b) Magnitude of the conventional MDCT spectrum. (c) Magnitude of the TW-MDCT spectrum.

Figure 8 shows an example of the energy compaction property of the TW-MDCT for one block of a pitch-varying tonal signal.

3.9. Transition between Core Coding Modes

The hard switching between the mutually exclusive core coding schemes puts high demands on the design of the mode transitions, which are usually prone to reduce coding efficiency and may be a cause for artifacts. To complicate matters, the two coding schemes operate in different domains: The LPD coding scheme operates in the (weighted) LPC filtered domain whereas the FD coding scheme operates solely in the MDCT domain, which includes time domain aliasing (TDA) [22], that spreads over

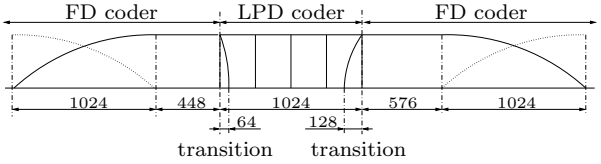

Fig. 9: Transition scheme between core coder modes.

frame boundaries.

To address these issues, new transition windows for the FD coding mode were designed [23], since the FD coding mode is most flexible in terms of bit allocation and number of spectral coefficients to transmit. These new windows are similar to AAC's "STOP" or "START" transitional windows, with one long window slope of length 1024 and a short window slope of 128 or 64 samples resp. (see Fig. 9). The short window slopes act as crossfade areas to smoothen the transitions. If necessary, an artificial TDA is generated on the decoded LPD signal portions to cancel the TDA components contained in the FD coded signal.

For the transitions from FD to LPD coding mode a shorter overlap of just 64 samples is used to circumvent the TDAC process. Due to the start-up of the internal filters in the LPD coder, the first reconstructed signal samples are usually inaccurate. Folding these samples, as needed for TDA cancelation, would propagate this error.

For the transition from LPD to FD coding mode the window is extended from 2048 to 2304 samples in order to meet the requirement of a constant framing. This window corresponds to 1152 MDCT coefficients, which are completely encoded by the FD coding scheme.

The minimum granularity for switching between the FD and LPD coder is dictated by the framing of the FD core, which is identical to one legacy AAC frame (corresponds to one so-called "superframe" in AMR-WB+). One LPD mode frame can hold up to four ACELP or one to four wLPT frames, or any corresponding permutation thereof.

4. RESULTS

To assess the quality of the novel coding scheme, formal listening tests at nine different operating points

were carried out. A set of a total of twelve test items was used, which could be categorized into three subsets, covering speech, music and mixed content. The mixed signal category contains two signal classes, speech and music played out simultaneously or shortly after each other. The items were encoded using AMR-WB+ and HE-AACv2 as quality references and the novel coding scheme (USAC). Additionally, a hidden reference (HR) and two low-pass anchors with a bandwidth of 3.5 kHz (LP35) and 7.0 kHz (LP70) were included in the test. At 64 kbps, no reference to AMR-WB+ was included, since it does not support this operating mode.

The listening tests were carried out using MUSHRA methodology [24]. Per test, a minimum number of 39 experienced listeners participated. In total, nine listening tests were carried out, one per operating point.

The operating points covered the envisioned application scenario, i. e. low and medium bitrates: Four operating points focusing on monophonic signals, with bitrates from 12 kbps to 24 kbps, and five operating points focusing on stereophonic signals, with bitrates from 16 kbps to 64 kbps. The test results are summarized in Figures 10 - 13. All figures show mean values and confidence intervals (95% level of significance) per operating point.

Firstly it can be observed that the included quality references, i. e. AMR-WB+ and HE-AACv2, exhibit uneven performance depending on operating mode and content type. For music signals, HE-AACv2 performs significantly better compared to AMR-WB+ for six out of eight operating points. The increased gap at stereo signals points at suboptimal stereo coding capabilities of AMR-WB+. Conversely, for speech signals AMR-WB+ performs significantly better compared to HE-AACv2 for seven out of eight operating points. For the mixed category, AMR-WB+ performs significantly better for the two lower mono bitrates, while HE-AACv2 performs significantly better for all stereo operating modes. Averaging over all categories, AMR-WB+ performs significantly better compared to HE-AACv2 for all mono operating modes, whereas HE-AACv2 performs significantly better for all stereo operating modes.

Comparing these results with the performance of the novel coding scheme, it can be seen that for each content type and for each operating mode, the novel coding scheme performs at least as good as the reference quality, i. e. the better of AMR-WB+ and HE-AACv2. For many operating points, the performance of the novel coding scheme even exceeds the reference quality, e. g. for music at 12 kbps and 16 kbps mono, and for speech signals at 16 - 32 kbps stereo. Averaging over all categories, the novel coding scheme performs significantly better than the reference quality for eight out of nine operating points. For 64 kbps stereo the performance of HE-AACv2 is matched, pointing at a good scalability towards higher bitrates.

5. CONCLUSION

In this paper, the reference model for the MPEG standardization process on "MPEG-D Unified Speech and Audio Coding" was presented. This activity aims to standardize a new codec with a consistent high quality for speech, music and mixed signals over a broad range of bitrates. To efficiently encode both, speech and general audio content, the presented technology switches between two schemes based on state of the art speech and audio codecs, such as AMR-WB+, HE-AACv2 and MPEG Surround. The technology components were further enhanced and innovatively complemented to match the requirements of the envisioned application scenario without sacrificing scalability and performance for higher bitrates.

The performance of the novel codec was evaluated at nine operating points, ranging from 12 kbps mono to 64 kbps stereo for speech, music and mixed signals. Whereas the reference quality in the market is represented depending on signal category and operating point by either AMR-WB+ and HE-AACv2, the novel coding scheme exhibits a consistent quality and performs for each category and operating point at least as good as the better of AMR-WB+ and HE-AACv2. On average over all categories, it performs significantly better than both on eight of nine operating points. For higher bitrates, a high quality scaling towards transparency can be expected, since the codec converges to an AAC-like operating mode.

6. ACKNOWLEDGEMENT

The authors would like to thank the following people for their invaluable contribution to this project.

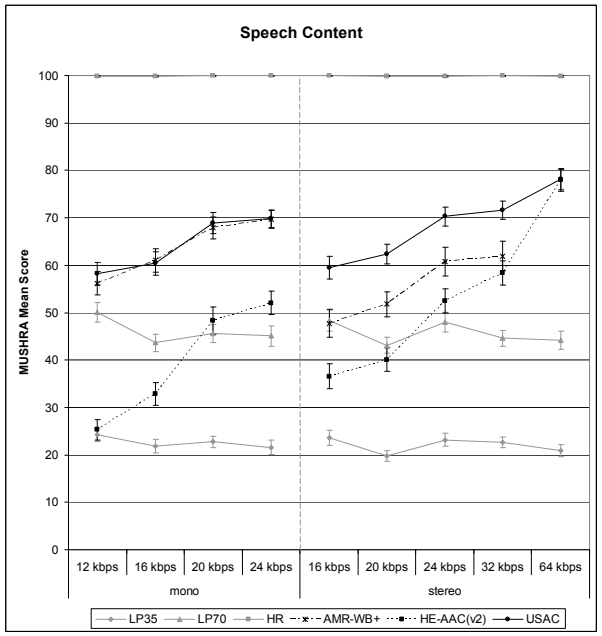

Fig. 10: Listening tests results for speech content (three items).

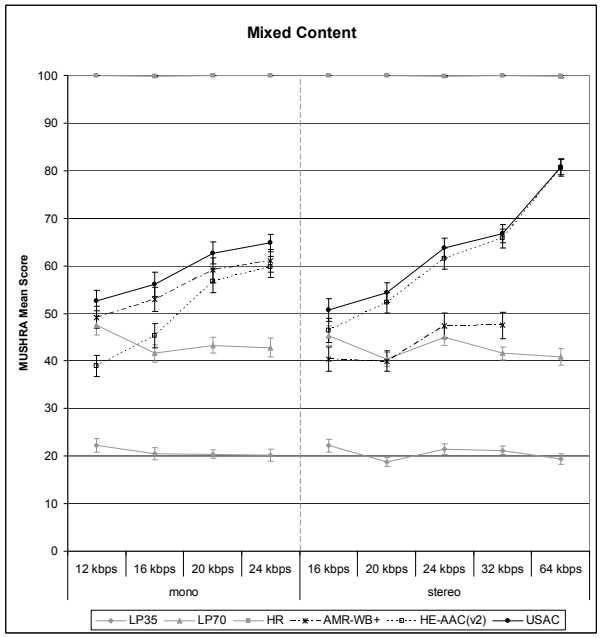

Fig. 12: Listening tests results for mixed content (three items).

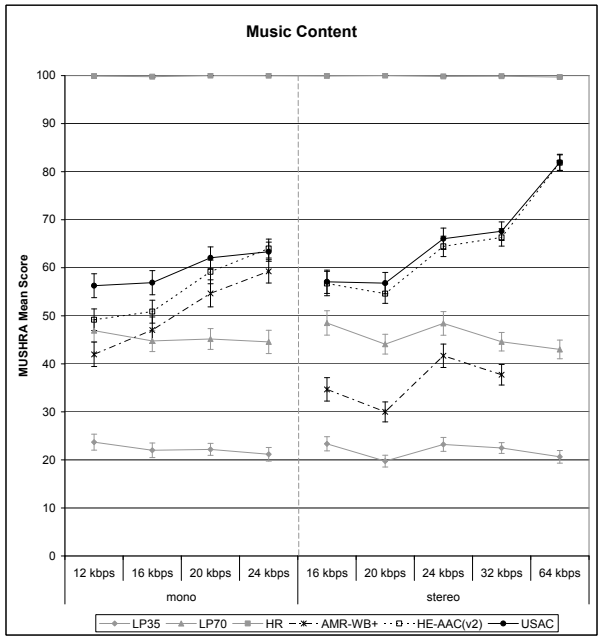

Fig. 11: Listening tests results for music content (three items).

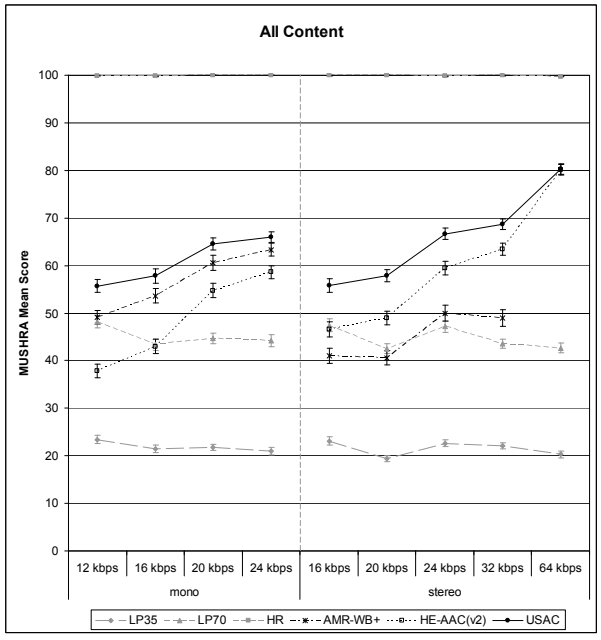

Fig. 13: Listening tests results for all content (twelve items).

Their persistence, expertise and dedication made the development of the presented system as enjoyable as it was successful: S. Disch, B. Edler, J. Herre, J. Hirschfeld, U. Krämer, J. Lapierre, M. Neusinger, C. Spenger, M. L. Valéro, S. Wabnik, Y. Yokotani.

7. REFERENCES

[1] ISO/IEC 14496-3:2009, "Coding of Audio-Visual Objects, Part 3: Audio," 2009.

[2] 3GPP, "Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General description," 2002, 3GPP TS 26.171.

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

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

[5] Martin Wolters, Kristofer Kjörling, Daniel Homm, and Heiko Purnhagen, "A closer look into MPEG-4 High Efficiency AAC," in *115th AES Convention*, New York, NY, USA, Oct. 2003, preprint 5871.

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

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

[8] B. Bessette, R. Salami, C. Laflamme, and R. Lefebvre, "A wideband speech and audio codec at 16/24/32 kbit/sec using hybrid acelp/tcx," in *Proc. IEEE Workshop Speech Coding*, May 1999, pp. 7–9.

[9] Ludovic Tancerel, Stéphane Ragot, Vesa T. Ruoppila, and Roch Lefebvre, "Combined speech and audio coding by discrimination," *Proc. 20th Biennal Symposium on Communications*, May 2000.

[10] Sean A. Ramprashad, "The multimode transform predictive coding paradigm," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 2, pp. 117–129, March 2003.

[11] Jari Mäkinen, Bruno Bessette, Stefan Bruhn, Pasi Ojala, Redwan Salami, and Anisse 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.

[12] 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.

[13] Martin Dietz, Lars Liljeryd, Kristofer Kjörling, and Oliver Kunz, "Spectral Band Replication, a Novel Approach in Audio Coding," in *112th AES Convention*, Munich, Germany, May 2002, preprint 5553.

[14] Andreas Ehret, Martin Dietz, and Kristofer Kjörling, "State-of-the-art audio coding for broadcasting and mobile applications," in *114th Convention*, Amsterdam, Mar. 2003, Audio Eng. Soc., preprint 5834.

[15] Mark Dolson, "The phase vocoder: A tutorial," *Computer Music Journal*, vol. 10, no. 4, pp. 14–27, 1986.

[16] Eberhard Zwicker and Hugo Fastl, *Psychoacoustics: facts and models*, Springer series in information sciences. Springer, Berlin, 2nd edition, 1999.

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

[18] Jürgen Herre and Donald Schulz, "Extending the MPEG-4 AAC Codec by Perceptual Noise Substitution," in *104th Convention of the AES*, Amsterdam, May 1998, preprint 4720.

[19] Nikolaus Meine and Bernd Edler, "Improved quantization and lossless coding for subband audio coding," in *118th AES Convention*, Barcelona, Spain, May 2005, preprint 6468.

[20] Khalid Sayood, *Introduction to Data Compression*, Morgan Kaufmann Publishers Inc., San Francisco, 2nd edition, 2000.

[21] Bernd Edler, Sascha Disch, Stefan Bayer, Guillaume Fuchs, and Ralf Geiger, "A Time-Warped MDCT Approach to Speech Transform Coding," in *126th AES Convention*, München, Germany, May 2009.

[22] 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.

[23] M. Neuendorf, P. Gournay, M. Multrus, J. Lecomte, B. Bessette, R. Geiger, S. Bayer, G. Fuchs, J. Hilpert, N. Rettelbach, R. Salami, G. Schuller, R. Lefebvre, and B. Grill, "Unified Speech and Audio Coding Scheme for High Quality at Low Bitrates," ICASSP'09, Taipei, Taiwan.

[24] International Telecommunication Union, "Method for the subjective assessment of intermediate sound quality (MUSHRA)," 2001, ITU-R, Recommendation BS. 1543-1, Geneva, Switzerland.