MPEG-4 Enhanced Low Delay AAC - a new standard for high quality communication

Markus Schnell[1], Markus Schmidt[1], Manuel Jander[1], Tobias Albert[1], Ralf Geiger[1],
Vesa Ruoppila[2], Per Ekstrand[2], Manfred Lutzky[1], Bernhard Grill[1]

[1]*Fraunhofer IIS, Erlangen, Germany*

[2]*Dolby, Stockholm/Sweden, Nuremberg/Germany*

Correspondence should be addressed to Markus Schnell (`Markus.Schnell@iis.fraunhofer.de`)

ABSTRACT

The MPEG Audio group has recently concluded the standardization process for the MPEG-4 Enhanced Low Delay AAC (AAC-ELD) codec. This codec is a new member of the MPEG Advanced Audio Coding family. It represents the efficient combination of the AAC Low Delay codec and the Spectral Band Replication (SBR) technique known from HE-AAC. This paper provides a complete overview of the underlying technology, presents points of operation as well as applications and discusses MPEG verification test results.

1 INTRODUCTION

In recent years AAC-LD, the low latency member of the ISO/MPEG-2/4 AAC audio codec familiy, has established itself as the de facto audio standard for video conferencing applications. Unlike traditional communication codecs, an audio codec does not only deliver good speech quality but is able to transmit any input signal without altering its natural sound. This creates an experience of virtual presence which in tests has lead to situations where people coming into the room were looking for a person who actually was at a location several thousand kilometers away. As far as delay is concerned, the 20 ms figure of AAC-LD is lower than the latency of most modern speech codec standards. The only drawback of AAC-LD is the somewhat higher data rate required when compared to modern speech coders. Recently, however, ISO has published a new amendment to MPEG-4 Audio which standardizes an enhanced version of AAC-LD, called Enhanced AAC-LD (AAC-ELD). AAC-ELD increases the bitrate efficiency of AAC-LD by combining it with Spectral Band Replication (SBR) technology well known from High-Efficiency AAC (HE-AAC) while maintaining a low end-to-end algorithmic delay of around 30 ms.

This paper provides an overview of the key features of this new technology, describes the underlying technology, presents results of listening tests performed at MPEG in the course of a final verification test, outlines application scenarios and offers suggestions on efficient implementation.

2 ELD FEATURES

AAC-ELD is a full-band audio codec for general content providing an algorithmic delay low enough to establish bidirectional communication. The standard allows a variable configuration of the codec with respect to the needs of the application. A list of typical points of operation of this codecs is given in Table 1.

Bit rate (mono)	Delay	Quality[a]
32 kbps	31.3 ms	Good
48 kbps	22.5..31.3ms	Good / Excellent
64 kbps	15..31.3ms	Excellent

[a]The quality refers to the MPEG verification test outlined in Section 5

Table 1: Points of operation for AAC-ELD

In general, these points of operation can be extended up to a transparent audio quality, as shown for the AAC technology in [1]. The delay of AAC-ELD leaves enough room to compensate additional delay sources occuring in a complete communication chain such as echo cancelling, transport or jitter buffer to assure a total end-to-end delay below 100 ms, required for highly interactive tasks [2].

The AAC-ELD codec allows encoding with constant bit rate modes as well as with variable bit rate modes. Although AAC-ELD is designed as a communication codec, it is not restricted to mono signals only. It supports stereo and all common multichannel configurations, i.e. 3.0, 4.0, 5.0, 5.1 and 7.1 .

The codec uses error resilience (ER) syntax which allows transmission bit errors to be detected, corrected or concealed [3, 4]. For packet-based transmission channels, e.g. IP, where the loss of complete packages represents a problem, AAC-ELD offers several advantages in its design. Namely the codec does not utilize any tool producing frame dependencies, as for example inter-frame prediction. In the case of frame loss, such tools would prevent the decoder from reconstructing the audio data although the bit stream data is valid for the subsequent frames.

The standard of AAC-ELD does not determine the behaviour of the codec for the case of frame loss or bit stream errors but several techniques for the concealment are described in [5, 6, 7, 8].

The AAC-ELD bit stream can be transmitted over a non packet based channel while providing random access by using the MPEG-4 LOAS transport format. Alternatively for packet-based transmission, the MPEG-4 access units can be transmitted using the real-time transport protocol RTP [9] in combination with the session initiation protocol SIP [10].

The AAC-ELD syntax allows the embedding of additional data packets, which is useful, for instance, to send control data to attached devices.

Due to the unified time-frequency transformation of AAC-ELD, the codec permits the mixing of audio data in the bit stream domain. A complete decoding and reencoding can be avoided for this purpose [11].

3 BACKGROUND: MPEG-4 TECHNOLOGY AND LOW DELAY FILTER BANKS

In this section, the MPEG-4 state-of-the-art general audio codecs and the low delay filter banks that form the basis of the Enhanced Low Delay AAC coder are reviewed briefly. Based upon this, the technical aspects of AAC-ELD are presented.

3.1 MPEG-4 AAC LD

While AAC Low Complexity (AAC-LC), the low-complexity subset of AAC [12, 13], provides high audio quality, its algorithmic delay of at least 55 ms (1024 samples per frame, 48 kHz) is evidently too high for bidirectional communication. Derived from AAC-LC, a low-delay general audio coder was introduced within MPEG-4 [13] as MPEG-4 ER AAC LD [14, 15].

With a reduced transform size, an introduced new low-overlap window and the deactivation of the block switching mechanism, the codec achieves an optimized algorithmic delay of down to 20 ms. Very good audio quality can be reached starting from 48 kbit/s per channel.

3.2 MPEG-4 HE-AAC

The next milestone in MPEG-4 towards low bit rate coding was the introduction of SBR, a generic parametric coding tool for high frequencies. The combination of SBR and AAC-LC was standardized in 2003 in the MPEG-4 High-Efficency (HE-AAC) [16] and achieves FM quality at bit rates of as low as 16 kbit/s per channel.

In order to limit the perceptible coding artifacts of common audio coding systems to a subjectively acceptable level, the entropy of the source has to be limited and the coding gain has to be optimized. This is generally achieved by reducing the coded audio bandwidth and the sampling frequency. To overcome this limitation, the SBR decoder reconstructs higher frequency components with the help of the low-frequency base band and a very compact parametric description of the high band [17, 18]. The low-frequency base band of the signal is coded by a conventional core coder. In addition to that, the high band is dealt with by a Quadrature Mirror Filterbank (QMF) with 64 channels from which the SBR data is derived. Figure 1 illustrates the coding process. A detailed description can be found in [17].

Naturally, a combination of the abilities of HE-AAC and AAC-LD appears quite appealing in order to achieve a low bit rate and low delay coding system with high audio quality. In such a combination the SBR part of HE-AAC acts only in part as a parallel structure to the core coder, thus adding far too high an amount of unnecessary delay. In addition, also the delay of the core coder is doubled by operating it at half the original sampling rate ('dual-rate'). Both mentioned facts are illustrated in Figure 3. In summary, the simple approach of combining AAC-LD and SBR leads to a total algortihmic delay of 60 ms [19]. To overcome this drawback, the technology of the low delay filter bank was adopted. This technology is briefly described in the following.

3.3 Low delay filter banks

It is a key design feature of traditional filter banks, like the TDAC filter banks [20] such as the MDCT, to use symmetric functions for the windowing step of their modulation calculations. Due to that fact, such transformations cause a system delay of *window length minus one* samples. The goal in designing low delay filter banks is to reduce the reconstruction delay independently of the filter length. Maintaining the perfect reconstruction property is paramount regarding the design process. With [21, 22] some of the first low delay filter banks are presented. They are set in the context of a generalized system delay, i.e. the system delay is no longer rigidly dependent on the filter length. [21] describes a direct design method via numerical optimization. This approach does not guarantee perfect reconstruction and a fast implementation is only obtainable with considerable effort. [22] describes an optimization method for cosine modulated filter banks. While this leads to a substantially more efficient implementation, perfect reconstruction still can not be achieved.

The design method used here was first described in [23, 24], and later in [25, 26] presenting a combination of the desired properties. The resulting filter banks utilize the same cosine modulation function as the traditional MDCT. However, they can make use of longer window functions that can be non-symmetric, with a generalized or low reconstruction delay.

3.3.1 Mathematical description

Although the design method allows an extension of the MDCT in both directions, only the extension of E blocks towards past samples is applied here, with each block consisting of M samples.

Analysis: The frequency coefficient X of band k and block i inside an M-channel filter bank is defined as

$$X_{i,k} = -2 \sum_{n=-E\cdot M}^{2M-1} p_A(n)\cdot x(n) \cos[\frac{\pi}{M}(n+\frac{1}{2}-\frac{M}{2})(k+\frac{1}{2})] \tag{1}$$

for $0 \le k < M$, with n a sample index and p_A an analysis window function.

Synthesis: The demodulated vector z is defined as

$$z_{i,n} = -\frac{1}{M} \sum_{k=0}^{M-1} p_S(n)\cdot X_{i,k} \cos[\frac{\pi}{M}(n+\frac{1}{2}-\frac{M}{2})(k+\frac{1}{2})] \tag{2}$$

for $0 \le n < M(2+E)$, with p_S a synthesis window compatible with p_A.

Overlap add: The reconstructed signal $\hat{x}$ is obtained by

$$\hat{x}_{i,n} = \sum_{j=-(E+1)}^{0} z_{i+j,n-j\cdot M} \tag{3}$$

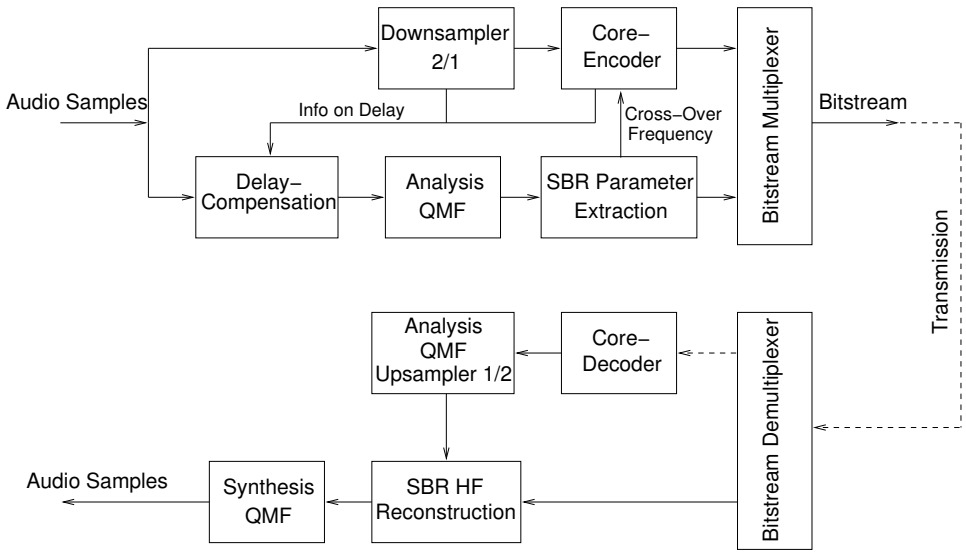


Fig. 1: Overview of SBR Codec in combination with a core coder, as specified in [16].

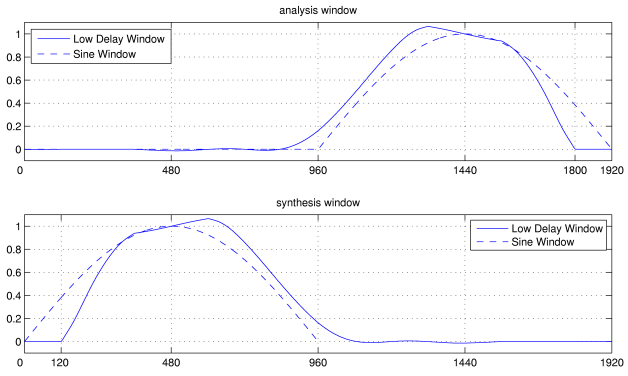


Fig. 2: Impulse response of low delay windows

4 TECHNICAL OVERVIEW AAC-ELD

With the techniques standardized within Enhanced Low Delay AAC it is possible to reduce the amount of delay in such an extent that it becomes possible to use the bit rate saving capabilities of SBR for uses within bidirectional communication scenarios. In the following, the core technologies of AAC-ELD are reviewed briefly. For more details see [19] and [27].

4.1 Low delay filter bank in AAC-LD core

The MDCT of the AAC-LD core is replaced by a low delay version, called LD-MDCT. The symmetrical shape of the windowing functions of the MDCT is changed into an asymmetrical one which allows to reduce the overlap towards future values. The impulse response is extended at the same time towards past samples.

With the LD-MDCT the filter bank delay is, due to the reduced overlap of 120 samples towards the future and assuming a frame length of $M = 480$, reduced from 959 samples $(2M-1)$ to 719 $(2M-\frac{M}{2}-1)$ samples. The impulse response is extended to the past by 960 samples $(E = 2)$. Figure 2 shows the new analysis and synthesis window functions in comparison with the sine window which is common in MPEG audio coding. Note that the analysis window is the time-reversed version of the synthesis window, i.e. $p_A(n) = p_S(4M-1-n)$.

The overall delay reduction of 240 samples is a result from the fact that in both the analysis and the synthesis window the respective parts that access future input values (and thus would cause delay) are reduced by 120 samples. The extended overlap of the 'tail' of the window does not result in any additional delay, as it only involves adding values from the past.

In [19] a comparison of the traditional AAC-LD windows and the LD-MDCT window is offered. It is shown

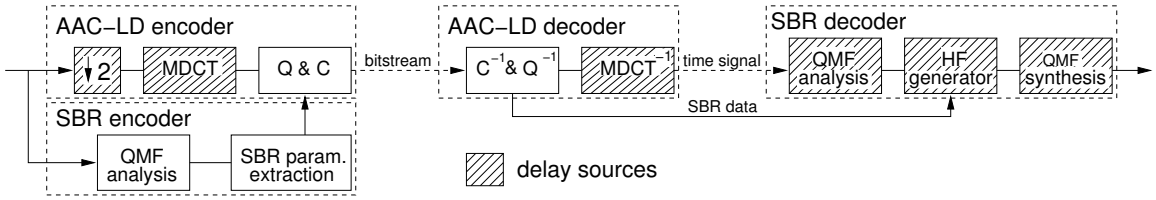


Fig. 3: Sources of delay in the encoder/decoder process of AAC-LD in combination with SBR

that a frequency response similar to that of the AAC-LD sine window is achieved for the LD-MDCT. In order to deal with pre-echo artifacts, i.e. the spreading of the quantization noise before the onset of a transient in the signal, AAC-LD offers a low-overlap window [14]. This technique works best in combination with Temporal Noise Shaping (TNS). While the window function of the LD-MDCT offers the same property, it exhibits a superior frequency response. The low delay window therefore replaces both traditional AAC-LD windows, rendering a dynamic window shape adaption obsolete.

4.2 Modification of SBR framing and HF-generator

The number of samples that the SBR module processes at once is reduced to match the number of samples the AAC-LD core uses (480 resp. 512). The flexibility of the high-frequency generator of the SBR module, which implies 384 samples of delay, is restricted in order to minimize the delay. Removing the additional delay leads to a frame-locked time grid comprising synchronized SBR data with respect to the borders of AAC-LD frames.

4.3 Complex low delay filter bank for SBR

The QMF inside the SBR decoder is replaced by a complex low delay filter bank (CLDFB). By leaving the number of bands (64) and the length of the impulse response (640) unchanged and by using a similar complex modulation, the CLDFB stays compatible to the SBR framework.

In Figure 4 the CLDFB prototype filter is compared with that of the original SBR QMF. As illustrated, the delay of modulated filter banks can be determined as consisting of two parts: the overlap delay introduced by the prototype filter, and the framing delay of the modulation core (i.e. a $\mathrm{DCT_{IV}}$ of length M). It can be observed that the CLDFB prototype introduces an overlap delay of only 32 samples

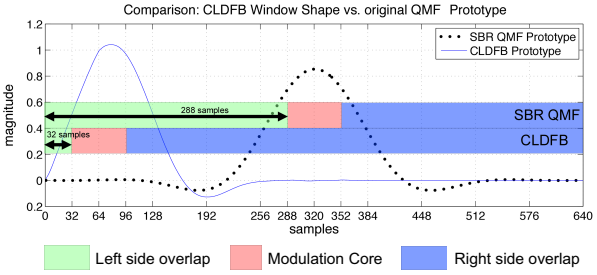


Fig. 4: Impulse responses of synthesis prototypes, CLDFB vs. SBR QMF [27]

for the same situation in which the prototype of the SBR QMF causes an overlap delay of 288 samples.

The CLDFB is, as the LD-MDCT, formulated in accordance with the design principle briefly outlined in paragraph 3.3. The extension of the function prototype towards past samples is $E = 8$. However, the delay reduction is achieved solely by a shift of the modulation core. The impulse response of the MDCT core is, unlike that of the LD-MDCT, not truncated. This results in an overlap delay of 64 samples with $M = 64$.

The presented techniques work in both a complex and a low-power version. The complex version for the normal SBR mode is obtained by adding a sine modulation to the given cosine one. In the low-power mode only the real valued part of the CLDFB is used.

4.4 Signaling of AAC-ELD in MPEG-4 and bit stream

The bit stream format is based on the ER syntax of MPEG-4 which is also used for AAC-LD. For AAC-ELD, the syntax was revised to remove all unnecessary elements in order to achieve a more compact representation. The revised syntax saves up to one kbps without losing any information.

The audio object type (AOT) 39 is allocated for signaling the AAC-ELD codec inside the *AudioSpecificConfig* (ASC) of the MPEG-4 framework. As a part of it, the *ELDSpecificConfig* is designed to signal the used coding and bitstream parameters, e.g frame length and ER tools.

Hence, the low delay SBR is fully integrated in AAC-ELD, the *ELDSpecificConfig* contains all parameters of this tool, i.e. presence of SBR, dual/single-rate mode and the complete SBR header necessary for the initialization of the decoder module. Due to the merge of all relevant data, a decoder can start up at once and is able to provide the full audio quality without any delay.

In addition, the *ELDSpecificConfig* offers a mechanism which allows the signaling and the transmission of configuration data of further extensions to AAC-ELD, potentially developed in the future. At the same time, backwards compatibility is always assured. Thus, future tools will always make use of explicit signaling methods which makes an ambiguous initialization of decoders obsolete.

5 MPEG VERIFICATION TEST

5.1 Overview

The Audio Subgroup of MPEG conducted a verification test on the AAC-ELD codec for characterizing the new technology [28]. This verification test comprised two sets of listening tests, technology-driven and application-driven, arranged in a total of six experiments that were conducted using the MUSHRA test methodology specified in ITU-R Recommendation BS.1534 [29]. The experiments covered the bit rates 24, 32, 48 and 64 kbps in single channel operation with a total of 32 items from speech, mixed content and music categories. Each experiment was conducted at least by two independent listening test laboratories. Five companies contributed to this exercise as listening test laboratories with a grand total of 152 subjects. According to the MUSHRA test methodology, low-pass filtered versions of the test items with cut-off frequencies at 3500 Hz (LP35) and 7000 Hz (LP70) were included in all experiments. In addition, a copy of the item under test was included as the hidden reference (HR).

5.1.1 Technology-driven test

The technology-driven test was designed to characterize the performance of selected AAC-ELD codec configurations across a broad bit rate range with a well-established set of critical test items as specified in Table 2, using two current state of the art MPEG-4 audio codecs, HE-AAC and AAC-LD, as references. The ITU-T Recommendation G.722.1 Annex C, referred to as G.722.1-C, was added as an additional reference. The AAC-LD and G.722.1-C codecs are broadly-used, state of the art communication codecs that operate at bit rates, an algorithmic delay and an audio bandwidth in the same dimensions as the AAC-ELD. The test consisted of two experiments, Experiment T1 that covered the bit rates 24 and 32 kbps and Experiment T2 that covered the bit rates 48 and 64 kbps as detailed in Table 3.

Item name	Category	Description
es01	Speech	Suzanne Vega, solo
es02	Speech	Male German speech
es03	Speech	Female English speech
sc01	Complex signal	Trumpet
sc02	Complex signal	Orchestra
sc03	Complex signal	Pop music
si01	Single instrument	Harpsichord
si02	Single instrument	Castanets
si03	Single instrument	Pitch pipe
sm01	Single instrument	Bag pipe
sm02	Single instrument	Glockenspiel
sm03	Single instrument	Plucked strings

Table 2: Items for technology-driven test

Codec ID	Codec	Bit rate	Delay	Bandwidth
Experiment T1				
1C-24	G.722.1-C	24 kbps	40.0 ms	14.0 kHz
1C-32	G.722.1-C	32 kbps	40.0 ms	14.0 kHz
LD-32	AAC-LD	32 kbps	42.7 ms	11.6 kHz
ELD-24	AAC-ELD	24 kbps	36.3 ms	12.8 kHz
ELD-32	AAC-ELD	32 kbps	33.3 ms	14.3 kHz
HE-24	HE-AAC	24 kbps	129.3 ms	15.4 kHz
Experiment T2				
1C-48	G.722.1-C	48 kbps	40.0 ms	14.0 kHz
LD-48	AAC-LD	48 kbps	32.0 ms	14.1 kHz
LD-64	AAC-LD	64 kbps	21.3 ms	14.5 kHz
ELD-48	AAC-ELD	48 kbps	33.3 ms	16.9 kHz
ELD-64	AAC-ELD	64 kbps	33.3 ms	20.3 kHz
ELD-64-S	AAC-ELD	64 kbps	16.0 ms	14.5 kHz
HE-32	HE-AAC	32 kbps	129.3 ms	16.9 kHz

Table 3: Setup of experiment T1 and T2

5.1.2 Application-driven test

The application-driven test was designed to evaluate some AAC-ELD configurations that are considered to be attractive for high quality communication and professional broadcasting applications. The reference codecs and the test items were chosen by taking into account the targeted use scenario. The speech-focused experiments cover three clean speech, three reverberant clean speech and six reverberant speech with office background noise conditions. The music-focused experiments cover music and mixed content. The AAC-LD codec and the G.722.1-C codec were included as references.

The speech items were designed to assess the performance of the AAC-ELD codec in high quality communication scenarios for which this technology is primarily designed. For generating these speech items, the source material of clean speech items consisted of various languages as listed in Table 4 served as a base. Several office background noise samples and software utilities were used for building the test items out of clean speech and background noise items in a way such that the test items exhibit the desired reverberation and background noise characteristics. Additionally, the different sources had been level adjusted before mixing and after reverberation. For generating clean speech items, the processing according to Figure 5(a) was applied. The processing for reverberant speech is described in Figure 5(b) and for reverberant speech with background office noise in Figure 5(c).

The tools used for the various processing steps are part of the ITU-T Software Tool Library [30]. The level adjustment was applied using the tool 'Speech Voltmeter'. Depending on the input content, the mode was switched between the speech activity detection mode (P.56) and the RMS mode. The reverberation was performed by using the REVERB tool with a full-band room impulse response sampled at 48 kHz.

Item name	Description
dan_speech	Danish male speech
es04	English male speech
es05	German female speech
jap_male_speech	Japanese male speech
nadib01	German male speech
nadib08	Tajik male speech
nadib09	German female speech
nadib10	English female speech
nadib13	French female speech
nadib14	Chinese female speech

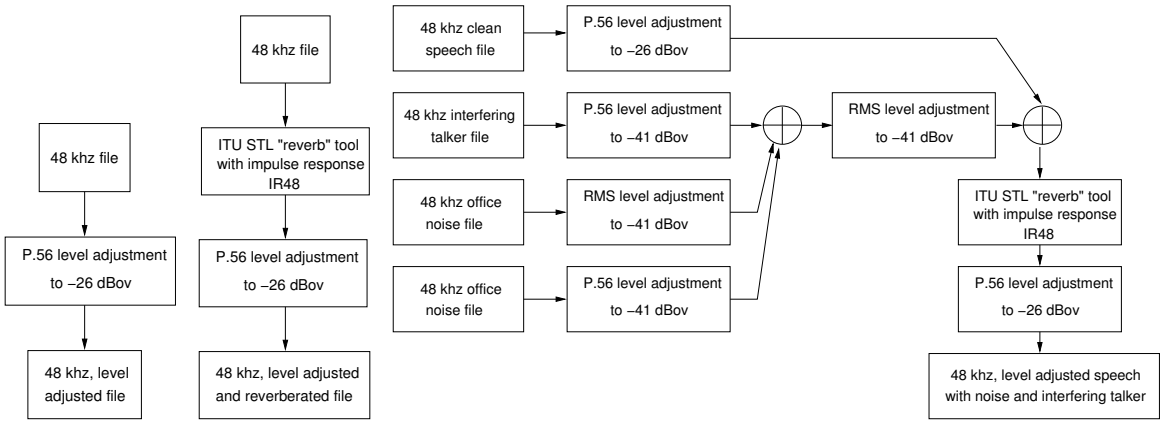
Table 4: Items of speech set

Codec ID	Codec	Bit rate	Delay	Bandwidth
Experiment A1 Speech				
1C-32	G.722.1-C	32 kbps	40.0 ms	14.0 kHz
1C-48	G.722.1-C	48 kbps	40.0 ms	14.0 kHz
LD-32	AAC-LD	32 kbps	40.0 ms	11.6 kHz
LD-48	AAC-LD	48 kbps	30.0 ms	14.1 kHz
ELD-32	AAC-ELD	32 kbps	31.3 ms	14.3 kHz
ELD-48	AAC-ELD	48 kbps	31.3 ms	16.9 kHz
Experiment A1 Speech				
1C-48	G.722.1-C	48 kbps	40.0 ms	14.0 kHz
LD-48	AAC-LD	48 kbps	30.0 ms	14.1 kHz
LD-64	AAC-LD	64 kbps	20.0 ms	14.5 kHz
ELD-48	AAC-ELD	48 kbps	31.3 ms	16.9 kHz
ELD-64-S	AAC-ELD	64 kbps	15.0 ms	14.5 kHz

Table 5: Setup of speech experiment

5.2 Results and observations

The data from various test sites was pooled to enable the most accurate observations and therefore to get the narrowest possible 95% confidence intervals on the mean score. The following plots display the mean values (bar) and associated 95% confidence intervals (vertical tick) averaged over all items for every system.



(a) clean speech (b) reverberant speech (c) speech with interfering talkers and office background noise

Fig. 5: Pre-Processing of speech items

5.2.1 Technology-driven test

Figure 6 shows the results from the technology-driven test for low bit rates (T1) as well as for the high bit rates (T2).

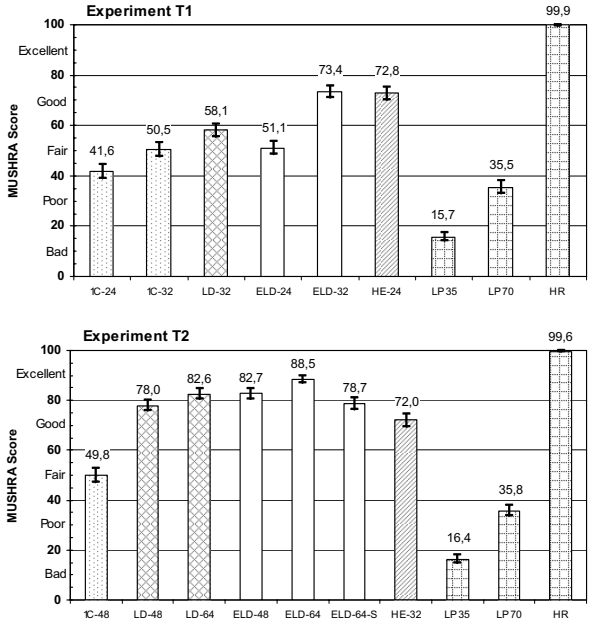


Fig. 6: Technology-driven tests

The following observations can be made concerning the results of this experiment:

- The performance of the AAC-ELD codec at 24 kbps is comparable to that of the G.722.1-C codec at a bit rate of 32 kbps.
- The mean score for the AAC-ELD codec at 32 kbps is over 15 points higher than the mean score for the G.722.1-C codec (1C-32) or the AAC-LD codec (LD-32) at the same bit rate. At this bit rate, AAC-ELD is the only low delay coding system in this experiment that scores in the *Good* range, regardless of having 6 ms (or 22%) lower algorithmic delay than AAC-LD and G.722.1-C.
- The AAC-ELD at a bit rate of 32 kbps performs as good as the HE-AAC codec at a bit rate of 24 kbps. Thus, a low delay cost factor can be estimated by this observation, which is around 8 kbps.
- The mean score for the AAC-ELD and AAC-LD codecs at a bit rate of 48 kbps is significantly higher than the mean score for the G.722.1-C codec at the same bit rate.
- The score of the AAC-ELD codec at 48 kbps is significantly higher than that of the AAC-LD codec at the same bit rate and is comparable to that of the AAC-LD operating at 64 kbps. Furthermore, AAC-ELD is the only system having a mean score in the *Excellent* range at 48 kbps.

- The mean score for the AAC-ELD codec utilizing SBR at a bit rate of 64 kbps is sigificantly higher than that of the AAC-LD codec at the same bit rate.
- The AAC-ELD codec with disabled SBR at 64 kbps (ELD-64-S) showed no significant difference compared to the AAC-LD codec at the same bit rate, although this AAC-ELD configuration operates at 5 ms (or 25%) shorter delay of only 15 ms.

5.2.2 Application-driven test

Figure 7 shows the results of the speech focused experiments from the application-driven test for low bit rates (A1) as well as for the high bit rates (A2).

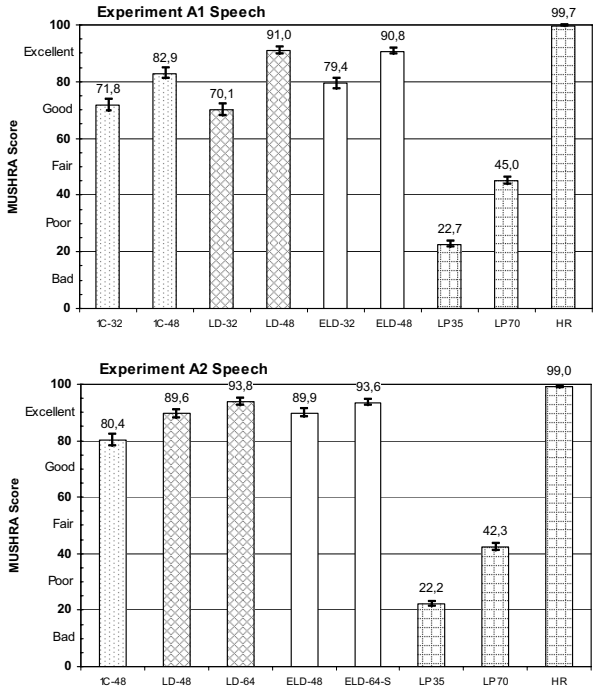


Fig. 7: Experiment speech

The following observations can be made concerning the results of this experiment:

- The mean score for the AAC-ELD codec at a bit rate of 32 kbps is significantly higher than the mean score for the G.722.1-C codec and the AAC-LD codec at the same bit rate. Thus, AAC-ELD is the best coding system regarding audio quality in this test for a bit rate of 32 kbps while offering 8 ms (or 22%) lower algorithmic delay compared to these two references.
- The mean score for the AAC-ELD codec at a bit rate of 32 kbps is, at the 95% level of significance, not different from the mean score for the G.722.1-C codec that operates at a bit rate of 48 kbps, at 50% higher bit rate.
- The mean scores for the AAC-ELD and AAC-LD codecs at a bit rate of 48 kbps are significantly higher than the mean score for the G.722.1-C codec at the same bit rate.
- The AAC-ELD codec with disabled SBR operating at 64 kbps (ELD-64-S) was not different from the AAC-LD codec at the equal bit rate, although this AAC-ELD configuration offers a 5 ms (or 25%) lower algorithmic delay of only 15 ms.

5.2.3 Summary

The AAC-ELD technology showed good performance in the verification tests with consistent and significant improvement in compression efficiency at the lowest tested bit rates, and comparable or better performance at the highest tested bit rates compared the previous low delay codec standard specified by MPEG. This gain in compression efficiency is attained with a substantially lower algorithmic delay that can be expected to contribute to better end-to-end performance in conversational applications compared to current state of the art high quality communication codecs.

As a further observation, the AAC-LD technology showed also good performance throughout the experiments. The performance of the AAC-LD codec was always equal or better than that of the G.722.1-C codec at the same bit rate.

5.2.4 Narrow-band and wide-band codecs

Trying to establish a link of the results to the current situation on the communication market, the 3.5 kHz low pass filtered anchor and the 7.0 kHz counterpart can be interpreted as *perfect* narrow-band and wide-band communication codecs. Narrow-band codecs cover per definition an audio bandwidth from 300 to 3400 Hz and their wide-band counterparts the range from 50 to 7000 Hz.

Popular representatives of these two categories are the G.711 [31] for narrow-band codecs and the G.722 [32] for the wide-band sector, both operate at a bitrate of 64 kbps. By observing the listening test results from the *Experiment A2, Speech* in Figure 7, a direct comparison of AAC-ELD to G.711 as well as to G.722 at a bitrate of 64 kbps can be estimated. As a result, AAC-ELD would score at least 70 points higher than a narrow-band codec, such as G.711, and at least 50 points higher than a wide-band codec, such as G.722.

In this context, the outcome of the verification test shows that by providing the full audio bandwidth, the perceived quality can be enhanced significantly compared to wide-band quality. This observation is confirmed over all speech focused experiments where clean speech, reverberant speech and speech with office background noise was under test.

6 IMPLEMENTATION ASPECTS

6.1 From HE-AAC to AAC-ELD

Since the AAC-ELD codec design has many components that are similar to an HE-AAC codec, a elaborated AAC-ELD implementation can be obtained very quickly by starting from an HE-AAC code code base. Many optimization efforts that have been invested into implementing highly optimized embedded implementations of HE-AAC codecs, can be reused to a large extent.

On the encoder side, tuning knowledge required to achieve optimum audio quality, is obviously a decisive factor on the quality of the implementation. But the experience gained by designing and implementing a high quality HE-AAC encoder obviously proves to be quite handy here as well.

The most notable differences between AAC-ELD and HE-AAC, from an implementor's point of view, are:

- Revised signalling methods and bit stream syntax, see 4.4
- Updated SBR filter bank, see Section 4.3
- SBR's internal time grid identical to time grid of SBR filter bank
- The core coder's framing and filter bank changed, see 4.1

Processor	AAC-ELD	HE-AAC
ADI Blackfin	103 / 111	100 / 100
ARM XScale	107 / 110	100 / 100
TMS320C6424	109 / 108	100 / 100

Table 6: Real time performance comparison in percent (encoder/decoder)

- No block switching or window shape adaption needed, see 4.1
- Framing overlap of the SBR decoder removed, see 4.2
- Delay adjustment between SBR module and core coder differs

The first four aspects require some effort, but the latter three are just changes to parameters that should be directly accessible in a reasonable fashion in existing implementations. If the dependencies of the configuration parameters were laid out correctly, it is possible to just change these parameters and the whole coder will operate as required for AAC-ELD.

The algorithm of the AAC-ELD codec is similar to that of an HE-AAC codec, and therefore their hardware requirements are comparable. As known from [33], the implementation of such an algorithm poses no challenge for existing embedded devices.

6.2 Computational complexity

Empirical measurements show that AAC-ELD has a slightly higher CPU workload (around 10 %) than HE-AAC. The reason is the fact that AAC-ELD uses a higher time resolution for the SBR grid, resulting in more computations for the same amount of processed audio.

As the AAC-ELD codec offers a low power operating mode for the decoder, about 20% - 30% workload savings can be achieved on the decoder side included in the following measurements. In [19] some preliminary computation complexity comparisons of AAC-ELD and HE-AAC were included. Current measurements show very similar results (see Table 6).

6.3 Memory requirements

With regard to the memory consumption, it can be observed in Table 7 that the "ram" and "const" requirements are slightly higher. This is because the prototypes

section	AAC-ELD	HE-AAC
ram	118 / 58	110 / 55
const	44 / 31	36 / 35
text	145 / 90	156 / 103

Table 7: Memory comparison in KiB (encoder/decoder) for one mono audio channel

of both filter banks, i.e. LD-MDCT and CLDFB, are asymmetric. Thus, the symmetry of regular windows cannot be exploited for storage purposes. In addition, the overlap-add buffers of the LD-MDCT are, due to the extension of the LD-MDCT window, slightly increased.

The "text" section is smaller, reflecting the more straightforward design of AAC-ELD. There is, among other things, no block switching and the SBR envelope description is simplified, thus reducing the code size.

7 USE CASES AND APPLICATIONS

ITU-T G.114 [2] recommends in the context of narrowband communication a one way end to end delay of below 150 ms. Under such conditions, "most applications, both speech and non-speech, will experience essentially transparent interactivity". Applications of highly interacive character "may be affected by delays below 100 ms". Obviously, superwide-band or full-band communication will at least not relax on this demand but rather tighten it. Therefore, it is especially important to keep the codec delay as low as possible.

The AAC-ELD is basically characterized as a communication codec. This is due to its low delay and its low bit rate ability (see Section 2). Additionally, the AAC-ELD provides a full audio bandwidth which means an extra improvement compared to common, bandwidth restricted, speech codecs which can be derived from the MPEG verification test data as outlined in Section 5.2.4. These abilities enable the communication participants to interact with each other over bit rate limited channels with a perception comparable to a face-to-face situation including a natural sound of voice and ambience in combination with transparent interactivity.

The problem in communication scenarios up to recent times was, that established speech coding schemes were not able to provide more than speech quality. Speech coding schemes in conferencing systems do not perform at a sufficient quality level when they have to deal with complex signals, like music or ambient sound. On the other hand audio coding schemes imply a delay that is usually far too high to supply a natural communication with a sufficiently low delay. With the launching of AAC-LD, the gap between established speech coding schemes and high quality audio coding schemes was closed. Since then, coding highly complex audio material at low bit rates with a low delay was not a hurdle anymore. But more importantly, keeping signal quality near perceptual transparency under these conditions was now possible.

Now, through the combination of AAC-LD and SBR within AAC-ELD, a communication codec is provided with the ability to perform with a lower algorithmic delay and at lower bit rates compared to AAC-LD while the level of quality is maintained or even improved, as can be seen in Section 5. Hence, AAC-ELD embodies an outstanding improvement for conferencing scenarios and other use cases which are described in the following.

Video conferencing The video conferencing market can be classified into stand alone units and desktop PC software. Stand alone units are dedicated video conferencing equipment and consist of a central, normally DSP driven, unit with a large screen and external loudspeakers and microphones. AAC-LD has been adopted as defacto standard in this market for high quality audio. These applications offer up to three video screens of up to 65 inch size (e.g. Cisco Telepresence). A video quality similar to tv-broadcasting on the rather large screens let users expect an audio quality that is similar to this experience. The same holds for multimedia presentations via video conferencing systems where it is essential to transmit integrated sounds or music in high quality. This requires an audio quality up to perceptual transparency, full audio bandwidth and at least stereo or even multichannel output. As AAC-ELD offers advantages in delay and audio quality it is the logical next step to adopt this codec.

It is also a benificial feature for a communication scenario, that the AAC-ELD is able to do delayless mixing operations in multipoint control units (MCU). In these units several end-point streams are mixed, encoded and distributed. Delayless mixing due to mixing in the frequency domain implies that tandem coding and its inherent quality loss is avoided. [11] shows an interesting item of how the AAC-ELD mixing technology can be utilized effectively.

Another comfortable feature is the ancillary data container in AAC-ELD. This feature allows a transmission

of encapsulated data like control data, text files or any other file format embedded inside the communication stream.

VoIP/teleconferencing In the near future the transmission of telephone signals will become more and more IP based. IP based phone calls can either be transmitted over public internet using dedicated clients, e.g. Skype or iChat, or dedicated infrastructure allocated by telecom service providers. In both fields of application, there is an ongoing trend to higher audio quality and bandwidth which leads now to the introduction of wide-band codecs and clearly afterwards to full audio bandwidth. Here, AAC-ELD competes against ITU-T codecs which lately have been standardized as superwide-band or full-band codecs [34, 35]. Due to the need for massive transcoding to legacy networks, complexity becomes an important design criteria. Furthermore robustness against packet loss and the ability to transmit telephone typewriter (TTY), dual-tone multi-frequency (DTMF) and FAX signals are essential features.

Next generation communication "Room to Room" communication is an additional use case for AAC-ELD. "Room to Room" communication means to transmit a whole room audio scene to another room and vice versa including as many details as possible. Both rooms are melting into each other and allow immersive audio interaction between groups of users. As part of the EU-funded TA2 project (Together Anywhere Together Anytime) [36], technologies for recording and rendering as well as for transport are being developed and are going to be integrated together with AAC-ELD coding to form the Audio Communication Engine. Within this project several concept demonstrators are being developed e.g. for remote playing of family games to cultivate family to family relationships.

IP audio gateways Broadcasters use IP audio gateways for connections to studios accomplishing live interviews and live reports. In this field of application, compromises on audio quality are not acceptable. The flexibility in data rates from 24 up to 96 kbps per channel and higher in combination with the support of time stretching allows to adapt the codec's parameters to the IP network load and thus, keeping delay as low and audio quality as high as possible [8].

8 CONCLUSIONS

MPEG-4 Enhanced Low Delay AAC (AAC-ELD), a new codec for high quality communication, has recently been finalized. It provides a further enhancement of the well-established MPEG-4 AAC-LD codec, both in terms of quality and in terms of delay. Results of the MPEG verification test demonstrate the improvements of AAC-ELD compared to the previous state-of-the-art super-wideband communication codecs. Based on coding techniques which are well-established in the AAC codec family, an implementation of AAC-ELD proves to be straight-forward. Applications of AAC-ELD include traditional tele and video conferencing, IP audio gateways and any application where high audio quality in combination with interactive communication is desired even for channels with low capacity.

9 ACKNOWLEDGEMENTS

We gratefully acknowledge the valuable assistance the development of AAC-ELD has received from all participating listeners of the verification test at Fraunhofer IIS, Dolby, ETRI, LGE and Thomson.

10 REFERENCES

[1] C. Bürgel, R. Bartholomäus, W. Fiesel, J. Hilpert, A. Hölzer, K. Linzmeier, M. Weishart, "Beyond CD-Quality: Advanced Audio Coding (AAC) for High Resolution Audio with 24 bit Resolution and 96 kHz Sampling Frequency", *111th AES Convention*, New York, USA, preprint 5476, Sept. 2001

[2] International Telecommunication Union, "One-way transmission time", ITU-T Recommendation G.114, 2003

[3] R. Sperschneider, D. Homm, L.-H. Chambat, "Error Resilient Source Coding with Differential Variable Length Codes and its Application to MPEG Advanced Audio Coding", *112th AES Convention*, Munich, Germany, preprint 5555, May 2002

[4] R. Sperschneider, "Error Resilient Source Coding with Variable Length Codes and its Application to MPEG Advanced Audio Coding", *109th AES Convention*, Los Angeles, USA, preprint 5271, Sept. 2000

[5] P. Lauber, R. Sperschneider, "Error Concealment for Compressed Digital Audio", *111th AES Convention*, New York, USA, preprint 5460, Sept. 2001

[6] S.-U. Ryu, K. Rose, "A Frame Loss Concealment Technique for MPEG-AAC", *120th AES Convention*, Paris, France, preprint 6662, May 2006

[7] S.-U. Ryu, K. Rose, "Frame Loss Concealment for Audio Decoders Employing Spectral Band Replication", *121th AES Convention*, San Francisco, USA, preprint 6962, Oct. 2006

[8] J. Issing, N. Färber, M. Lutzky, "Adaptive Playout for VoIP based on the Enhanced Low Delay AAC Audio Codec", *124th AES Convention*, Amsterdam, The Netherlands, preprint 7395, May 2008

[9] H. Schulzrinne, S. L. Casner, R. Frederick, V. Jacobson, "RFC 3550 - RTP: A Transport Protocol for Real-Time Applications", 2003, URL `http://www.ietf.org/rfc/rfc3550.txt`

[10] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, E. Schooler, "RFC 3261 - SIP: Session Initiation Protocol", 2002, URL `http://www.ietf.org/rfc/rfc3261.txt`

[11] M. Schnell, M. Schmidt, P. Ekstrand, T. Albert, D. Przioda, M. Lutzky, R. Geiger, V. Ruoppila, F. Henn, E. Tårnes, "Delayless mixing – on the benefits of MPEG-4 AAC-ELD in high quality communication systems", *124th AES Convention*, Amsterdam, The Netherlands, preprint 7337, May 2008

[12] ISO/IEC International Standard 13818-7, "Information Technology - Generic Coding of Moving Pictures and Associated Audio, Part 7: Advanced Audio Coding", 1997

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

[14] E. Allamanche, R. Geiger, J. Herre, T. Sporer, "MPEG-4 Low Delay Audio Coding Based on the AAC Codec", *106th AES Convention*, Munich, Germany, preprint 4929, May 1999

[15] J. Hilpert, M. Gayer, M. Lutzky, T. Hirt, S. Geyersberger, J. Hoepfl, R. Weidner, "Real-Time Implementation of the MPEG-4 Low Delay Advanced Audio Coding Algorithm (AAC-LD) on Motorola DSP56300", *108th AES Convention*, Paris, France, Feb. 2000

[16] ISO/IEC 14496-3:2001/Amd.1:2003, "Coding of Audio-Visual Objects - Part 3: Audio, Amendment 1: Bandwidth extension", 2003

[17] M. Dietz, L. Liljeryd, K. Kjörling, O. Kunz, "Spectral Band Replication, a Novel Approach in Audio Coding", *112th AES Convention*, Munich, Germany, preprint 5553, Apr. 2002

[18] P. Ekstrand, "Bandwidth Extension of Audio Signals by Spectral Band Replication", *Proc. 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio*, Leuven, Belgium, Nov. 2002

[19] M. Schnell, R. Geiger, M. Schmidt, M. Jander, M. Multrus, G. Schuller, J. Herre, "Enhanced MPEG-4 Low Delay AAC - Low Bitrate High Quality Communication", *122nd AES Convention*, Vienna, Austria, preprint 6998, May 2007

[20] J. P. Princen, A. W. Johnson, A. B. Bradley, "Subband/Transform Coding Using Filter Bank Designs Based on Time Domain Aliasing Cancellation", *IEEE ICASSP*, pp. 2161–2164, 1987

[21] K. Nayebi, T. P. Barnwell, M. J. T. Smith, "Design of Low Delay FIR Analysis-Synthesis Filter Bank Systems", *Proc. Conf. on Info. Sci. and Sys. (CISS)*, Mar. 1991

[22] T. Q. Nguyen, "A Class of Generalized Cosine-Modulated Filter Bank", *International Symposium on Circuits and Systems (ISCAS)*, San Diego, USA, pp. 943–946, 1992

[23] G. Schuller, M. J. T. Smith, "A General Formulation for Modulated Perfect Reconstruction Filter Banks with Variable System Delay", *NJIT 94 Symposium on Appl. of Subbands and Wavelets*, Mar. 1994

[24] G. Schuller, M. J. T. Smith, "Efficient Low Delay Filter Banks", *DSP Workshop*, Yosemite, CA, USA, Oct. 1994

[25] G. Schuller, M. J. T. Smith, "New Framework for Modulated Perfect Reconstruction Filter Banks", *IEEE Transactions on Signal Processing*, vol. 44, no. 8, pp. 1941–1954, Aug. 1996

[26] G. Schuller, T. Karp, "Modulated Filter Banks with Arbitrary System Delay: Efficient Implementations and the Time-Varying Case", *IEEE Transactions on Signal Processing*, vol. 48, no. 3, Mar. 2000

[27] M. Schnell, R. Geiger, M. Schmidt, M. Multrus, M. Mellar, J. Herre, G. Schuller, "Low Delay Filterbanks for Enhanced Low Delay Audio Coding", *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, pp. 235–238, Oct. 2007

[28] ISO/IEC JTC1/SC29/WG11, "Report on the Verification Test of MPEG-4 Enhanced Low Delay AAC", MPEG2008/N10032, July 2008, URL `http://www.chiariglione.org/mpeg/quality_tests.htm`

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

[30] International Telecommunication Union, "Software tools for speech and audio coding standardization", ITU-T, Recommendation G.191, Geneva, Switzerland, 2005

[31] International Telecommunication Union, "Pulse code modulation (PCM) of voice frequencies", ITU-T Recommendation G.711, 1988

[32] International Telecommunication Union, "7 kHz audio-coding within 64 kbit/s ", ITU-T Recommendation G.722, 1988

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

[34] International Telecommunication Union, "Low-complexity coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss", ITU-T Recommendation G.722.1 Annex C, Geneva, Switzerland, 2005

[35] International Telecommunication Union, "Low-complexity full-band audio coding for high-quality conversational applications", ITU-T Recommendation G.719, Geneva, Switzerland, 2008

[36] "TA2 project - 'Together Anywhere, Together Anytime'", 2008, URL `http://www.ta2-project.eu/`