

International Telecommunication Union

ITU-R
Radiocommunication Sector of ITU

Report ITU-R BS.2161
(11/2009)

Low delay audio coding for broadcasting applications

BS Series
Broadcasting service (sound)



International
Telecommunication
Union

Foreword

The role of the Radiocommunication Sector is to ensure the rational, equitable, efficient and economical use of the radio-frequency spectrum by all radiocommunication services, including satellite services, and carry out studies without limit of frequency range on the basis of which Recommendations are adopted.

The regulatory and policy functions of the Radiocommunication Sector are performed by World and Regional Radiocommunication Conferences and Radiocommunication Assemblies supported by Study Groups.

Policy on Intellectual Property Right (IPR)

ITU-R policy on IPR is described in the Common Patent Policy for ITU-T/ITU-R/ISO/IEC referenced in Annex 1 of Resolution ITU-R 1. Forms to be used for the submission of patent statements and licensing declarations by patent holders are available from <http://www.itu.int/ITU-R/go/patents/en> where the Guidelines for Implementation of the Common Patent Policy for ITU-T/ITU-R/ISO/IEC and the ITU-R patent information database can also be found.

Series of ITU-R Reports

(Also available online at <http://www.itu.int/publ/R-REP/en>)

Series	Title
BO	Satellite delivery
BR	Recording for production, archival and play-out; film for television
BS	Broadcasting service (sound)
BT	Broadcasting service (television)
F	Fixed service
M	Mobile, radiodetermination, amateur and related satellite services
P	Radiowave propagation
RA	Radio astronomy
RS	Remote sensing systems
S	Fixed-satellite service
SA	Space applications and meteorology
SF	Frequency sharing and coordination between fixed-satellite and fixed service systems
SM	Spectrum management

Note: This ITU-R Report was approved in English by the Study Group under the procedure detailed in Resolution ITU-R 1.

Electronic Publication
Geneva, 2010

© ITU 2010

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without written permission of ITU.

REPORT ITU-R BS.2161

Low delay audio coding for broadcasting applications

(2009)

1 Operational requirements of a low delay audio coding**1.1 Requirements for digital wireless microphones**

For wireless microphones, it is essential to reduce latency so that the sound of voice such as speech or vocals is reproduced from loudspeakers simultaneously.

In this section, the required latency in each operation of a wireless microphone is described. The requirements were formulated by broadcast audio experts in Japan. Table 1 presents a list of requirements for digital wireless microphones.

1.1.1 Studio

In the studio, particularly for live broadcasting, the maximum acceptable delay time is approximately 1 ms to ensure smooth conversation over talk-back.

In other cases, for programme production in a studio for recording, the relative delay between picture and audio should be minimized and not vary although it can be adjusted later through the editing process. In addition, there are multiple audio sources in most cases, and differences of delays among the sources should be minimized. As some of the sources are from wired microphones, the delay time of wireless microphones should be less than approximately 1 ms.

It is also desirable for operators to know the actual delay time.

1.1.2 ENG and outside broadcasting

In outside broadcasting, the acceptable delay time is the same as in the studio case. In live sports programmes, the maximum acceptable delay time is approximately 5 ms for the sound of a ball and other sounds made in a game to achieve good synchronization with the picture. The delay time should not vary.

On the other hand, when it is not live broadcasting, the acceptable delay could be relaxed to approximately 25 ms as a trade-off to get robustness against interference.

It is also desirable for operators to know the actual delay time.

1.1.3 Talk-back

Speakers or singers find it difficult to speak or sing if their talk-back voice has significant latency, so very strict delay time management is required. The delay of both a wireless microphone and a talk-back circuit should be taken into consideration. A delay of less than 1 ms for studio use and less than 5 ms for outside broadcasting are required. The delay time should not vary.

It is also desirable for operators to know the actual delay time.

1.1.4 Concerts

On the stage, various delays are generated depending on the allocation of speakers and microphones. For example, a 3-ms delay corresponds to a distance of 1 m. It is considered that professional players can detect a 2-ms delay.

The relative delay between wireless microphone and wired microphone or other electronic musical instruments should be minimized. The maximum acceptable delay is approximately 2 ms, or 1 ms if possible.

1.1.5 Musicals and plays

To express fine vocals and music performance, the delay time should be as small as possible. A good singer would not wish to use a microphone if the delay exceeds 3 ms, and more than 5 ms is unacceptable. As this value is the total delay time in the audio system from microphone to loudspeaker, the maximum acceptable delay of a wireless microphone is 2 ms.

1.1.6 In-ear monitor

Music players play the music picking up the beat through this monitor. The maximum acceptable delay is 1 ms.

TABLE 1
Requirements for digital wireless microphones

Application	Studio	ENG and outside broadcasting	Talk-back	Concerts	Musicals and plays	In-ear monitor
Content	Voice	Voice	Voice and broadcast programme	Voice and musical instruments	Voice and musical instruments	Voice and musical instruments in stereo
Audio frequency	20 Hz-20 kHz	20 Hz-20 kHz (50 Hz-10 kHz by trade-off with interference)	100 Hz-10 kHz (100 Hz-7 kHz by trade-off with interference or latency)	20 Hz-over 20 kHz	20 Hz-over 20 kHz	20 Hz-15 kHz
Audio dynamic range	More than 100 dB (preferably 20-bit linear PCM and more than 120 dB)	More than 100 dB	More than 70 dB	More than 100 dB	90 dB	95-100 dB
Maximum sound pressure level of microphone	More than 130 dBSPL	More than 140 dBSPL	–	140 dBSPL	130 dBSPL	–
Maximum acceptable latency	1 ms	5 ms (25 ms by trade-off with interference)	5 ms	2 ms	2 ms	1 ms
Audio interface	AES/EBU output at receiver		AES/EBU input at transmitter	AES/EBU output at receiver		AES/EBU input at transmitter

2 Delay calculation for audio codecs

It is essential to consider the delay introduced by the codecs when comparing the achieved audio quality. Therefore, the underlying assumptions of the delay calculation are described before the particular overall delays are defined in the corresponding codec §§ 2.1 to 2.4. For the codec described in § 2.5 the end-to-end delay was measured.

There are three main categories of delay in communication systems:

- Algorithmic delay – the part of the latency introduced by the algorithm which is independent from the properties of the transmission channel and the speed of the digital signal processor.
- Transmission delay – the part of the latency introduced by sending the bit reduced audio data from the encoder to the decoder.
- Processing delay – the part of the latency dependent on the processing speed of the digital signal processor.

As the focus of this document is a broadcasting environment, the transmission delay is important. We assume a restricted channel capacity (constant bit rate) for the data transfer, which equals the bit rate of the signal or exceeds it slightly. The additional delay in milliseconds caused by this so-called “continuous transmission” is equal to the average bit-stream frame-length divided by the network transmission clock θ . The “continuous transmission” mode (with a fixed upper limit for the bit rate) implies furthermore that the bit reservoir, which is implemented in several audio codecs, have to be incorporated into the delay calculation of the algorithmic delay.

Besides the aforementioned bit reservoir, other potential sources for algorithmic delay are the framing delay, the filter bank delay and the look-ahead delay for a potential block switching decision.

Assuming an unlimited hardware computation capacity for the encoding/decoding process, any additional delays caused by time for computation is therefore ignored. This is reasonable because the processing delay in general is small compared to the other two factors and is getting smaller with technical progress in microelectronics.

The algorithmic delay and the transmission delay sum up to the total delay of a codec.

2.1 ISO/IEC 11172-3 (MPEG-1 Audio) Layer II and MPEG-1 Layer III

2.1.1 Overview

The international standard ISO/IEC 11172 was introduced November 1992. The MPEG/Audio-Group developed the audio part of the standard, which is described in ISO/IEC 11172-3. This part consists of three perceptual coders, called Layers, with their complexity and performance increasing from Layers I to III.

Layer I can be thought of as a coding scheme for applications that do not require very low bit rates, e.g. for home recording on Digital Compact Cassette.

Layer II is used for broadcasting purposes. More precisely, Layer II is the coding scheme for contribution, distribution and emission applications in the broadcasting domain.

Layer III is more complex than the other layers, but achieves a higher compression performance.

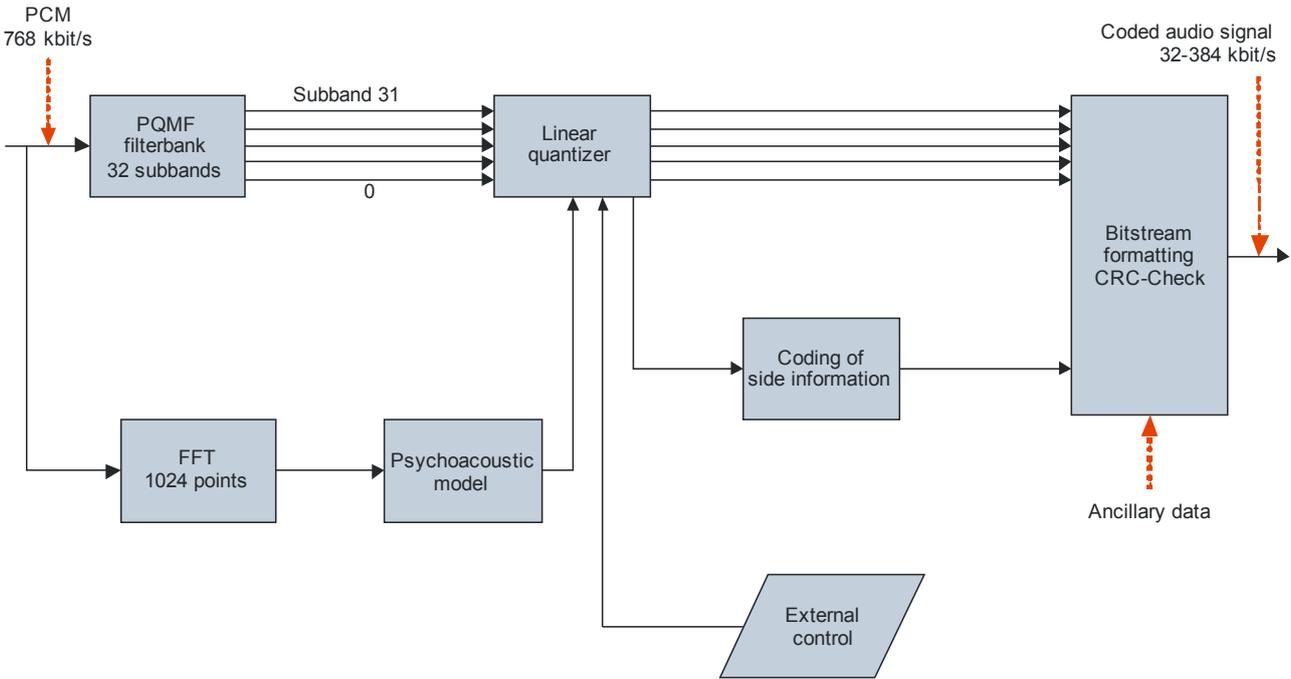
In the ISO standard, only the decoder structure and the bit stream are exactly defined. The encoding structure given in the standard can be seen as a minimum requirements version. Proprietary enhancements in the encoder, e.g. a better psychoacoustic model can be implemented.

Both Layers II and III are referenced in Recommendation ITU-R BS.1115.

2.1.2 Structure and delay of Layer II

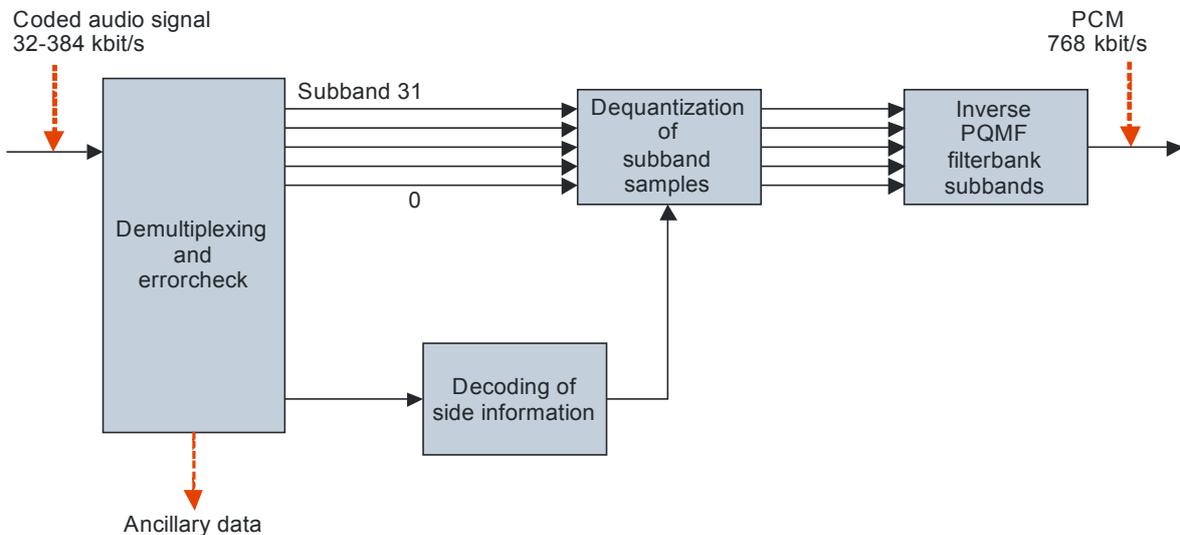
MPEG-1 Layer II uses a 32-channel polyphase filter bank (PQMF) with a length of 512 taps to map the time domain audio signal samples into the time-frequency domain. This filter bank provides a frequency resolution of 500 Hz at a sampling rate of 32 kHz and a time resolution of 1 ms. The Layer II encoder is shown in Fig. 1 and the decoder is shown in Fig. 2.

FIGURE 1
The system structure of the Layer II encoder



Report BS.2161-01

FIGURE 2
The system structure of the Layer II decoder



Report BS.2161-02

A psychoacoustic model is used to determine the masking threshold which is used to find the just noticeable noise level of each band in the filter bank. The actual quantizer level in each frequency sub-band of every time block is calculated by allocating the available bits depending on the difference between the maximum signal level and the masking threshold.

Before the psychoacoustic model is computed, the time domain signal has to be transformed, in parallel to the PQMF, by a 1 024 point FFT which causes no additional delay. Those two transforms work independently of each other.

The psychoacoustic model in Layer II delivers an accurate approximation of the masking threshold. It controls the quantization of the 32-channel PQMF sub-bands through its blockwise adaptive bit allocation by adjusting the quantization step size. The bit allocation is done in an iterative process and is based on the principle of minimizing the “total noise-to-mask ratio over the frame with the constraint that the number of bits used does not exceed the number of bits available for that frame” 0. This process has to be repeated until the needed bits for coding the samples, the scale factors and the bit allocation information are as close as possible to the number of available bits for the whole frame without exceeding the available bits.

The block length of one time domain frame is 1 152 samples which corresponds to 36 ms at 32 kHz. With 32 sub-bands this corresponds to 36 samples per sub-band in the time-frequency domain.

The quantized sub-band samples and the side information have to be multiplexed and transmitted. The Layer II codec is operating at bit rates between 32 kbit/s and 384 kbit/s.

In Layer II the overall delay adds up to 2 815 samples. 511 samples for the filter bank delay plus 1 152 samples for the transmission of the sub-band samples plus 1 152 samples for the framing.

This corresponds to an algorithmic delay of 59 ms at a sampling rate of 48 kHz, and an algorithmic delay of 88 ms at a sampling rate of 32 kHz.

2.1.3 Structure and delay of Layer III

In Layer III, there is a 32-channel PQMF filter bank cascaded with an 18-point MDCT. The cascading results in $32 \cdot 18 = 576$ spectral lines with a frequency resolution of 27.77 Hz and a time resolution of 6 ms at 32 kHz. The Layer III encoder is shown in Fig. 3 and the decoder is shown in Fig. 4.

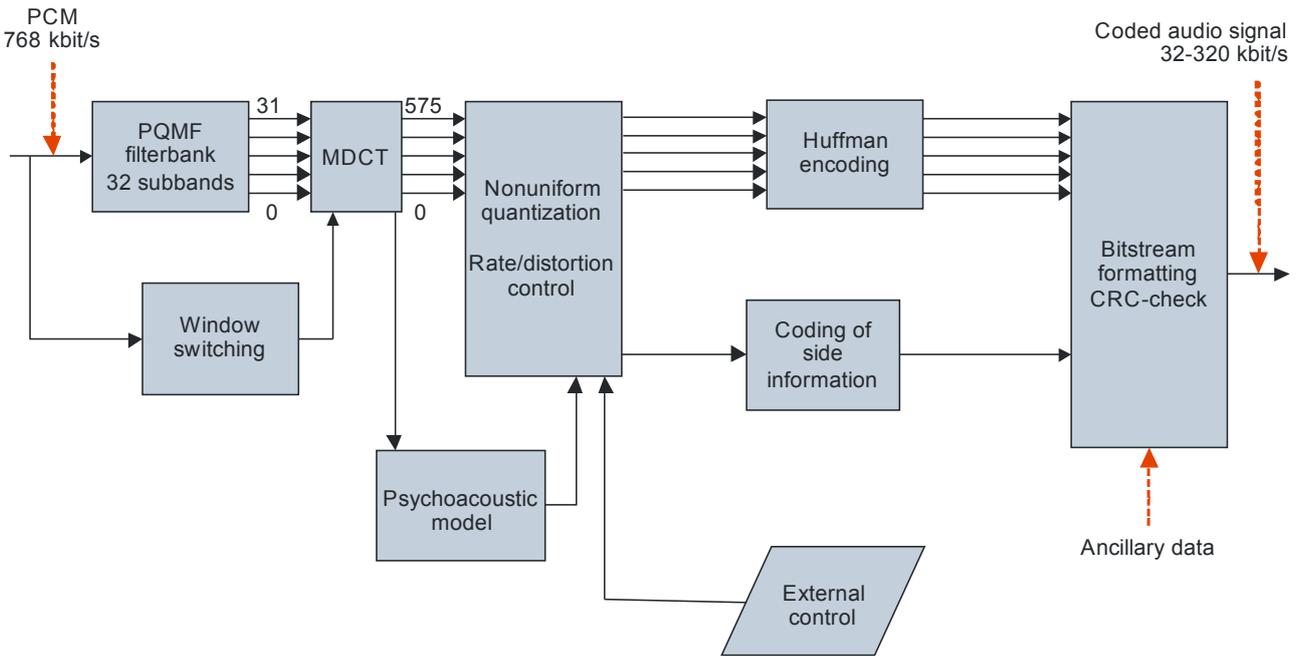
The psychoacoustic model processes the MDCT frequency domain output samples in order to provide the masking threshold and to estimate the tonality of the signal by using a magnitude-phase prediction scheme. The calculation causes no additional delay.

The window switching decision for the MDCT filter bank is done in parallel to the PQMF filtering in the time domain and produces a delay of 144 samples.

The Layer III frame length is equal to the Layer II frame length, which is 1 152 samples or 36 ms at a sampling rate of 32 kHz.

FIGURE 3

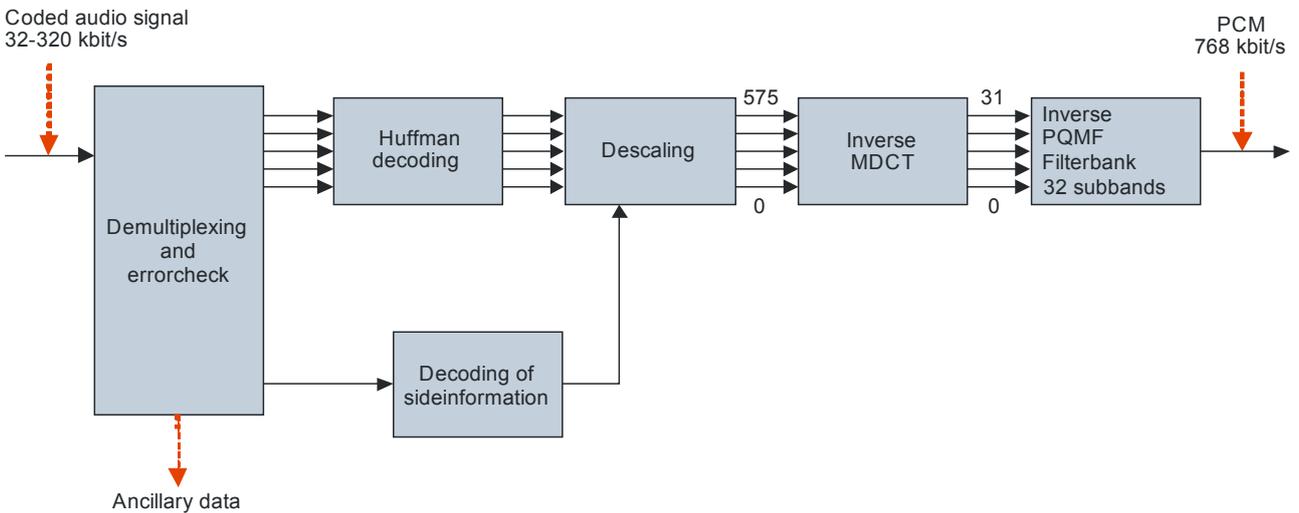
The system structure of the Layer III encoder



Report BS.2161-03

FIGURE 4

The system structure of the Layer III decoder



Report BS.2161-04

Layer III uses a non-uniform quantizer and a Huffman entropy coder with static tables to compress the sub-band samples. For sequences of zeroes a run-length-coder reduces the bit demand further. A bit reservoir technique allows for saving bits from frames which did not exploit the target bit rate. Those bits can be used to code subsequent frames of a higher complexity which would otherwise exceed the maximum channel bit rate.

The Layer III codec is operating at bit rates between 32 and 320 kbit/s and reaches excellent audio quality at 96 kbit/s/channel.

In Layer III the overall delay sums up to 5 809 samples. 144 samples delay are caused by the window switching look-ahead. 481 samples delay are caused due to the PQMF filter bank plus an MDCT delay of 576 samples. Furthermore, there are 1 152 samples delay for the transmission and 1 152 for the framing delay. And finally the delay due to the use of a bit reservoir has to be added which is 2 304 samples at 96 kbit/s.

This corresponds to an algorithmic delay of 121 ms at a sampling rate of 48 kHz and to an algorithmic delay of 181 ms at a sampling rate of 32 kHz.

2.2 ISO/IEC 14496-3 (MPEG-4 Audio) AAC-LD

2.2.1 Overview

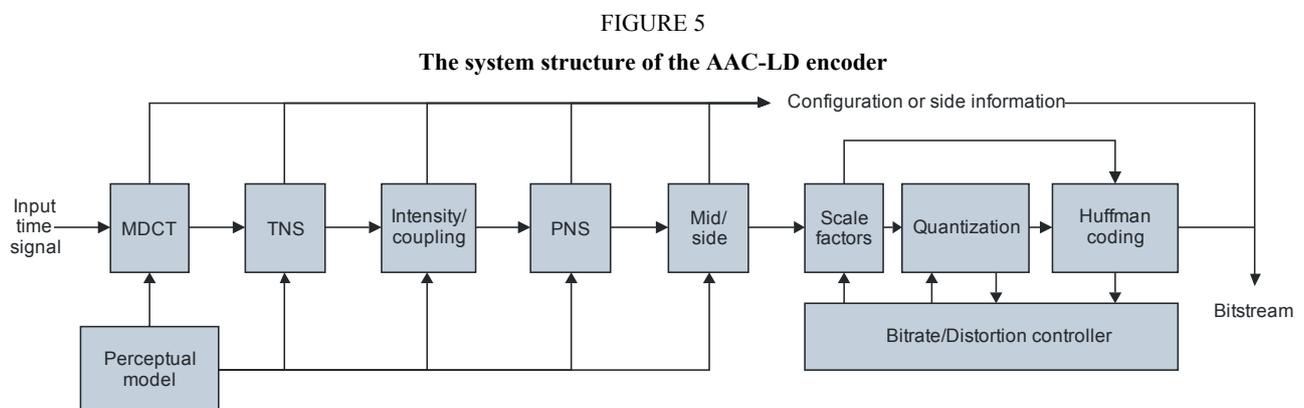
ISO/IEC 13818-7 (MPEG-2 Advanced Audio Coding (AAC)) was developed with the aim to meet the quality requirements as specified by “at a data rate of 384 kbit/s for five full-bandwidth channel audio signals”. An extended version of MPEG-2 AAC is also used for MPEG-4 general purpose audio coding. Operating modes of AAC fulfilling the requirements for broadcasting are listed in the appendices of Recommendation ITU-R BS.1548.

The low delay version AAC-LD is derived from the MPEG-4 AAC. The AAC Low Delay audio codec (MPEG-4 ER AAC-LD) combines the advantages of perceptual audio coding with the low delay necessary for two-way communication. AAC-LD was initially defined in “Version 2” of the MPEG-4 Audio specification and is part of the MPEG-4 standard. In MPEG-4, AAC-LD is standardized in an error robust form to cope with transmission errors often present in communication scenarios.

AAC-LD is a low-delay, a full-bandwidth and high-quality communications codec which is not bound to the limitations usual to speech coders (focus on single-speaker speech material, bad performance for music signals and background noise etc.). The codec is already used in many video/teleconferencing and other communication applications. AAC-LD delivers a good audio quality at a bit rate of 64 kbit/s/channel and operates at bit rates from 32 kbit/s/channel up to 320 kbit/s/channel.

2.2.2 Structure of the AAC-LD

The structure of the AAC-LD encoder is shown in Fig. 5:



AAC-LD uses an MDCT filter bank to transform the time domain input samples into a frequency domain representation. For AAC-LD the size of the MDCT analysis window can either be, 960 or 1 024 time domain samples. In the following, we choose a window length of 960 samples. For this window length the MDCT filter bank has a frequency resolution of 50 Hz and a time resolution of 10 ms at 48 kHz sampling rate. At a sampling rate of 32 kHz the frequency resolution is 66.66 Hz and the time resolution is 15 ms.

Temporal Noise Shaping is used in the AAC-LD to handle transient signals and allow the coder to exercise control over the temporal fine structure of the quantization noise, thus avoiding coding artefacts known as pre-echoes.

Intensity Coupling and Mid/Side Stereo increase the coding gain for encoding a stereo channel pair compared to encoding two mono channels separately.

Perceptual Noise Substitution (PNS) uses a parametric representation of noise-like frequency bands for an efficient transmission. Instead of transmitting single frequency lines, only the average power of the noise is transmitted to the decoder which then uses a random noise generator to regenerate these noise-like frequency bands.

A bit-rate control loop is used to adjust the non-uniform quantization of the spectral components in order to maintain a target bit rate and to meet the demands of the psychoacoustic model.

For AAC-LD a bit reservoir can be used optionally. Since the size of this bit reservoir influences the overall algorithmic delay, usually a minimized bit reservoir is designated for a continuous transmission. For the conducted listening tests, which are described in § 4, the AAC-LD operated with a 100 bit/channel bit reservoir.

2.2.3 Delay of AAC-LD

The overlap of the MDCT filter bank generates a delay of 480 samples, the framing adds another 480 samples, and the transmission also causes a delay of 480 samples. To these 1 440 samples an amount of 50 samples for the bit reservoir delay has to be added. The resulting overall algorithmic delay is 1 490 samples. This corresponds to 31 ms at a sampling rate of 48 kHz and at a sampling rate of 32 kHz. Regarding the delay calculation described in § 1, the delay of the AAC-LD is independent of the input signal sampling rate, since the input signal is internally resampled to 48 kHz.

2.3 ISO/IEC 14496-3 (MPEG-4 Audio) AAC-ELD

2.3.1 Overview

Enhanced Low Delay AAC (AAC-ELD) [Schnell *et al.*, 2008] is the successor of MPEG AAC-LD incorporating MPEG HE-AAC technology. It is standardized in ISO/IEC 14496-3:2005/Amd 9:2008. It adds the Spectral Bandwidth Replication (SBR) technology to the AAC-LD (see Chapter 2.2) enabling better quality at low bit rates. The AAC-ELD operates with a lower delay than AAC-LD at the same codec configuration. AAC-ELD is optimized for bit rates from 24 to 64 kbit/s per channel and sampling rates from 32 kHz to 48 kHz. Supported channel configurations are mono, stereo, and multi-channel. The codec can operate at block lengths of 480 or 512 samples.

2.3.2 Structure of AAC-ELD

AAC-ELD is a combination of an AAC-LD core coder with lower filterbank delay and a low delay version of the bandwidth extension technology Spectral Band Replication (SBR), well known from MPEG-4 High-Efficiency AAC (HE-AAC). The general structure is outlined in the following Fig. 6. In the encoder the input signal is sent to the SBR encoder and a down-sampled version is

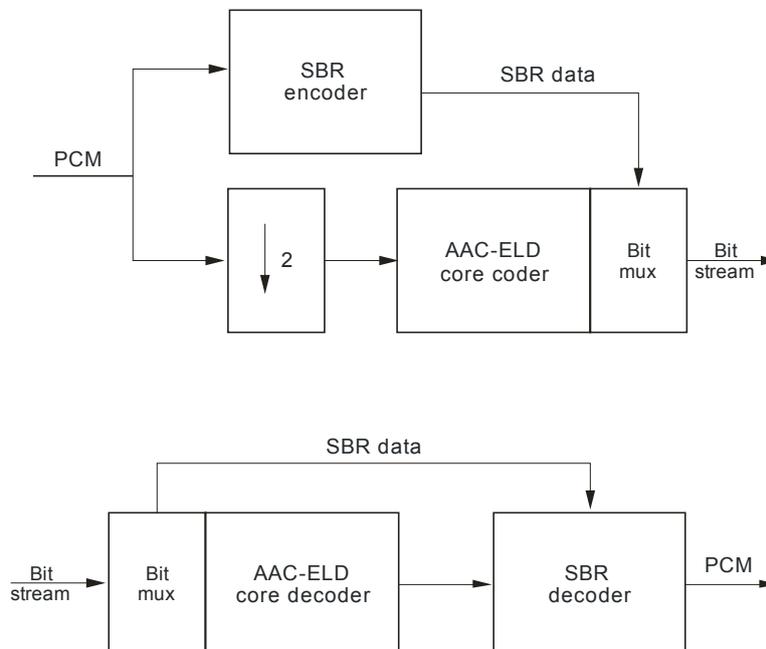
sent to the AAC-ELD core coder. In the decoder the signal decoded by the AAC-ELD core decoder is sent to the SBR decoder, where it is up-sampled and the bandwidth extension is applied.

The AAC-ELD core coder is derived from AAC-LD, but the MDCT 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.

Also the SBR module is optimized with respect to delay. The QMF inside the SBR decoder of HE-AAC is replaced by a complex low delay filter bank (CLDFB), which, similar to the LD-MDCT, reduces the delay by using asymmetrical window functions.

For lower bitrates the SBR module helps to improve the audio quality compared to AAC-LD. For higher bitrates this module and the corresponding down-sampling is not necessary, and thus AAC-ELD can also run in a mode comparable to AAC-LD, but with lower delay.

FIGURE 6
Structure of AAC-ELD encoder and decoder



Report BS.2161-06

2.3.3 Delay of AAC-ELD

The algorithmic delay of AAC-ELD varies depending on sampling rate and block length. At 480 samples block length the AAC core features a delay of 720 samples whereas SBR adds 64 samples delay.

TABLE 2
**Delay values in ms of typical AAC-ELD configurations
 compared to AAC-LD**

Bit rate (kbit/s)	AAC-ELD (ms)	AAC-LD (ms)
64	15 ⁽¹⁾	20 ⁽¹⁾
48	23 ⁽²⁾ /16 ⁽⁴⁾	30 ⁽²⁾
32	32 ⁽¹⁾ /16 ⁽⁴⁾	40 ⁽³⁾
24	32 ⁽¹⁾	–

⁽¹⁾ Sampling rate: 48 kHz.

⁽²⁾ Sampling rate: 32 kHz.

⁽³⁾ Sampling rate: 24 kHz.

⁽⁴⁾ Down sampled mode: sampling rate AAC core and SBR 48 kHz.

2.4 Fraunhofer ultra low delay coder

2.4.1 Overview

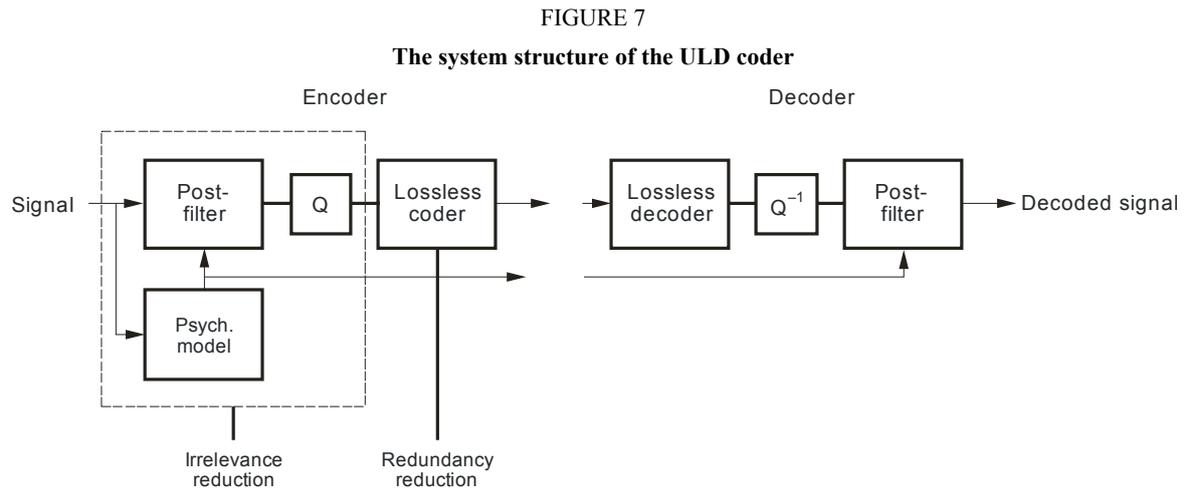
The goal of the Fraunhofer ultra low delay (ULD) coder is to obtain a very low encoding/decoding delay with a perceptual audio coder. With the ULD coder this is achieved by using the predictive coding principle, instead of sub-band coding as is generally more common for perceptual coders. Because of this different approach, the ULD coder can achieve delays which are unusually low for perceptual audio coders. For instance, for a 32 kHz sampling rate an algorithmic encoding/decoding delay of about 8 ms is achieved (excluding delay of a bit-rate limited channel). The usual operating point for this delay is in the range of 72 to 96 kbit/s. If desired, this delay can even be further reduced (for instance down to 2 ms), at the cost of a slight increase of the required bit rate for a comparable audio quality. In the case of the 2 ms delay, the resulting range of usual operating points is 128 to 192 kbit/s.

Another advantage of the ULD coder is that it features comparable quality for a variety of audio signals, including both music and speech, because of the predictive coding principle. Speech items are frequently a problem with sub-band based audio coders, due to the long frame lengths of sub-band coders.

The encoding/decoding delays are low enough to accommodate even very delay critical applications, like musicians playing together remotely over long distances, wireless speakers when there is a mix with wired speakers, wireless microphones, in ear monitoring for artists, wireless connections for hearing aids, communications between production studios and so on.

2.4.2 Structure of the ULD coder

The ULD coder separates the steps for irrelevance and redundancy reduction into independent functional units. This contrasts with traditional audio coding where both steps are integrated within the same sub-band decomposition. The separation allows for the independent optimization of the irrelevance and redundancy reduction units. Both reductions are obtained by adaptive filtering and predictive coding in order to minimize the system delay. While a psycho-acoustically controlled adaptive linear filter is used for the irrelevance reduction, the redundancy reduction employs a predictive lossless coding scheme that outperforms other competing lossless coders in terms of compression ratios and delay, as applied to the pre-filtered signal. The general system structure is shown in Fig. 7.



Report BS.2161-07

2.4.3 Characteristic parameters

The range of usable bit rates is between 32 and above 100 kbit/s/channel.

The delay depends on sampling rate and some other parameters. The following table gives some delay examples:

TABLE 3
Delay values of the ULD

Sampling-rate (kHz)	Low delay psych	Algorithmic delay (ms)	Real-time delay/hardware	Minimum bit rate per channel (kbit/s)
32	No	8	17 ms/demonstrator on evaluation board TI 320C6713	72
44.1	No	5.8		96
48	No	5.3	10 ms/suitable hardware with 100% workload for encoder	96

With the use of a low delay psychoacoustic module the following delay can be achieved:

TABLE 4
Delay values of the ULD with a low delay psychoacoustic module

Sampling-rate (kHz)	Low delay psych	Algorithmic delay (ms)	Estimated minimum real-time delay (ms)	Minimum bit rate per channel (kbit/s)
32	Yes		Below 3	128
44.1	Yes	1.45		160
48	Yes	1.3		192

Generally, there is a three-way trade-off between delay, audio quality at a certain bit rate and computational complexity, i.e. each of these characteristics can be improved at the cost of the others. As an example, the low delay psychoacoustic model called “Low delay psych” in Table 4 can modify the algorithm in a way that a further reduction of delay can be achieved by spending more bit rate.

By adding the mentioned transmission delay (blocks of size 128 at a 32 kbit/s/channel) to the ULD algorithmic delay, we get an overall delay of 384 samples respective 12 ms at 32 kHz sampling rate.

2.5 Sub-band ADPCM

A sub-band ADPCM codec has been developed for use in digital wireless microphones. The codec delay was measured for an actual codec, and was 3.7083 ms and 5.5625 ms depending on the time length of transmission frame.

TABLE 5

Delay values and coding parameters of a sub-band ADPCM codec

No	Bit rate (kbit/s)	Delay ⁽¹⁾ (ms)	Sampling frequency (kHz)
1	192	3.7083	48
2	128	3.7083	32 ⁽²⁾
3	128	5.5625	32
4	96	5.5625	32

⁽¹⁾ Total delay including codec and transport. The amount of delay depends on the time length of transmission frame.

⁽²⁾ 48 kHz sampling was not good in audio quality at this bit rate and latency.

3 Subjective quality evaluation of audio codecs

Sound quality, required bit rate and coding delay are the basic parameters to determine the possibility of using an audio codec for intended applications. In this section, two subjective listening tests of audio codecs are described.

3.1 AAC-LD and ULD in comparison with MPEG-1 Layer II and MPEG-1 Layer III

In the first test described, two audio codecs with reduced delay (AAC-LD and ULD) are compared to schemes recommended in Recommendations ITU-R BS.1115 and ITU-R BS.1548 (MPEG-1 Layer II and MPEG-1 Layer III). The subjective listening test was conducted to evaluate the performance of the presented codecs relative to each other.

3.1.1 Test procedure

The listening test was conducted using the test method described in Recommendation ITU-R BS.1534 (MUSHRA). The goal of the test was to evaluate the subjective performance of both the AAC-LD and the ULD codec in comparison with the established codecs MPEG-1 Layer II and MPEG-1 Layer III.

In the MUSHRA (**M**ulti **S**timulus test with **H**idden **R**eference and **A**ncor) test on each trial the uncompressed original at full bandwidth is presented to the listener as an open reference signal.

Each listener has to compare a hidden reference signal at full bandwidth, two anchor signals (3.5 kHz and 7.0 kHz band-limited versions of the original) and the unmarked codecs under test with the open reference, and grade the quality of each item on a scale ranging from bad to excellent.

The subjective listening test was conducted in an acoustically isolated listening room fulfilling the requirements of Recommendation ITU-R BS.1116.

The playback was controlled by a 2 GHz Linux computer with a RME Hammerfall digital sound output interface. The interface was connected to a Lake People DAC A54 D/A converter via a Kurzweil Digital Multitrack Interface and a Z-Sys z.32.32r digital audio router. STAX Lambda Pro headphones driven by a Stax driver unit with diffuse field equalizer were used for reproduction.

All codecs were available as software implementations. The test was implemented using only mono signals. The bit rate for the coders in the test was 96 kbit/s.

The test was accomplished by a total number of 12 experienced listeners who are all trained in listening to degraded audio material and skilled at detecting coding artefacts. The listeners ranged from 28 to 33 years in age and had normal hearing.

To reject non-reliable listeners, it was checked whether the subjects were able to detect the hidden reference. One subject has been excluded from the statistics because he clearly failed to indicate the reference for five of the eight testes cases. Another subject was excluded from the statistics because he severely failed to indicate the reference two times.

3.1.2 Audio item description

The test should evaluate the performance of the codecs under worst case conditions. Therefore, an effort was made to find material that would best reveal the limitations of each codec. Each of the tested codecs has been included in subjective tests before. A pre-selection of critical audio material was made by analysing the results of the previously conducted tests and selecting the test items that have shown to be most critical. This pre-selection was processed through each of the four codecs (AAC-LD, MPEG-1 Layer II, MPEG-1 Layer III (MP3) and ULD) and independently evaluated by eight expert listeners. Regarding a balance of the types of artefacts created by the codecs, for every codec the three items that were found to be most critical were selected. The final test set consisted of the eight items described in Table 5.

TABLE 6
List of audio items

Name	Length	Description	Original source
Glockenspiel	8 s	Glockenspiel melodious phrase	EBU/SQAM
Radio Jingle	8 s	Radio jingle with French male voice over background music	CCETT private recording
Ice Hockey	12 s	Scene from an ice hockey match with chanting fans and a radio commentary (English female voice)	IRT private recording
Pop Music	11 s	Pop music	“Rock You Gently” by Jennifer Warnes from the CD “The Hunter”
Harpsichord	17 s	Harpsichord arpeggio	EBU/SQAM
Castanets	15 s	Castanets	EBU/SQAM
Bagpipe	22 s	Bagpipe melodious phrase	BBC private recording
Pitch Pipe	28 s	Pitch pipe	Dolby private recording

3.1.3 Test results

The results of the listening test are shown in Fig. 7 as mean values with associated 95% confidence intervals. The overall delay of every codec is given next to its name (a detailed overview of the algorithmic and overall delays of all codecs is given in Table 6).

The MPEG-1 Layer II codec is the only scheme rated with an overall mean value clearly in the “good” range. Compared to the other codecs testes, the MPEG-1 Layer II scheme shows a performance which is worse in a statistically significant sense, since the confidence interval of this codec does not overlap with the confidence interval of any of the other codecs.

AAC-LD, MPEG-1 Layer III and ULD are rated with an overall mean value in the “excellent” range. The results indicate that the performance of the AAC-LD and MPEG-1 Layer III can be said to be equally high. The ULD shows a slightly but not statistically significant worse performance compared to AAC-LD and MPEG-1 Layer III.

The results show that the quality of the AAC-LD, the MPEG-1 Layer III, and the ULD codec is very high at the tested bit rate of 96 kbit/s for mono signals.

The overall mean ratings of those three codecs indicate that the impairments introduced at the tested bit rate (96 kbit/s/channel) are small enough to consider an evaluation using the test method described in Recommendation ITU-R BS.1116.

FIGURE 8
 Test results at 96 kbit/s mono (10 listeners)
 Average and 95% Confidence Intervals

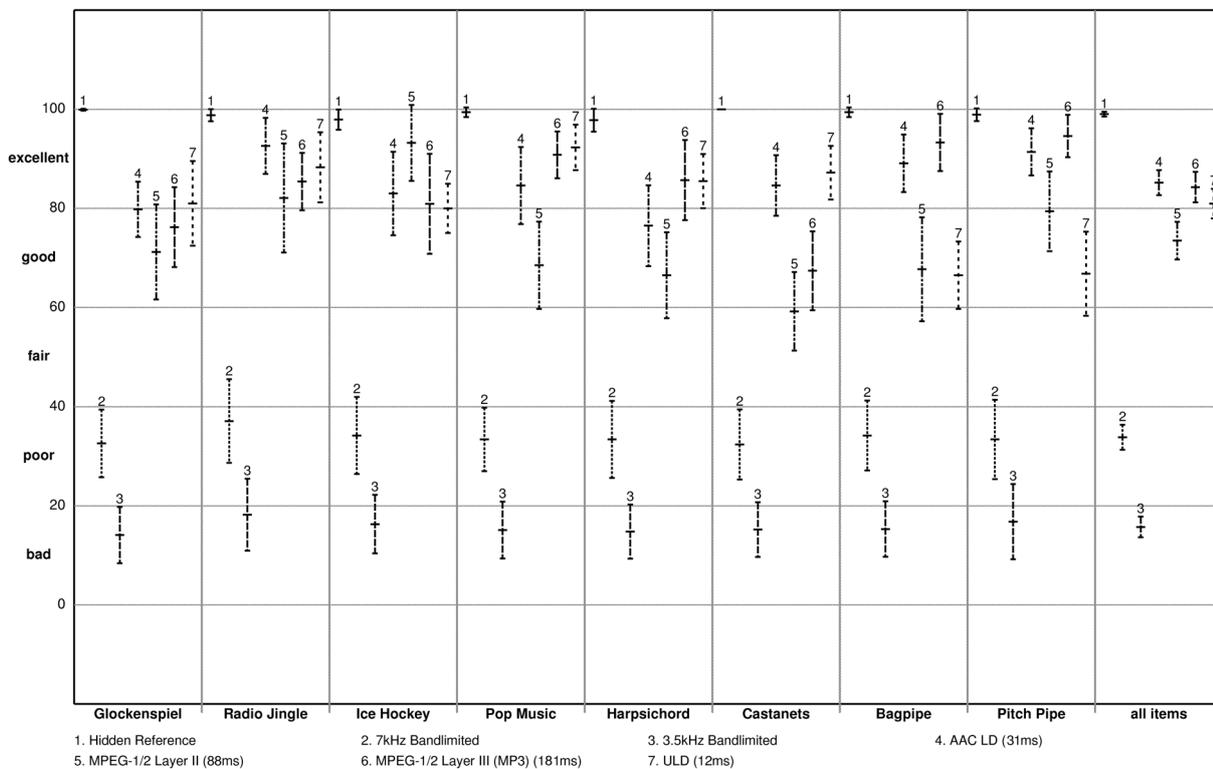


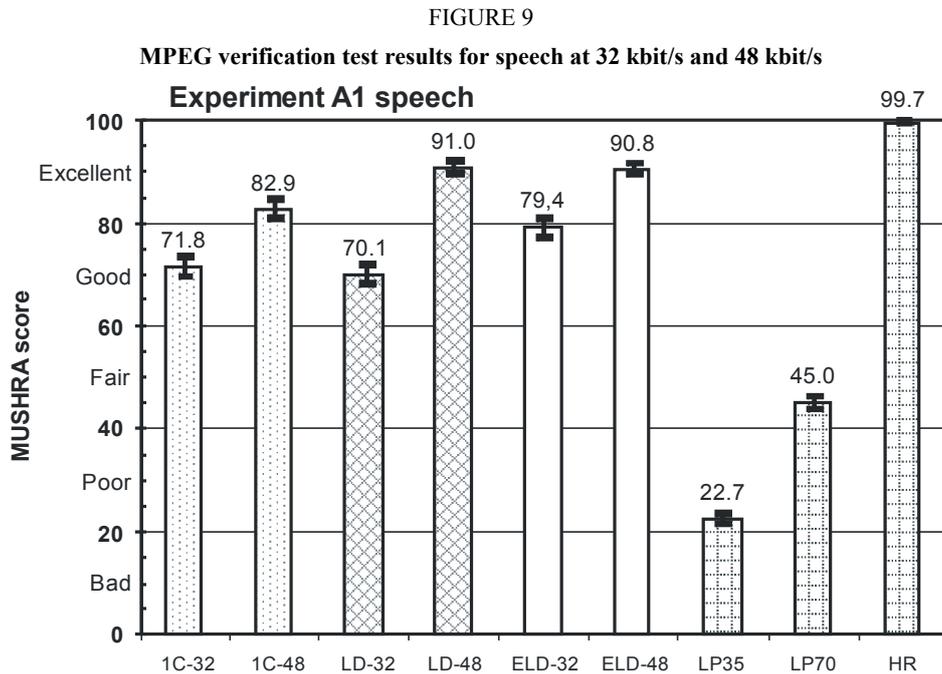
TABLE 7
**Algorithmic and overall (algorithmic + transmission delay)
 delays of the tested codecs⁽¹⁾**

	Algorithmic delay (ms)		Algorithmic + transmission delay (ms)	
	32 kHz	48 kHz	32 kHz	48 kHz
Layer II	52	35	88	59
Layer III	116	77	181	121
AAC-LD	21 ⁽¹⁾	21	31 ⁽¹⁾	31
ULD	8	5.3	12	8

⁽¹⁾ See explanation in § 2.2.3.

3.2 AAC-ELD listening test results

ISO/MPEG has performed listening tests to verify the audio quality of AAC-ELD. The following shows a part of the results which are documented¹.



Report BS.2161-09

¹ ISO/IEC JTC1/SC29/WG11 MPEG2008/N10032 Report on the Verification Test of MPEG-4 Enhanced Low Delay AAC.

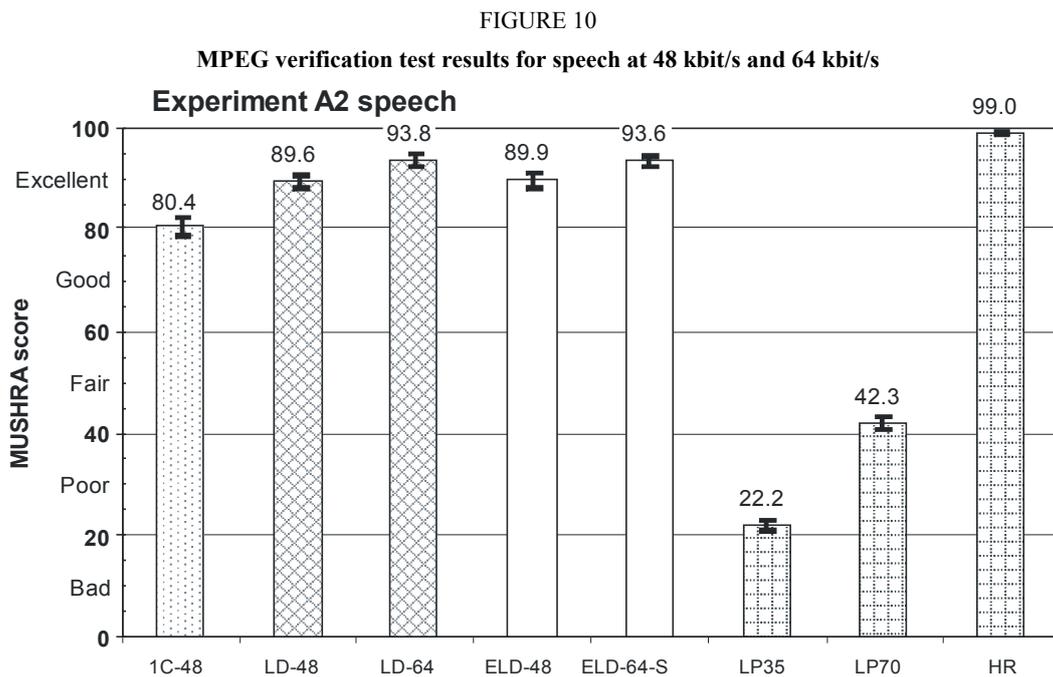
The Fig. 9 shows the performance of AAC-ELD (ELD-32, ELD-48), AAC-LD (LD-32, LD-48) and ITU-T G.722.1-C (1C-32, 1C-48) operating at 32 and 48 kbit/s. LP35 and LP70 are the 3.5 and 7 kHz low pass filtered anchors. HR is the hidden reference.

These tests included clean speech, reverberant clean speech, and speech with office background noise test conditions complemented with music and mixed content conditions. 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.

In all these experiments, the AAC-ELD codec showed good performance.

The mean scores for the AAC-ELD codec at a bit rate of 32 kbit/s were in the Good or Excellent range on the MUSHRA scale, and consistently, at the 95% level of significance, higher than those of the AAC-LD codec and the ITU-T Recommendation G.722.1-C codec at the same bit rate in all experiments. The AAC-ELD codec attained this performance gain with over 6 ms lower algorithmic delay compared to these two reference codecs that have an algorithmic delay of 40 ms.

For a bit rate of 48 kbit/s, the AAC-ELD codec was the only system under test that consistently had the mean score in the Excellent range of the MUSHRA scale in all experiments across all listening test laboratories. At this bit rate, both the AAC-ELD and AAC-LD codecs consistently outperformed, at the 95% level of significance, the ITU-T Recommendation G.722.1-C codec in all experiments.



The Fig. 10 shows that at a bit rate of 64 kbit/s, the AAC-ELD codec features, at the 95% level of significance, a mean score equal to that of AAC-LD, but has 5 ms (or 25%) lower algorithmic delay than this reference codec, which has an algorithmic delay of 20 ms.

3.3 Sub-band ADPCM

3.3.1 Test procedure

A subjective evaluation of audio quality was conducted using five audio sources which included human voice and the playing of musical instruments (see Table 8). Eighteen audio experts evaluated the audio quality in accordance with Recommendation ITU-R BS.1116-1 using the five-grade impairment scale shown in Table 9.

TABLE 8
Audio sources

No	Sources
1	Female vocal (ISO16: Suzanne Vega “Tom’s dinner”)
2	Male speech
3	Female speech
4	Electrical bass solo
5	Drum solo

TABLE 9
The five-grade impairment scale
(based on Recommendation ITU-R BS.1116-1)

Grade	Impairment
5	Imperceptible
4	Perceptible, but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

3.3.2 Test results

The test result of sound quality is shown in Table 10 and Figs 11, 12 and 13.

With the codec of 3.7083 ms latency, at the bit rate of 192 kbit/s and the sampling frequency of 48 kHz, the result shows that the quality difference between the coded audio and the hidden reference is almost within ± 0.5 in the five-grade scale for all the sources. This means that it achieves broadcast sound quality.

With the codec of 5.5625 ms latency, at the bit rate of 128 kbit/s and the sampling frequency of 32 kHz, the result shows that the quality difference between the coded audio and the hidden reference is almost within ± 1.0 in the five-grade scale for the vocal and speech sources. For “Electrical bass solo” the quality is not satisfactory. This means that it is possible to use this codec for specific purposes. The particular impairment of “Electrical bass solo” seems to be caused by the 32 kHz sampling that reduces second- and higher-order sound which characterize this sound.

Compared with the previous two test conditions, the codec of 3.7083 ms latency at 128 kbit/s and the codec of 5.5625 ms latency at 96 kbit/s are inferior in sound quality.

TABLE 10

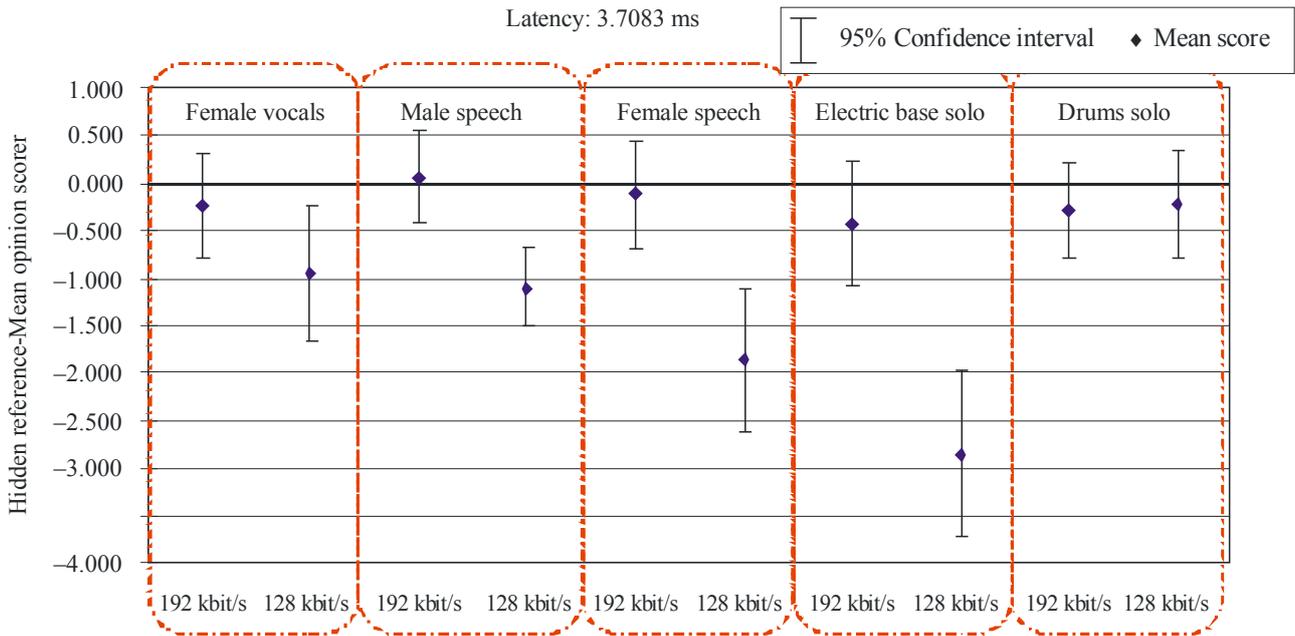
Analysis of the estimated sound quality and its confidence interval

Sources	Conditions of codec		Mean opinion score ⁽¹⁾	Confidence interval (95%, half side)
	Latency (ms)	Bit rate (kbit/s)		
Female vocals	3.7083	192	-0.231	0.549
Male speech			0.069	0.489
Female speech			-0.115	0.564
Electrical bass			-0.423	0.652
Drums solo			-0.285	0.485
Female vocals		128	-0.954	0.720
Male speech			-1.085	0.419
Female speech			-1.854	0.766
Electrical bass			-2.846	0.880
Drums solo			-0.215	0.561
Female vocals	5.5625	128	-0.600	0.470
Male speech			-0.115	0.606
Female speech			0.000	0.546
Electrical bass			-1.577	0.752
Drums solo			-0.592	0.478
Female vocals		96	-1.608	0.535
Male speech			-1.208	0.571
Female speech			-2.615	0.577
Electrical bass			-2.992	0.464
Drums solo			-0.900	0.603

⁽¹⁾ A bold letter in Table 10 shows that the mean opinion score is within ± 0.5 .

FIGURE 11

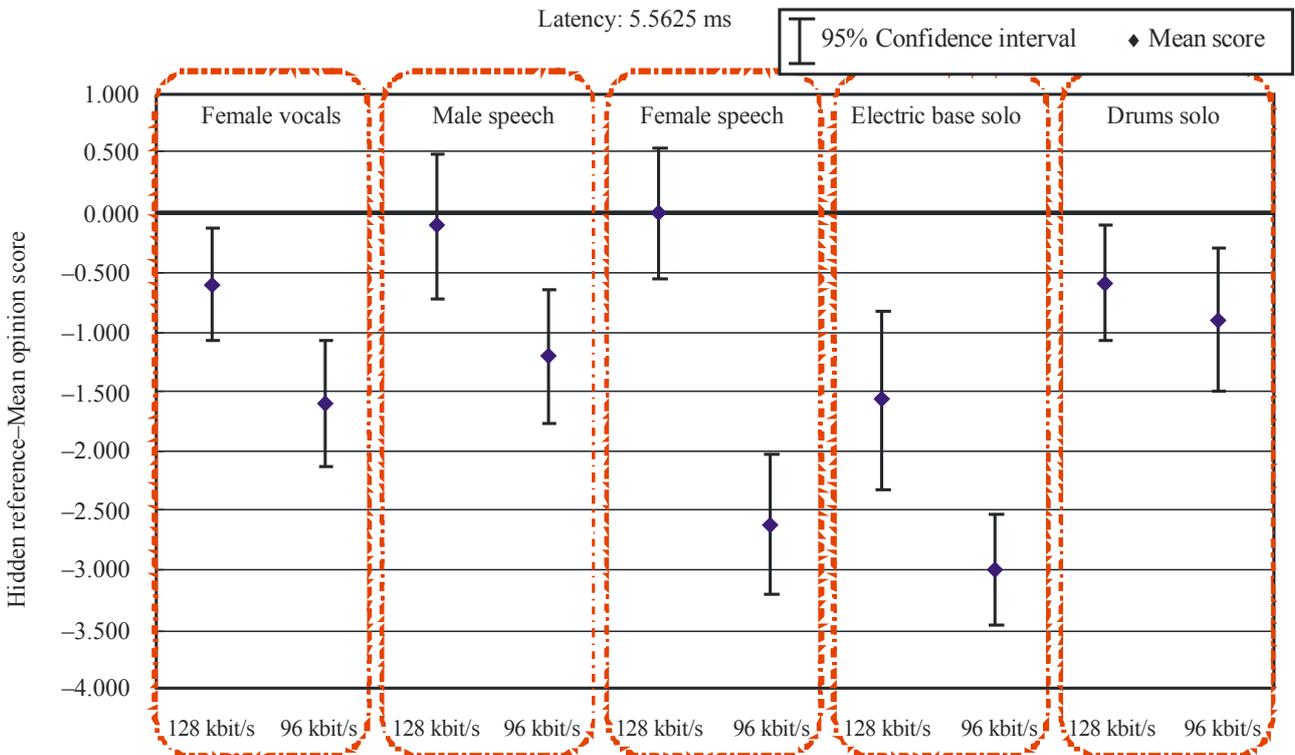
Evaluation result (latency = 3.7083 ms, at 192 and 128 kbit/s)



Report BS.2161-11

FIGURE 12

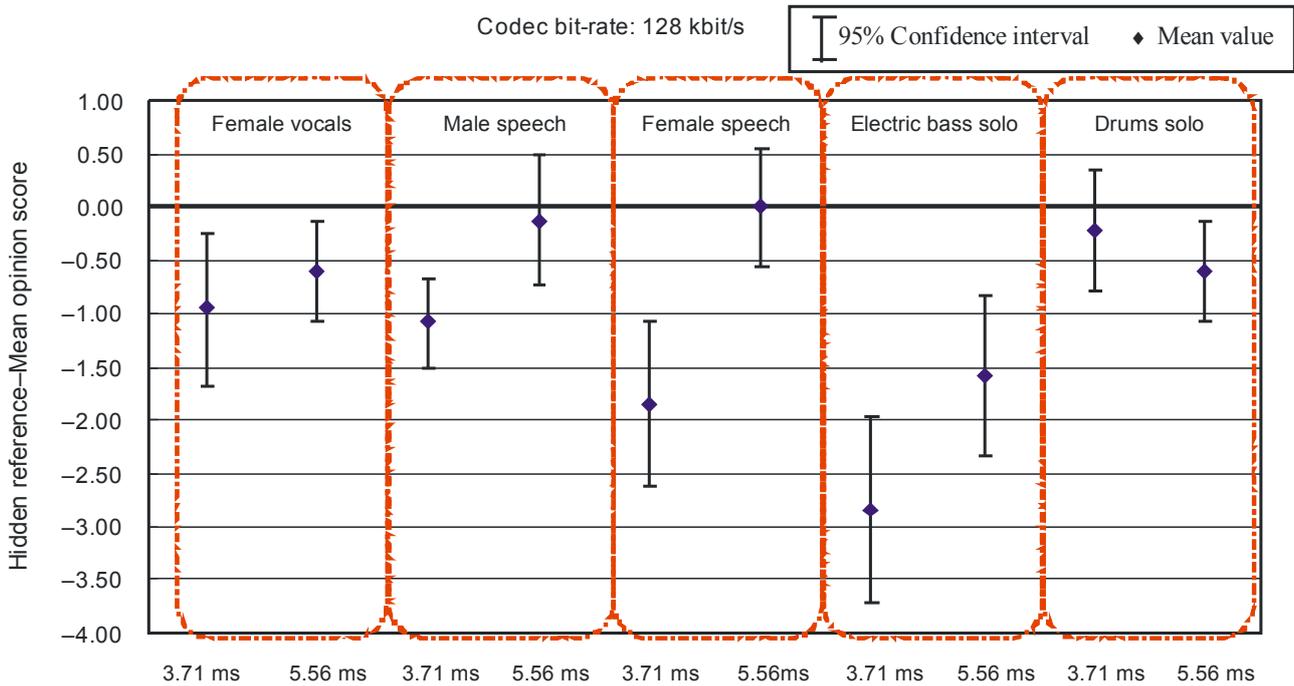
Evaluation result (latency = 5.5625 ms, at 128 and 96 kbit/s)



Report BS.2161-12

FIGURE 13

Evaluation result (latency = 3.7083 and 5.5625 ms at 128 kbit/s)



Report BS.2161-13

References

- MANFRED, L., SCHELLER, G., GAYER, M., ULRICH, K., WABNIK, S. [8-11 May 2004] A guideline to audio codec delay. 116th AES Convention, Berlin, Germany.
- ALLAMANCHE, E., GEIGER, R., HERRE, J., SPORER, T. [8-11 May 1999] MPEG-4 low delay coding based on the AAC Codec. 106th AES Convention, Munich, Germany.
- BOLAND, S., DERICHE, M., SRIDHARAN, S. [26-28 April 1995] High quality audio coding: An overview. 5th AES Australian Regional Convention, Sydney, Australia.
- STOLL, G. [28-30 June 1993] Status and future activities in standardization of low bit-rate audio codecs. 12th AES International Conference, Copenhagen, Denmark.
- ISO/IEC JTC1/SC29/WG11 (MPEG), International Standard ISO/IEC 11172-3 – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s.
- BRANDENBURG, K., BOSI, M. [January/February 1997] Overview of MPEG Audio: Current and future standards for low-bit-rate audio coding. JAES, Vol. 45, 1/2.
- [ITU-R, 28 October 1991] (Doc. TG10-2/3-E) –Basic audio quality requirements for digital audio bit-rate reduction systems for broadcast emission and primary distribution.
- [ISO/IEC14496-3] ISO/IEC JTC1/SC29/WG11 (MPEG), International Standard ISO/IEC 14496-3 – Coding of audio-visual objects: Audio.
- BOSI, M., GOLDBERG, R. [2003] Introduction to digital audio coding and standards. Kluwer Academic Publishers, Boston, United States of America.

DIMINO, G., PARLADORI, G. [6-9 October 1995] Entropy reduction in high quality audio coding. 99th AES Convention, New York, United States of America.

ISO/IEC JTC1/SC29/WG11 MPEG2008/N10032 Report on the Verification Test of MPEG-4 Enhanced Low Delay AAC.

MARKUS, S., MARKUS, S., Manuel J., Tobias, A., GEIGER, R., VESA RUOPPILA, PER EKSTRAND, MANFRED, L., BERNHARD GRILL [2-5 October 2008] MPEG-4 Enhanced Low Delay AAC – a new standard for high quality communication. 125th AES Convention, San Francisco, California.

