



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

Low delay audio coding for broadcasting applications

BS Series
Broadcasting service (sound)



Documents
Document Sharing Hub

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 Δ . 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.

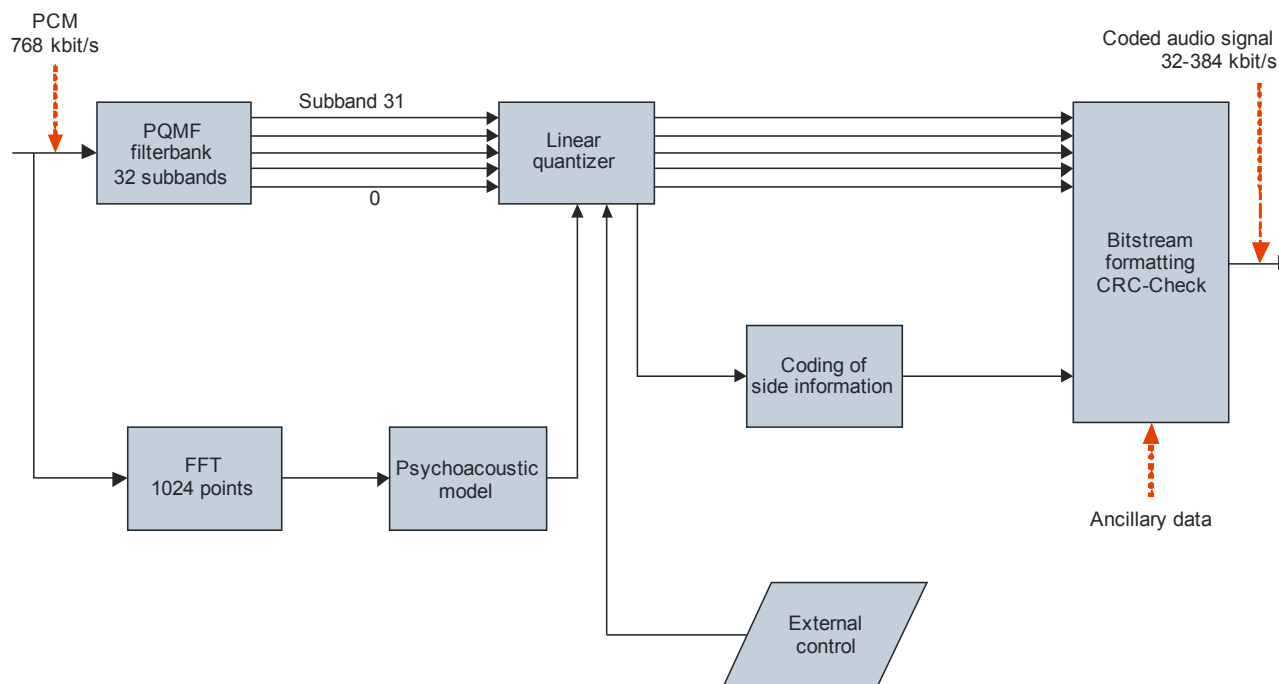
Both Layers II and III are referenced in Recommendation ITU-R BS.1115.

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

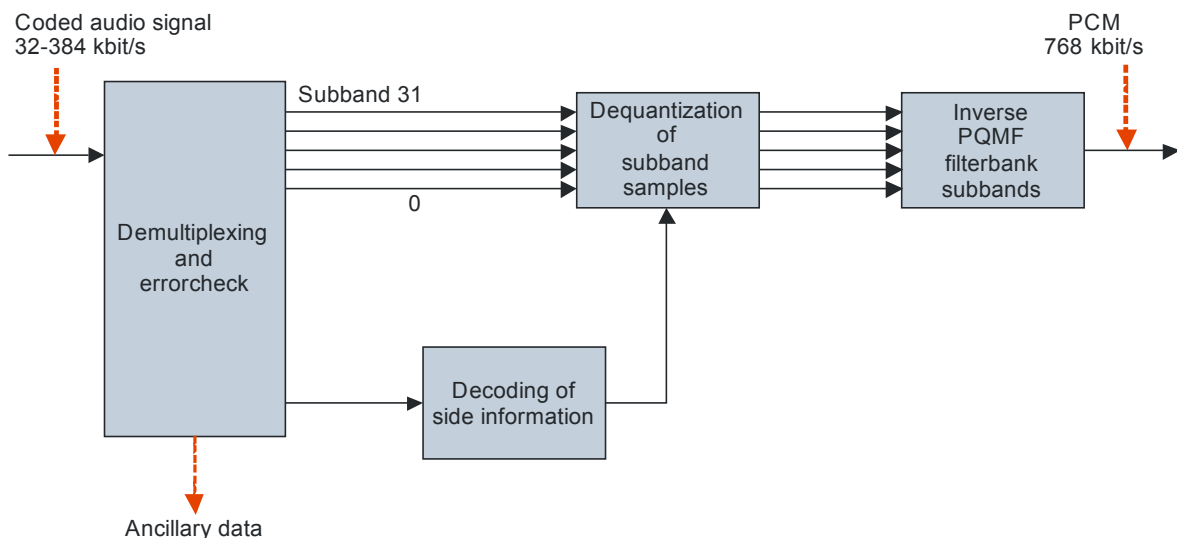
The system structure of the Layer II encoder



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” [10]. 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 $32 \times 18 = 576$ 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.