Information and Coding

Armando J. Pinho

Departamento de Electrónica, Telecomunicações e Informática

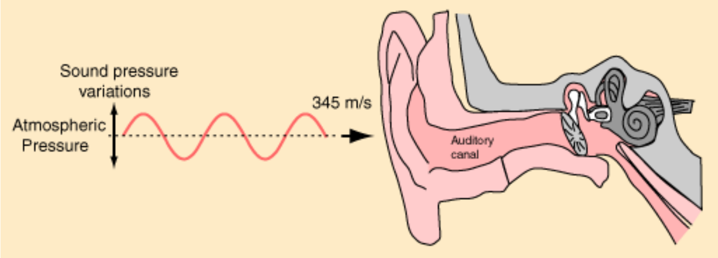
Universidade de Aveiro

ap@ua.pt

Contents

The human auditory system

- Humans perceive sound by the sense of hearing. By sound, we commonly mean the vibrations that travel through air and are audible to humans.



The human auditory system

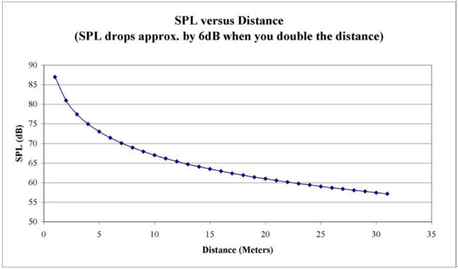
- Audio is the electrical representation of sound.
- Generally, humans can perceive variations in sound pressure from 16-20 Hz to 20-22 kHz.
- However, our capacity for perceiving sounds of very small amplitude varies according to frequency, being maximum between 2 and 4 kHz.
- The human voice produces frequencies approximately between 200 Hz and 8 kHz. Telephone communications limit this range from 300 Hz to 3.4 kHz (200 Hz to 3.2 kHz in the USA) .

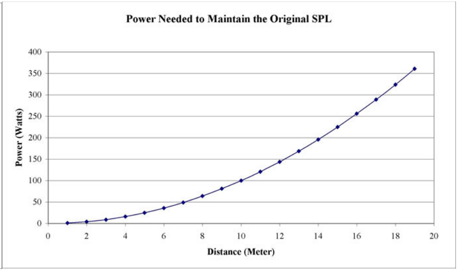
The human auditory system

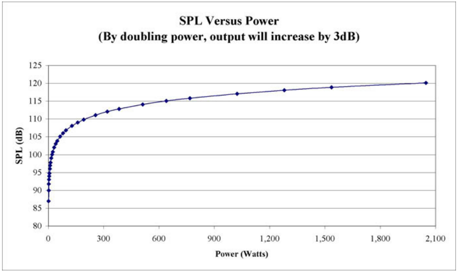
- Normally, the amplitude range that we can hear is about 100 dB:

Source of sound	Sound pressure level (dB)
Jet engine at 30 m	150
Jet engine at 100 m	140
Threshold of pain	125–130
Hearing damage (short-term exposure)	120
Maximum output of some MP3 players	110
Hearing damage (long-term exposure)	100
Major road at 10 m	80–90
TV (at home level) at 1 m	60
Normal talking at 1 m	40–60
Very calm room	20–30
Calm breathing	10
Auditory threshold at 2 kHz	0

The human auditory system

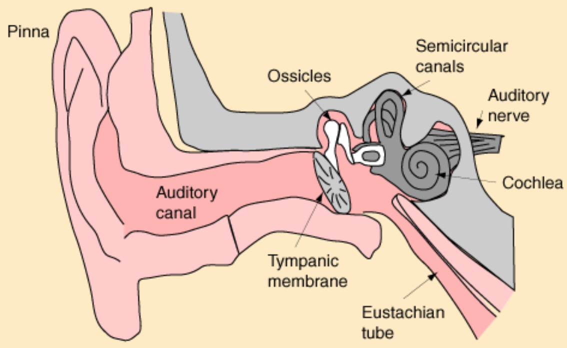






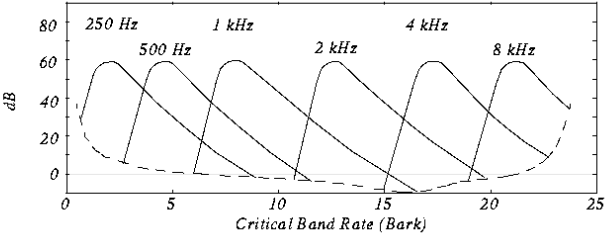
The human auditory system

- The auditory system can roughly be described as a bandpass filter-bank, consisting of strongly overlapping bandpass filters.



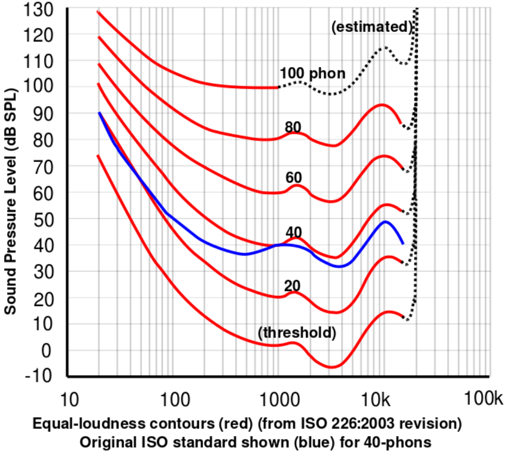
The human auditory system

- These “filters” have bandwidths in the order of 50 to 100 Hz for signals below 500 Hz and up to 5000 Hz for signals at high frequencies.



- They are called critical bands.
- Twenty-five critical bands, covering frequencies of up to 20 kHz, are normally taken into account.

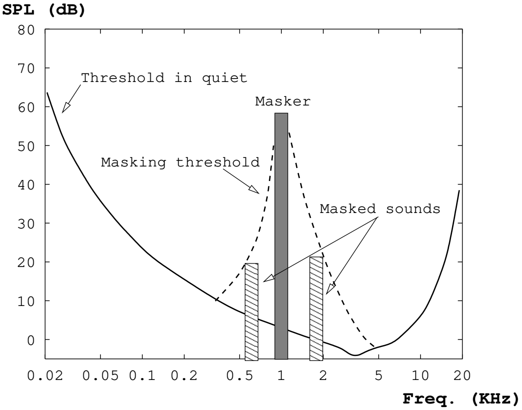
The human auditory system



The **phon** number is the SPL (dB) of a sound at 1 kHz that sounds just as loud. So, these are equal loudness curves.

The human auditory system

- Simultaneous masking is a frequency domain phenomenon where a low-level signal (maskee) can be made inaudible (masked) by a simultaneously and close in frequency stronger signal (masker).



The human auditory system

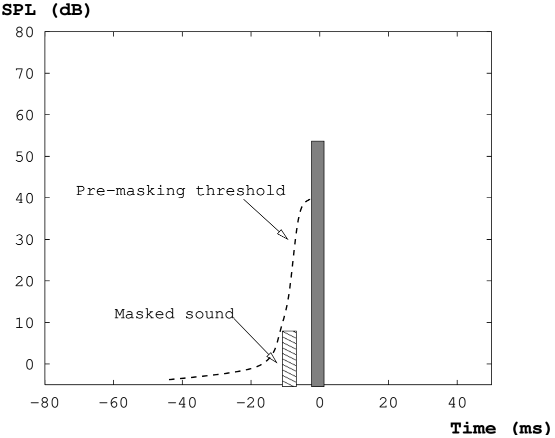
- Such masking is greatest in the critical band in which the masker is located, and it is effective to a lesser degree in neighboring bands.
- The masked signal can consist of:
  - Low-level signal contributions.
  - Quantization noise.
  - Aliasing distortion.
  - Transmission errors.

The human auditory system

- In addition to simultaneous masking, the time domain phenomenon of temporal masking plays an important role in human auditory perception.
- Temporal masking may occur when two sounds appear within a small interval of time.
- Depending on the individual sound pressure levels, the stronger sound may mask the weaker one, even if the maskee precedes the masker…

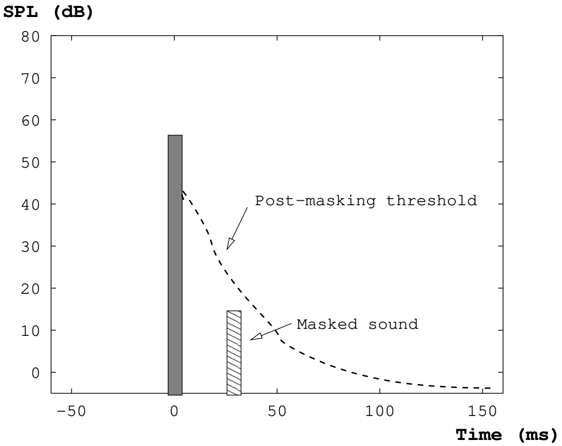
The human auditory system

- The pre-masking has a duration of about 5 to 20 ms:



The human auditory system

- The post-masking has a duration of about 50 to 200 ms:



Quality assessment of audio

- The audio quality may be evaluated using subjective or objective measures.
- One of the scales used for subjective evaluation of wide band audio codecs is the ITU-R five grade impairment scale:

5.0	Imperceptible
4.0	Perceptible, but not annoying
3.0	Slightly annoying
2.0	Annoying
1.0	Very annoying

- Regarding objective measures, the signal-to-noise-ratio (SNR) is the most used.

Contents

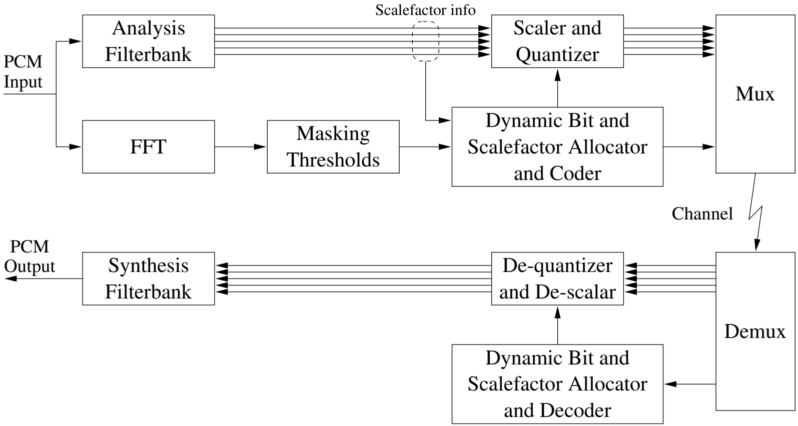
MPEG–1

- MPEG–1 audio coding is organized in three layers, I, II and III, with increasing performance, but also complexity and delay.
- It allows sampling frequencies of 32, 44.1 and 48 kHz, and bitrates between 32 kb/s (mono) and 448 kb/s (Layer I), 384 kb/s (Layer II) and 320 kb/s (Layer III).
- In terms of transparent CD (stereo) quality, the bitrates and compression rates are, approximately,

Layer	Bitrate	Compression rate
I	384 kb/s	4
II	192 kb/s	8
III	128 kb/s (VBR)	12

MPEG–1

MPEG–1 layer I and II



MPEG–1

- The analysis filterbank has 32 subbands, equally spaced in frequency.
- Each block is formed of 384 audio samples (8 ms for $f_s = 48$ kHz), meaning that each subband contains 12 samples.
- For $f_s = 48$ kHz, the width of each subband is 750 Hz.
- Usually, the bits are dynamically assigned to the coefficients of the subbands according to a psychoacoustic model (how to do this is not part of the standard).
- For each block of 12 coefficients (subband), an uniform quantizer (from 15 available) is selected, according to predefined levels of quality and compression.

MPEG–1

- The 12 coefficients of each block (subband) are divided by a scale factor, normalizing it to a maximum value of one.
- The main differences of MPEG–1 layer II with respect to layer I are:
  - Use of super-blocks obtained by grouping 3 consecutive (in time) blocks of 12 coefficients of a subband (24 ms, for $f_s = 48$ kHz).
  - In this case, bit assignment is performed on a super-block basis.
  - The scale factors are still calculated for each block of 12 samples, but they are encoded together for reducing the redundancy.

MPEG–1

MPEG–1 layer III

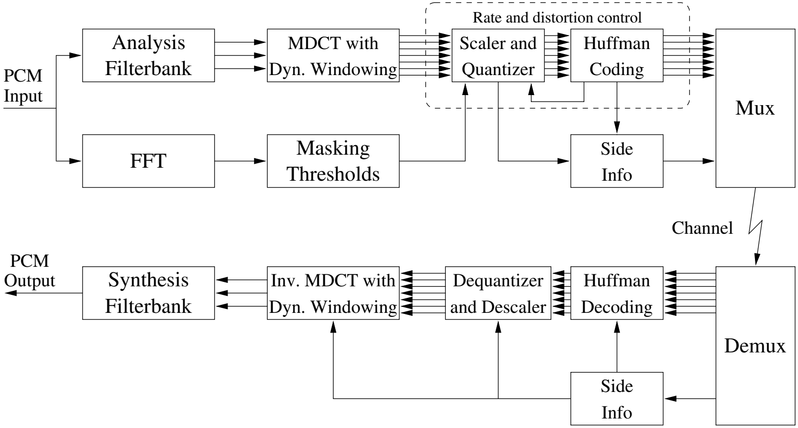
- MPEG–1 layer III has several differences in relation to the other two layers, being much more complex.
- It is based on hybrid coding: subband and transform.
- It allows variable bitrate coding (VBR).
- It relies on a technique designated analysis by synthesis for dynamic bit assignment.
- It uses an advanced pre-echo control.
- It uses non-uniform quantization and statistical coding.

MPEG–1

- For increasing the frequency resolution (for a better critical band approximation), each of the 32 subbands is transformed using a modified DCT of 6 or 18 points, with 50% overlapping.
- In this case, the maximum number of frequency components is $32 \times 18 = 576$, each one representing a 41.67 Hz band (for $f_s = 48$ kHz).
- The 18 point transform is used when higher frequency resolution is required.
- The 6 point transform is used for better temporal resolution, for example for preventing pre-echos.

MPEG–1

MPEG–1 layer III



MPEG–2

- The MPEG–2 audio coding standard includes the MPEG–1 audio and introduces extensions for multi-channel configurations.
- Regarding the multi-channel configurations, we have:
  - MPEG–1 allows audio mono, stereo, dual, two separate channels and joint stereo.
  - Besides those, MPEG–2 allows 3/2 stereo (L, R, C, LS and RS), as well as other combinations of these 5 channels.
- MPEG–2 provides two audio coding standards: one forward and backward compatible with MPEG–1, the other incompatible.
- Forward compatibility means that a multi-channel decoder understands MPEG–1 mono and stereo streams.
- Backward compatibility means that a MPEG–1 decoder is able to extract stereo audio from a MPEG–2 multi-channel stream.
- The second standard is MPEG–2 AAC (Advanced Audio Coding).

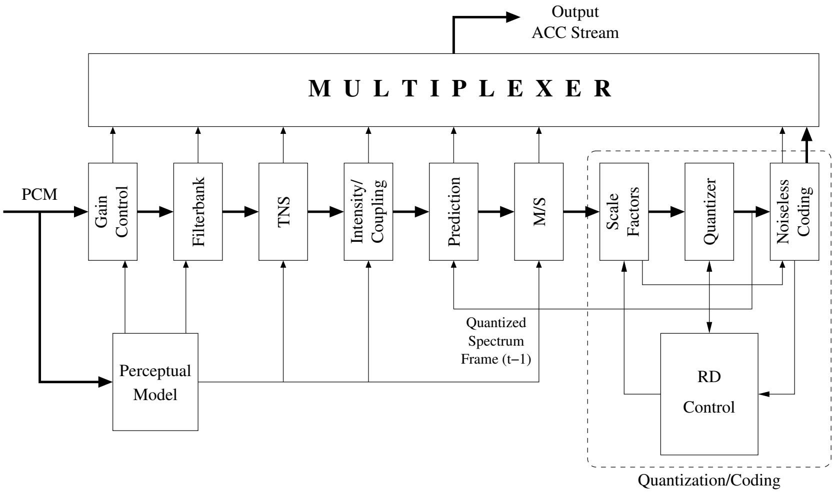
MPEG–2 AAC

- MPEG–2 AAC was developed after the introduction of the MPEG–2 multi-channel standard (compatible with MPEG–1).
- The standardization process was concluded in 1997 (MPEG–2 Part 7, ISO/IEC 13818-7).
- Main objective: to achieve transparent quality at 384 kbit/s or less for 5–channel audio.
- Almost transparent quality for 256 to 360 kbit/s for 5 channels and between 96 and 128 kbit/s for stereo.
- Excellent results for low bitrates (16 kbits/s).

MPEG–2 AAC

- Some parts are identical to those of MPEG–1/2 layer III:
  - Filterbank with dynamic windowing.
  - Non-uniform quantizers.
  - Huffman coding.
- MPEG–2 AAC defines 3 coding profiles:
  - The low complexity profile.
  - The main profile.
  - The scalable sampling rate profile.

MPEG–2 AAC


Output
ACC Stream
M U L T I P L E X E R
PCM
Gain Control
Filterbank
TNS
Intensity/ Coupling
Prediction
M/S
Scale Factors
Quantizer
Noiseless Coding
Perceptual Model
Quantized Spectrum Frame (t−1)
RD Control
Quantization/Coding

MPEG–2 AAC

Filterbank

- High resolution filterbank based on the modified discrete cosine transform (MDCT).
- It uses 50% overlapping between consecutive windows (2048 samples sliding 1024 samples).
- The filterbank produces 1024 spectral coefficients (frequency resolution of 23.4 Hz for $f_s = 48$ kHz).
- In zones where signal transitions occur, the resolution of the filterbank is reduced 8 times (producing only 128 coefficients per frame). This improves the temporal resolution.

MPEG–2 AAC

Quantization

- Quantization is performed according to

$$x_q = \mathrm{sign}(x)\,\mathrm{round}\left[\left(\frac{|x|}{\left(\sqrt[4]{2}\right)^s}\right)^{3/4} - \alpha\right]$$

  where α is a small constant and s is a scale factor associated to the resolution of the quantization.
- The scale factors are calculated for groups of spectral coefficients multiples of 4, at most 32, covering approximately 1/2 bark.
- The scale factors are differentially encoded in relation to the scale factors of adjacent bands, using Huffman codes.

MPEG–2 AAC

Statistical coding

- Huffman coding is provided by 11 code-tables (plus one pseudo-table indicating that all coefficients are zero).
- The bands are grouped with the aim of reducing the code size (including the auxiliary information).
- For each group, information regarding the identification of the Huffman code-table, the number of bands grouped and the corresponding codewords is sent to the decoder.
- The Huffman code-tables allow joint coding of 2 or 4 coefficients in a single codeword.

MPEG–2 AAC

Prediction

- The spectral coefficients of the frame being encoded are estimated using the spectral coefficients of the two previous frames.
- The estimator is adapted using a minimum mean squares algorithm.
- Because it is a backward adaptive process, it is not required to send auxiliary information to the decoder.
- This is the dual process of another technique used in AAC: Temporal Noise Shaping (TNS). In this case, it is performed temporal prediction, whereas in TNS it is spectral prediction.

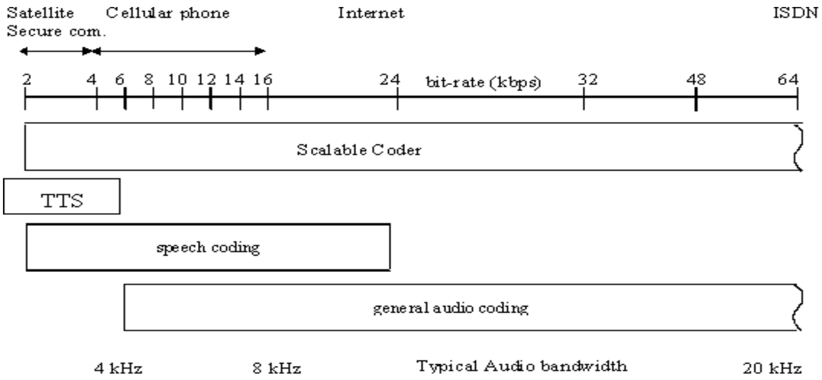
MPEG–4

- The MPEG–4 coding standard provides tools for coding audio objects, such as natural audio (for example, speech and music) and synthetic audio, aiming several applications:
  - Telephone over the Internet.
  - High quality music.
  - Text-to-speech conversion.
  - Synthesized music.
  - ...
- The synthesized audio can be obtained through text (TTS) or instrumental descriptions.

MPEG–4

- The encoding of the natural audio relies on several techniques:
  - Harmonic vector excitation coding (HVXC), for $f_s \leq 8$ kHz, and bitrates between 2 and 4 kbp/s (until 1.2 kbp/s, for VBR).
  - Code excited linear predictive (CELP), for $8 \leq f_s \leq 16$ kHz, and bitrates between 4 and 24 kbp/s.
  - Transform-domain weighted interleave vector quantization (TwinVQ) and AAC, for $f_s \geq 8$ kHz, and bitrates greater than 6 kbp/s.
- There are also means for:
  - Error resilience.
  - Low-delay audio coding.
  - Large-step scalability (MPEG–4 v1), and fine-grain scalability (FGS) (MPEG–4 v2).

MPEG–4


Satellite
Secure com.
Cellular phone
Internet
ISDN
2
4
6
8
10
12
14
16
24
bit-rate (kbps)
32
48
64
Scalable Coder
TTS
speech coding
general audio coding
4 kHz
8 kHz
Typical Audio bandwidth
20 kHz

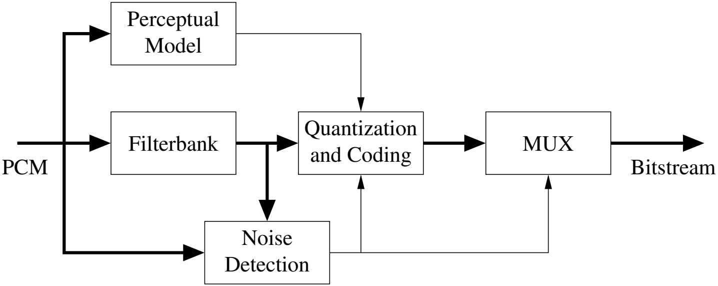
MPEG–4

- Besides the techniques provided by AAC, MPEG–4 includes several extensions in the time/frequency encoder:
  - Perceptual noise substitution (PNS).
  - Long time prediction (LTP).
  - Transform-domain weighted interleave vector quantization (TwinVQ).
  - Low-delay AAC.
  - Error resilience.
  - Scalable coding.

MPEG–4

Perceptual noise substitution

- Produces a perceptual equivalent signal, instead of trying to reproduce the original waveform.
- It is used for audio components similar to noise.
- If this type of signal is detected in a certain band, then instead of coding the coefficients, it is encoded their total power.



MPEG–4

Low-delay coding

- Mode used in bi-directional, real-time, communications, where long delays are not acceptable (the “standard” encoder may introduce delays of several hundreds of ms).
- It uses windows with half the normal size.
- It does not use dynamic window adaptation (this implies long delays). The pre-echos are controlled only by the TNS.
- The bit reservoir is minimized or even eliminated.
- Even with all these restrictions, this encoder only requires a bitrate increase of about 8 kbit/s in comparison to the “standard” encoder. In fact, it offers better quality than a MP3 encoder at 64 kbit/s/ch.

MPEG–4

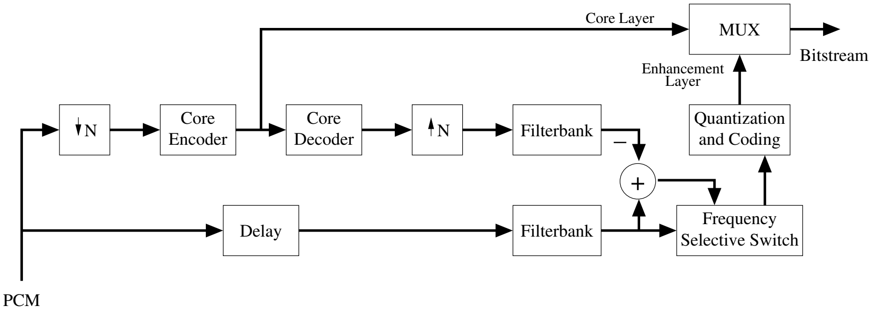
Scalable coding

- Scalable coding is important when the transmission channels have variable characteristics.
- MPEG–4 is the first standard allowing scalable audio coding.
- MPEG–4 uses the concept of hierarchical scalable coding:
  - A base encoder produces the base layer of the scalable bitstream.
  - Then, the difference between the original and the signal encoded by the base layer is further encoded, producing an enhancement layer.

MPEG–4

Scalable coding

- The MPEG–4 allows a limited number of layers (typically 2 to 4). This is the large-step scalability.



MPEG–4

Scalable coding

- The large-step scalability is not efficient for large numbers of enhancement layers.
- For fine-grain scalability (FGS), MPEG–4 uses a method named bit-sliced arithmetic coding (BSAC), that substitutes the block "Quantization and Coding" in the codec.