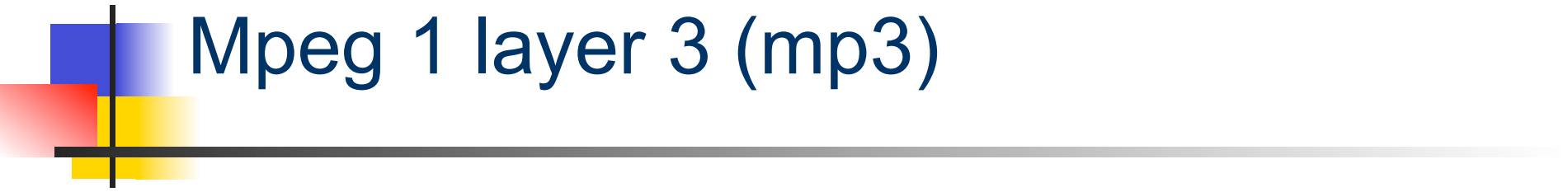
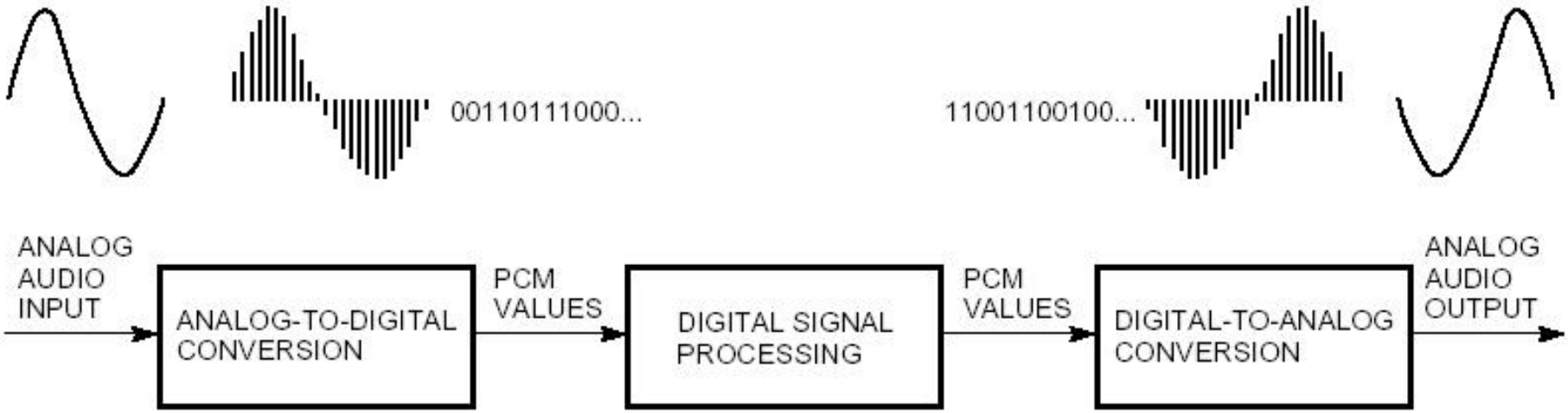


Mpeg 1 layer 3 (mp3)

general overview

Digital Audio

- CD Audio:
  - 16 bit encoding
  - 2 Channels (Stereo)
  - 44.1 kHz sampling rate



2 * 44.1 kHz * 16 bits = 1.41 Mb/s **+** Overhead (synchronization, error correction, etc.)

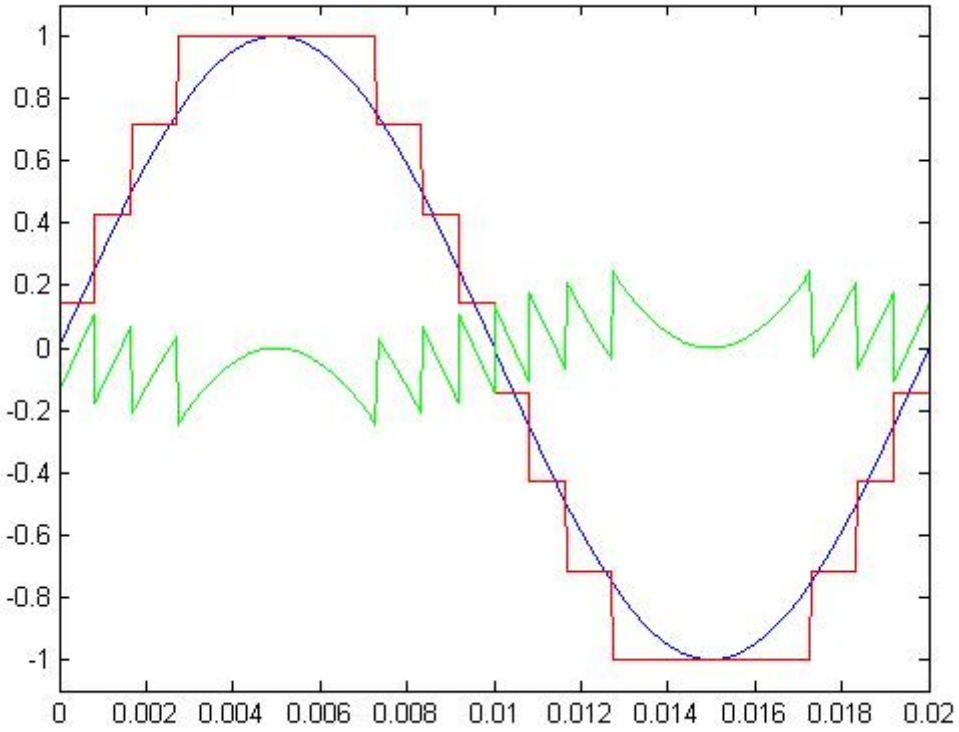
CD Audio = 4.32 Mb/s

Bit rate reduction

Audio signal	Frequency range in Hz	Sampling rate in kHz	PCM (bit/sample)	PCM bit rate (kbit/sec)
Telephone speech	300 -3,400	8	8	64
Wideband speech	50-7,000	16	8	128
Wideband audio (stereo)	10-20,000	48	2×16	2×768
CD	10-20,000	44.1	2×16	2×705.6

Quantization

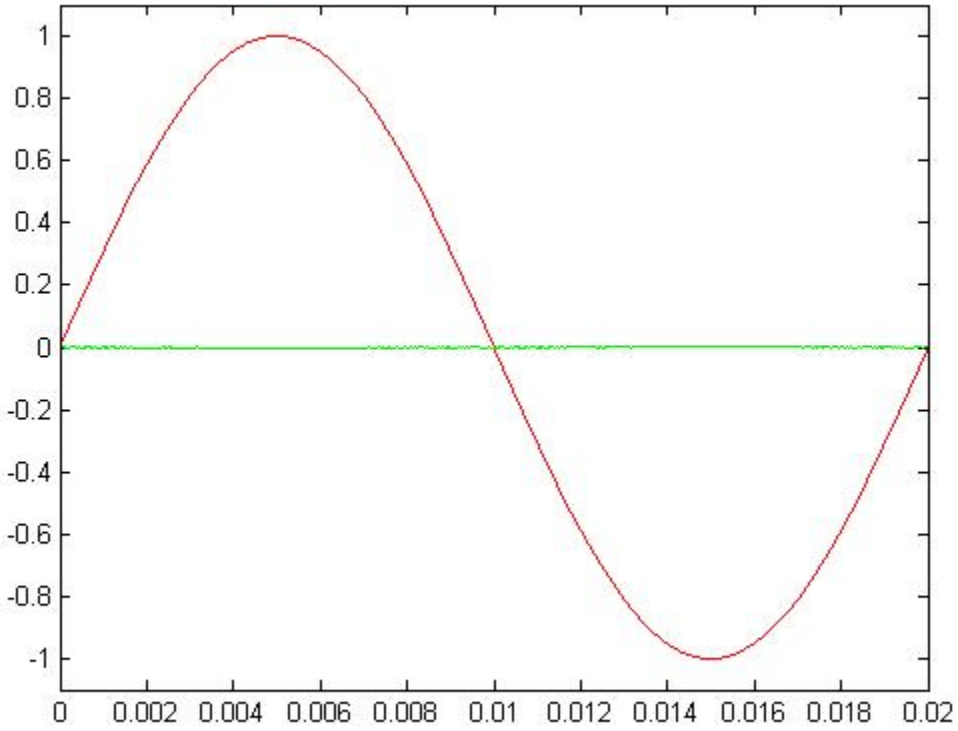
- **Quantization Noise**
  - the difference between the analog signal and the digital representation,
  - a result of the error in the quantization of the analog signal.



N Bits => 2^N levels

Quantization

- With each increase in the bit level,
  - the digital representation of the analog signal increases in fidelity,
  - the quantization noise becomes smaller.



Bits	Levels
3	8
4	16
5	32
8	256
16	65536

Compression

- High data rates, such as CD audio (4.32 Mb/s), are incompatible with internet & wireless applications.
- Audio data must somehow be compressed to a smaller size (less bits), while not affecting signal quality (minimizing quantization noise).
- Perceptual Audio Encoding is the encoding of audio signals, incorporating psychoacoustic knowledge of the auditory system, in order to reduce the amount of bits necessary to faithfully reproduce the signal.
    - MPEG-1 Layer III (aka mp3)
    - MPEG-2 Advanced Audio Coding (AAC)

MPEG

- MPEG = Motion Picture Experts Group
- MPEG is a family of encoding standards for digital multimedia information

- MPEG-1: a standard for storage and retrieval of moving pictures and audio on storage media (e.g., CD-ROM).
  - Layer I
  - Layer II
  - Layer III (aka MP3)

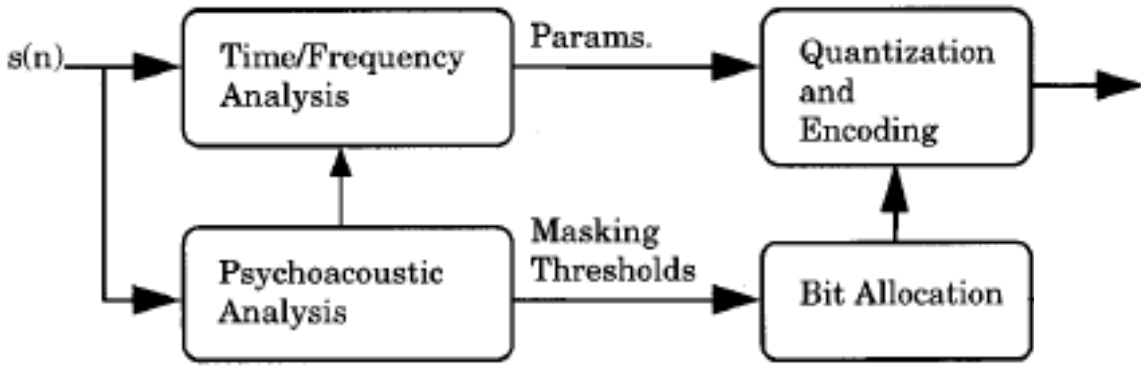
- MPEG-2: standard for digital television, including high-definition television (HDTV), and for addressing multimedia applications.
  - Advanced Audio Coding (AAC)

- MPEG-4: a standard for multimedia applications, with very low bit-rate audio-visual compression for those channels with very limited bandwidths (e.g., wireless channels).

- MPEG-7: a content representation standard for information search

9.1.2 - Overview of Perceptual Encoding

- General Perceptual Audio Encoder
  - Psychoacoustic analysis => masking thresholds
    - **irrelevancy** is identified and then removed
  - Basic principle of Perceptual Audio Encoder:
    - use masking pattern of stimulus to determine the least number of bits necessary for each frequency sub-band,
    - to prevent the quantization noise from becoming audible.

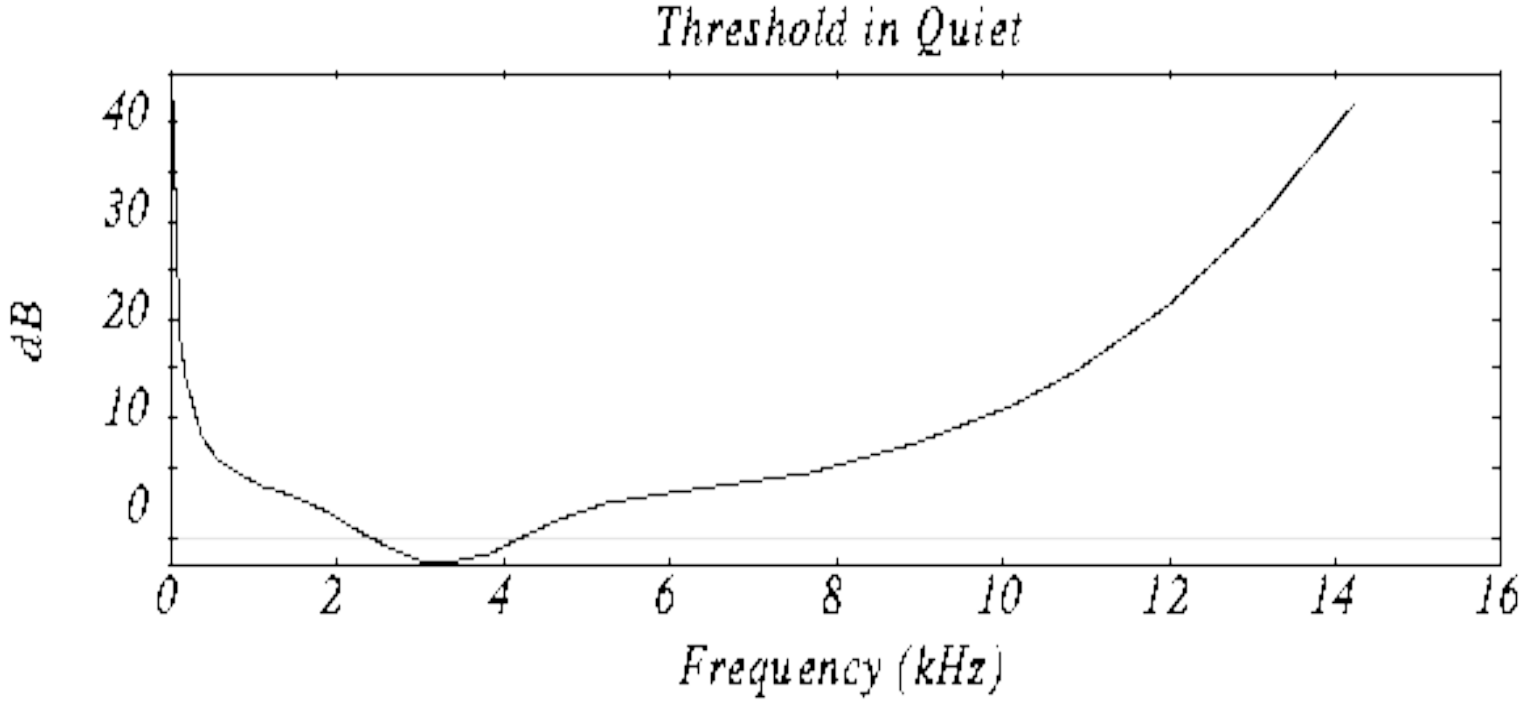


Masking

- Two major types of masking take place for encoding:
  - **Frequency Masking**:
    - areas of the spectrum that have high energy and are close together can be filtered to eliminate “spikes” in the spectrum
  - **Temporal Masking**:
    - loud noises that occur close to each other in time (about 3 to 5 milliseconds) can be approximated by a single loud noise

Threshold in quiet

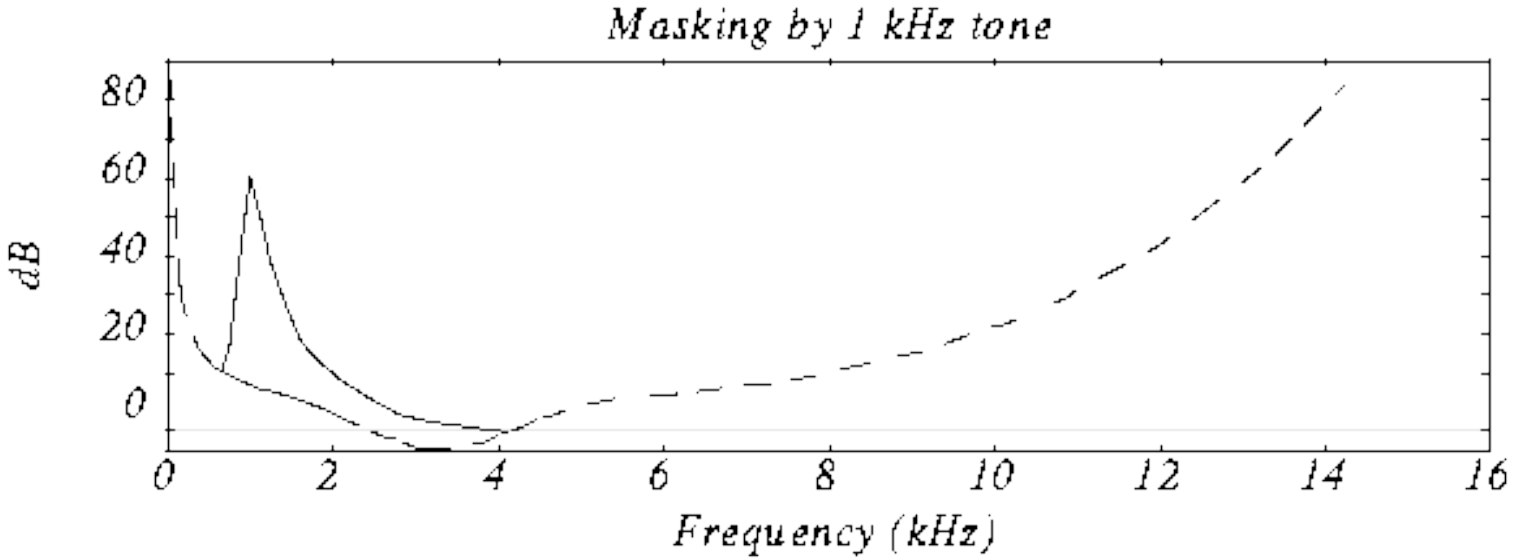
- Human auditory system has limitations
  - Frequency range: 20 Hz to 20 kHz, sensitive at 2 to 4 KHz.
  - Dynamic range (quietest to loudest) is about 96 dB



- Moreover, based on psycho-acoustic characteristics of human hearing, algorithms perform some tricks to further reduce data rate

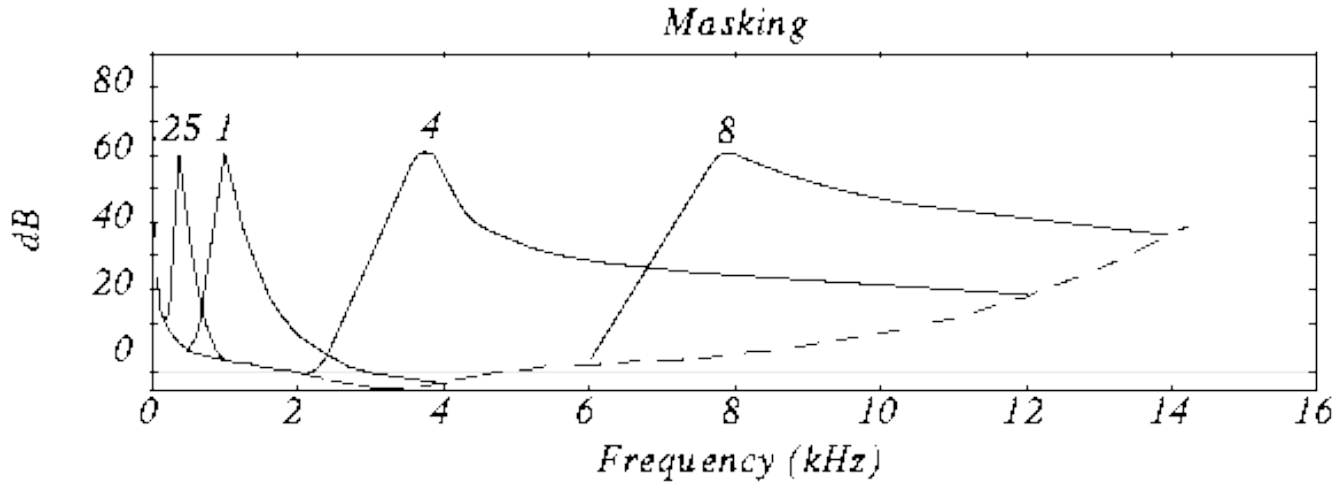
Masking Effects: **Frequency Masking**

- **Frequency Masking**: If a tone of a certain frequency and amplitude is present, then other tones or noise of similar frequency cannot be heard by the human ear
- the louder tone (masker) maskes the softer tone (maskee)
  - => no need to encode and transfer the softer tone
- Different masking when masking and maskee are tones or noise
  - tone masking noise => strong
  - noise masking tone => weak

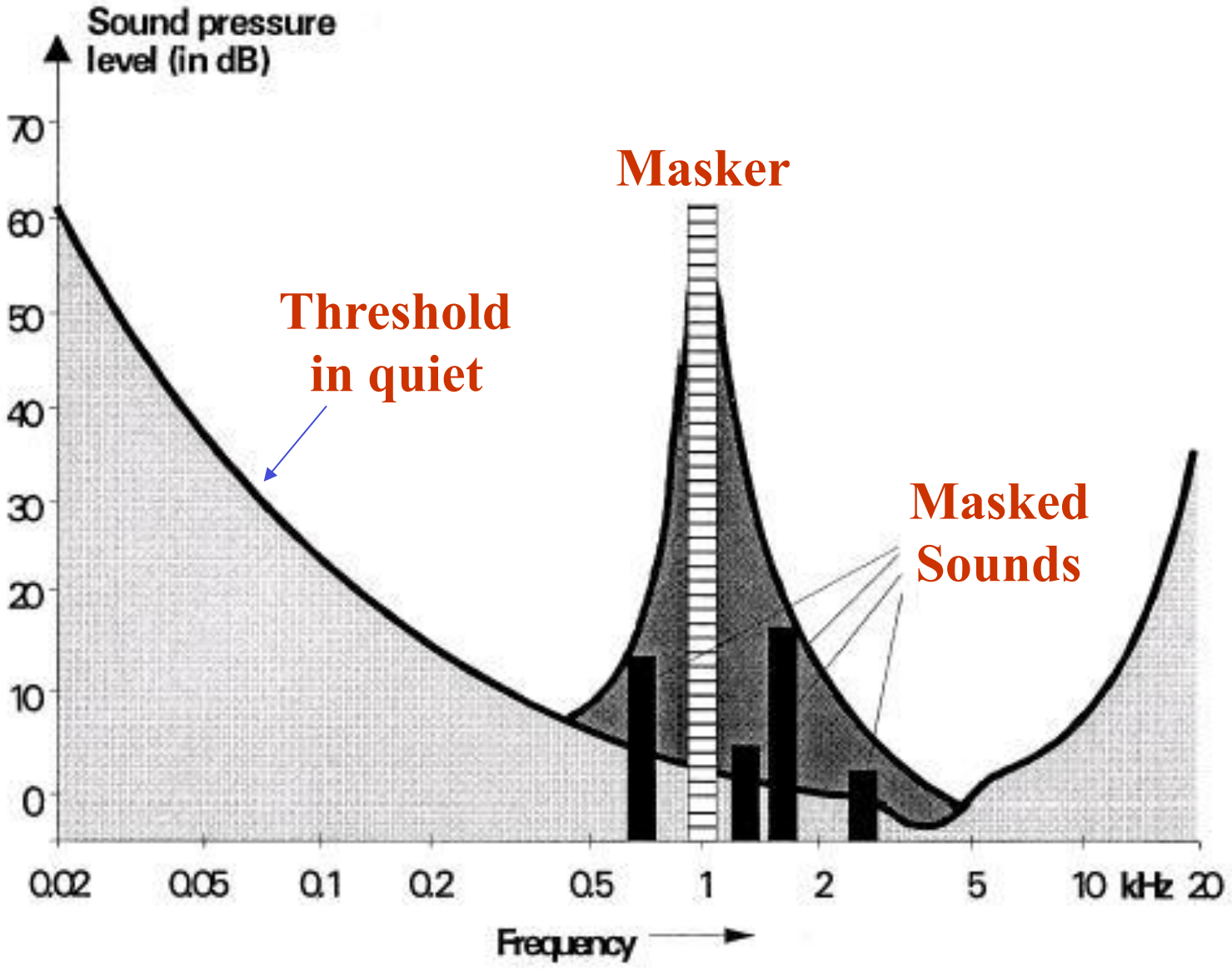


Masking Effects (Cont…)

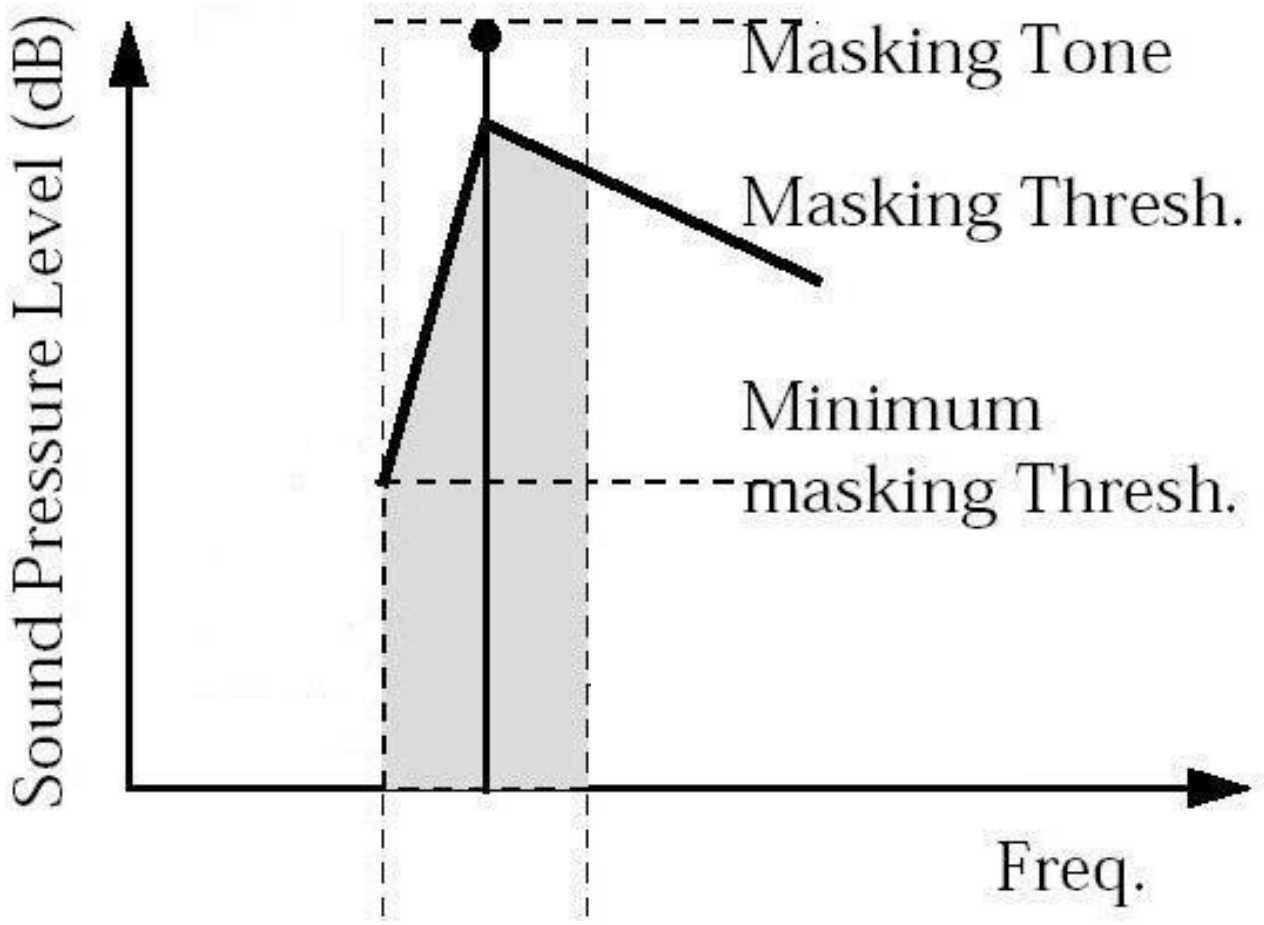
- Simultaneous masking changes with frequency
- Repeat for various frequencies of masking tones
- Masking Threshold: Given a certain masker, the maximum non-perceptible amplitude level of the softer tone



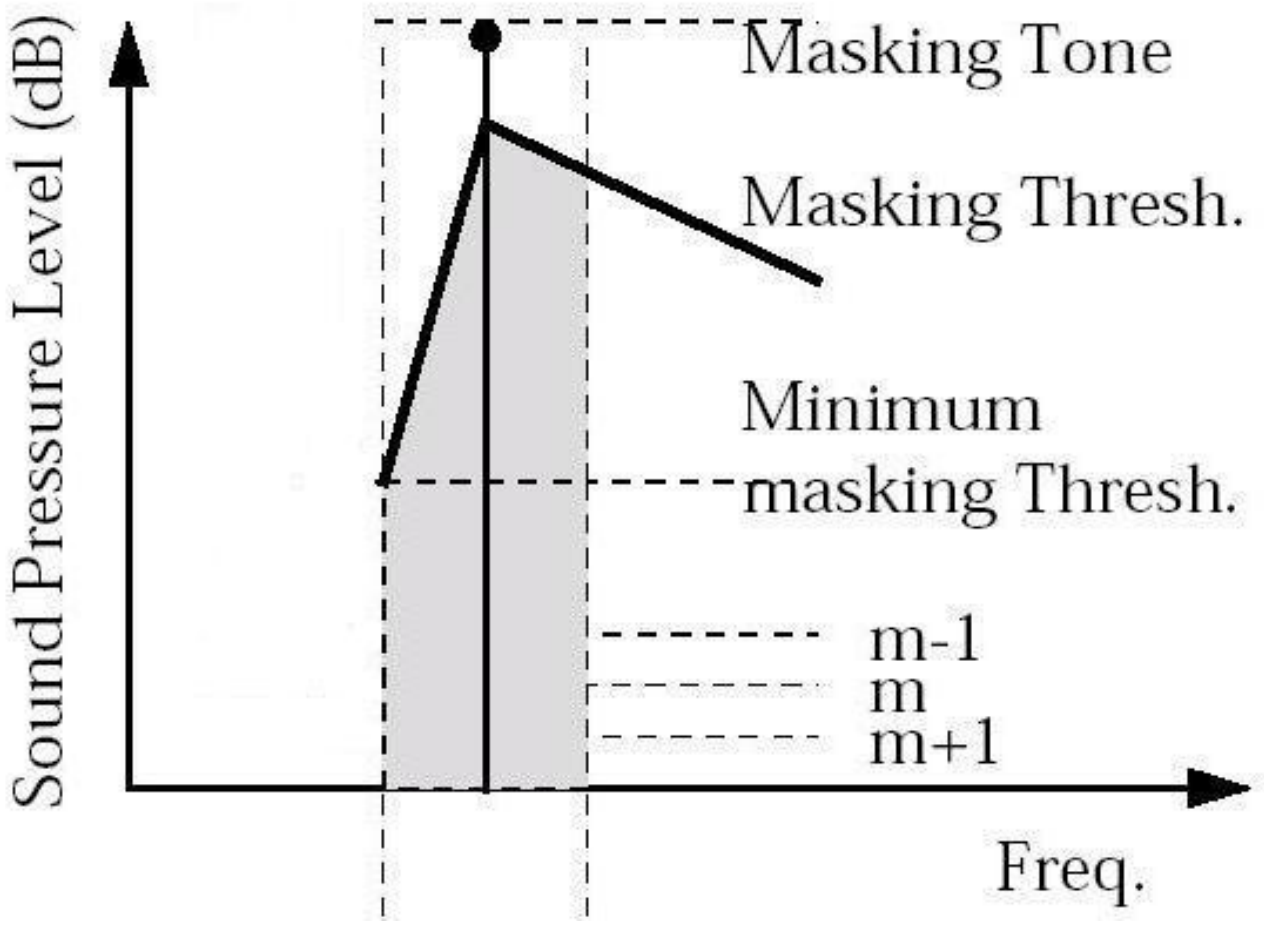
Masking Pattern



Masking


Sound Pressure Level (dB)
Masking Tone
Masking Thresh.
Minimum
masking Thresh.
Freq.

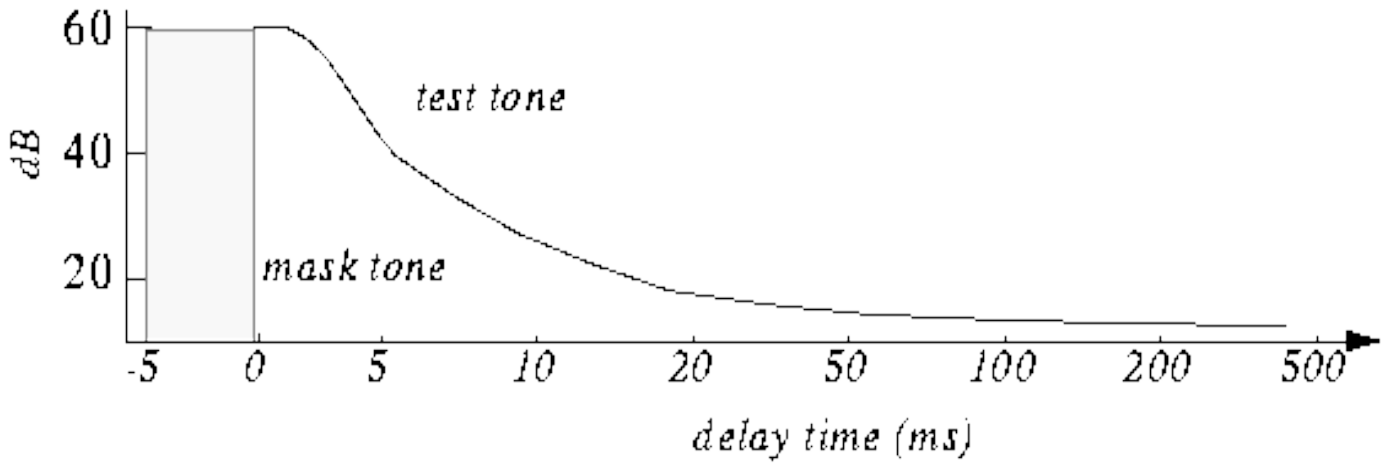
Masking/Bit Allocation



The number of bits used to encode each frequency sub-band is equal to the least number of bits with a quantization noise that is below the minimum masking threshold for that sub-band.

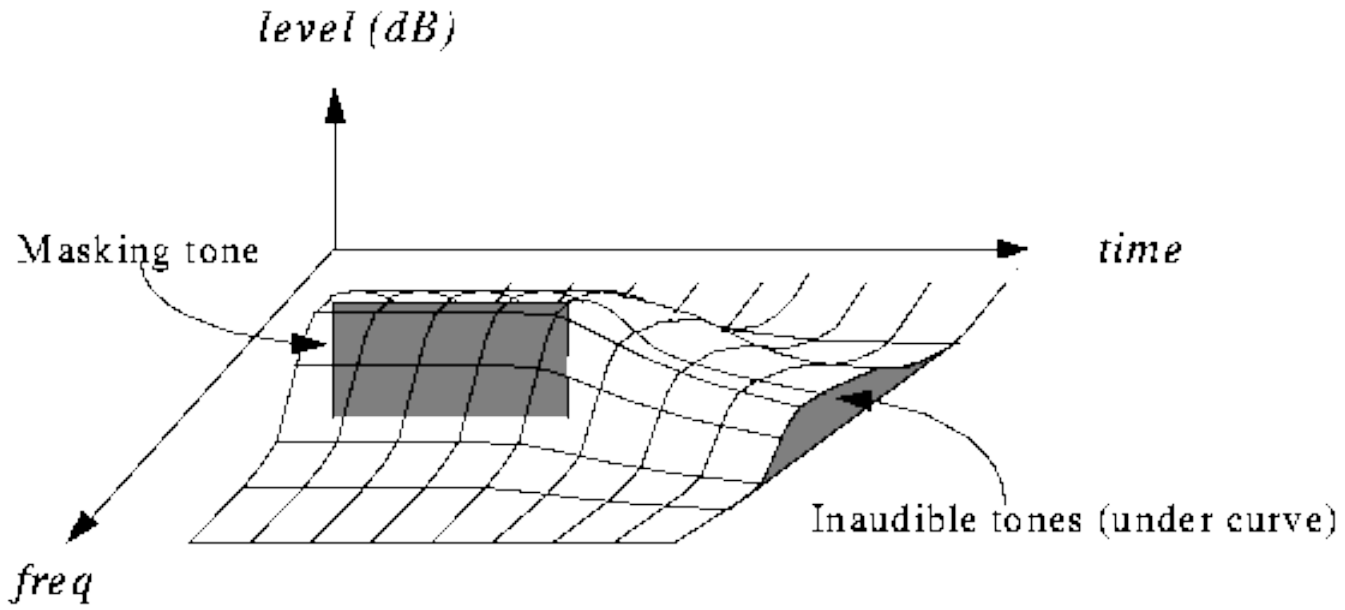
Masking Effects – temporal masking

- **Temporal Masking**: If we hear a loud sound, then it stops, it takes a little while until we can hear a soft tone nearby.



- The **Masking Threshold** is used by the audio encoder to determine the maximum allowable quantization noise at each frequency to minimize noise perceptibility: remove parts of signal that we cannot perceive

Net effect of masking:



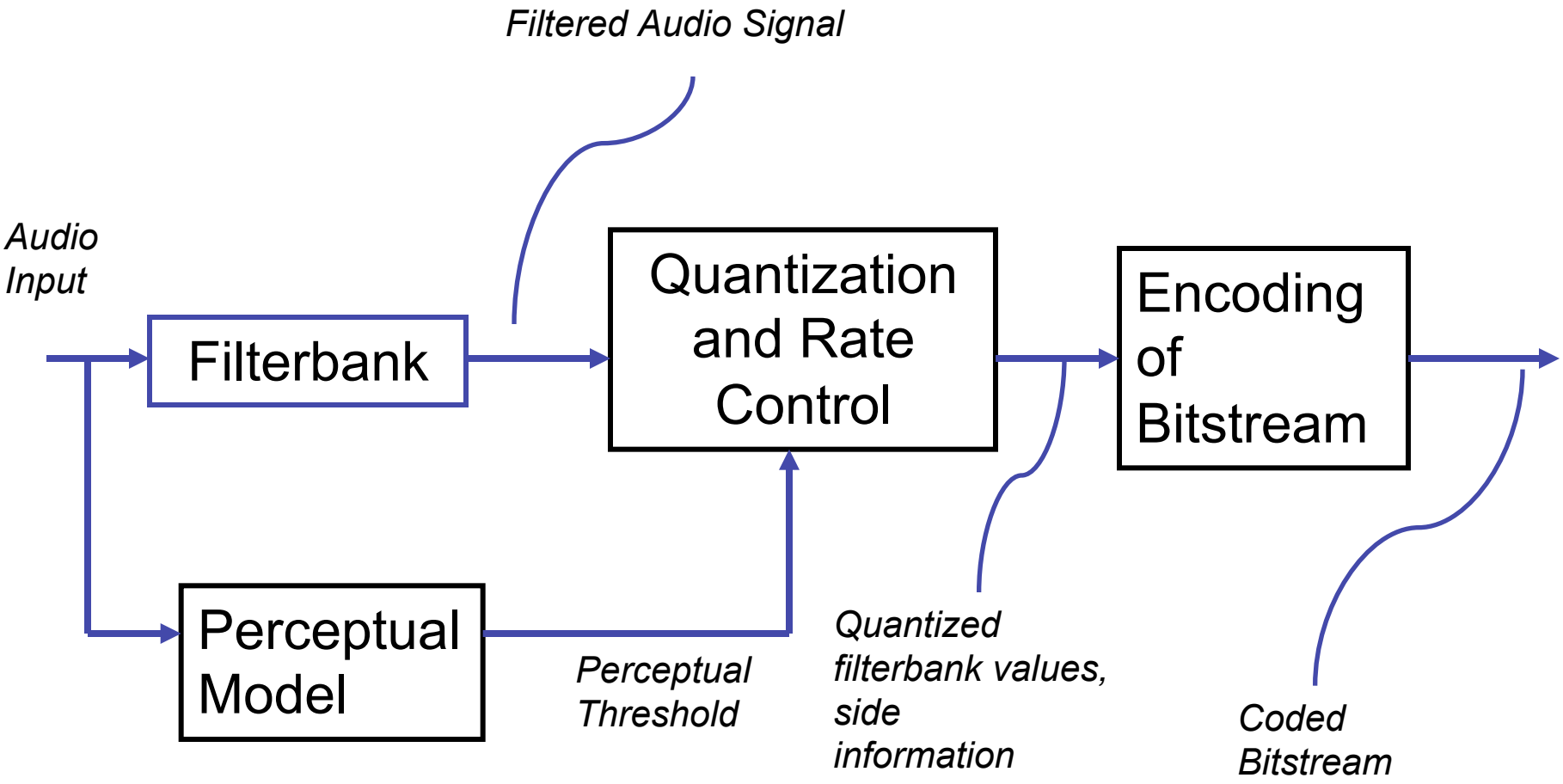
MPEG 1 Layer 3 audio coding

9.1.3

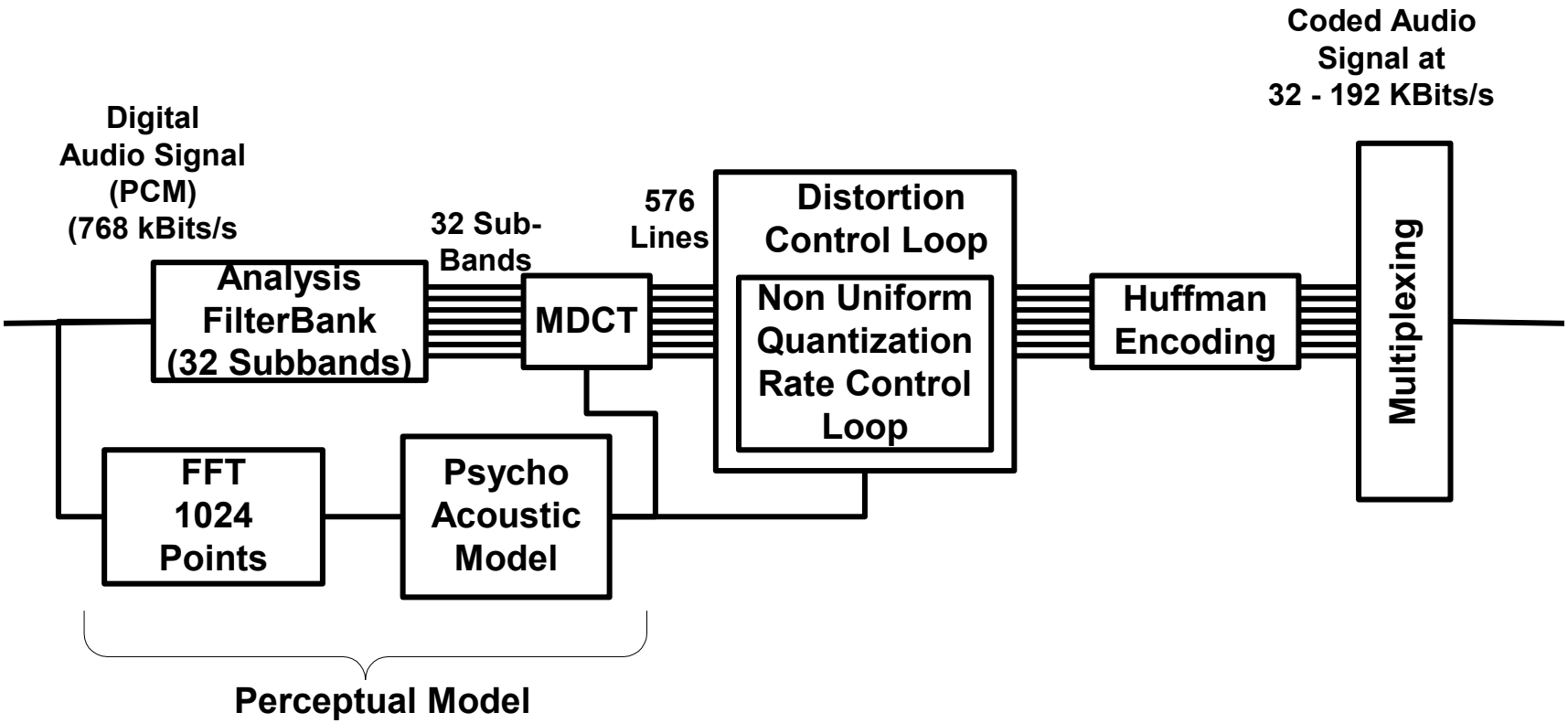
Layers of MPEG Audio

- Many different standard sampling rates:
  - 16kHz, 22.05kHz, 24kHz, 32 kHz, **44.1 kHz, 48 kHz**
- Layer I
  - 12 samples per sub-band (384 total samples)
  - Compression ratio: approx 4:1
  - Around 384kbps (depends on chosen sampling rate)
- Layer II
  - 36 samples per sub-band (1152 total samples)
  - Compression ratio: approx 6:1 to 8:1
  - Around 256kbps to 192kbps
- Layer III
  - 12 samples per sub-band (384 total samples)
  - Compression ratio: approx 10:1 to 12:1

Perceptual Audio **Coder**



MPEG 1 Layer 3 (MP3) Encoder



MP3 Components

- **Perceptual model**: An estimate of the actual (time and frequency dependent) masking threshold is computed by using rules known from psychoacoustics.
- **Filter bank**: A hybrid polyphase / MDCT filter bank is used to decompose the input signal into sub-sampled spectral components. Together with the corresponding inverse filter bank in the decoder it forms an analysis/synthesis system.
- **Quantization and coding**: The spectral components are quantized and coded with the aim of keeping the noise introduced by the quantization below the masking threshold.
  - Distortion Control Loop
  - Non-uniform Quantization Control Loop
  - Huffman Coding
- **Multiplexing**: A bit stream formatter is used to assemble the bit stream, which consists of the quantized and coded spectral coefficients and some side information, e.g. bit allocation information.

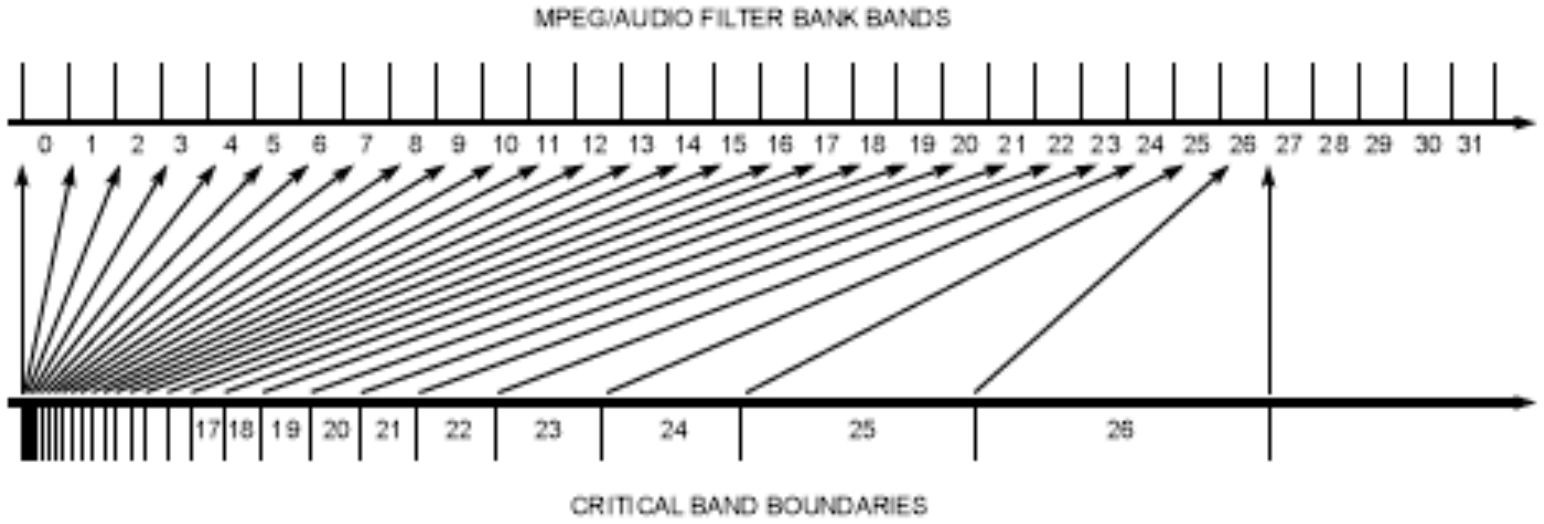
Perceptual Model

- The perceptual model consists of outputs values for the masking threshold or allowed noise for each coder partition.
- In Layer-3, these coder partitions are roughly equivalent to the critical bands of human hearing.
  - The the compression result should be indistinguishable from the original signal If the quantization noise can be kept below the masking threshold for each coder partition

Psychoacoustic Model

- Time align audio data
  - The psychoacoustic model must account for both the delay of the audio data through the filter bank and a data off-set so that the relevant data is centered within its analysis window
- Convert audio to spectral domain
  - The psychoacoustic model uses a time-to-frequency map-ping such as a 512- or 1,024-point Fourier transform
  - A standard Hanning window, applied to audio data before Fourier transformation, conditions the data to reduce the edge effects of the transform window.
- Partition spectral values into critical bands
  - To simplify the psychoacoustic calculations, the model groups the frequency values into perceptual quanta

MPEG Audio Filter Bank Boundaries



Finer resolution at lower frequencies

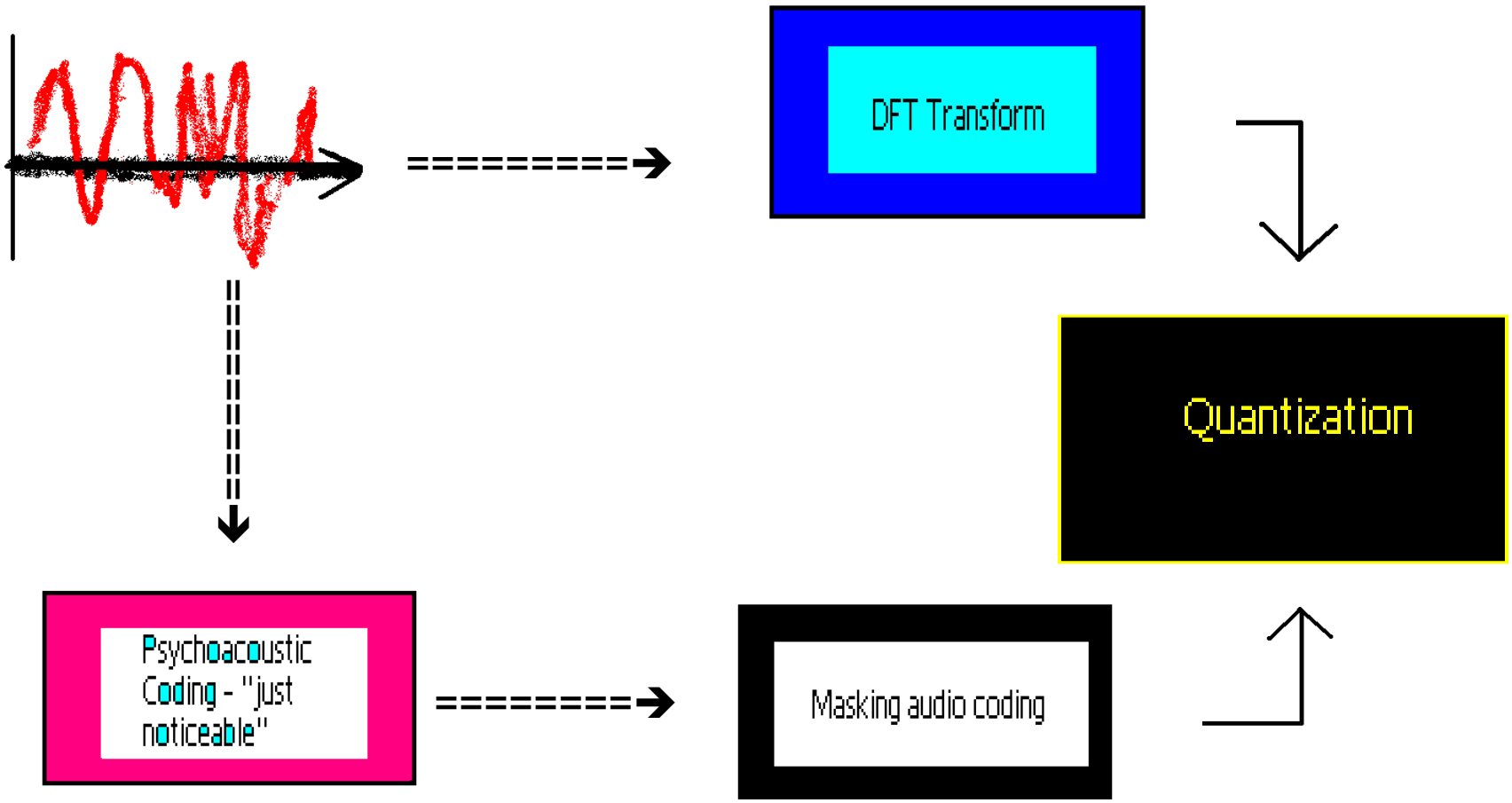
Psychoacoustic Model Functions

- Incorporate threshold in quiet
  - This threshold is the lower bound for noise masking and is determined in the absence of masking signals
- Separate into tonal and non-tonal components
  - The model must identify and separate the tonal and noiselike components of the audio signal
- Apply spreading function
  - The model determines the noise-masking thresholds by applying an empirically determined masking or spreading function to the signal components

Psychoacoustic Model Functions

- Find the minimum masking threshold for each sub-band
  - The psychoacoustic model calculates the masking thresholds with a higher-frequency resolution than provided by the filter banks.
  - Where the filter band is wide relative to the critical band (at the lower end of the spectrum), the model selects the minimum of the masking thresholds covered by the filter band.
  - Where the filter band is narrow relative to the critical band, the model uses the average of the masking thresholds covered by the filter band.

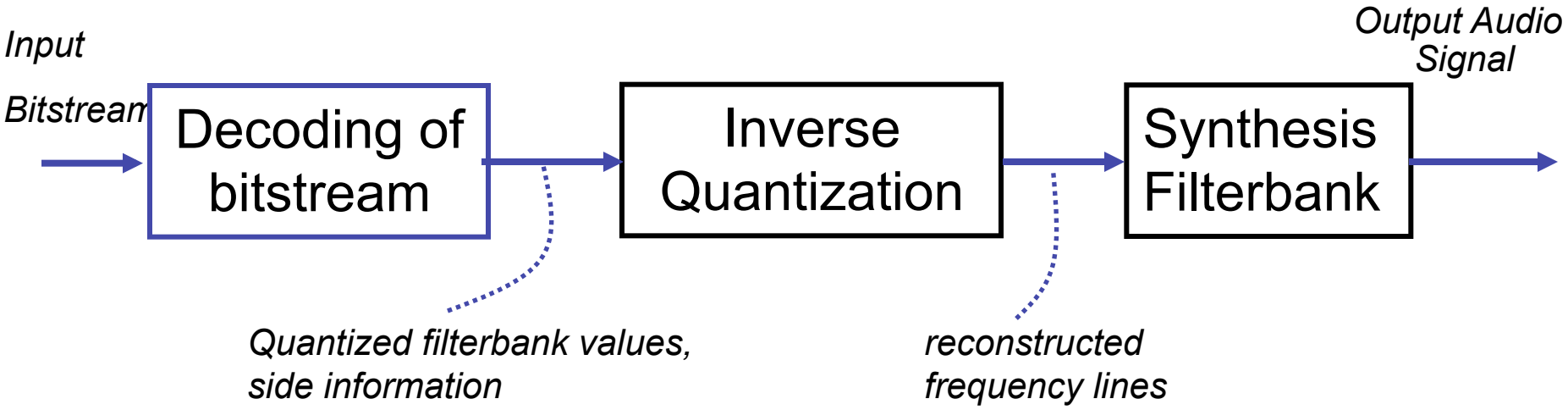
Encoding model for Layer I



Determine Masking Threshold

- algorithm
  1. Time align audio data
  2. Convert audio to spectral domain
  3. Partition spectral values into critical bands
  4. Incorporate threshold in quiet
  5. Separate into tonal and non-tonal components
  6. Apply spreading function
  7. Find the minimum masking threshold for each sub-band
  8. Calculate signal-to-mask ratio

Perceptual Audio **Decoder**



Stereo coding: problems

- Binaural Masking Level Depression (BLMD)
  - low frequency effect
  - phase of interaural signals taken into account
  - ➔ a noise image and a tone image can be in different places.
  - reduce the masking threshold by up to 20dB

- Image distortion or elimination
  - high frequency effect
  - a signal with a distorted high-frequency envelope may not provide the same imaging effects in the stereo

Stereo coding strategies

- Left-Right (L/R) or "Simple" Stereo (SS)
  - each band coded separately

- Intensity Stereo coding (IS)
  - (left + right) signal (L+R)
  - and left and right gains gL gR
  - in high frequency bands
  - for lower quality coding, equivalent of a pan-pot.

- Middle/Side (M/S) stereo coding
  - $M = L + R$
  - $S = L - R$
  - good for signals:
    - with strong central images
    - with a strong surround component

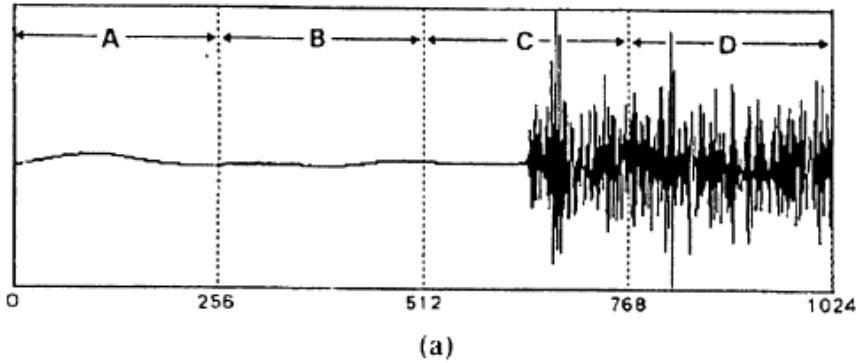
Discussion on mp3 coding

- Features
  - *Open standard*
  - *Availability of encoders and decoders*
  - *Supporting technologies*
  - *Normative versus Informative*

- Quality Considerations:
  - *Common types of artifacts*
  - *Loss of bandwidth*
  - *Pre-echoes*
  - *Roughness, double-speak*

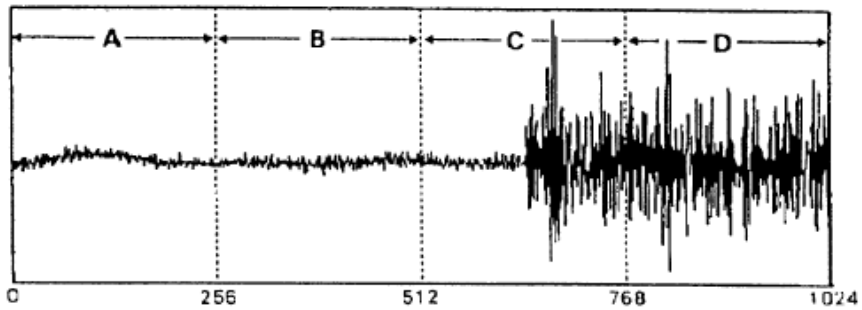
Pre-echoes

- Source signal



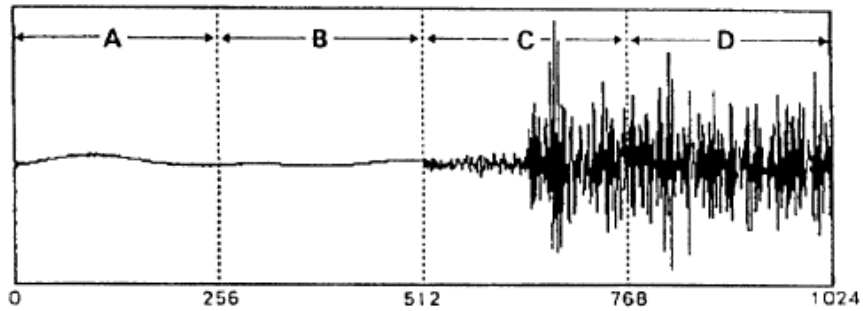
(a)

- Reconstructed signal
  - with block size 1024



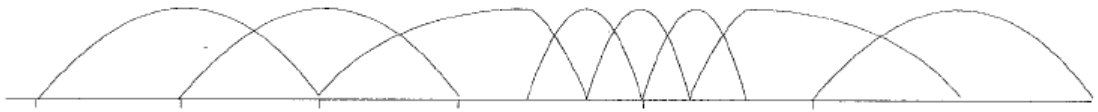
(b)

  - with block size 256



(c)

- Variable length window
  - better frequency resolution where it is needed
  - Mixes require transition blocks



Bit reservoir

- **Problem**:
  - A frame with little audio interest may require few bits to encode
    - with a constant bitrate these bits may be unused
  - A frame with substantial audio interest may require more bits to encode
    - with a constant bitrate the audio quality may decrease
- **Solution**:
  - Allow frames to give to or take from a reservoir
  - Each frame that save spaces, allows subsequent frames to store bits if needed
    - typical situation of a silence (few bits) followed by an attack (many bits)

Confronti

original 

- Qualità telefonica: 96:1 (2.5 kHz / mono / 8 kbps)
- 
- Radio AM: 24:1 (7.5 kHz / mono / 32 kbps)
  - 24 kbps, mono
- 
- Radio FM: 26...24:1 (11 kHz / stereo / 56...64 kbps)
  - 64 kbps, stereo
- quasi-CD: 16:1 (15 kHz / stereo / 96 kbps)
- 
- CD: 14..12:1 (>15 kHz / stereo / 112..128 kbps)
  - approx. 1MB/minuto di spazio hard-disk
  - 128 kbps, stereo
- Oltre: 8…4:1 per registrazioni moderne ad alta qualità