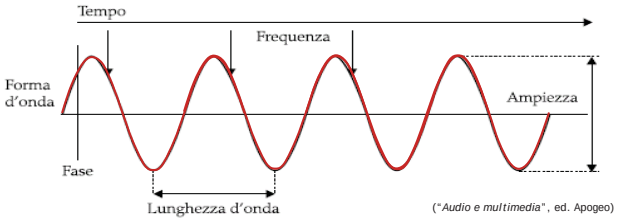
# Audio: fundamentals

The sound is a longitudinal pressure wave that propagates through a transmission medium. The range of human hearing is between 16 Hz and 22 kHz.



("*Audio e multimedia*", ed. Apogeo)

***Amplitude*** (dB) → volume, the intensity of the sound

***Frequency*** (Hz) → "higher" or "lower"

***Waveform*** → timbre, it allows us to recognize different types of sound production



Mobile Programming and Multimedia

# Audio Formats

Prof. Ombretta Gaggi
University of Padua

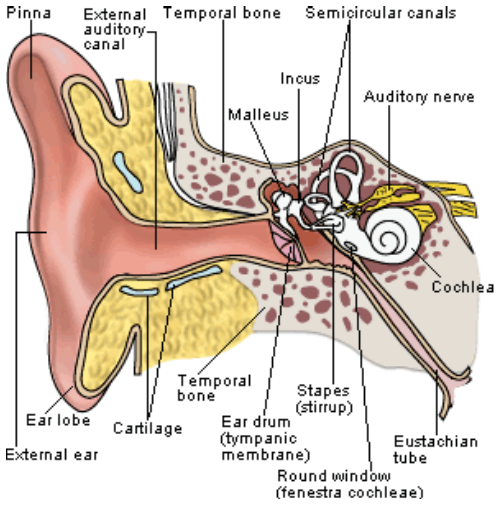
# Intensity levels

Approximate sound levels and intensities within human hearing range

| Source of sound | Intensity level (dB) | Intensity (W m$^{-2}$) | Perception |
|---|---|---|---|
| jet plane at 30 m | 140 | 100 | extreme pain |
| threshold of pain | 125 | 3 | pain |
| pneumatic drill | 110 | $10^{-1}$ | very loud |
| siren at 30 m | 100 | $10^{-2}$ | |
| loud car horn | 90 | $10^{-3}$ | loud |
| door slamming | 80 | $10^{-4}$ | |
| busy street traffic | 70 | $10^{-5}$ | noisy |
| normal conversation | 60 | $10^{-6}$ | moderate |
| quiet radio | 40 | $10^{-8}$ | quiet |
| quiet room | 20 | $10^{-10}$ | very quiet |
| rustle of leaves | 10 | $10^{-11}$ | |
| threshold of hearing | 0 | $10^{-12}$ | |

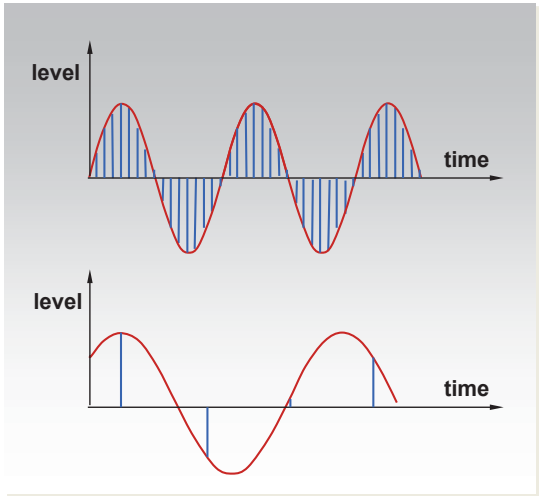


# Internal structure of the ear




(Reference: https://www.mydr.com.au/ear-anatomy/)

# Analogical/digital conversion of audio signal

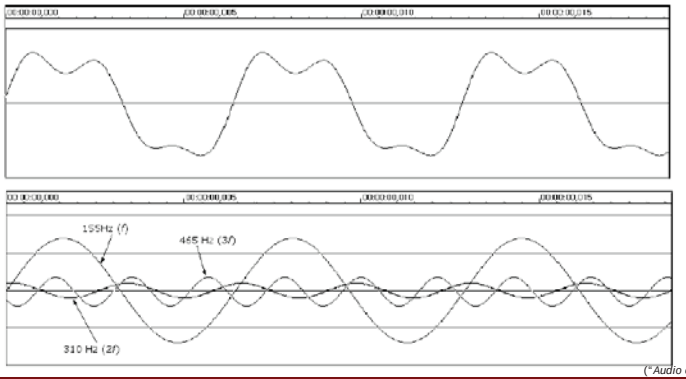



If a periodic signal contains no frequencies higher than *N* hertz, it can be completely reconstructed if *2N* samples per second are used (*Nyquist Theorem*). So the a sampling-rate o f *2N* samples per second I sufficient.



# Audio: Fourier analysis

*"A periodic signal can be broken down into a series of harmonically related sinusoids, each one with its amplitude and phase, and frequencies that are harmonics of the fundamental frequency of the signal."*




("*Audio e multimedia*", ed. Apogeo)



# Signal to noise ratio

Analogical signals are altered by ***noise***, a random fluctuation of the signal determined by electronic phenomena

- The *signal/noise ratio* (*SNR*, *signal to noise ratio*) is a measure of signal quality

$$SNR = 10\log\frac{V_{signal}^2}{V_{noise}^2} = 20\log\frac{V_{signal}}{V_{noise}}$$

In digital systems, the noise is the difference between the real signal and the quantized signal

$$SQNR = 20\log\frac{V_{signal}}{V_{quant-noise}} = 20\log\frac{2^{N-1}}{1/2} = 6.02N\ (dB)$$

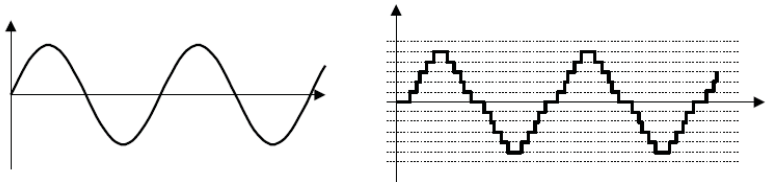


# Digital representation of the sound

The first step to elaborate a signal is to digitalize it (A/D transformation)

- ***sampling***: time division (Hz)
- ***quantization***: discrete representation of the signal level (measured in bits of precision)
- ex. Audio CD is digitalized at 44.1 kHz, 16 bits

What is the effect of digitalization on sound quality?



## Audio: size & transfer time

Uncompressed audio of good quality usually exceeds the transmission capacity of conventional networks

- Radio FM > 640 Kbit/sec, audio CD > 1.2 Mbit/sec

High dimensions

- Coded signal takes a lot of space
- Length is not limited (live audio)

It is not possible to transfer the whole audio file before playback

- ***streaming***: playback while receiving



## Audio quality vs. size

| *Quality* | *Frequencies interval Hz* | *Sampling kHz* | *Bits for sample* | *Mono/stereo* | *Bit rate kbit/s* |
|---|---|---|---|---|---|
| Mobile network | 200-4000 | 8.0 | 8 | mono | 8 |
| Radio AM | 100-6500 | 11.025 | 8 | mono | 11 |
| Radio FM | 20-12000 | 22.050 | 16 | stereo | 705.6 |
| Audio CD | 20-20000 | 44.1 | 16 | stereo | 1411.2 |
| DAT | 20-20000 | 48.0 | 16 | stereo | 1536 |
| DVD audio | 20-20000 | 192.0 | 24 | stereo | 9216 |



## Audio formats (1)

UNIVERSITÀ DEGLI STUDI DI PADOVA

**WAV**, Waveform Audio File

- Developed together by Microsoft and IBM
- standard *de-facto* for audio encoding on PCs
- Not compressed

**AIFF**, Audio Interchange File Format

- Developed by Apple Computer
- audio standard for Macintosh
- not compressed (there is a compressed version)

**µ-LAW**

- Standard audio format for Unix
- Telephonic standard in the USA (8KHz, 8 bits)

**A-LAW**

- European version of µ-LAW



## Problems with audio encoding

Digital encoding and decoding of audio signals has more problems than image encoding (and video)

- Temporal structure of the audio cannot be modified (frequency)
- Audio information varies over time (“audio stop” does not exist)
- Required reproduction quality is usually higher than simple understandability

# Audio compression (2)

## Silence compression

- Silence is a consecutive set of samples under a defined threshold
- Similar to RLE (*run-length encoding*) compression

## Adaptive Differential Pulse Code Modulation

- Encodes the difference between consecutive samples
- Difference is quantized, hence there is loss of information

## Linear Predictive Coding

- Adapts the signal to a human speech model
- Transmits the parameters of the model and the differences of the real signal to the model



# Audio formats (2)

## **MPEG-1** Audio

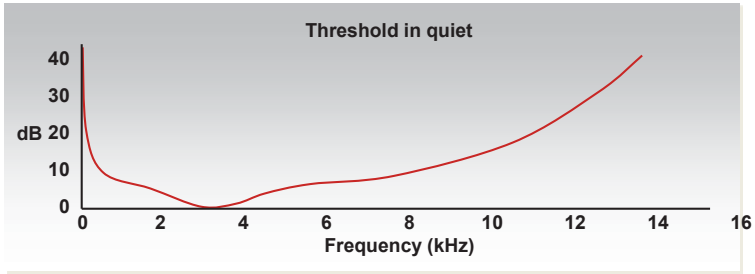
- encodes audio tracks in MPEG-1 videos
- Compressed format for variable quality encoding
- Compression algorithm with several steps, based on psychoacoustic principles
- Three different encoding levels with three different bit-rates
- cross-platform standard
- Several applications for the *consumer market* and commercially important



# Psycoachoustics elements

## Human ear sensibility varies along the audio spectrum

- Maximum sensibility is around 2-3 kHz, and decreases at spectrum extremities
- Ear sensibility strongly changes depending on several personal factors (e.g., age)






# Audio compression (1)

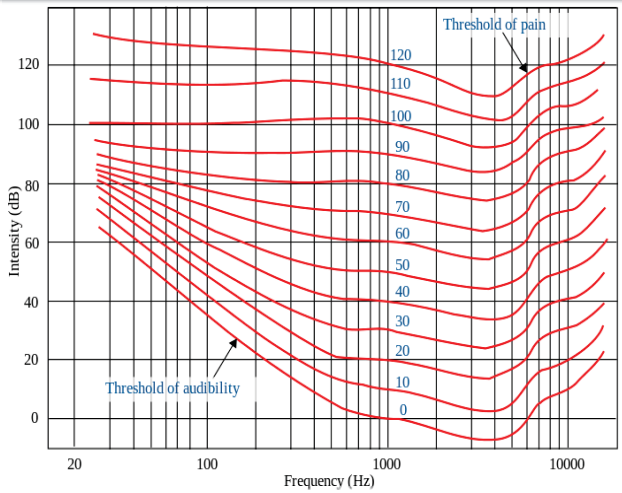
## Lossless compression has low performances

- Audio data are extremely variable
- Recurrent configurations are rare

## Necessary to use ***lossy compression***

- Audio information is frequently redundant
- Compression quality can be controlled
- Human ear does not have a linear behavior

# Fletcher-Munson Curves



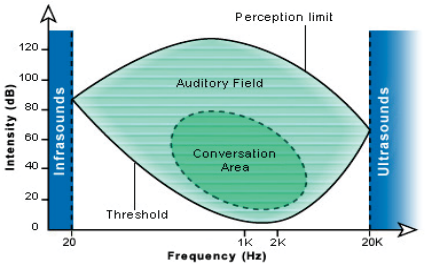

$$T_q(f) = 3{,}64\,(f/1000)^{-0.8} - 6{,}5\,e^{-0.6(f/1000-3.3)^2} + 10^{-3}\,(f/1000)^4$$



# Human sound perception (1)

## How we perceive sound and voice

- human hearing interval: ~ 20 Hz - 20 kHz
- The recognizable ***dynamic interval*** is the interval between the weakest and the strongest sound, is ~ 96 dB
- The human voice has frequencies in the interval ~ 500 Hz (vocals) - 2 kHz (consonants)




(Fonte: Institut Universitaire de Recherche Clinique - Montpellier)

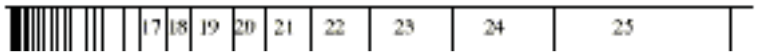


# Critical bands

The frequencies where there is a uniform perception of the amplitude of the sound are grouped in ***critical bands***

- Each band has an amplitude from 100 Hz to 4 kHz
- The entire spectrum of the audible frequencies is divided into 25 critical bands

The human hearing can be considered, broadly speaking, as a series of overlapping ***band-pass filters***



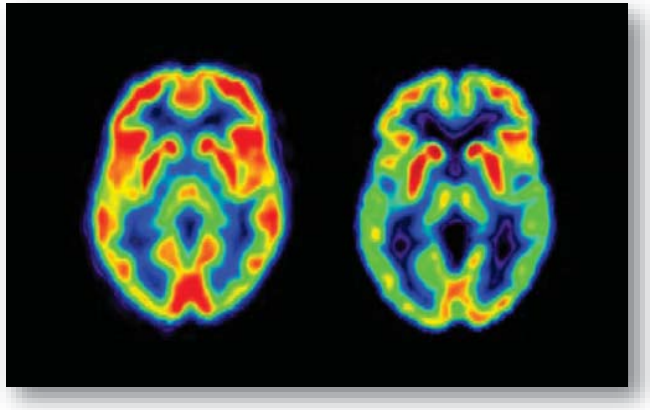

- 1 Bark = amplitude of a critical band (in honor of German physicist **Heinrich Georg Barkhausen)**

  Per frequencies lower than 500 Hz: $1\ Bark \approx freq/100$

  Per frequencies higher than 500 Hz: $1\ Bark \approx 9 + 4\log(freq/1000)$

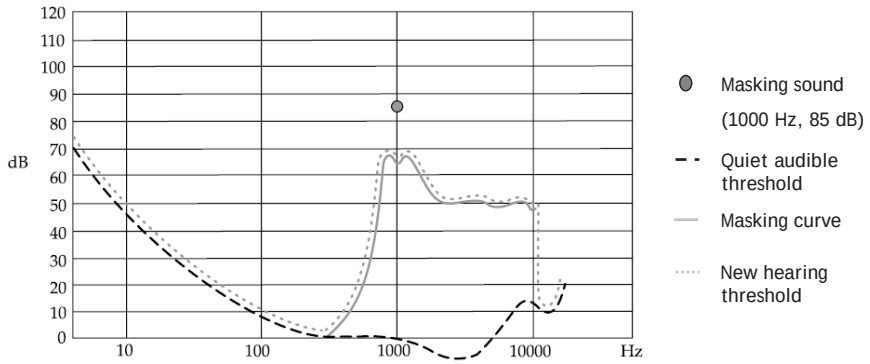


# Human sound perception (2)



**Female** **Male**

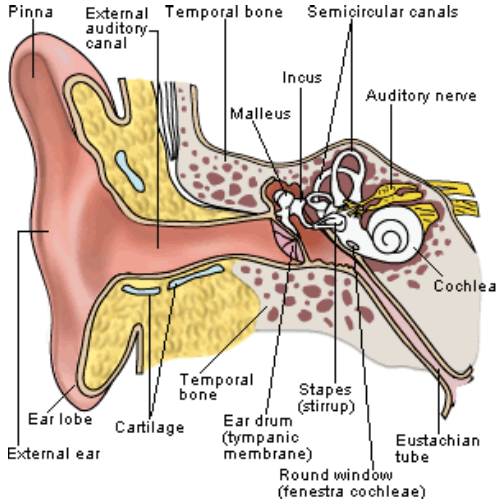
# Frequency or tonal masking - 2




(Reference: Vincenzo Lombardo e Andrea Valle, "*Audio e multimedia*", ed. Apogeo – pages 148-150)



# Internal structure of the ear



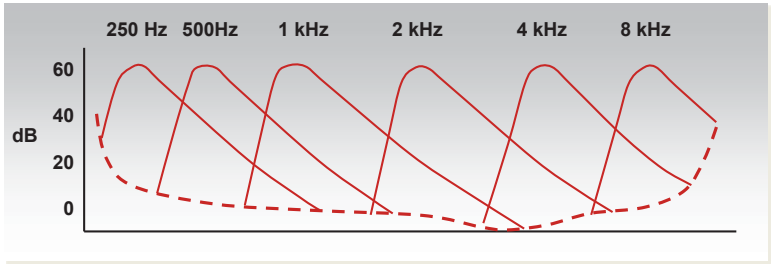

(Reference: https://www.mydr.com.au/ear-anatomy/)



# Frequency or tonal masking - 3

Frequency masking:

- Different for each frequency
  - ✓ Can be defined for each critical band
- Depends on the sound amplitude

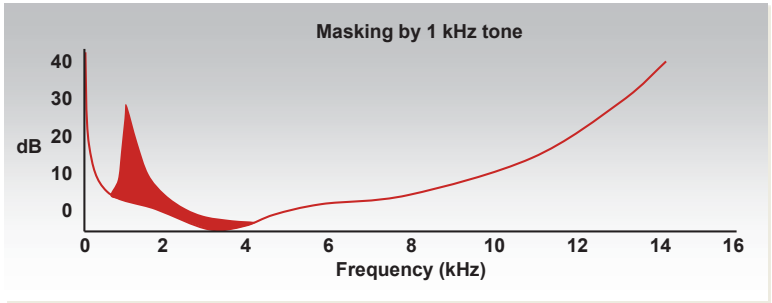





# Frequency or tonal masking - 1

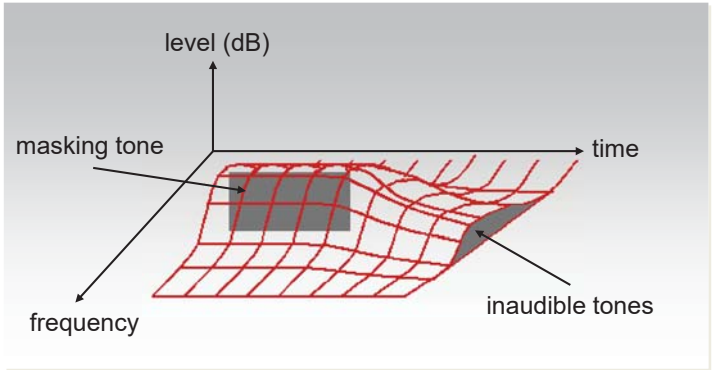
A pure sound can mask another one with near frequency and lower volume

- During the playback of a sound at 1 kHz, other simultaneous sounds in the masking interval cannot be perceived



# Combined masking

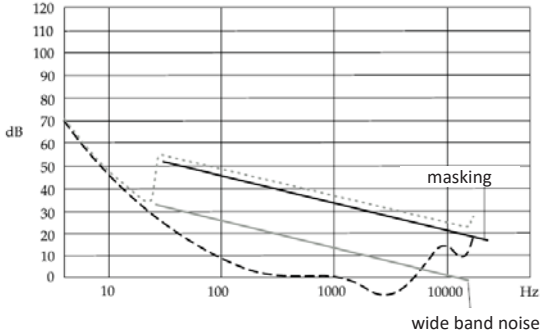
A combination of frequency and temporal masking






# Non-tonal masking

Non-tonal masking happens when the masking sound is a wide band signal (i. e. a noise) where it is not possible to find a specific tone




Different from tonal masking, non-tonal masking involves sounds with a higher intensity than the masking sound.

(Fonte: Vincenzo Lombardo e Andrea Valle, "*Audio e multimedia*", ed. Apogeo – pagg. 148-150)



# MPEG audio properties

MPEG-1 layer 3 (MP3) is the current standard for high-quality audio (music) with high compression

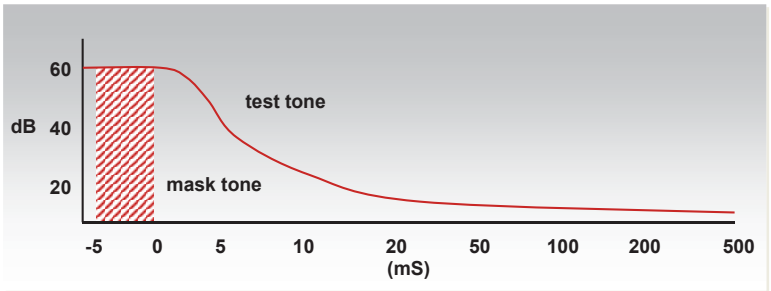
- The most common *bit-rate* for MPEG standard is from 48kbit/sec to 384 kbit/sec (audio CD not compressed >1.4 Mbit/sec)
- Compression level in the interval 2.7 – 24
- Compression level of 6:1 (256 kbit/sec) is almost identical to the original signal
- From 96 to 128kbit/sec, it represents the best quality for consumer applications
- Different sampling frequencies (32, 44.1 and 48 kHz)
- Monophonic, dual, stereo, joint stereo signal



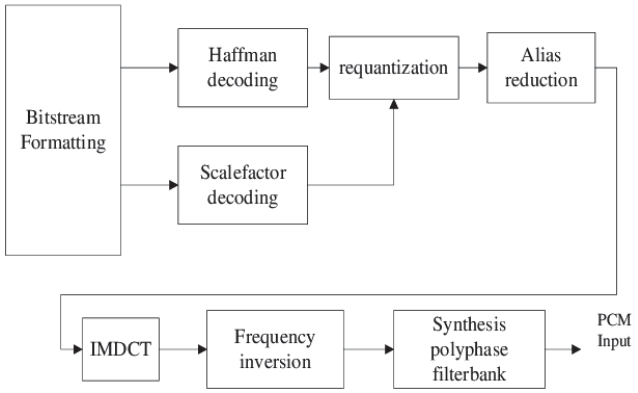
# Temporal masking

A sound can mask another one for a small interval of time

- the *pre-masking* hides preceding sounds between 5ms and 40ms before
- the *post-masking* hides the weakest sounds after the masking sound between 50ms to 200ms



# MPEG Audio decoder

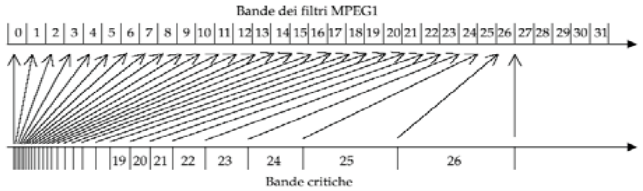





# MPEG Audio Compression Algorithm

## Four different steps based on the psychoacoustic model

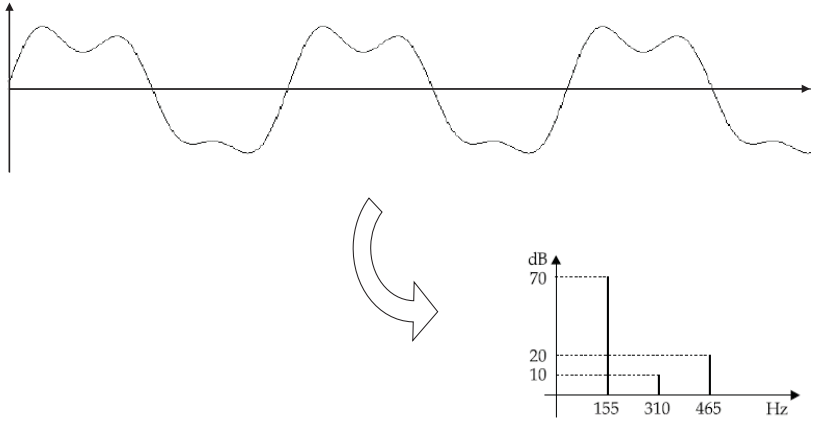
- It divides the audio signal into 32 frequency sub-bands
- For each sub-band, it calculates the masking level
- If the amplitude of the signal in the sub-band is lower than the masking threshold, the signal is not encoded
- Otherwise, it calculates the number of bits necessary to represent the signal (from 0 to 15) such that the quantization noise is lower than the masking threshold (1 bit ~ 6 dB of noise)
- Creates the bitstream following a standard format for transmission





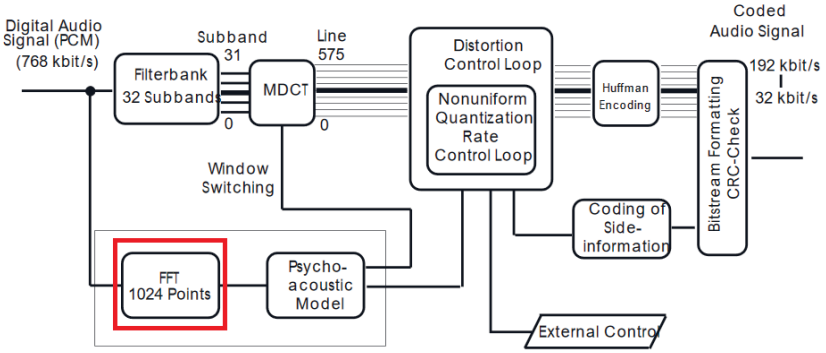

# Audio DCT

## From the time domain to frequencies domain:






# MPEG Audio Encoder



# Another masking application: the watermarking

*Watermarking* is the inclusion of digital information (source, destination, copyright information, access information, etc...) hidden inside multimedia data (images, videos, audios, texts, animations)

The watermarks (information):

- Cannot be modified
- Do not have to modify the enclosing data
- Must survive to all the operations done on the signal
- Must be directly connected to the data (not in the header)
- Must be statistically invisible



# Audio MPEG: example

The level at band 8 is 60 dB

- Masking is 12 dB on band 7, 15dB on band 9

The level at band 7 is 10 dB (< 12 dB ), it is ignored

The level at band 9 is 35 dB (> 15 dB ), it is encoded

Only the difference between the signal and the masking threshold is encoded

- Using 4 bits instead of 6 (2 bits = 12 dB)

| Band | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Level (dB) | 0 | 8 | 12 | 10 | 6 | 2 | 10 | 60 | 35 | 20 | 15 | 2 | 3 | 5 | 3 | 1 |

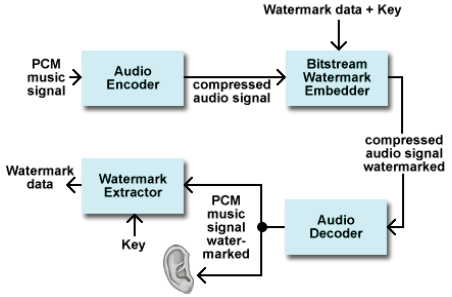


# Watermarking bitstream

Watermarks insertion inside an audio bitstream works with a sort of masking technique that is ''opposite'' to the MPEG algorithm.

The watermark is inserted near high-level signals, such that it will be masked by the latter. In this way, it is not distinguishable from the original one.

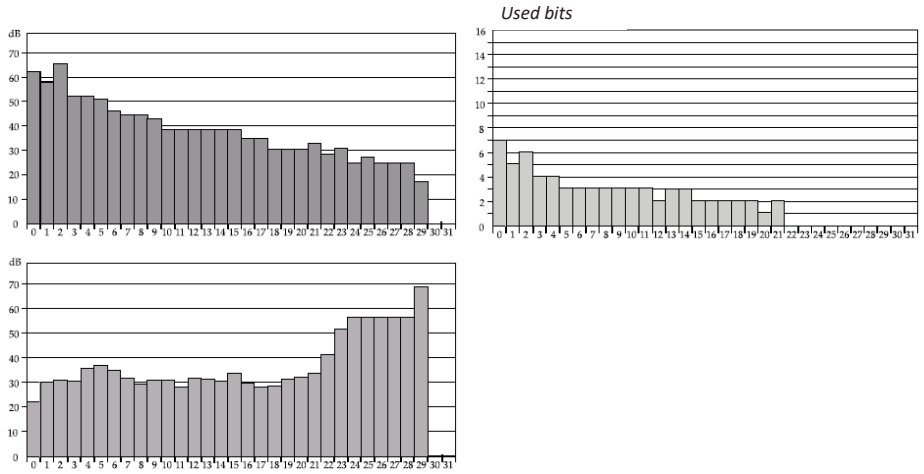
Subsequent MPEG encoding would delete the watermark. Other techniques insert the watermark with a frequency outside the human hearing range.




(fonte: Fraunhofer Institut Integrierte Schaltungen)

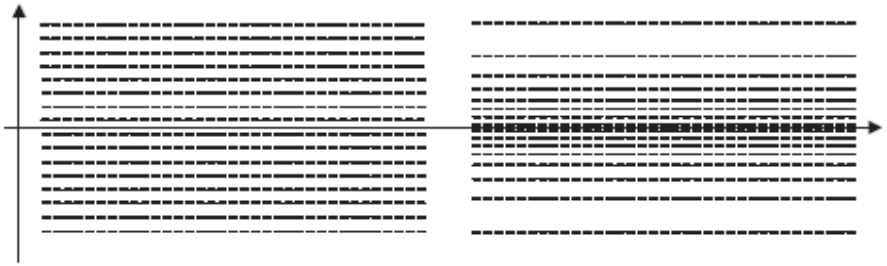


# Bits allocation for each band




(V. Lombardo, A.Valle, "*Audio e multimedia*", Apogeo)

# Non uniform quantization





# MPEG audio encoding layers (1)

Layer 1 (bitrate higher than 128 Kbps): DCT filter with only one *frame* and equal distribution of the frequencies in the sub-bands

- Psychoacoustic model uses only frequency masking
- Each frame has 32 blocks of 12 samples, a header, an error-detection code (CRC, Cyclic Redundancy Check), and optional additional information

Layer 2 (bitrate equal to 128 Kbps): works with three *frames* during the filtering process (previous, current, next, 1152 samples in total)

- Partially works with temporal masking
- Uses a more compact representation of additional information (header, number of bits for each band, … )



# MPEG audio quality

| Layer | Target bit rate | Compression | Quality at 64 kb | Quality at 128 kb | Delay |
|---|---|---|---|---|---|
| Layer I | 192 kb/s | 4:1 | -- | -- | 19 msec |
| Layer II | 128 kb/s | 6:1 | < 3 | 4+ | 35 msec |
| Layer III | 64 kbit/s | 12:1 | < 4 | 4+ | 59 msec |

Quality factor: 5 - perfect, 4 – barely noticeable, 3 – slightly annoying, 2 - annoying, 1 - very annoying



# Layer 3 of MPEG audio encoding (MP3)

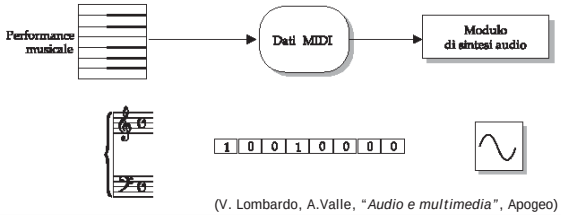
Layer 3 (bitrate at 64 Kbps): divides the frequencies spectrum into different sub-bands with nonequal amplitudes, more comparable to the critical sub-bands in the lower frequencies

- Psychoacoustic model with temporal masking
- It considers stereo redundancy
- Variable bitrate:
  - Huffman compression with pairs of values
  - Uses a bits reservoir with bits from frames that do not need them and can be allocated to the frames that do

# Musical Instruments Digital Interface

MIDI protocol (1983) provides a standard and efficient way to describe musical events

- It enables a computers, synthesizers, keyboards, and other musical devices to communicate each other
- MIDI is a scripting language – it codes ''events'' that stand for the production of sounds
- Sound generation is local to the synthesizers
- Messages describe the type of instruments, the notes to play, the volume, the speed, the effects, …




(V. Lombardo, A.Valle, "*Audio e multimedia*", Apogeo)



# Following MPEG audio formats

MPEG2 (November 1994)

- Is the standard for DVDs
- It was aimed at transparent sound reproduction for theaters. Works with five channels (left, center, right, left-surround, right-surround), plus a *low-frequency enhancement* (LFE) channel for very low frequencies (subwoofer)
- Works at 16 kHz, 22.05 kHz, or 24 kHz plus MP3 rates

MPEG4 (December 1999)

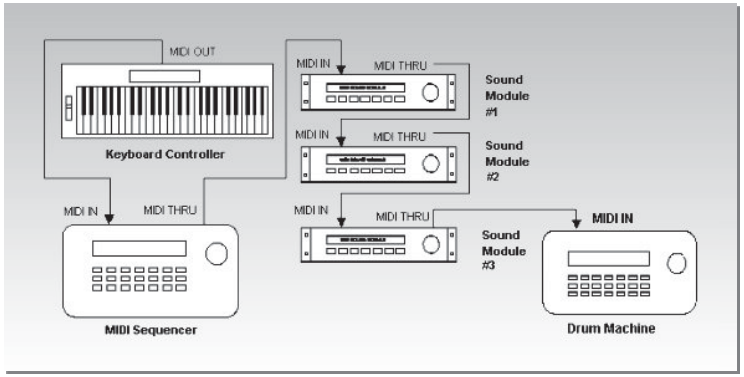
- Audio (and images) are considered a composition of different objects. The user can decide to listen to a concert situated in different places/environments or to emphasize some sounds over others



# MIDI systems



MIDI systems can be very complex …

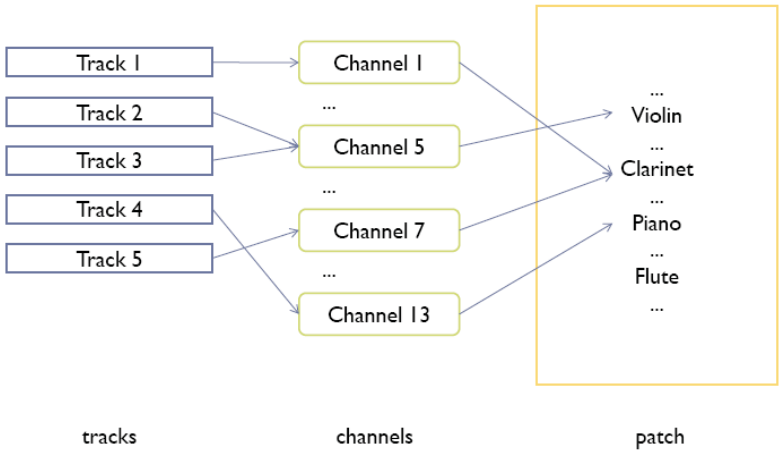



… but most of the soundcards come with all the necessary hardware



# MIDI, Musical Instruments Digital Interface

# MIDI music representation





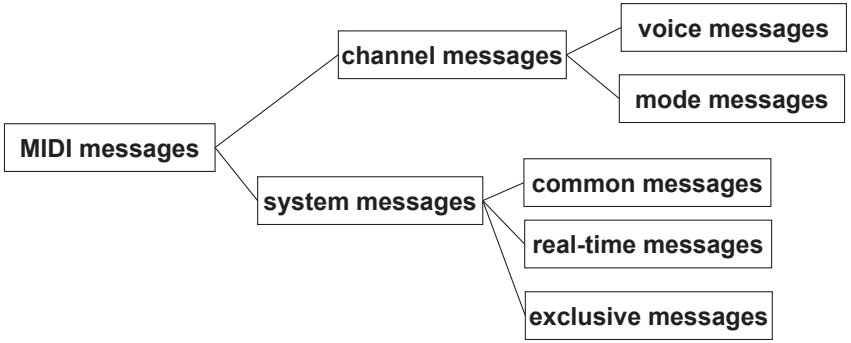

# The MIDI sequencer

## MIDI sequencer

- A recording and execution system for storing and editing a sequence of musical events, in the form of MIDI data
- It receives data from the input device, allows editing, and creates the music sending data to the synthesis device (example, sound card)
- It does not influence the quality of the sound. The quality totally depends on the synthesis device (or the synthesizer)



# MIDI messages




*Channel messages* describe which note to play (*voice*) and how to play it (*mode*)

*Systems messages* define set-up and synchronization information

All MIDI messages are a sequence of 10 bits (1 byte of useful data)



# MIDI Channels and Tracks

## Channels

- They allow to send and receive music data
- Method to differentiate timbres and send independent information: different channels for different instruments
- MIDI protocol provides 16 channels numbered from 1 to 16

## Tracks

- A track is a structured autonomous flow of MIDI messages
- Example: in a piano song, there are two tracks, the melody and the arrangement
- It can be considered as a messages container that can be assigned to different channels

## Patch

- It specifies the timbre produced by the generator
- MIDI can contain up to 128 different patches

# Channel messages

They contain the number of the channel through which the information is sent

*Voice messages* define what an instrument plays:

- Which note to play(*Note on*)
- Which note to turn off(*Note Off*)
- Potential controller effects (ex. vibrato) (*Pitch Bend Change*)
- Force measure for the keys on a specific channel(*Channel pressure*)
- …

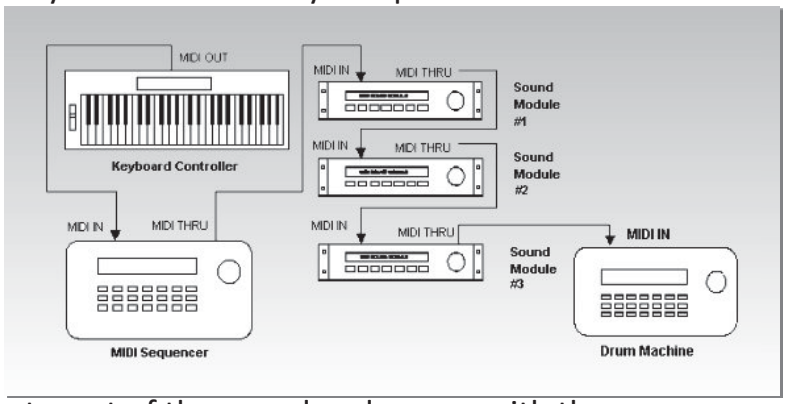
*Mode messages* describe how the instrument behaves when a voice message arrives

- Omni On/Off
- Poly/Mono
- General MIDI Mode



# MIDI systems

MIDI systems can be very complex …




… but most of the soundcards come with the necessary hardware



# System messages (1)

They do not use a channel because they are meant for commands that are not channel-specific.

Each device responds only to the messages it is enabled to answer

*System common messages*

- Carry out general functions that involve the entire system (ex., song synchronization when played by different devices)
- Set up a common clock
- Positioning inside a song (*Song Position Pointer*)
- Track selection (*Song Select*)



# MIDI Messages

Channel Messages

| 1st nibble | | | | 2nd nibble | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | x | x | x | n | n | n | n | Channel |

Message type

System messages

| 1st nibble | | | | 2nd nibble | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | x | x | x | x | Message type |

Following bytes

| 1st nibble | | | | 2nd nibble | | | |
|---|---|---|---|---|---|---|---|
| 0 | x | x | x | x | x | x | x |

# System messages (2)

## *System real-time messages*

- Related to real-time synchronization of the different modules of a system
- Device synchronization based on a relative time (24 messages every quarter)
- Start or stop the playback of a song (*Start*/*Stop*/*Continue*)
- Reset functions

## *System exclusive messages*

- Allow to manufacturers to extend the MIDI standard, sending messages that apply to their own product.



# MIDI: why & when (1)

The MIDI standard is an efficient way to encode musical sounds inside Web documents

- MIDI files are compact and have temporization information (there are no *hard real-time* constraints)
- Sound encoding is based on predefined discrete events, not on the waveform of the sound
- Complex musical songs take a small amount of storage

Particularly suitable for background music



# MIDI: why & when (2)

But…

- Only traditional western music can be encoded (tonal scale)
- It is not possible to represent sounds like noise, voice, other acoustic phenomena
- Computers and/or devices must have appropriate soundcards
- Quality depends on MIDI equipment (synthesizer)
- Channels and messages coding is not completely standard (ex. Roland, Yamaha, …)