INT 307
Multimedia Security System

Audio Representation and Compression

Sichen.Liu@xjtlu.edu.cn

XJTLU

Xi'an Jiaotong-Liverpool University
西交利物浦大学

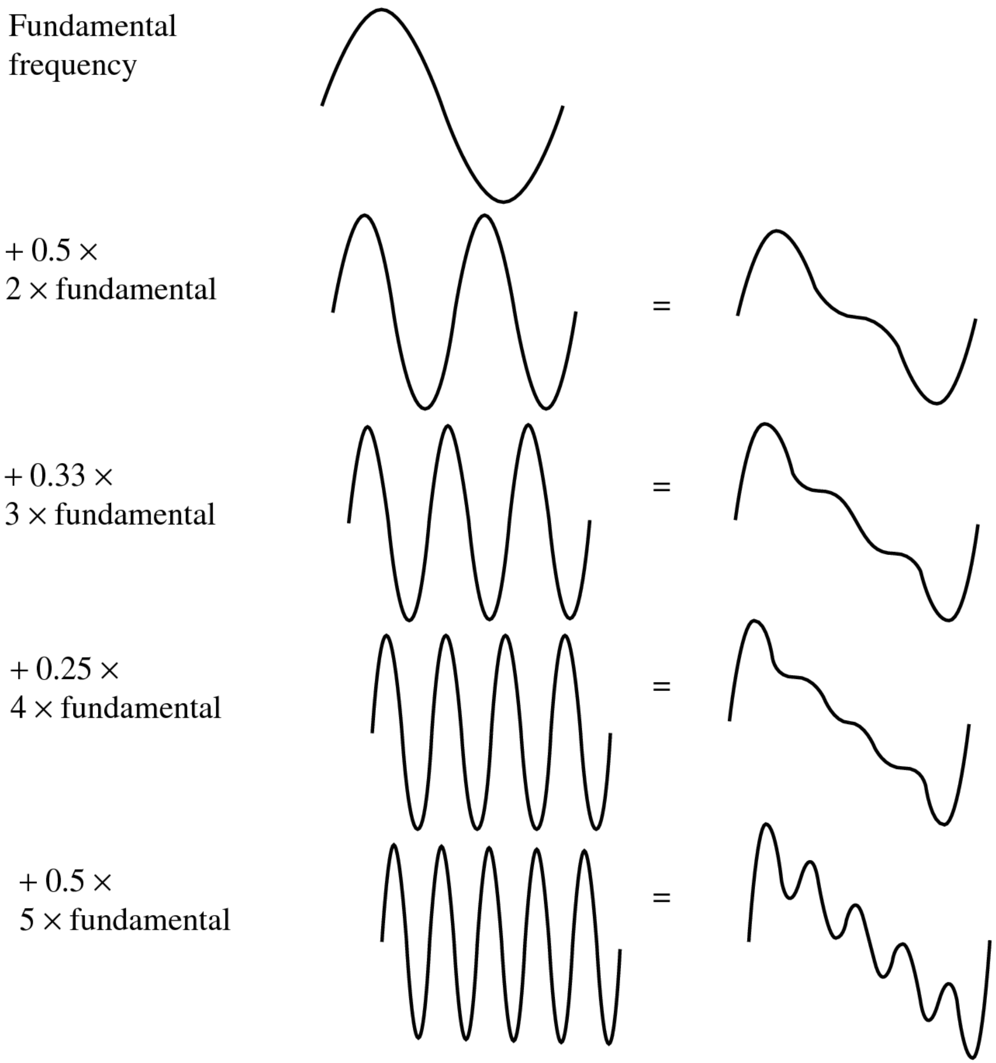
Aims

Master how audio is represented by computer system

Understand how people perceive audio

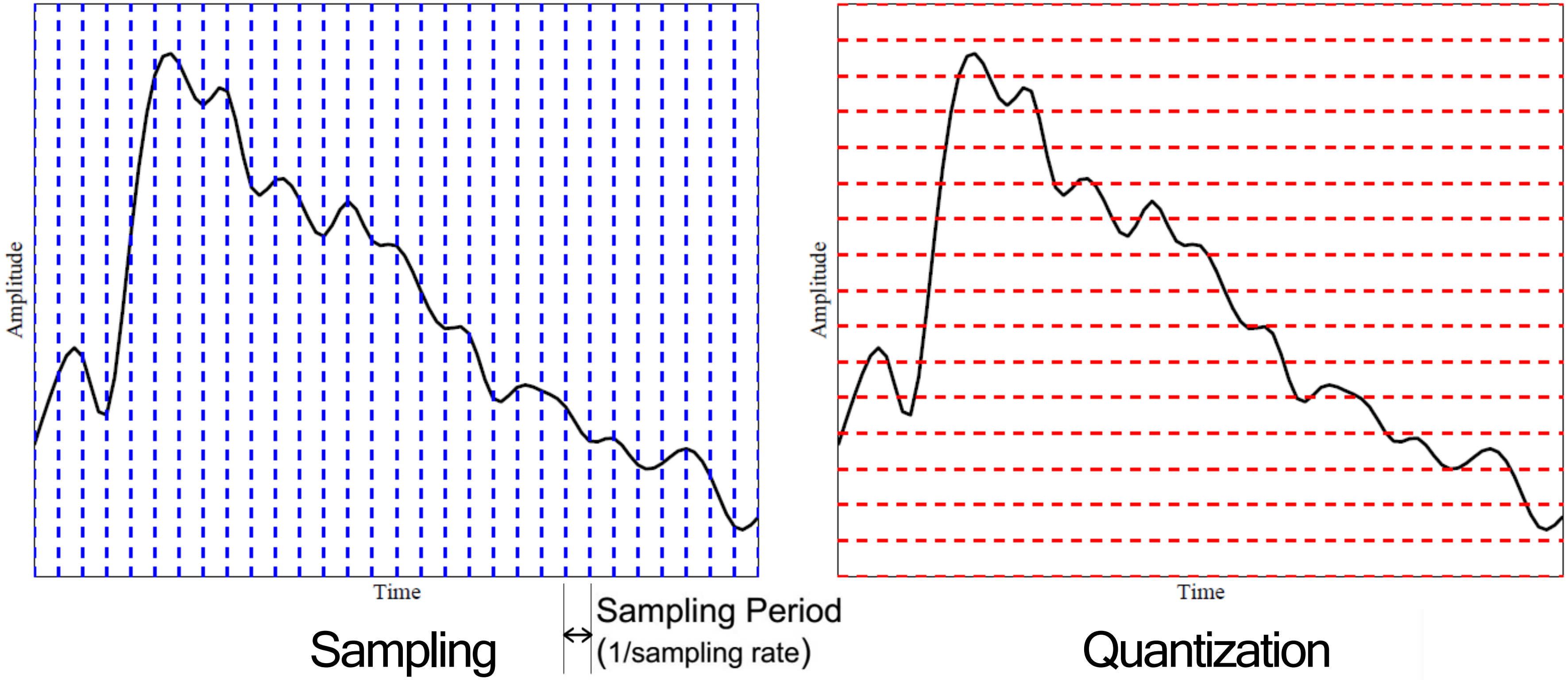
Understand how audio representation is compressed by computer system

What is Sound?



- Sound is a pressure wave. It takes on continuous values, as opposed to digitized ones.
- If we wish to use a digital version of sound waves, we must form digitized representations of audio information.

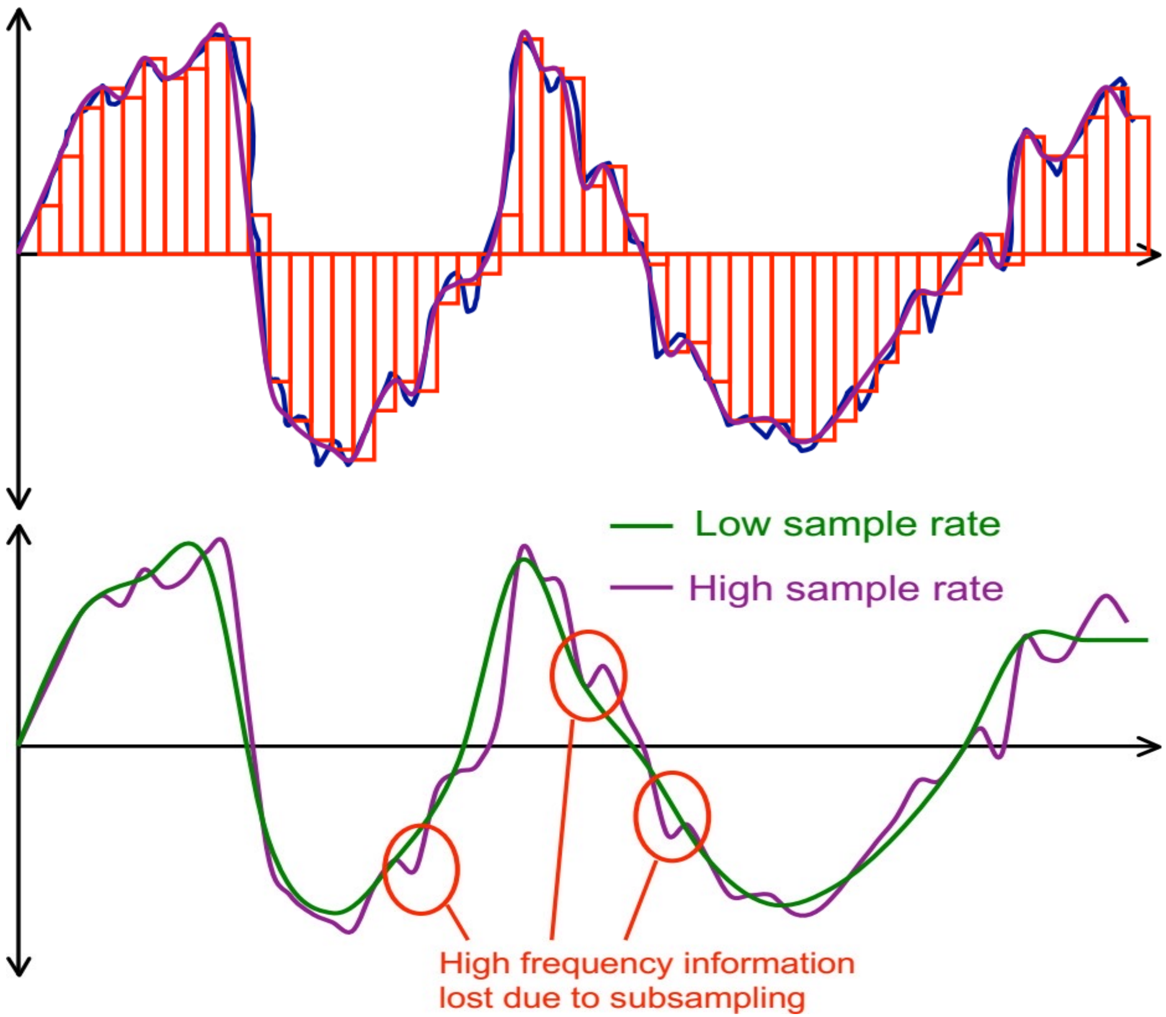
Sampling and Quantization



- Sampling the analog signal in the time dimension.

- Quantization is sampling the analog signal in the amplitude dimension.

Sample Rate



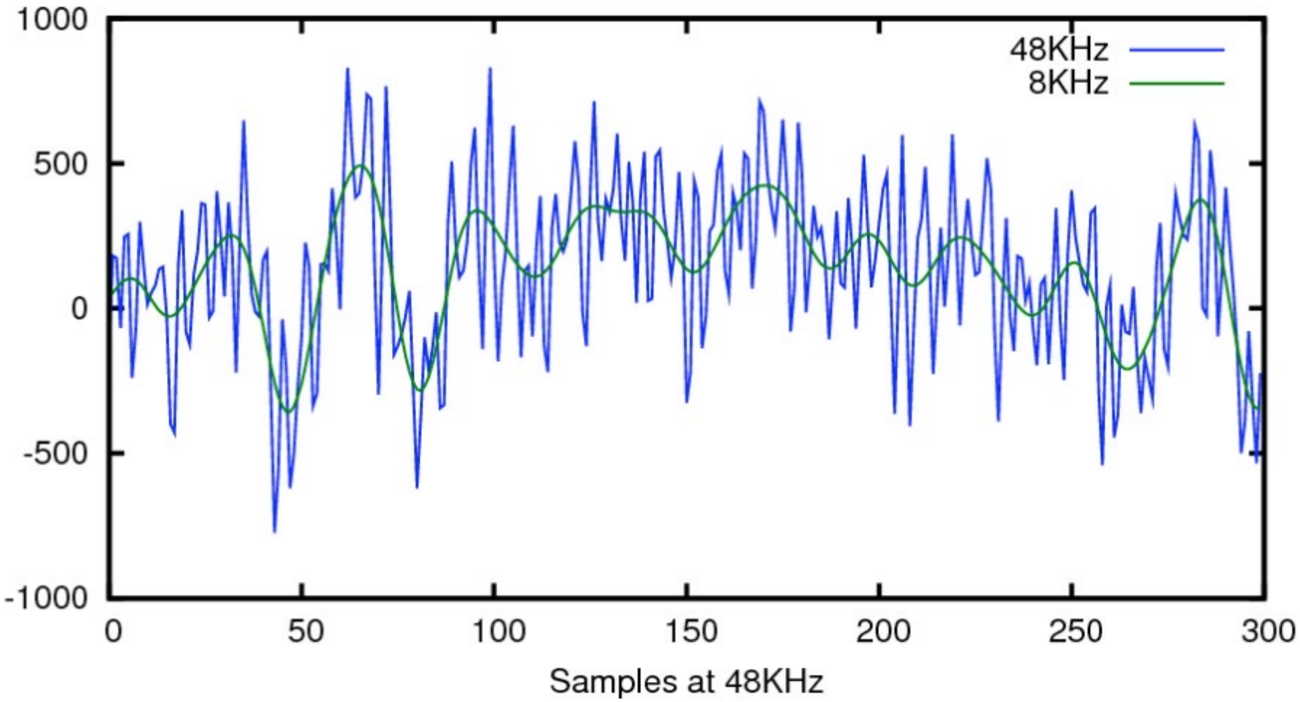
- Sample Rate:

  The number of samples per second. Also known as sampling frequency.
- Telephone:

  8000 Hz.
- CD (Compact Disc):

  44100 Hz

Real Music Example



How fast to sample?

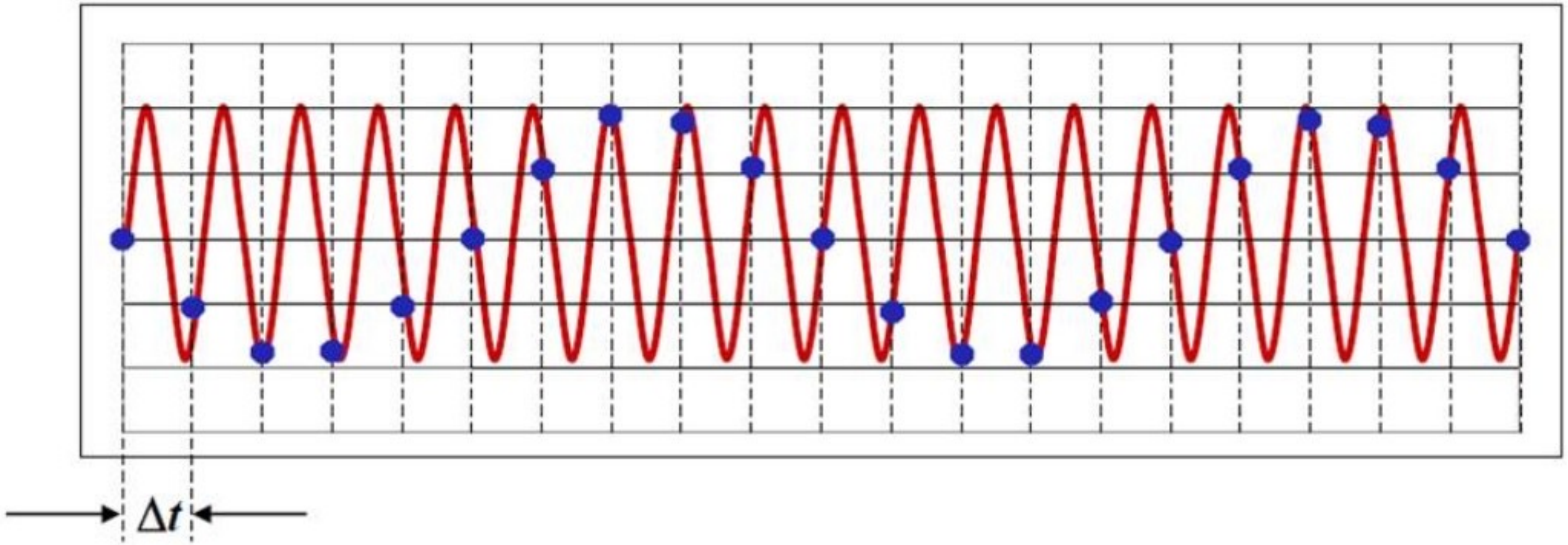
- Sampling with the times of the period can not detect the variation of the phase
- Sampling with half period can detect the rotation of the wheel, but can not judge the rotation direction (clock-wise or counter-clockwise).
- Sampling with less than half period can detect the variation of the phrase
- Sampling frequency > 2 * maximum signal frequency

Sampling Theorem

- Nyquist-Shannon sampling theorem
  - Formulated by Harry Nyquist in 1928 (“Certain topics in telegraph transmission theory”)
  - Proved by Claude Shannon in 1949 (“Communication in the presence of noise”).
- For no loss of information
- Sampling frequency > 2 * maximum signal frequency
- For a particular sampling frequency
  - Nyquist frequency = Sampling frequency/2
  - Nyquist frequency (or rate) is the highest frequency that can be accurately represented.
- Example
  - Limit of human hearing: 20KHz
  - By Nyquist, sample rate must be ≥ 40,000 samples/sec.
  - CD sample rate: 44,100 samples/sec.

Aliasing

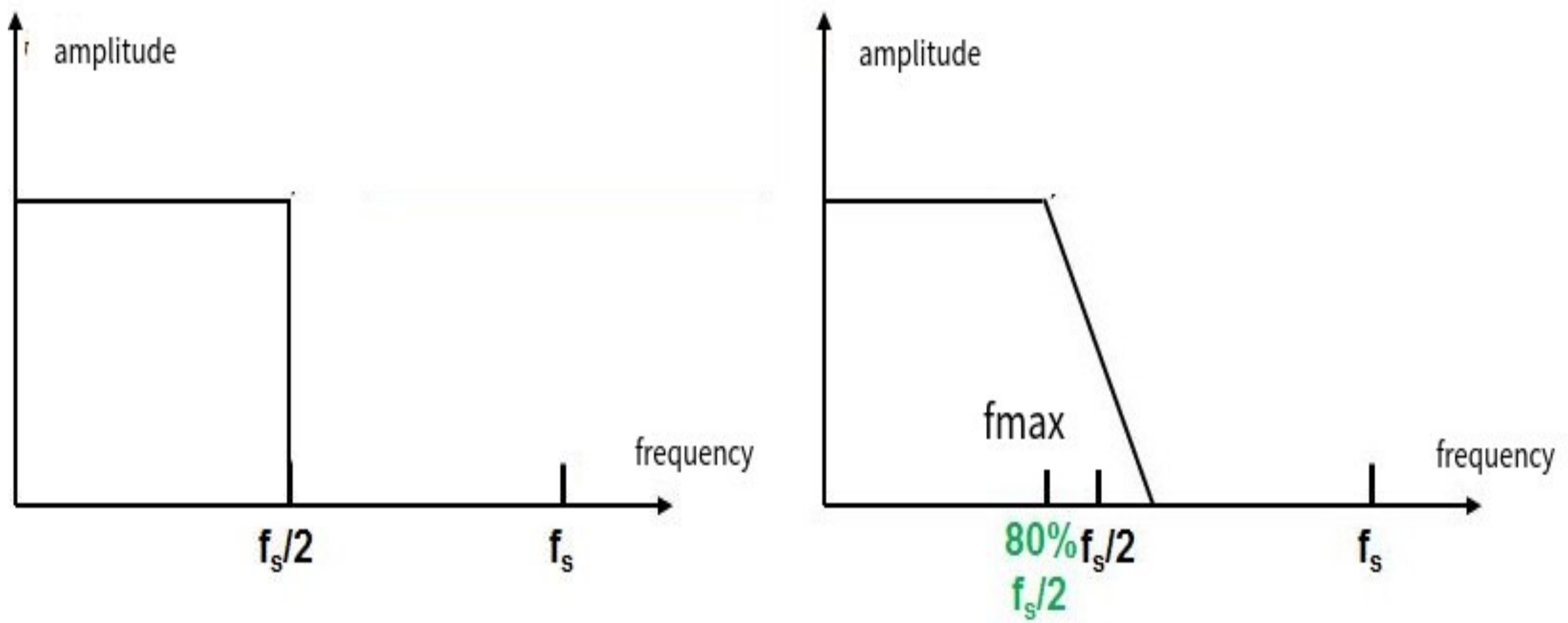
- What happens to all those higher frequencies you can't sample?
- They add noise to the sampled data at lower frequencies
- The signal with red color contains 18 periods, but the sampling signal with the blue color only has 2 periods.



- The high-frequency signal is aliased to the low-frequency one.

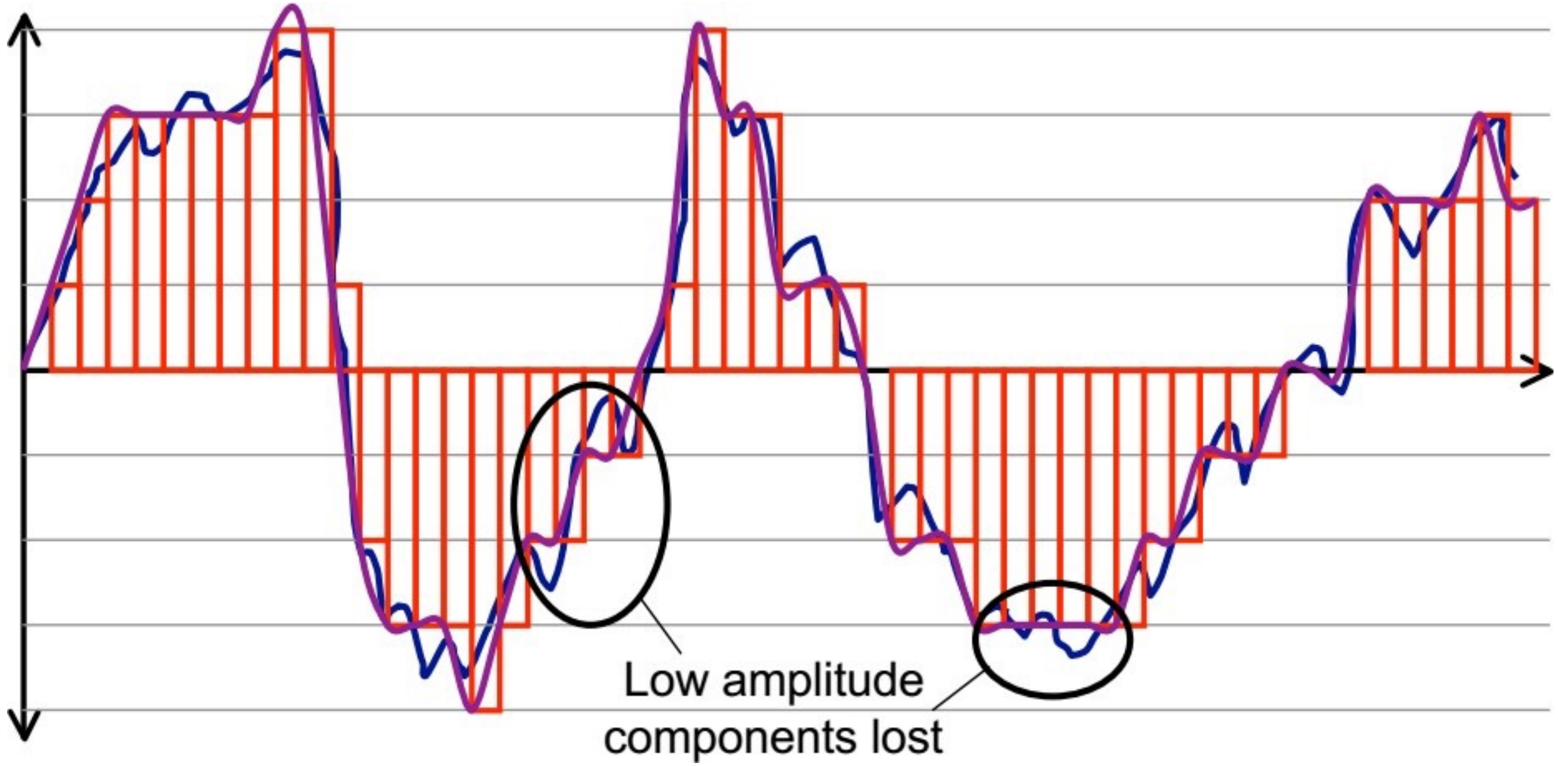
Solve Aliasing: Filter

- If Sampling frequency < = 2 * maximum signal frequency, then sampling signal will be aliased.
- Low-pass filter: before sampling, the frequencies above the Nyquist frequency component should be filtered out
- The area with 80% of the bandwidth is alias-free.



Quantization

- Sampled analog signal needs to be quantized (digitized).
- How many discrete digital values?
- Simplest quantization: linear
  - 8-bit linear, 16-bit linear



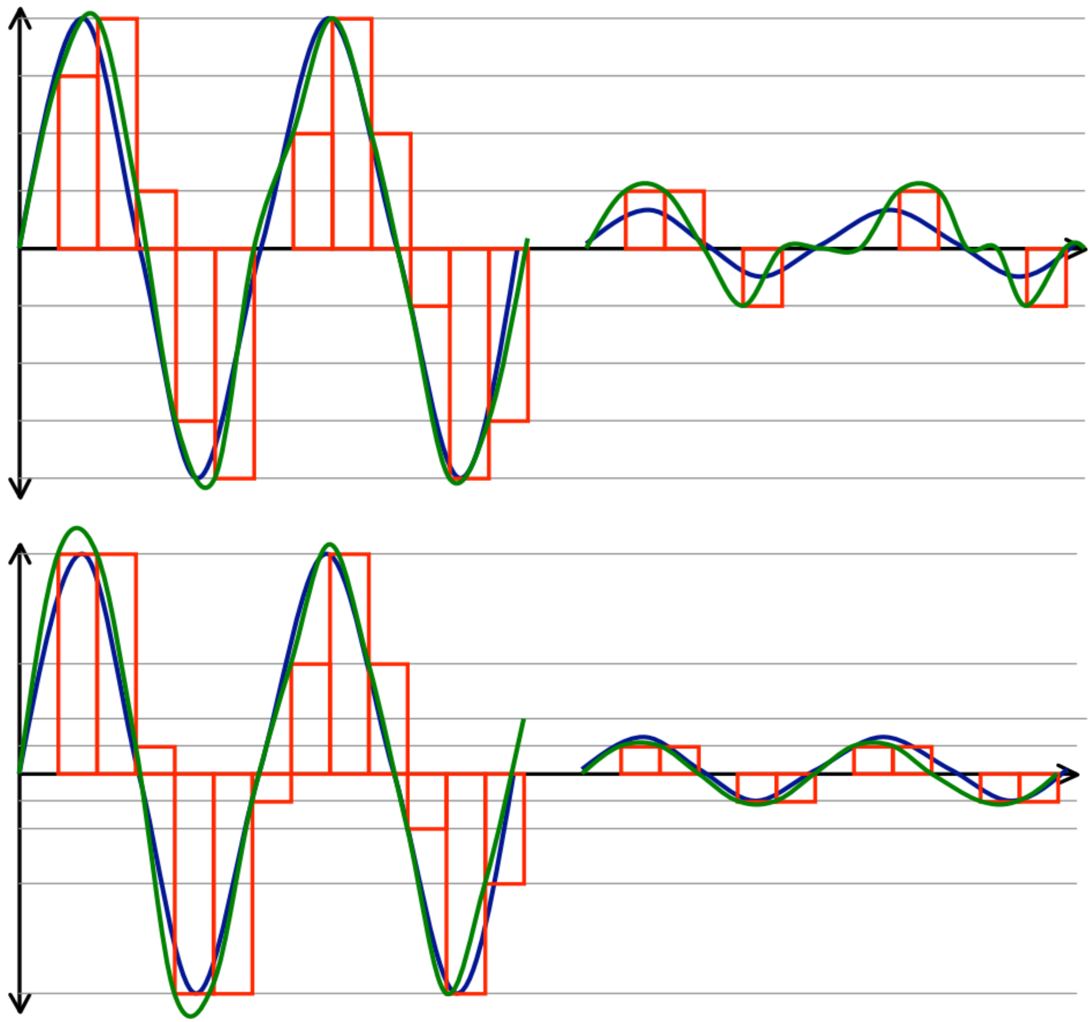
How many levels?

- 8 bits (256 levels) linear encoding would probably be enough if the signal always used the full range.
- But signal varies in loudness.
  - If full range is used for loud parts, quiet parts will suffer from bad quantization noise (only a few levels used).
  - If full range is used for quiet parts, loud parts will clip, resulting in really bad noise.
- CD uses 16-bit linear encoding (65536 levels).
  - Pretty good match to dynamic range of human ear.
- Solution: use 8 bits with an “logarithmic” encoding.
  - Goal is that quantization noise is a fixed **proportion** of the signal, irrespective of whether the signal is quiet or loud.

Linear Encoding and Logarithmic Encoding

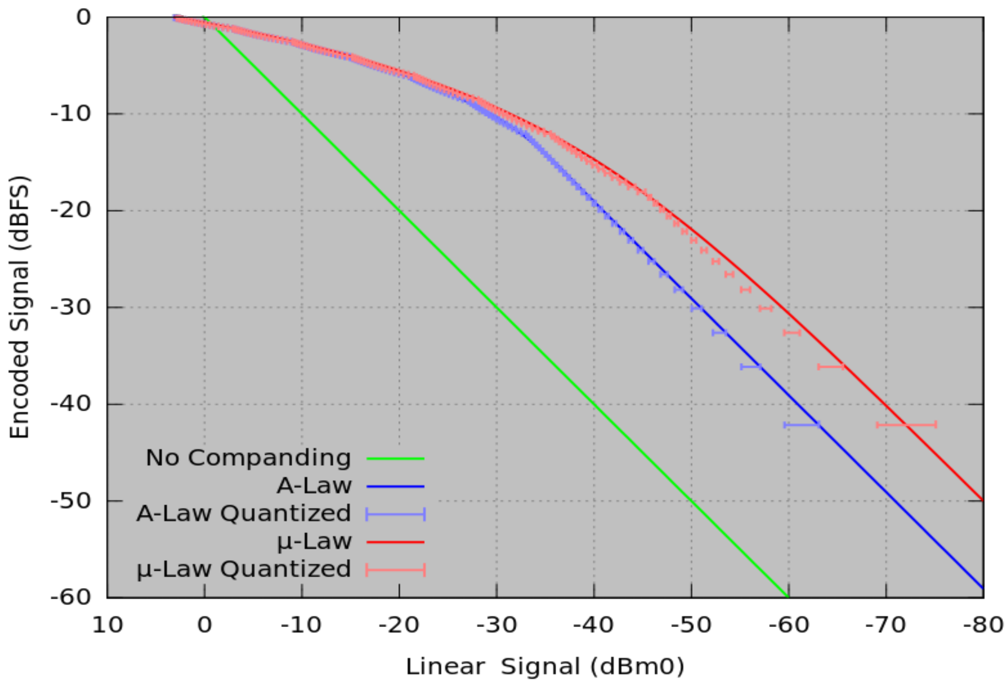
Linear Encoding

Logarithmic Encoding



μ-law vs A-law

- 8-bit μ-law used in US for telephony
- 8-bit A-law used in Europe for telephony
  - Similar, but a slightly different curve.
  - Both give similar quality to 12-bit linear encoding.
  - A-law used for International circuits.
- Both are linear approximations to a log curve.



Calculation Question

– Suppose we have a piece of audio lasting for 1 hour with sampling rate of 44.1 kHz. How many bits are needed to record the audio with 16-bit depth? How many bits are needed per second?

Why Audio Perception?

- We need to compress audio files
- Traditional lossless compression (such as entropy coding, Huffman coding) can achieve a compression rate of 50% at most
- For better compression rate, we need to compress the piece of audio in a lossy way
- Lossy compression means that we remove the redundancy information that cannot be perceived
- Hence we need to understand **auditory perception** first

Psychoacoustics

- The range of human hearing is about 20 Hz to about 20 kHz

- The frequency range of the voice is typically only from about 500 Hz to 4 kHz

- The dynamic range, the ratio of the maximum sound amplitude to the quietest sound that humans can hear, is on the order of about 120 dB

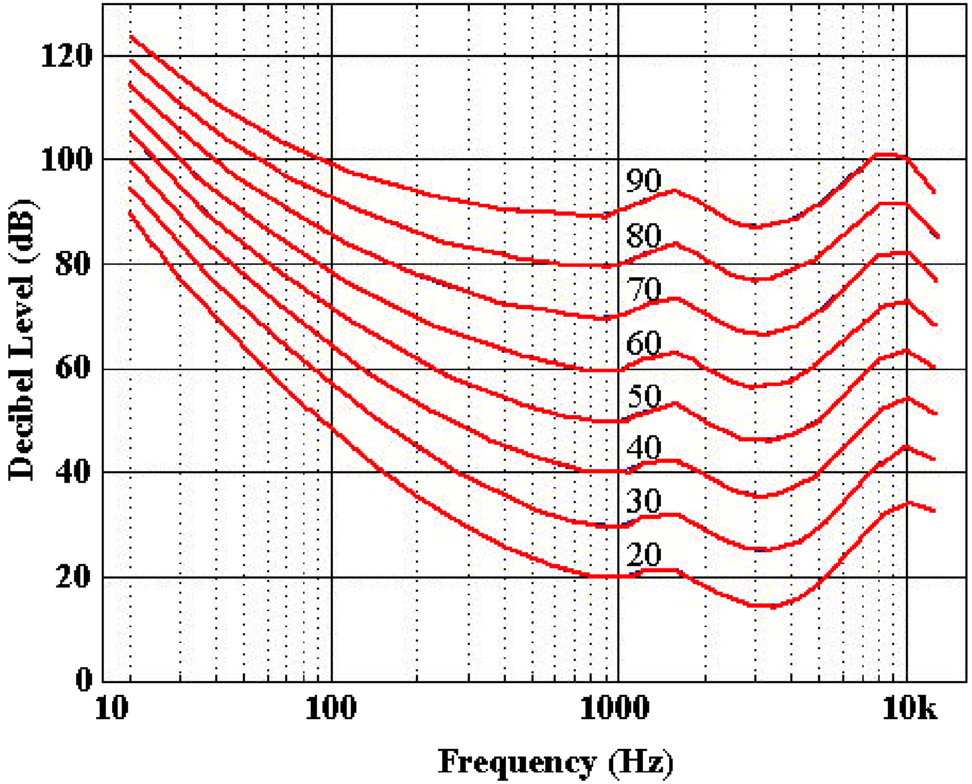
Equal-Loudness Relations

- Decibel
  - A ratio with a standardized threshold of hearing intensity
- Phons
  - Equal intensity is not equal loudness
  - 60 Phons means "as loud as a 60 dB of a 1000 Hz sound"
- Equal loudness curves that display the relationship between perceived loudness (Phons) for a given stimulus sound volume (Sound Pressure Level), as a function of frequency
- Fletcher-Munson Curves

1000Hz

50Hz

Equal-Loudness Relations



- The ear's perception of equal loudness
- The bottom curve shows what level of pure sound stimulus is required to produce the perception of a 10 dB sound
- All the curves are arranged so that the perceived loudness level gives the same loudness as for that loudness level of a pure sound at 1 kHz

Frequency Masking

- Lossy audio data compression methods, such as MPEG/Audio encoding, remove some sounds which are masked anyway

- The general situation regarding masking is as follows:

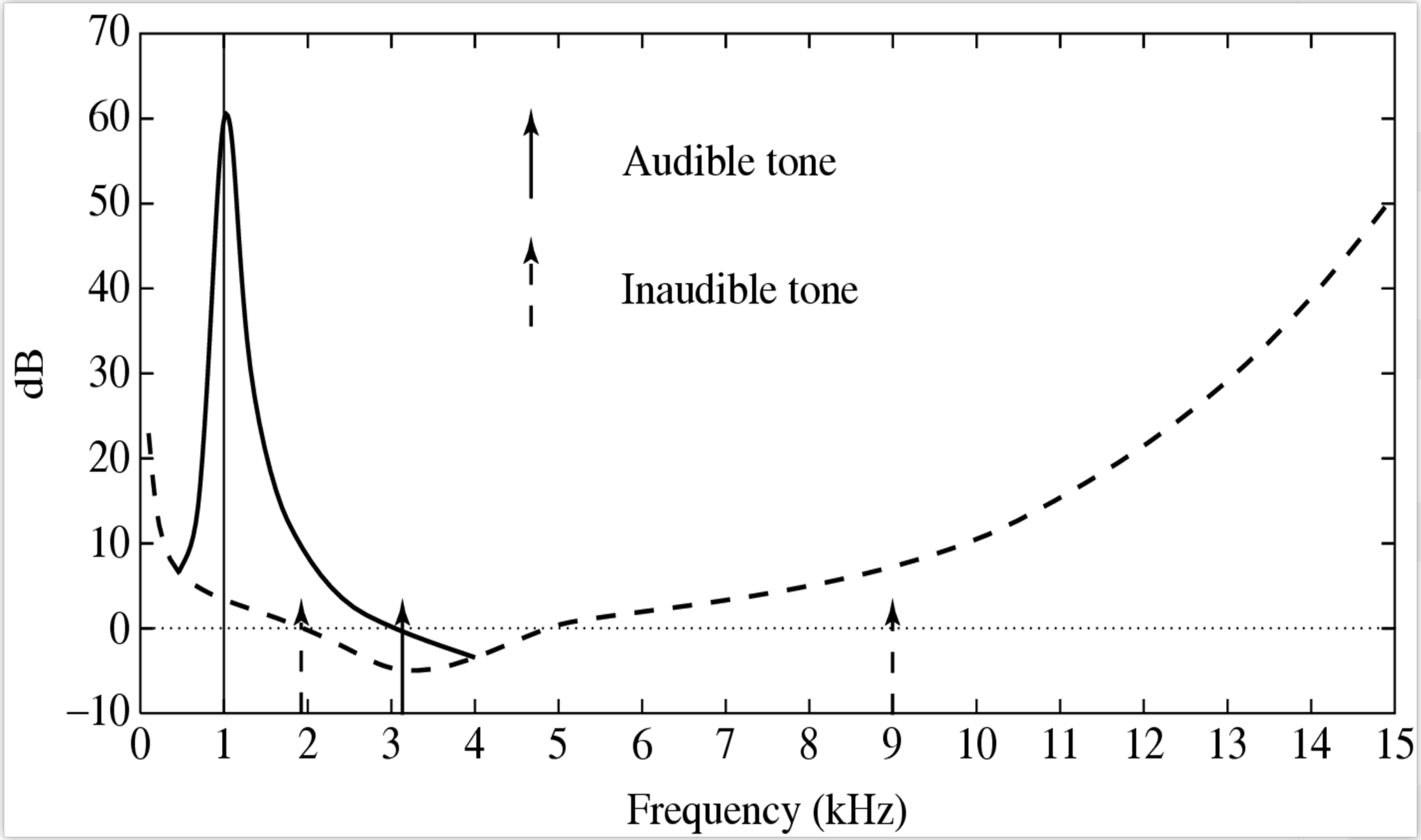
  - A lower sound can effectively mask (make us unable to hear) a higher sound

  - The reverse is not true - a higher sound does not mask a lower sound well

  - The greater the power in the masking sound, the wider is its influence - the broader the range of frequencies it can mask

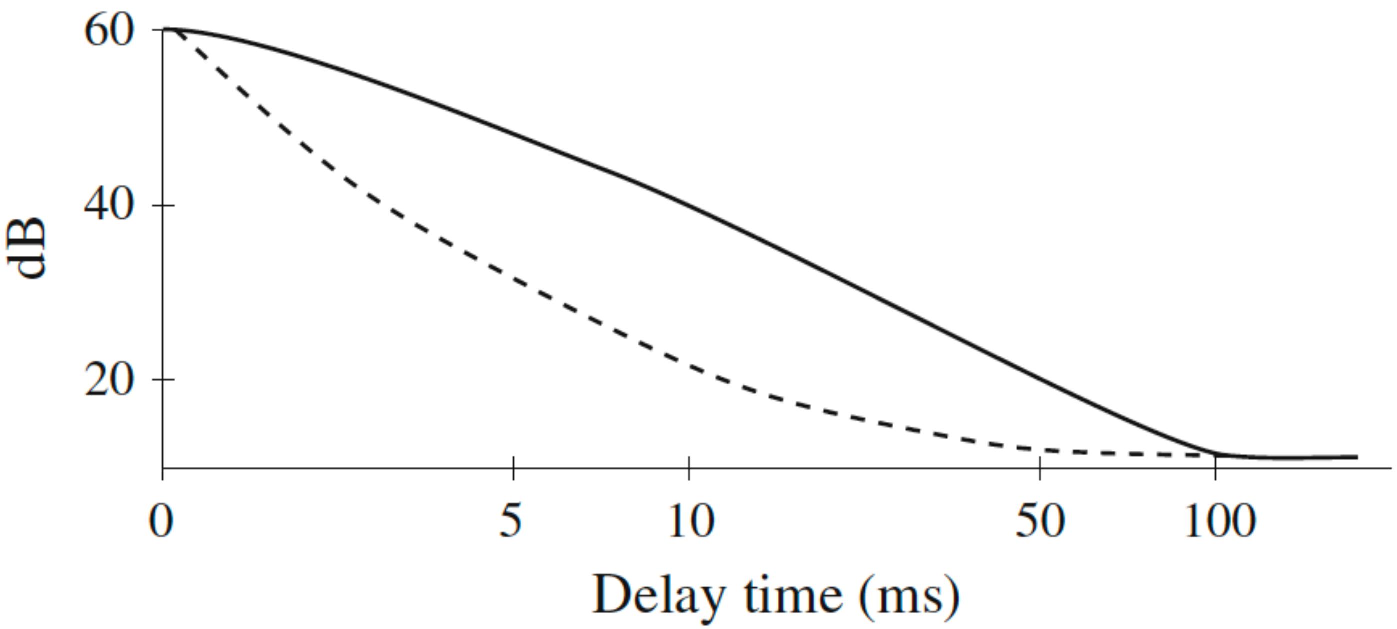
  - Therefore, if two sound are widely separated in frequency then little masking occurs

Frequency Masking Curves

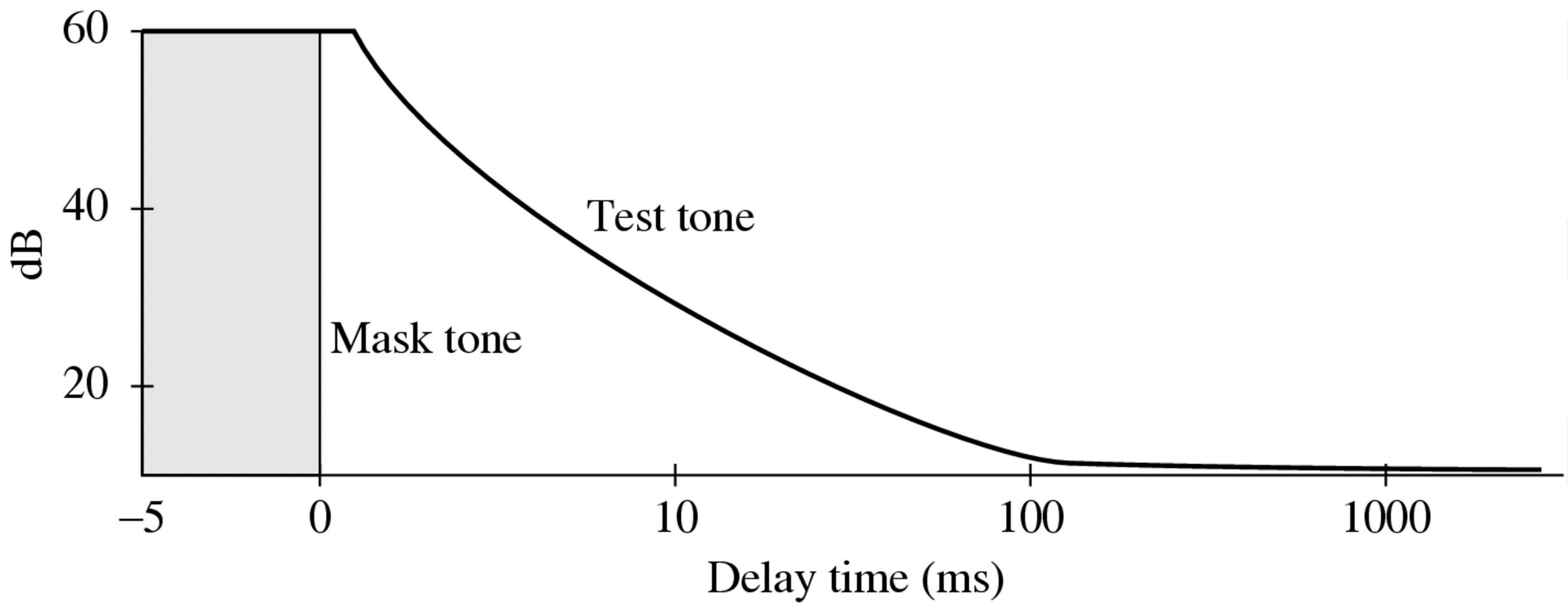


Temporal Masking

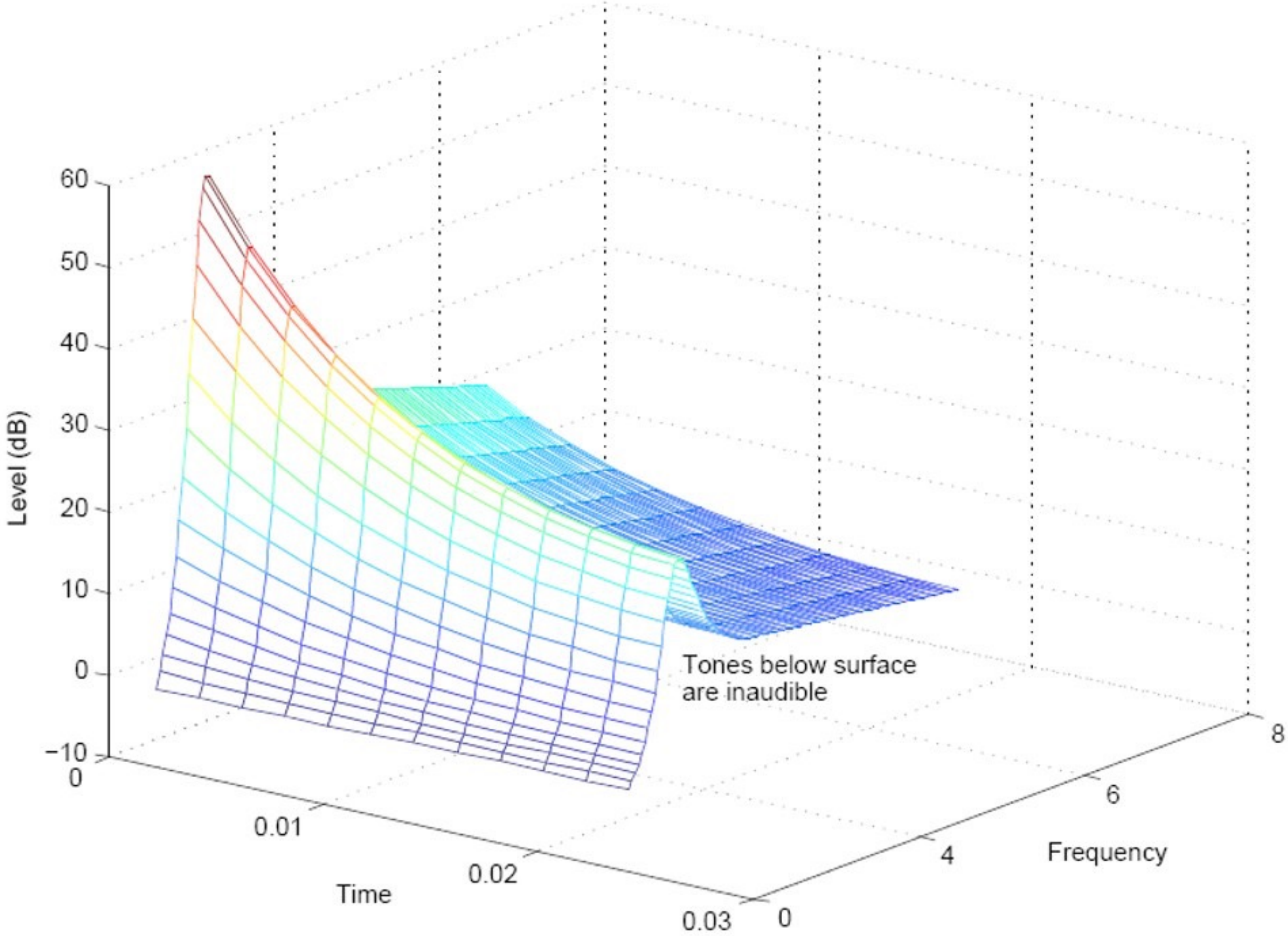
- Phenomenon: any loud sound will cause the hearing receptors in the inner ear to become saturated and require time to recover
- For a masking sound that is played for a longer time, it takes longer before a test sound can be heard



Temporal Masking



Effect of temporal and frequency masking

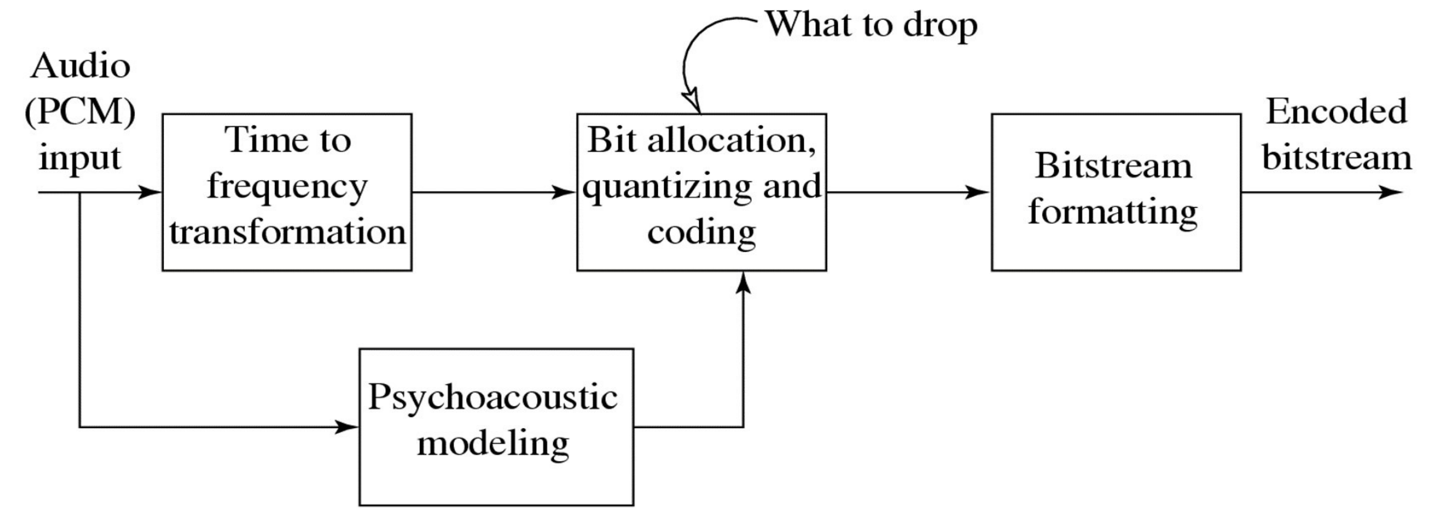


Audio Compression

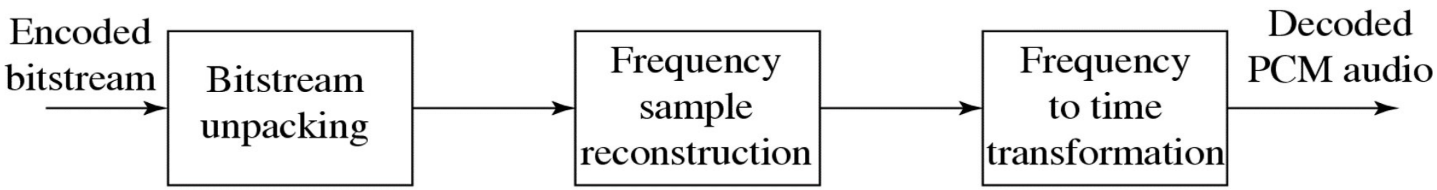
MPEG Audio Strategy

- MPEG approach to compression relies on
  - Quantization
  - Make use of masking effects on loudness / frequency / temporal
- Frequency masking: by using a psychoacoustic model to estimate the just noticeable noise level
  - Encoder balances the masking behaviour and the available number of bits by discarding inaudible frequencies
  - Scaling quantization according to the sound level that is left over, above masking levels
- May consider the actual width of the critical bands
  - For practical purposes, audible frequencies are divided into 25 main critical bands
  - To keep simplicity, adopts a uniform width for all frequency analysis filters, using 32 overlapping sub bands

MPEG Audio Compression Algorithm

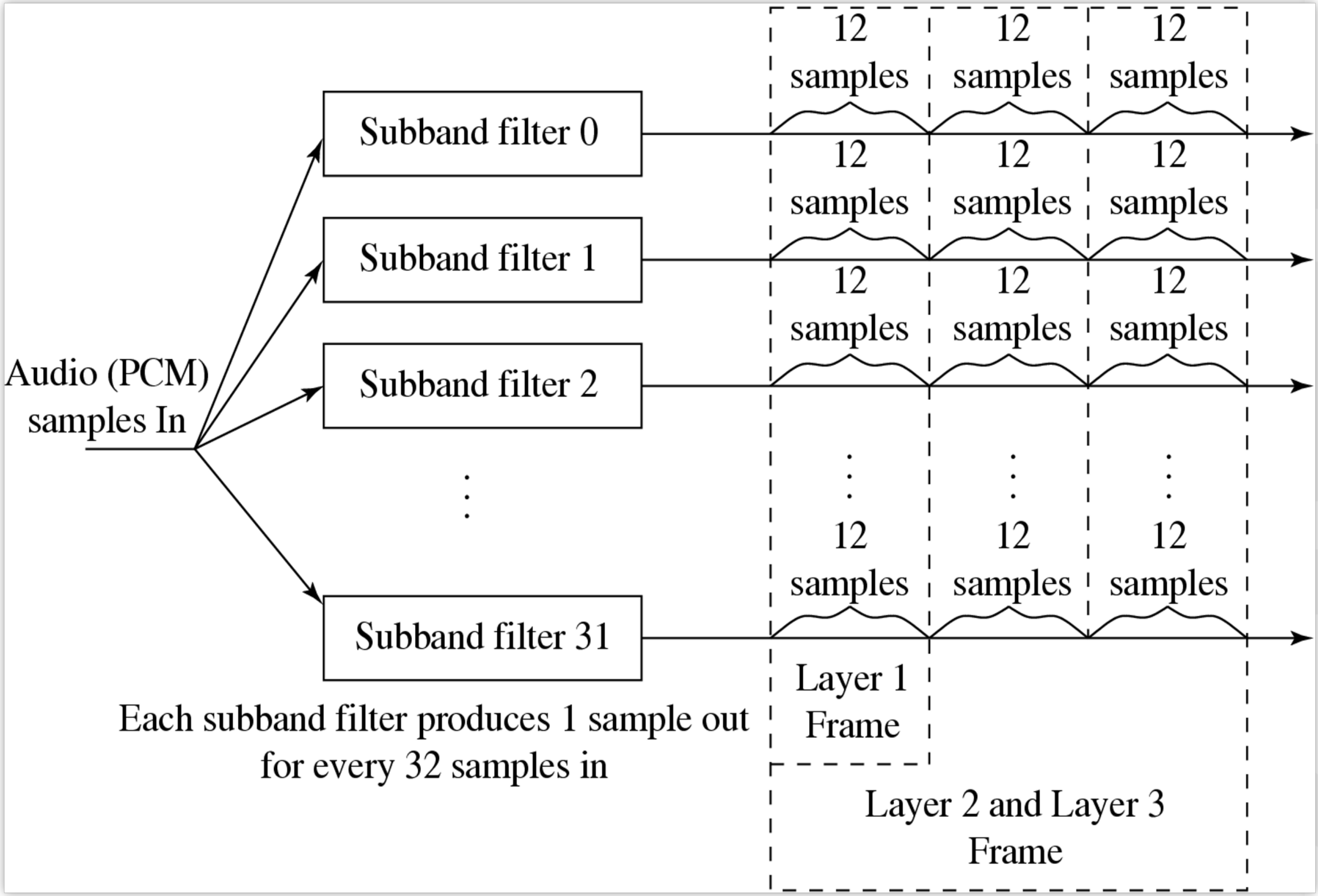


(a) MPEG Audio Encoder



(b) MPEG Audio Decoder

Filter Bank



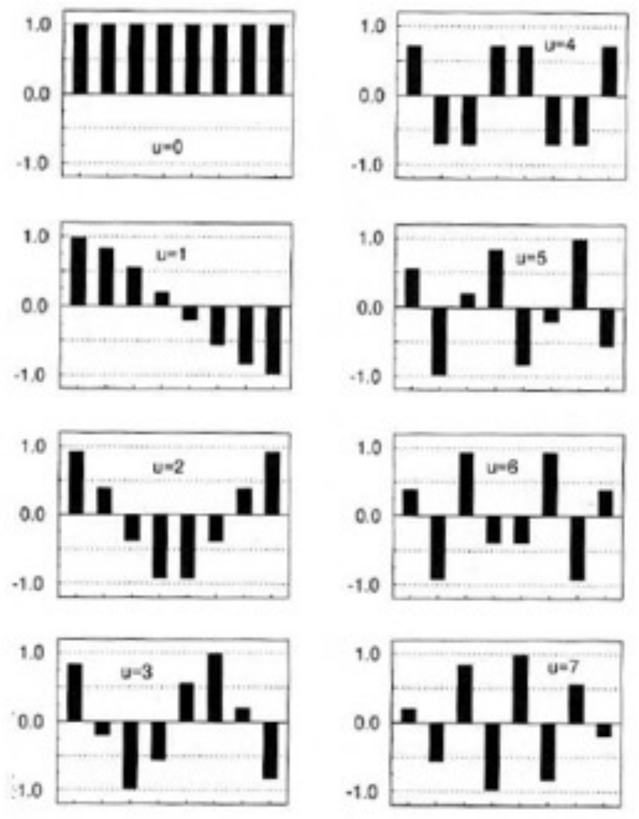
Audio (PCM) samples In
Subband filter 0
Subband filter 1
Subband filter 2
Subband filter 31
12 samples
12 samples
12 samples
Layer 1 Frame
Layer 2 and Layer 3 Frame
Each subband filter produces 1 sample out for every 32 samples in

Filter Bank

Inversible Transform

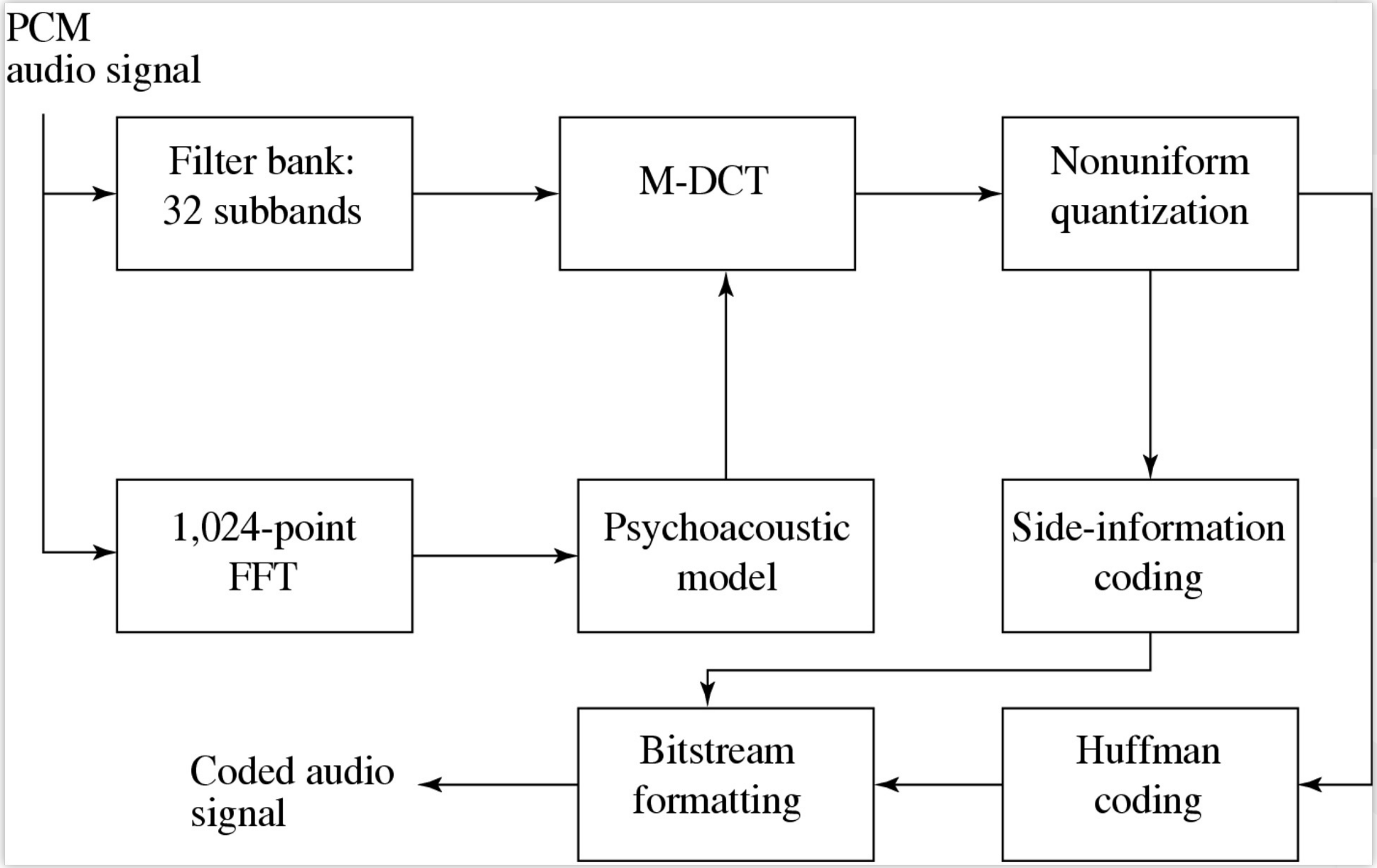
$$F(\mu) = \frac{C(\mu)}{2} \sum_{x=0}^{7} f(x) \cos[(2x+1)\mu\pi/16]$$

$$C(\mu) = \begin{cases} \frac{1}{\sqrt{2}}, \mu = 0 \\ 1, \mu > 0 \end{cases}$$



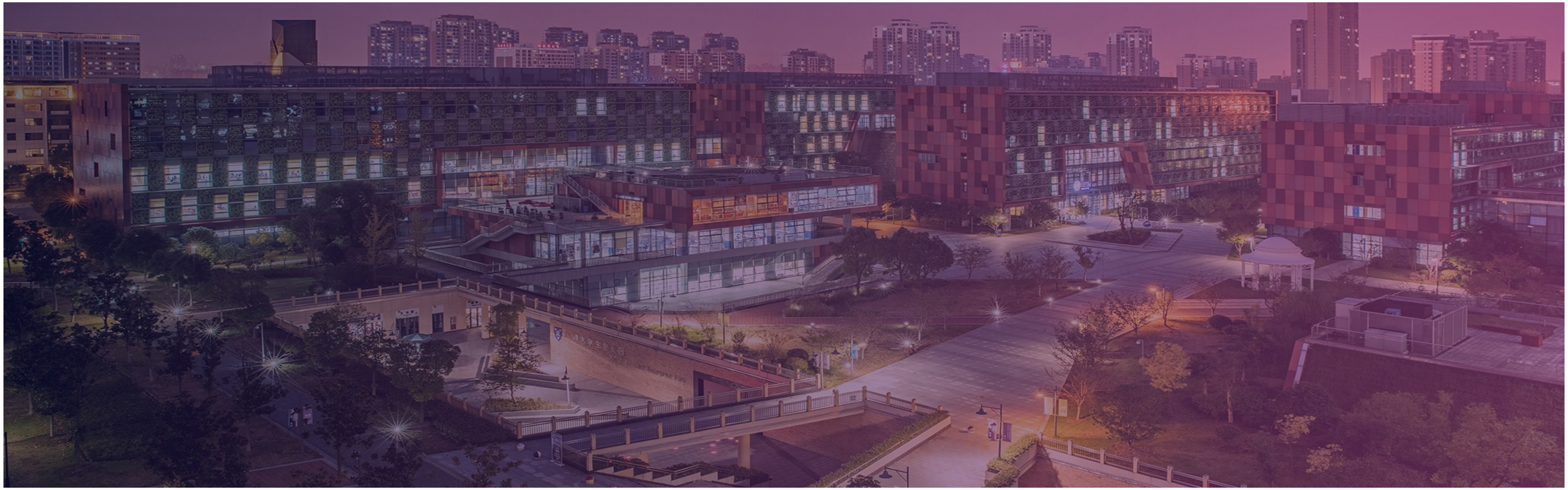
DCT bases

MPEG Audio Framework



Industrial Standards

- MPEG-2 AAC (Advanced Audio Coding)
  - The standard vehicle for DVDs
  - Aimed at transparent sound reproduction for theaters
  - Also capable of delivering high-quality stereo sound at bit-rates below 128 kbps
  - Supports three different "profiles"
- MPEG-4 Audio
  - Integrates several different audio components into one standard: speech compression, perceptually based coders, text-to-speech and MIDI
- Others: Dolby AC-2, Dolby AC-3, Sony ATRAC



THANK YOU

VISIT US

WWW.XJTLU.EDU.CN

FOLLOW US

@XJTLU

Xi'an Jiaotong-Liverpool University
西交利物浦大学

XJTLU | SCHOOL OF FILM AND TV ARTS