INT307 Multimedia Security System

Audio Compression

Shengchen Li

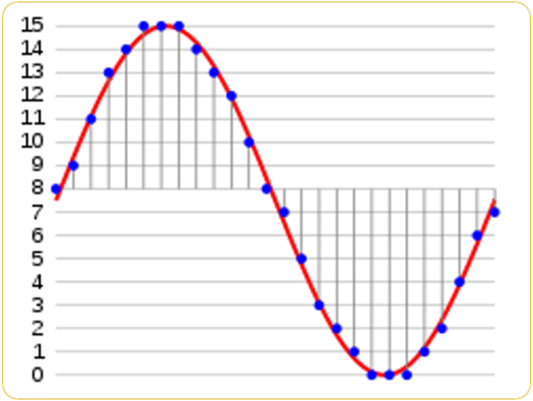
Xi'an Jiaotong-Liverpool University

23rd Aug 2022

Aims

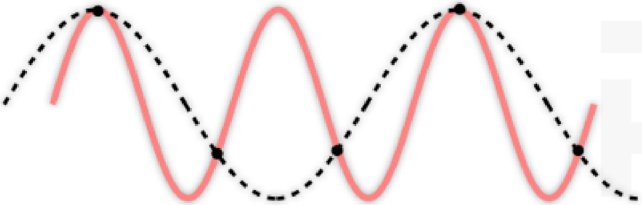
- Master how audio is represented by computer system
- Understand how people perceive audio
- Understand how audio representation is compressed by computer system

Audio Representation



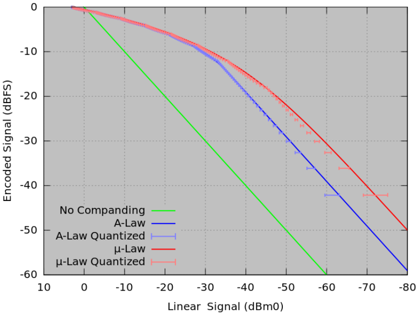
Analogue Signal $\rightarrow$ Discrete Signal $\rightarrow$ Digital Signal

Sampling



- If the sampling rate is too low, the signal will distort such aliasing happens

Quantisation



- Bit depth: the number of bits used to represent each sample
- Usually, the raw waveform is transformed by A-law or μ-law to optimise the dynamic range
- A-law and μ-law introduce non-linear quantisation effectively

A-Law

The function of A-law is

$$F(x) = \operatorname{sgn}(x) \begin{cases} \dfrac{A|x|}{1+\ln(A)}, & |x| < \dfrac{1}{A} \\[2ex] \dfrac{1+\ln(A|x|)}{1+\ln(A)}, & \dfrac{1}{A} \leq |x| \leq 1. \end{cases} \tag{1}$$

The inverse function of A-law is

$$F^{-1}(y) = \operatorname{sgn}(y) \begin{cases} \dfrac{|y|(1+\ln A)}{A}, & |y| < \dfrac{1}{1+\ln A} \\[2ex] \dfrac{e^{-1+|y|(1+\ln A)})}{A}, & \dfrac{1}{1+\ln A} \leq |y| \leq 1. \end{cases} \tag{2}$$

A is a number.

μ-Law

The function of μ-law is

$$F(x) = \operatorname{sgn}(x)\frac{\ln 1 + \mu|x|}{\ln 1 + \mu}, -1 \leq x \leq 1. \tag{3}$$

The inverse function of μ-law is

$$F^{-1}(y) = \operatorname{sgn}(y)\frac{(1 + \mu)^{|y|} - 1}{\mu}, -1 \leq y \leq 1. \tag{4}$$

μ is a number.

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

- Fletcher-Munson Curves

- Equal loudness curves that display the relationship between perceived loudness (“Phons”, in dB) for a given stimulus sound volume (“Sound Pressure Level”, also in dB), as a function of frequency

Decibel

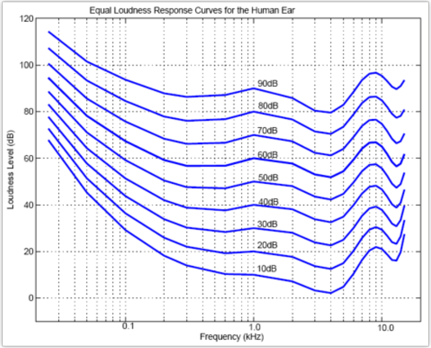
A ratio with a standardised threshold of hearing intensity

$$I_0(\text{dB}) = 10^{-12}\text{watt}/m^2 \tag{5}$$

$$P_0 = 2 \times 10^{-5} N/m^2 \tag{6}$$

$$I(\text{dB}) = 10\log_{10}[\frac{I}{I_0}] = 10\log_{10}[\frac{P^2}{P_0^2}] = 20\log_{10}[\frac{P}{P_0}] \tag{7}$$

Equal-Loudness Relations



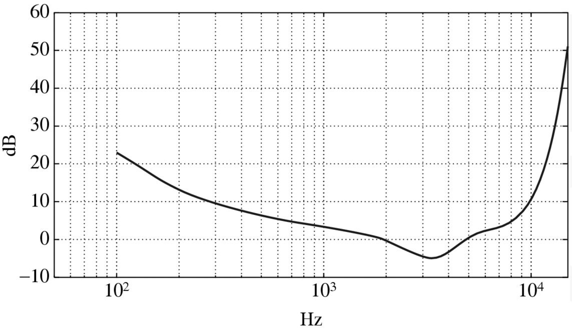
- The ear's perception of equal loudness
- The bottom curve shows what level of pure tone stimulus is required to produce the perception of a 10 dB sound
- All the curves are arranged so that the perceived loudness level gives the same loudness as for that loudness level of a pure tone at 1 kHz

Loudness Measurements

- Phons
  - Equal intensity is not equal loudness
  - 60 Phons means "as loud as a 60 dB of a 1000 Hz tone"
  - Orchestral music is usually between 40 to 100 phones
- Sones
  - Equal intensity is not equal loudness
  - $L_N = 40 + 10 \log_2 N$
  - Orchestral music is usually between 1 to 64 sones

Threshold of Hearing

A plot of the threshold of human hearing for a pure tone



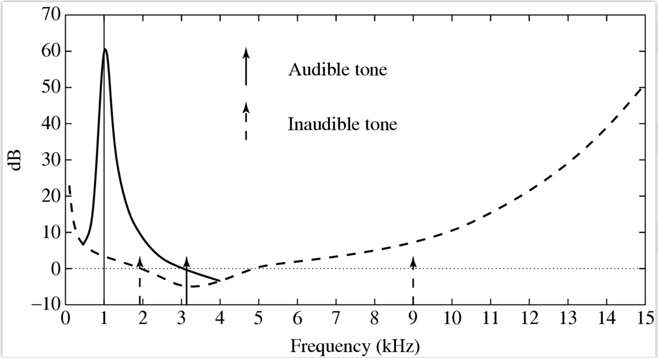
Frequency Masking

- 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 tone can effectively mask (make us unable to hear) a higher tone
  - The reverse is not true – a higher tone does not mask a lower tone well
  - The greater the power in the masking tone, the wider is its influence – the broader the range of frequencies it can mask
  - Therefore, if two tones are widely separated in frequency then little masking occurs

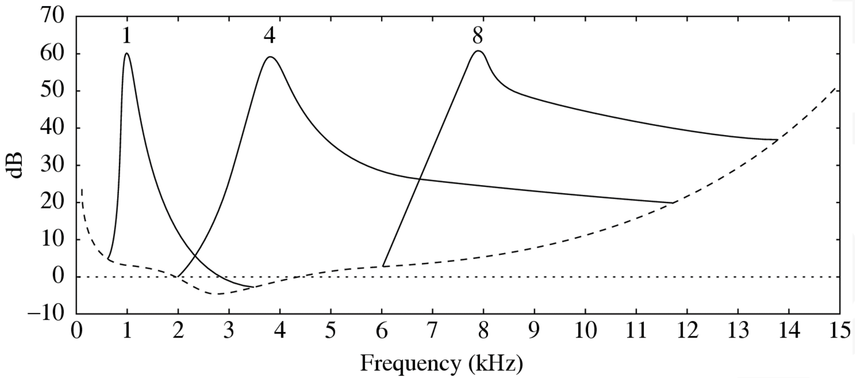
Frequency Masking Curves

- Frequency masking is studied by playing a particular pure tone, say 1 kHz again, at a loud volume, and determining how this tone affects our ability to hear tones nearby in frequency
- One would generate a 1 kHz masking tone, at a fixed sound level of 60 dB, and then raise the level of a nearby tone, e.g., 1.1 kHz, until it is just audible

Effect on threshold for 1 kHz masking tone



Effect of masking tone at three different frequencies



Critical Bands

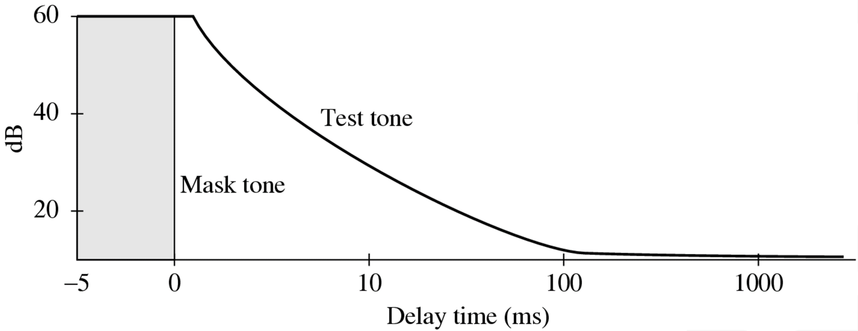
- Critical bandwidth represents the ear's resolving power for simultaneous tones or partials
- At the low-frequency end, a critical band is less than100 Hz wide, while for high frequencies the width can be greater than 4 kHz
- Experiments indicate that the critical bandwidth:
  - for masking frequencies $<$ 500 Hz: remains approximately constant in width (about 100 Hz)
  - for masking frequencies $>$ 500 Hz: increases approximately linearly with frequency

Temporal Masking

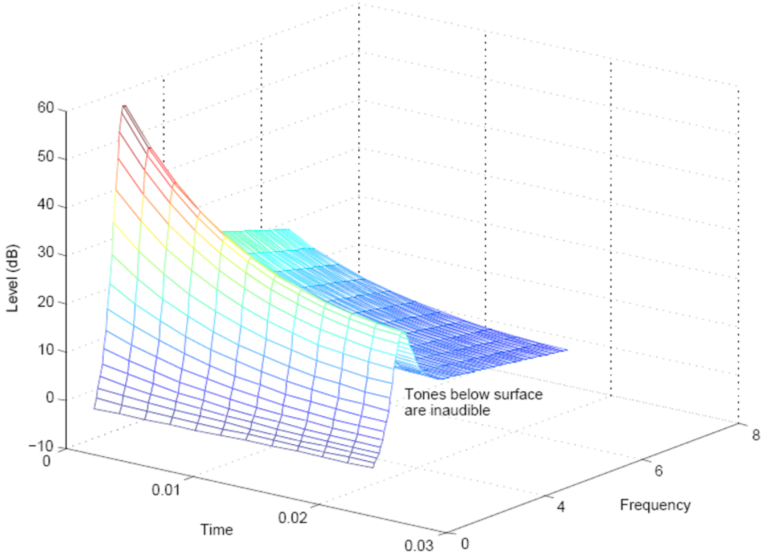
- Phenomenon: any loud tone will cause the hearing receptors in the inner ear to become saturated and require time to recover

- For a masking tone that is played for a longer time, it takes longer before a test tone can be heard

Temporal Masking



Effect of temporal and frequency masking



Audio Compression

MPEG Audio Strategy

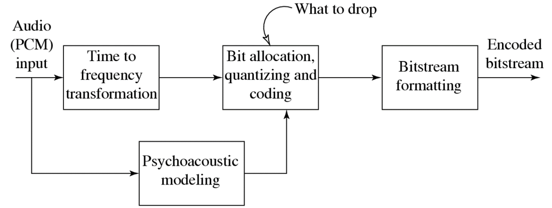
- MPEG approach to compression relies on
  - Quantisation
  - Make use of masking effects on loudness / frequency / temporal
- MPEG encoder employs a bank of filters to
  - Analyse the frequency (“spectral”) components of the audio signal by calculating a frequency transform of a window of signal values
  - Decompose the signal into sub-bands by using a bank of filters

Audio Compression

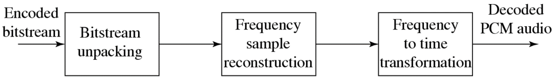
MPEG Audio Strategy

- 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

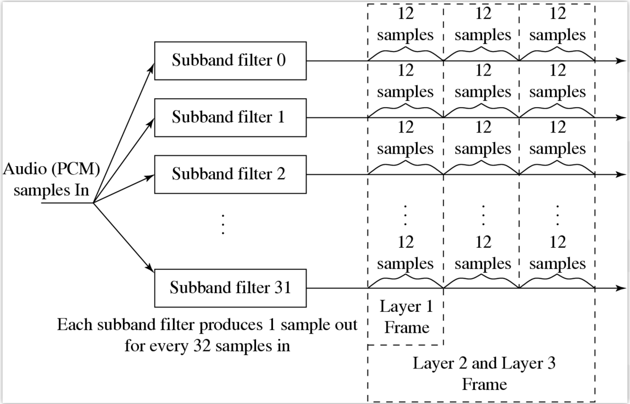


(a) MPEG Audio Encoder



(b) MPEG Audio Decoder

Filter Bank

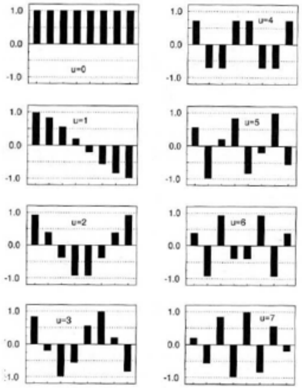

Audio (PCM)
samples In
Subband filter 0
Subband filter 1
Subband filter 2
Subband filter 31
12
samples
Layer 1
Frame
Layer 2 and Layer 3
Frame
Each subband filter produces 1 sample out
for every 32 samples in

Filter Bank

Inversible Transform

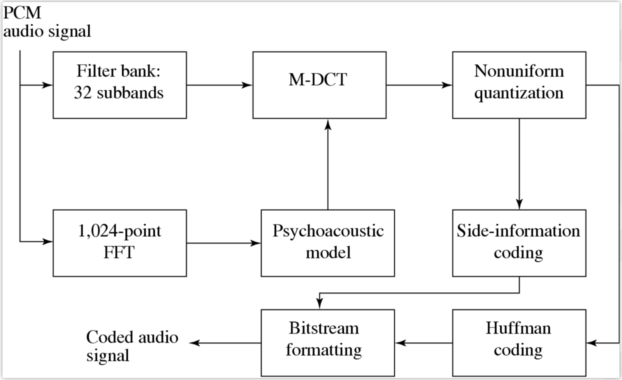
$$F(\mu) = \frac{C(\mu)}{2} \sum_{x=0}^{7} f(x) \cos[(2x+1)\mu\pi/16] \tag{8}$$

$$C(\mu) = \begin{cases} \frac{1}{\sqrt{2}}, \mu = 0 \\ 1, \mu > 0 \end{cases} \tag{9}$$



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