CpSc 863: Multimedia Systems and Applications

Audio Compression

James Wang

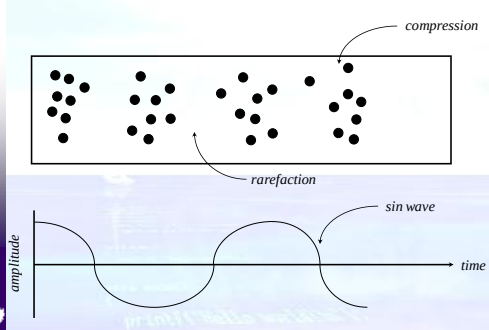
Chapter 13, 14 of book "Fundamentals of Multimedia" by Li and Drew
Chapter 6 of book "Multimedia: Computing, Communications and Applications" by Steinmetz and Nahrstedt
Some content from lecture notes by Prof. Lawrence A. Rowe

Audio

- **Acoustics is the study of sound**
  - Generation, transmission, and reception of sound waves
  - Sound wave - energy causes disturbance in a medium
- **Example is striking a drum**
  - Head of drum vibrates => disturbs air molecules close to head
  - Regions of molecules with pressure above and below equilibrium
  - Sound transmitted by molecules bumping into each other

2

Sound Waves



3

Sending/Receiving

- **Receiver**
  - A microphone placed in sound field moves according to pressures exerted on it
  - Transducer transforms energy to a different form (e.g., electrical energy)
- **Sending**
  - A speaker transforms electrical energy to sound waves

4

Signal Fundamentals

- **Pressure changes can be periodic or aperiodic**
- **Periodic vibrations**
  - *cycle* - time for compression/rarefaction
  - *cycles/second* - frequency measured in hertz (Hz)
  - *period* - time for cycle to occur (1/frequency)
- **Frequency ranges**
  - barametric pression is 10^{-6} Hz
  - cosmic rays are 10^{22} Hz
  - human perception [0, 20kHz]

5

Wave Lengths

- **Wave length is distance sound travels in one cycle**
  - 20 Hz is 56 feet
  - 20 kHz is 0.7 inch
- **Bandwidth is frequency range**
- **Transducers cannot linearly produce human perceived bandwidth**
  - Frequency range is limited to [20 Hz, 20 kHz]
  - Frequency response is not flat

6

Measures of Sound

- **Sound volume related to pressure amplitude**
  - sndpres = instantaneous sndpres - equilibrium sndpres
  - usually very small (e.g., normal conversation 10^{-6} in)
- **Sound level is a logarithmic scale**
  - SPL = 10 log (*pressure*/reference) decibels (dB)
  - where reference is $2*10^{-4}$ dyne/cm^2
  - 0 dB SPL - essentially no sound heard
  - 35 dB SPL - quiet home
  - 70 dB SPL - noisy street
  - 120 dB SPL - discomfort

7

Sound Phenomena

- **Sound is typically a combination of waves**
  - Sin wave is fundamental frequency
  - Other waves added to it to create richer sounds
  - Musical instruments typically have fundamental frequency plus overtones at integer multiples of the fundamental frequency
- **Waveforms out of phase cause interference**
- **Other phenomena**
  - Sound reflects off walls if small wave length
  - Sound bends around walls if large wave lengths
  - Sound changes direction due to temperature shifts

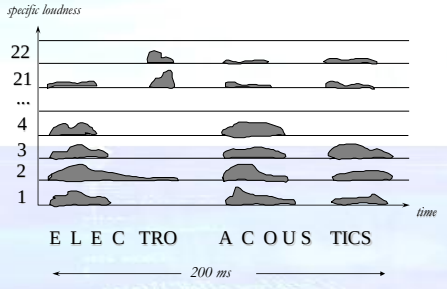
8

Human Perception

- **Speech is a complex waveform**
  - Vowels and bass sounds are low frequencies
  - Consonants are high frequencies
- **Humans most sensitive to low frequencies**
  - Most important region is 2 kHz to 4 kHz
- **Hearing dependent on room and environment**
- **Sounds masked by overlapping sounds**

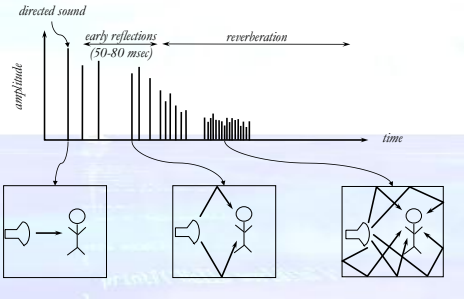
9

Critical Bands



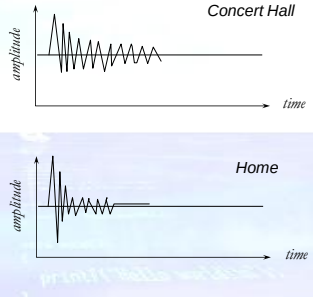
10

Sound Fields



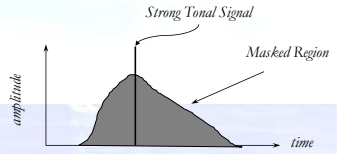
11

Impulse Response



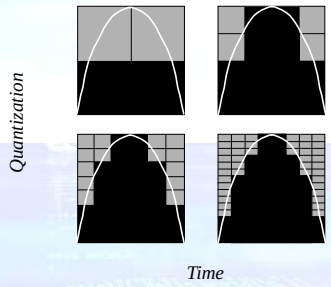
12

Audio Noise Masking



13

Audio Sampling



14

Audio Representations

Optimal sampling frequency is twice the highest frequency to be sampled (Nyquist Theorem)

Format	Sampling Rate	Bandwidth	Frequency Band
Telephony	8 kHz	3.2 kHz	200-3400 Hz
Teleconferencing	16 kHz	7 kHz	50-7000 Hz
Compact Disk	44.1 kHz	20 kHz	20-20,000 Hz
Digital Audio Tape	48 kHz	20 kHz	20-20,000 Hz

15

Jargons/Standards

- **Emerging standard formats**
  - 8 kHz 8-bit U-LAW mono
  - 22 kHz 8-bit unsigned linear mono and stereo
  - 44 kHz 16-bit signed mono and stereo
  - 48 kHz 16-bit signed mono and stereo
- **Actual standards**
  - G.711 - A-LAW/U-LAW encodings (8 bits/sample)
  - G.721 - ADPCM (32 kbs, 4 bits/sample)
  - G.723 - ADPCM (24 kbs and 40 kbs, 8 bits/sample)
  - G.728 - CELP (16 kbs)
  - GSM 06.10 - 8 kHz, 13 kbs (used in Europe)
  - LPC (FIPS-1015) - Linear Predictive Coding (2.4kbs)
  - CELP (FIPS-1016) - Code excitied LPC (4.8kbs, 4bits/sample)
  - G.729 - CS-ACELP (8kbs)
  - MPEG1/MPEG2, AC3 - (16-384kbs) mono, stereo, and 5+1 channels

16

Audio Packets and Data Rates

- **Telephone uses 8 kHz sampling**
  - ATM uses 48 byte packets ➔ 6 msecs per packet
  - RTP uses 160 byte packets ➔ 20 msecs per packet
- **Need many other data rates**
  - ≤ 30 kbs ➔ audio over 28.8 kbs modems
  - 32 kbs ➔ good stereo audio is possible
  - 56 kbs or 64 kbs ➔ conventional telephones
  - 128 kbs ➔ MPEG1 audio
  - 256 - 384 kbs ➔ higher quality MPEG/AC3 audio

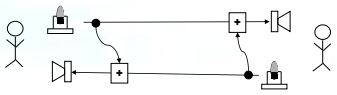
17

Discussion

- **Higher quality**
  - Filter input
  - More bits per sample (i.e. 10, 12, 16, etc.)
  - More channels (e.g. stereo, quadraphonic, etc.)
- **Digital processing**
  - Reshape impulse response to simulate a different room
  - Move perceived location from which sound comes
  - Locate speaker in 3D space using microphone arrays
  - Cover missing samples
  - Mix multiple signals (i.e. conference)
  - Echo cancellation

18

Interactive Time Constraints



- **Maximum time to hear own voice: *100 msec***
- **Maximum round-trip time: *300 msec***

19

Importance of Sound

- **Passive viewing (e.g. film, video, etc.)**
  - Very sensitive to sound breaks
  - Visual channel more important (ask film makers!)
  - Tolerate occasional frame drops
- **Video conferencing**
  - Sound channel is more important
  - Visual channel still conveys information
  - Some people report that video teleconference users turn off video
  - Need to create 3D space and locate remote participants in it

20

Producing High Quality Audio

- **Eliminate background noise**
  - Directional microphone gives more control
  - Deaden the room in which you are recording
  - Some audio systems will cancel wind noise
- **One microphone per speaker**
- **Keep the sound levels balanced**
- **Sweeten sound track with interesting sound effects**

21

Summary

- **Some people argue that sound is easy and video is hard because data rates are lower**
  - Not true ⇒ audio is every bit as hard as video, just different!
- **Computer Scientists will learn about audio and video just as we learned about printing with the introduction of desktop publishing**

22

Audio Compression

- **Traditional lossless compression methods (Huffman, LZW, etc.) usually don't work well on audio compression (the same reason as in image compression).**
- **Temporal masking and frequency masking are natural hearing experience.**

23

Simple Audio Compression Methods

- **Silence Compression - detect the "silence", similar to run-length coding**
- **Adaptive Differential Pulse Code Modulation (ADPCM) e.g., in CCITT G.721 -- 16 or 32 Kbits/sec.**
  - Encode the difference between two or more consecutive signals; the difference is then quantized --> *hence the loss*
  - Adaptive quantization
  - It is necessary to predict where the waveform is headed
  - Apple has proprietary scheme called ACE/MACE. A Lossy scheme that tries to predict where wave will go in next sample. Gives about 2:1 compression.
- **Linear Predictive Coding (LPC) fits signal to speech model and then transmits parameters of model. It sounds like a computer talking, 2.4 kbits/sec.**
- **Code Excited Linear Predictor (CELP) does LPC, but also transmits error term --> audio conferencing quality at 4.8 kbits/sec.**

24

Psychoacoustics

Human hearing and voice

- **Frequency range is about 20 Hz to 20 kHz, most sensitive at 2 to 4 KHz.**
- **Dynamic range (quietest to loudest) is about 96 dB**
- **Normal voice range is about 500 Hz to 2 kHz**
  - Low frequencies are vowels and bass
  - High frequencies are consonants

25

Psychoacoustics

Critical Bands

- **Human auditory system has a limited, frequency-dependent resolution.**
- **The perceptually uniform measure of frequency can be expressed in terms of the width of the *Critical Bands*.**
  - It is less than 100 Hz at the lowest audible frequencies, and more than 4 kHz at the high end.
  - Altogether, the audio frequency range can be partitioned into 25 critical bands.

26

Psychoacoustics

- **A new unit for frequency: *bark* (after Barkhausen) is introduced:**

1 Bark = width of one *critical band* (arf! arf!)

- **For frequency < 500 Hz,**

  it converts to *freq* / 100 **Bark,**
- **For frequency > 500 Hz,**

  it is $9 + 4 \cdot \log_2(freq/1000)$ **Bark.**
- **The size of the critical band is just under a tempered minor third at frequencies above about 500 Hz. The size of the critical band gradually grows larger at lower frequencies, becoming nearly an octave at 100 Hz, which explains why similar intervals can sound clear and harmonious at higher pitches but muddy or inharmonious at lower pitches.**

27

Psychoacoustic Model

How sensitive is human hearing?

To answer this question we look at the following concepts:

- Threshold of hearing

  Describes the notion of "quietness"
- Frequency Masking

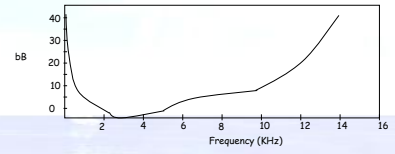
  A component (at a particular frequency) masks components at neighboring frequencies. Such masking may be partial.
- Temporal Masking

  When two tones (samples) are played closed together in time, one can mask the other.

28

Threshold of hearing

Experiment: ***Put a person in a quiet room. Raise level of 1 kHz tone until just barely audible. Vary the frequency and plot***

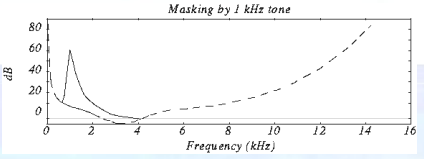


- The ear is most sensitive to frequencies between 1 and 5 kHz, where we can actually hear signals below 0 dB.
- Two tones of equal power and different frequencies will not be equally loud.
- Sensitivity decreases at low and high frequencies.

29

Frequency Masking

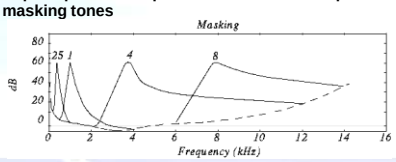
Experiment: ***Play 1 kHz tone (masking tone) at fixed level (60 dB). Play test tone at a different level (e.g., 1.1 kHz), and raise level until just distinguishable. Vary the frequency of the test tone and plot the threshold when it becomes audible:***



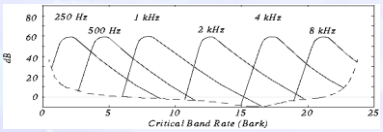
30

Frequency Masking (Contd.)

- **Repeat previous experiment for various frequencies of masking tones**



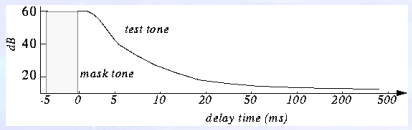
- **Frequency Masking on critical band scale:**



31

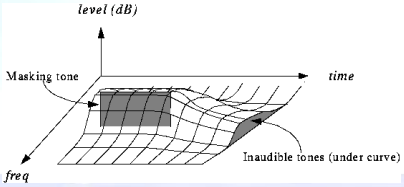
Temporal Masking

- ***If we hear a loud sound, and then it stops, it takes a little while until we can hear a soft tone nearby (in frequency).***
- ***Experiment:***
  - *Play 1 kHz masking tone at 60 dB, plus a test tone at 1.1 kHz at 40 dB. Test tone can't be heard (it's masked).*
  - *Stop masking tone, then stop test tone after a short delay.*
  - *Adjust delay time to the shortest time when test tone can be heard (e.g., 5 ms).*
  - *Repeat with different level of the test tone and plot:*



32

Net effect of masking:



Summary:

- ***If we have a loud tone at, say, 1 kHz, then nearby quieter tones are masked.***
- ***Best compared on critical band scale -- range of masking is about 1 critical band***
- ***Two factors for masking -- frequency masking and temporal masking***
- ***Question: How to use this for compression?***

33

MPEG Audio

Facts

- **The two most common advanced (beyond simple ADPCM) techniques for audio coding are:**
  - Sub-Band Coding (SBC) based
  - Adaptive Transform Coding based
- **MPEG audio coding is comprised of three independent layers. Each layer is a self-contained SBC coder with its own time-frequency mapping, psychoacoustic model, and quantizer.**
  - Layer I: Uses sub-band coding
  - Layer II: Uses sub-band coding (longer frames, more compression)
  - Layer III: Uses both sub-band coding and transform coding.
- **MPEG-1 Audio is intended to take a PCM audio signal sampled at a rate of 32, 44.1 or 48 kHz, and encode it at a bit rate of 32 to 192 kbps per audio channel (depending on layer).**

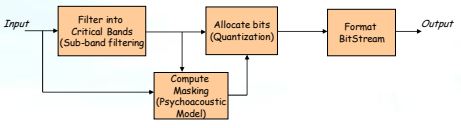
34

More Facts

- **MPEG-1: Bitrate of 1.5 Mbits/sec for audio and video About 1.2 Mbits/sec for video, 0.3 Mbits/sec for audio**
  - (Uncompressed CD audio is 44,100 samples/sec * 16 bits/sample * 2 channels > 1.4 Mbits/sec)
- **Compression factor ranging from 2.7 to 24.**
- **With Compression rate 6:1 (16 bits stereo sampled at 48 KHz is reduced to 256 kbits/sec)**
  - Under optimal listening conditions, expert listeners could not distinguish between coded and original audio clips.
- **Supports one or two audio channels in one of the four modes:**
  1. Monophonic -- single audio channel
  2. Dual-monophonic -- two independent channels, e.g., English and French
  3. Stereo -- for stereo channels that share bits, but not using Joint-stereo coding
  4. Joint-stereo -- takes advantage of the correlations between stereo channels

35

MPEG Coding Algorithm



1. **Use convolution filters to divide the audio signal (e.g., 48 kHz sound) into 32 frequency sub-bands. (*sub-band filtering*)**
2. **Determine amount of masking for each band caused by nearby band using the *psychoacoustic model* .**
3. **If the power in a band is below the masking threshold, don't encode it.**
4. **Otherwise, determine number of bits needed to represent the coefficient such that, the noise introduced by quantization is below the masking effect (One fewer bit of quantization introduces about 6 dB of noise).**
5. **Format bitstream**

36

Masking and Quantization (Example)

- **Say, performing the sub-band filtering step on the input results in the following values (for demonstration, we are only looking at the first 16 of the 32 bands):**

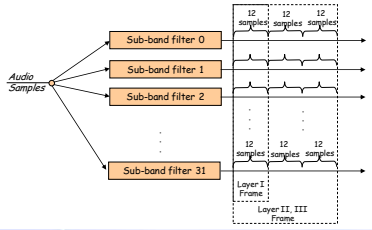
Band	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Level	0	8	12	10	6	2	10	60	35	20	15	2	3	5	3	1

- The 60dB level of the 8th band gives a masking of 12 dB in the 7th band, 15dB in the 9th. (according to the Psychoacoustic model)
- The level in 7th band is 10 dB (< 12 dB), so ignore it.
- The level in 9th band is 35 dB (> 15 dB), so send it.
  - We only send the amount above the masking level
  - Therefore, instead of using 6 bits to encode it, we can use 4 bits -- a saving of 2 bits (= 12 dB).
    - "determine number of bits needed to represent the coefficient such that, the noise introduced by quantization is below the masking effect" [noise introduced = 12bB; masking = 15 dB]

37

MPEG Coding Specifics

- **MPEG defines 3 layers for audio. Basic model is same, but codec complexity increases with each layer.**
- **Divides data into frames, each of them contains 384 samples, 12 samples from each of the 32 filtered subbands as shown below.:**



38

MPEG Coding Specifics

- **MPEG Layer I**
  - Filter is applied one frame (12x32 = 384 samples) at a time. At 48 kHz, each frame carries 8ms of sound.
  - Uses a 512-point FFT to get detailed spectral information about the signal. (sub-band filter). Uses equal frequency spread per band.
  - Psychoacoustic model only uses frequency masking.
  - Typical applications: Digital recording on tapes, hard disks, or magneto-optical disks, which can tolerate the high bit rate.
  - Highest quality is achieved with a bit rate of 384k bps.
- **MPEG Layer II**
  - Use three frames in filter (before, current, next, a total of 1152 samples). At 48 kHz, each frame carries 24 ms of sound.
  - Models a little bit of the temporal masking.
  - Uses a 1024-point FFT for greater frequency resolution. Uses equal frequency spread per band.
  - Highest quality is achieved with a bit rate of 256k bps.
  - Typical applications: Audio Broadcasting, Television, Consumer and Professional Recording, and Multimedia.

39

MPEG Coding Specifics

- **MPEG Layer III**
  - Better critical band filter is used
  - Uses non-equal frequency bands
  - Psychoacoustic model includes temporal masking effects, takes into account stereo redundancy, and uses Huffman coder.

Stereo Redundancy Coding:

- Intensity stereo coding -- at upper-frequency sub-bands, encode summed signals instead of independent signals from left and right channels.
- Middle/Side (MS) stereo coding -- encode middle (sum of left and right) and side (difference of left and right) channels.

40

Effectiveness of MPEG Audio

Layer	Target bit-rate	Ratio	Quality* at 64 kbps	Quality at 128 kbps
Layer I	**192 kbps**	**4:1**	--	--
Layer II	**128 kbps**	**6:1**	**2.1 to 2.6**	**4+**
Layer III	**64 kbps**	**12:1**	**3.6 to 3.8**	**4+**

Quality factor:

- 5 – perfect
- 4 - just noticeable
- 3 - slightly annoying
- 2 – annoying
- 1 - very annoying

41