Audio Fundamentals, Compression Techniques & Standards

Instructor: Hamid R. Rabiee
Spring 2013
Outline

✧ Audio Fundamentals
  ✧ Sampling, digitization, quantization
  ✧ μ-law and A-law

✧ Sound Compression
  ✧ PCM, DPCM, ADPCM, CELP, LPC

✧ Audio Compression

✧ MPEG Audio Codec

✧ Perceptual Coding

✧ MPEG1
  ✧ Layer1
  ✧ Layer2
  ✧ Layer3(MP3)
Sound, Sound Wave, Acoustics

- **Sound:** A continuous wave that travels through a medium
- **Sound wave:** Energy causes disturbance in a medium, made of pressure differences (measures pressure level at a location)
- **Acoustics:** The study of sound (generation, transmission, and reception of sound waves)
- **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
The Characteristics of Sound Waves

✧ **Frequency**
- The rate at which sound is measured
- Number of cycles per second or Hertz (Hz)
- Determines the pitch of the sound as heard by our ears
- The higher frequency, the clearer and sharper the sound => the higher pitch of sound

✧ **Amplitude**
- Sound’s intensity or loudness
- The louder the sound, the larger the amplitude.
- All sounds have a duration and successive musical sounds is called rhythm
The Characteristics of Sound Waves (cont.)
Audio Digitization

✧ To get audio or video into a computer, we must digitize it
✧ Audio signals are continuous in time and amplitude
✧ Audio signal must be digitized in both time and amplitude to be represented in binary form.

Voice

\[
\text{Sinusoid wave}
\]

Analog-Digital Converter (ADC)

10010010000100
How Digitized Audio Data

To decide how to digitize audio data we need to answer the following questions:

✧ What is the sampling rate?
✧ How finely is the data to be quantized, and is quantization uniform?
✧ How is audio data formatted? (File format)
Human Perception

✧ **Speech is a complex waveform**
  ✧ Vowels (a,i,u,e,o) and bass sounds are low frequencies
  ✧ Consonants (s,sh,k,t,...) 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**
  ✧ Temporal & Frequency masking
Audio Digitization

✧ **Step1. Sampling**

divide the horizontal axis (time dimension) into discrete pieces. Uniform sampling is ubiquitous (everywhere at once).
Nyquist Theorem

- Suppose we are sampling a waveform. How often do we need to sample it to figure out its frequency?
- Signals can be decomposed into a sum of sinusoids.
- The Nyquist theorem states how frequently we must sample in time to be able to recover the original sound.

- The **Nyquist Theorem**, also known as the sampling theorem, is a principle that engineers follow in the digitization of analog signals.

- For Analog-to-Digital conversion (ADC) to result in a faithful reproduction of the signal, slices, called *samples*, of the analog waveform must be taken frequently.

- *The number of samples per second* is called the *sampling rate* or *sampling frequency*. 
Nyquist Theorem (cont.)

黝 Nyquist rate

It can be proven that a bandwidth-limited signal can be fully reconstructed from its samples, if the sampling rate is at least twice the highest frequency in the signal ($f_{\text{max}}$).
**Quantization**

- **Step 2. Quantization** — Converts actual amplitude sample values (usually voltage measurements) into an integer approximation.
- Tradeoff between number of bits required and error
  - Human perception limitations affect allowable error
  - Specific application affects allowable error
- Once samples have been captured (Discrete in time by sampling – Nyquist), they must be made discrete in amplitude (Discrete in amplitude by quantization).
- Two approaches to quantization
  - Rounding the sample to the closest integer.
    - (e.g. round 3.14 to 3)
  - Create a Quantizer table that generates a staircase pattern of values based on a step size.
Reconstruction

✧ Analog-to-Digital Converter (ADC) provides the sampled and quantized binary code.

✧ Digital-to-Analog Converter (DAC) converts the quantized binary code back into an *approximation* of the analog signal (reverses the quantization process) by clocking the code to the same sample rate as the ADC conversion.
Quantization Error (Noise)

- After quantization, some information is lost
- Errors (noise) introduced
  - The difference between the original sample value and the rounded value is called the *quantization error*
- A *Signal to Noise Ratio* (SNR) is the ratio of the relative sizes of the signal values and the errors.
  - The higher the SNR, the smaller the average error is with respect to the signal value, and the better the fidelity.

Quantization is only an approximation.
Quantization (An Example)

Consecutive samples always differ by ±1

Digital Media Lab - Sharif University of Technology
μ-law, A-law

- Non-uniform quantizers: Difficult to make, Expensive.
- Solution: Companding => Uniform Q. => Expanding

\[ x[n] = \text{speech /song/} \]
\[ y[n] = C(x[n]) \]
Companded Signal

Segment of \( x[n] \)

Close View of the Signal

Segment of \( y[n] \)
Companded Signal
\[ y(n) = F(x(n)) = X_{\text{max}} \frac{\log \left| 1 + \mu \frac{|x(n)|}{X_{\text{max}}} \right|}{\log[1 + \mu]} \cdot \text{sign}(x(n)) \]
µ-law vs. A-law

- 8-bit µ-law used in US for telephony
- ITU Recommendation G.711
- 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.
- 8000 samples/sec * 8bits per sample = 64Kb/s data rate
AUDIO & SOUND COMPRESSION TECHNIQUES
Quantization and transformation of data are collectively known as coding of the data.

- For audio, the μ-law technique for compounding audio signals is usually combined with an algorithm that exploits the temporal redundancy present in audio signals.

- Differences in signals between the present and a past time can reduce the size of signal values and also concentrate the histogram of pixel values (differences, now) into a much smaller range.

- The result of reducing the variance of values is that lossless compression methods produce a bit stream with shorter bit lengths for more likely values.
In general, producing quantized sampled output for audio is called PCM (Pulse Code Modulation). The differences version is called DPCM (and a crude but efficient variant is called DM). The adaptive version is called ADPCM.
Compression

Every compression scheme has three stages:

✧ The input data is transformed to a new representation that is easier or more efficient to compress.

✧ We may introduce loss of information.
  ✧ (Quantization is the main lossy step).

✧ Coding.
  ✧ Assign a codeword (thus forming a binary bit stream) to each output level or symbol.
Sound Compression

✧ Some techniques for sound compression:
  ✧ PCM - send every sample
  ✧ DPCM - send differences between samples
  ✧ ADPCM - send differences, but adapt how we code them
  ✧ LPC - linear model of speech formation
  ✧ CELP - use LPC as base, but also use some bits to code corrections for the things LPC gets wrong.

✧ The techniques that code received sound signals
  ✧ PCM, DPCM, ADPCM

✧ The techniques that parameterize the sound model (model-based coding)
  ✧ LPC, CELP
Pulse-code Modulation (PCM)

✧ μ-law and a-law PCM have already reduced the data sent.
✧ However, each sample is still independently encoded.
✧ In reality, samples are correlated.
✧ Can utilize this correlation to reduce the data sent.

Diagram: PCM signal encoding and decoding.
Pulse-code Modulation (PCM)

✧ Digital Representation of an Analog Signal
✧ Sampling and Quantization

Parameters:
✧ Sampling Rate (Samples per Second)
✧ Quantization Levels (Bits per Sample)

![Original analog signal and its corresponding PCM signals.](image1)
![Decoded staircase signal.](image2)
![Reconstructed signal after low-pass filtering.](image3)
Differential PCM (DPCM)

- Makes a simple prediction of the next sample, based on weighted previous $n$ samples.
- Normally the difference between samples is relatively small and can be coded with less than 8 bits.
  - Typically use 6 bits for difference, rather than 8 bits for absolute value.
- Compression is lossy, as not all differences can be coded.
Adaptive DPCM (ADPCM)

- Different Coding techniques in comparison with DPCM

- 2 Methods
  - Adaptive Quantization
    - Quantization levels are adaptive, based on the content of the audio.
  - Adaptive Prediction

- Receiver runs same prediction algorithm and adaptive quantization levels to reconstruct speech.
Model-based Coding

- PCM, DPCM and ADPCM directly code the received audio signal.
- An alternative approach is to build a parameterized model of the sound source (i.e., Human voice).
- For each time slice (e.g. 20ms):
  - Analyze the audio signal to determine how the signal was produced.
  - Determine the model parameters that fit.
  - Send the model parameters.
- At the receiver, synthesize the voice from the model and received parameters.
Speech formation

- Voiced sounds: series of pulses of air as larynx opens and closes. Basic tone then shaped by changing resonance of vocal tract.
- Unvoiced sounds: larynx held open, turbulent noise made in mouth.
LPC

- Introduced in 1960s.
- Low-bit rate encoder:
  - 1.2Kb/s - 4Kb/s
- Sounds very synthetic
  - Basic LPC mostly used where bitrate really matters (eg in military applications)
  - Most modern voice codecs (eg GSM) are based on enhanced LPC encoders.
LPC

✧ Digitize signal, and split into segments (eg 20ms)

✧ For each segment, determine:
  
  ✧ Pitch of the signal (i.e. basic formant frequency)
  
  ✧ Loudness of the signal.
  
  ✧ Whether sound is voiced or unvoiced
    
    ✧ Voiced: vowels, “m”, “v”, “l”
    
    ✧ Unvoiced: “f”, “s”
  
  ✧ Vocal tract excitation parameters (LPC coefficients)
Limitations of LPC Model

- LPC linear predictor is very simple.
- For this to work, the vocal tract “tube” must not have any side branches (these would require a more complex model).
  - OK for vowels (tube is a reasonable model)
  - For nasal sounds, nose cavity forms a side branch.
- In practice this is ignored in pure LPC.
  - More complex codecs attempt to code the residue signal, which helps correct this.
Code Excited Linear Prediction (CELP)

- Goal is to efficiently encode the residue signal, improving speech quality over LPC, but without increasing the bit rate too much.
- CELP codecs use a codebook of typical residue values.
  - Analyzer compares residue to codebook values.
  - Chooses value which is closest.
  - Sends that value.
- Receiver looks up the code in its codebook, retrieves the residue, and uses this to excite the LPC formant filter.
Problem is that codebook would require different residue values for every possible voice pitch.

- Codebook search would be slow, and code would require a lot of bits to send.

One solution is to have two codebooks.

- One fixed by codec designers, just large enough to represent one pitch period of residue.
- One dynamically filled in with copies of the previous residue delayed by various amounts (delay provides the pitch)

CELP algorithm using these techniques can provide pretty good quality at 4.8Kb/s.
Enhanced LPC Usage

- GSM (Group Special Mobile)
  - Residual Pulse Excited LPC
  - 13Kb/s

- LD-CELP
  - Low-delay Code-Excited Linear Prediction (G.728)
  - 16Kb/s

- CS-ACELP
  - Conjugate Structure Algebraic CELP (G.729)
  - 8Kb/s

- MP-MLQ
  - Multi-Pulse Maximum Likelihood Quantization (G.723.1)
  - 6.3Kb/s
Audio Compression

✧ LPC-based codec model the sound source to achieve good compression.
  ✧ Works well for voice.
  ✧ Terrible for music.

✧ What if you can’t model the source?
  ✧ Model the limitations of the human ear.
  ✧ Not all sounds in the sampled audio can actually be heard.
  ✧ Analyze the audio and send only the sounds that can be heard.
  ✧ Quantize more coarsely where noise will be less audible.

✧ Audio vs. Speech
  ✧ Higher quality requirement for audio
  ✧ Wider frequency range of audio Sound Signal
Psychoacoustics Model

- Dynamic range is ratio of maximum signal amplitude to minimum signal amplitude (measured in decibels).
  - \[ D = 20 \log \left( \frac{A_{\text{max}}}{A_{\text{min}}} \right) \text{ dB} \]
  - Human hearing has dynamic range of \(~ 96\text{dB}\)

- Sensitivity of the ear is dependent on frequency.
  - Most sensitive in range of \(2\text{-}5\text{KHz}\)

![Threshold in Quiet graph](image)

Sensitivity of human ears

Most sensitive at \((2 \text{ kHz}, 5\text{kHz})\)
Psychoacoustics Model

✧ Amplitude Sensitivity

✧ Frequencies only heard if they exceed a sensitivity threshold:

Source: Halsall, p184
Psychoacoustics Model

✧ **Frequency Masking**

✧ The sensitivity threshold curve is distorted by the presence of loud sounds.

✧ Frequencies just above and below the frequency of a loud sound need to be louder than the normal minimum amplitude before they can be heard.
Psychoacoustics Model

✧ Temporal masking

✧ If we hear a loud sound, then it stops, it takes a little while until we can hear a soft tone nearby

✧ After hearing a loud sound, the ear is deaf to quieter sounds in the same frequency range for a short time.

![Diagram showing temporal masking]
Principle of Audio Compression

✧ Audio compression – Perceptual coding
  ✧ Take advantage of psychoacoustics model
  ✧ Distinguish between the signal of different sensitivity to human ears
    ✧ Signal of high sensitivity – more bits allocated for coding
    ✧ Signal of low sensitivity – less bits allocated for coding

✧ Exploit the frequency masking
  ✧ Don’t encode the masked signal (range of masking is 1 critical band)

✧ Exploit the temporal masking
  ✧ Don’t encode the masked signal
AUDIO CODING STANDARDS
Audio Coding 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)
- LPC (FIPS-1015) - Linear Predictive Coding (2.4kbs)
- CELP (FIPS-1016) - Code excited LPC (4.8kbs, 4bits/sample)
- G.729 - CS-ACELP (8kbs)
- MPEG1/MPEG2, AC3 - (16-384kbs) mono, stereo, and 5+1 channels
MPEG Audio Codec

- MPEG (Motion Picture Expert Group) and ISO (International Standard Organization) have published several standards about digital audio coding.
  - MPEG-1 Layer 1, 2, and 3 (MP3)
  - MPEG2 AAC
  - MPEG4 AAC and TwinVQ
- Exploits the psychoacoustic models.
  - Frequency masking is always utilized
  - More complex forms of MPEG also employ temporal masking
- They have been widely used in consumer electronics, digital audio broadcasting, DVD and movies etc.
Advantages of MPEG approach

✧ Complex psychoacoustic modeling only in coding phase
  ✧ Desirable for real time (Hardware or software) decompression
  ✧ Essential for broadcast purposes.
  ✧ Decompression is independent of the psychoacoustic models used
  ✧ Different models can be used
  ✧ If there is enough bandwidth no models at all.

✧ Although (data) lossy, MPEG claims to be perceptually lossless:
  ✧ Human tests (part of standard development), Expert listeners.
  ✧ Under Optimal listening conditions no statistically distinguishable difference between original and MPEG.
MPEG Audio Codec

- **Procedure of MPEG audio coding**
  - The audio signal is first sampled and quantized using PCM
  - The PCM samples are then divided up into 32 frequency sub-bands (sub-band filtering using DFT)
  - Use psychoacoustics model in bit allocation
    - If the amplitude of signal in band is below the masking threshold, don’t encode
    - Otherwise, allocate bits based on the sensitivity of the signal

- **Multiplex the output of the 32 bands into one bit stream**

![Diagram of the MPEG audio codec process](image)
MPEG defines 3 levels of processing layers for audio:

- Level 1 is the basic mode,
- Levels 2 and 3 more advance (use temporal masking).
- Level 3 is the most common form for audio files on the Web
  - Our beloved MP3 files
  - Strictly speaking these files should be called MPEG-1 level 3 files.

Each level:

- Increasing levels of sophistication
- Greater compression ratios.
- Greater computation expense (but mainly at the coder side)
MPEG1 Audio Layer 1

- MPEG 1 audio allows sampling rate at 44.1, 48, 32, 22.05, 24 and 16KHz.
- MPEG1 Divides data into frames
  - Each of them contains 384 samples,
  - 12 samples from each of the 32 filtered sub-bands.
- Psychoacoustic model only uses frequency masking.
- Optional Cyclic Redundancy Code (CRC) error checking.

```
Audio samples → Filtering And downsampling → 12 samples
                  ⏫                       ⏫
                  12 samples           12 samples
                  ⏫                       ⏫
Normalize By scale factor → Perceptual coder
```
MPEG1 Audio Layer 2

- Layer 2 is very similar to Layer 1, but codes audio data in larger groups:
  - Use three frames in filter: before, current, next, a total of 1152 samples.
  - This models a little bit of the temporal masking.
- Imposes some restrictions on bit allocation in middle and high sub-bands.
- More compact coding of scale factors and quantized samples.
- Better audio quality due to saving bits here so more bits can be used in quantized sub-band values
MPEG Audio Layer 3: MP3

- Targeted at bit rates of 64 kbits/sec per channel.
- Much more complex approach.
- Psychoacoustic model includes *temporal* masking effects,
- Better critical band filter is used (non-equal frequencies)
- Uses a modified DCT (MDCT) for lossless sub-band transformation.
- Greater frequency resolution accounts for poorer time resolution
- Uses Huffman coding on quantized samples for better compression.
MPEG Audio Layer 3: MP3

PCM audio signal

Filter bank: 32 subbands

1,024-point FFT

Psychoacoustic model

M-DCT

Nonuniform quantization

Side-information coding

Coded audio signal

Bitstream formatting

Huffman coding
MPEG Audio

<table>
<thead>
<tr>
<th>Layer</th>
<th>Application</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Digital Audio Cassette</td>
<td>Hi-fi quality at 192Kb/s per channel</td>
</tr>
<tr>
<td>2</td>
<td>Digital Audio and Digital Video Broadcasting</td>
<td>Near CD-quality at 128Kb/s per channel</td>
</tr>
<tr>
<td>3</td>
<td>CD-quality audio over low bitrate channels</td>
<td>CD-quality at 64Kb/s per channel</td>
</tr>
</tbody>
</table>