Audio Compression & Coding

Lecture 7:

1. Information, compression & quantization
2. Speech coding
3. Wide bandwidth audio coding

EE E6820: Speech & Audio Processing & Recognition

Spring 2006
Columbia University Dept. of Electrical Engineering

Dan Ellis <dpwe@ee.columbia.edu>
http://www.ee.columbia.edu/~dpwe/E6820/
Compression & Quantization

- How big is audio data? What is the bitrate?
  - \( F \times C \times B \) bits/second (e.g. 8000 or 44100 frames/second)

- How to reduce?
  - Lower sampling rate → less bandwidth (multified)
  - Lower channel count → no stereo image
  - Lower sample size → quantization noise

- Or: use data compression

```
bits / frame frames / sec 8000 8 \times \frac{8}{32} 44100 CD Audio 1.4 Mbps
Telephony 64 Kbps ≤ \frac{F \times C \times B}{8} ≤ 1.3 Kbps
Mobile
```

How big is audio data? What is the bitrate?
Data compression: Redundancy vs. Irrelevance

- Two main principles in compression:
  - Redundant info is implicit in remainder
  - Irrelevant info is unique but unnecessary

Examples:
- Redundant info is implicit in remainder
  - E.g., recording a microphone signal bandlimited at 20 kHz, but sample at 80 kHz.
  - In a bandlimited signal, the red samples can be exactly recovered by interpolation.

- Irrelevant info is unique but unnecessary
  - E.g., recording a microphone signal at 80 kHz sampling rate.

Graphical representation:
- The blue samples can be exactly recovered by interpolating the red samples.
- Can recover every other sample by interpolation.
- But sample at 80 kHz, not 20 kHz.

Redundancy vs. Irrelevance

- Redundant info is implicit in remainder
- Irrelevant info is unique but unnecessary
- Two main principles in compression:
Irrelevant data in audio coding

• For coding of audio signals, irrelevant means perceptually insignificant.

• Problem: separating salient & irrelevant

  * Dynamic properties - hard to characterize
  * Sinusoidal phase, detail of noise structure
  * 20 KHz bandwidth, 100 dB intensity

Reflect limits of human sensitivity:

• 16 bit linear samples for ~ 96 db peak SNR
• 44 KHz sampling for 20 KHz bandwidth

Compact disc standard is adequate:

• An empirical property

Irrelevant means perceptually insignificant

For coding of audio signals,
Quantization

- Represent waveform with discrete levels

\[ \frac{1}{2} \int_{-D/2}^{D/2} p(e[n]) \sigma_e^2 \, de = \frac{1}{2} \int_{-D/2}^{D/2} p(e[n]) e^2 \, de \]

\[ e[n] \sim \text{uncorrelated, uniform white noise} \]

\[ [u][e] + [[u][x]] \sigma = [u][x] \]

\[ e[n] = x[n] - Q[x[n]] \]

- Equivalent to adding error \( e[n] \)
Quantization noise (Q-noise)

- Uncorrelated noise has flat spectrum
- With a $B$ bit word and a quantization step $D$
- Max signal range $x = (2^B - 1) \cdot D$
- Quantization noise $e = \frac{D}{2} \cdot \frac{2^B}{2}$
- Best signal-to-noise ratio (power) $\frac{E_x}{E} = \frac{1}{\frac{e^2}{2}}$
- as dB $\approx 20 \cdot 10 \log_{10} B \cdot 6 \cdot B$ dB

```
\text{level / dB}
\begin{array}{cccccc}
0 & 1000 & 2000 & 3000 & 4000 & 5000 \\
\text{freq / Hz} & \text{level / dB}
\end{array}
```

Quantized at 7 bits

```
```

```
\begin{array}{cccccc}
\text{freq / Hz} & \text{level / dB}
\end{array}
```

```
\begin{array}{cccccc}
0 & 1000 & 2000 & 3000 & 4000 & 5000 \\
\text{freq / Hz} & \text{level / dB}
\end{array}
```
Redundant Information

- Problem: separating unique & redundant
- White noise sequence has no redundancy
- If $x[n] = x[n-1] + v[n]$
- $x[n]$ has a greater amplitude range → more bits
- $u[n]$ + $[1 - u]x = [u]x$
- E.g. 'white noise' sequence has no redundancy

Redundancy removal is lossless

Signal correlation implies redundant

Redundant Information
Optimal coding

- variable-length tokens ↔ Huffman coding
- nonuniform quantization for equiprobable tokens
- transform signal to have uniform pdf

\[ \text{Information in bits} = -\log_2(\text{probability}) \]

\begin{align*}
\text{A} & \quad \text{is expected} \\
\text{B} & \quad \text{is big news}
\end{align*}

\begin{align*}
\text{AAABBBBAABBBAAABB} & \quad \text{p(A)} = 0.5 \\
\text{BBABABABABABABAB} & \quad \text{p(B)} = 0.9
\end{align*}
Quantization for optimum bitrate

- Quantization should reflect pdf of signal.
- Quantization for optimum bitrate.
Vector Quantization

- Larger space \{x_1, x_2\} is easier for Huffman
- May help even if values are largely independent

Quantize mutually dependent values in joint space.
• As always, success depends on representation

Compression & Representation

• In right domain, irrelevance is easier to get at
- e.g. STFT to separate magnitude and phase

• Can reduce sampling rate without data loss
- e.g. vocal-tract-shape coefficients

• Appropriately domain may be naturally bandlimited

• As always, success depends on representation
Aside: Coding standards

- RA, WMA, Skype...
- More recent proprietary standards
  - MPEG 4 Synthesized-Natural Hybrid Codec
  - MPEG 2 Advanced Audio Codec
  - MPEG Audio Layers 1, 2, 3
  - MPEG

  - G.729 low delay CELP
  - G.726 ADPCM
  - ITU G.Series

  - FS1016: 4.8 Kbps CELP
  - FS1015e: LPC-10 2.4 Kbps

Federal Standards: Low bit-rate secure voice:

Standardization efforts are important

Coding only useful if recipient knows the code!

Aside: Coding standards
Outline

1. Information, compression & quantization
2. Speech coding
3. Wide bandwidth audio coding
   - CELP & friends
   - General principles
Speech coding

- **Standard voice channel:**
  - analog: 4 kHz slot (~ 40 dB SNR)
  - digital: 64 Kbps = 8 bit μ-law x 8 kHz

- **How to compress?**
  
  **Redundant**
  - signal assumed to be a single voice, not any possible waveform
  
  **Irrelevant**
  - need code only enough for intelligibility, speaker identification (c/w analog channel)

- **Specifically, source-filter decomposition**
  - vocal tract & fund. frequency change slowly

- **Applications:**
  - live communications  - offline storage
Channel Vocoder (1940s-1960s)

- Basic source-filter decomposition
- or: baseband + "filtering"...
  - pitch / noise model
  - Excitation?
  - Downsampling?

- 10-20 bands, perceptually spaced
- Transmit slowly-changing energy in each band
  - Filterbank breaks into spectral bands

Diagram:

- Encoder
  - Input
  - Bandpass(filter 1
  - Smoothing energy
  - Bandpass
  - V/O
  - Pitch
- Decoder
  - Output
  - Bandpass(filter 1
  - Smoothing energy
  - Bandpass
  - V/O
  - Pitch
The classic source-filter model

- Excitation can be represented in many ways.
- Filter parameters are slowly changing.

Compression gains:

\[
\begin{align*}
\text{Decoder} & \\
\text{Encoder} & \\
\text{The classic source-filter model} & \\
\text{LPC encoding} &
\end{align*}
\]

\[
H(z) = \frac{1}{\sum_{i=0}^{N} a_i z^{-i}}
\]
Encoding LP-C filter parameters

- For 'communications quality':
  - 10 LSPs x 3-4 bits / 30 ms = 1.1 Kbps
  - FS1016 (4.8 Kbps):
  - 8 LARs x 3-6 bits / 20 ms = 1.8 Kbps
  - GSM (13 Kbps):

- Bit allocation (filter):
  - LSPs - lovely!
  - guaranteed stable
  - reflection coefficients
  - can't interpolate
  - poor distribution
  - quantization
  - {\{a\}}

- update every 20-30 ms
- ~10th order LPC (up to 5 pole pairs)
- ~8 KHz sampling (4 KHz bandwidth)
  - For 'communications quality':
Line Spectral Pairs (LSPs)

- LSPs encode LPC filter by a set of frequencies
- Excellent for quantization & interpolation
- Definition: zeros of $A(z)$
- $\mathbb{Z}/[(z)d + (z)d] = (z)A$
- \begin{align*}
(z)A = \frac{P(z) + Q(z)}{2} \\
(z)d = P(z) - Q(z)
\end{align*}
- \begin{align*}
A(z) = 0 \rightarrow P(z) = 0 \quad Q(z) = 0 \\
A(z-1) = 0
\end{align*}

\begin{align*}
\angle \{A(z)\} &= \pm \angle \{A(z-1)\} \\
\angle \{P(z)\} &= \angle \{Q(z)\} \\
\angle \{P(z)\} &= \angle \{Q(z)\} = \angle \{A(z)\}
\end{align*}

- Reconstruct $P(z), Q(z)$ etc.
- \begin{align*}
\angle \{P(z)\} &= \angle \{Q(z)\} = \angle \{A(z)\}
\end{align*}
• Excitation already better than raw signal:
  - save several bits/sample, still > 32 Kbps

  - Crude model: U/V flag + pitch period

  - Band-limit then re-extend (RELP)
    - 7 bits / 5 ms = 1.4 Kbps \(\Rightarrow\) LP C10 \(\&\) 2.4 Kbps

  - 16 ms frame boundaries Pitch period values

Original signal
LPC residual
LPC Residual
5000
0
5000
0

Exitation

Exitation already better than raw signal:
Encoding excitation

- Something between full-quality residual (32 Kbps) and pitch parameters (1.4 Kbps)?

- Perceptual weighting discounts peaks.

- Analysis by synthesis loop:
  - Perceptual weighting
  - MSError minimization
  - Control

Excoding excitation

- Analysys by synthesis loop:
  - Filter coefficients
  - Pitch-cycle predictor
  - LPC filter

(32 Kbps) and pitch parameters (1.4 Kbps)

Something between full-quality residual
Multi-Pulse Excitation (MPE-LPC)

- Stylize excitation as $N$ discrete pulses

$$\text{Greedy algorithm places one pulse at a time:}$$

- Cross-correlate $y^n$ and $r^n$

$$\left(\frac{\lambda}{z}\right)V - \left(\frac{\lambda}{z}\right)U =$$

$$\left(\begin{array}{cc}
(\lambda/z)^V & (\lambda/z)^U \\
(\lambda/z)Y & (\lambda/z)X
\end{array}\right) = \text{dec} E$$

- Encode as $N \times (m \times n)$ pairs

<Diagram of multi-pulse excitation and residual signal>

- Stylize excitation as $N$ discrete pulses
• Represent excitation with codebook
e.g. 512 sparse excitation vectors

**FS1016 4.8 Kbps CELP (30ms frame = 144 bits):**

- Linear search for minimum weighted error?
Aside: CELP for nonspeech?

- CELP is sometimes called a 'hybrid' coder: originally based on source-filter voice model.
- CELP residual is waveform coding (no model).
- CELP does not break with multiple voices etc.
- just does the best it can.

LPC filter models vocal tract:
also matches auditory system?
- i.e. the 'source-filter' separation is good for relevance as well as redundancy?

Original (mrZebra-8k)
Outline

1 Information, compression & Quantization

2 Speech coding

3 Wide bandwidth audio coding
   - General principles
   - MPEG-Audio
Wide-Bandwidth Audio Coding

- Goals:
  - transparent coding i.e. no perceptible effect
  - general purpose - handles any signal

Simple approaches (redundancy removal):
- general purpose - handles any signal
- transparent coding i.e. no perceptible effect

Larger compression gains needs irrelevance
- e.g. Shorten - lossless LPC encoding
- as prediction gets smarter, becomes LPC

Adaptive differential PCM (ADPCM)

psychoeacoustic masking
- hide quantization noise with psychoacoustic masking

- Simple approaches (redundancy removal):
Noise shaping

• Plain Q-noise sounds like added white noise
  - actually, not all that disturbing
  - but worst-case for exploiting masking

• Have Q-noise scale
  - μ-law quantization
  - μ(\(x\))

• Transform Q-noise
  - Linear Q-noise

• Or: put Q-noise around peaks in spectrum
  - key to getting benefit of perceptual masking
  - μ-law quantization
  - μ(\(x\))

• Have Q-noise scale
  - linear μ-law

- for some center
- minimum distortion
- amplitude
- gets larger with
- i.e. quantizer step
- with signal level
Subband coding

- Idea: Quantize separately in separate bands
- Subband coding
- \textit{Critical sampling} \iff \frac{1}{M} \text{ of spectrum per band}
- \textit{Q-noise stays within band, gets masked}
- \textit{some aliasing inevitable}
- \textit{Trick is to cancel with alias of adjacent band}
- \textit{Trick is to cancel with alias of adjacent band}

- Bandpass filters
- Reconstruction filters
- Downsample
- Quantize
- Analysis filters
- Encoder
- Decoder

- normalized freq
- gain / dB
- alias energy
Basic idea: Subband coding plus

- Scale factors like LPC envelope?
  - Fixed bitrates 32 - 256 Kbps/chan (1-6 bits/samp)

- MPEG-Audio

• Basic idea: Subband coding plus psychoacoustic masking model
to choose dynamic Q-levels in subbands

- Scale indices: 32 equal bands = 690 Hz bandwidth

- 24 ms / 24 ms frames = 12 / 36 subband samples

- 8 / 24 ms frames = 12 / 36 subband samples

- 8 kHz ÷ 32 equal bands = 690 Hz bandwidth

- Fixed bitrates 32 - 256 Kbps/chan (1-6 bits/samp)
MPEG Psychoacoustic model

- Based on simultaneous masking experiments
- Complex, dynamic sounds not well understood
- Masking level nonlinear in frequency & intensity
- Noise energy masks ~10 dB better than tones
- Masking level nonlinear
- Procedure
- Pick 'tonal peaks' in NB FFT spectrum
- Remaining energy
- Sum all masking & threshold in power & spreading function
- Apply nonlinear 'spreading function'
- Non-tonal energy

Difficulties:

- Based on simultaneous masking experiments

MPEG Psychoacoustic model
MPEG Bit allocation

- Result of psychoacoustical model is maximum tolerable noise per subband

- Bands with no signal above masking curve can be skipped entirely

- Bit allocation procedure (fixed bit rate):
  - Pick channel with worst noise-masker ratio
  - Improve its quantization by one step
  - Repeat while more bits available for this frame

\[
\text{SNR} \sim 6 \cdot B
\]
Transform coder on top of subband coder.

Layer III

MPEG Audio

Power-law quantizer

Fixed Huffman tables optimized for audio data

- Scale factors now in band-blocks
- Tier control e.g. of aliasing, masking
- More redundancy in spectral domain

Blocks of 36 subband time-domain samples become 18 pairs of frequency-domain samples

Filterbank

32 Subbands

FFT

1024 Points

Psycho-
acoustic
Model

Line 575

Huffman
Encoding

Coding of
Side-
information

Distortion
Control Loop

Nonuniform
Quantization

Rate
Control Loop

Bitstream Formatting

Audio Signal

Coded

192 kbit/s

32 kbit/s

External Control

CRC-Check

Bitstream Formatting

128 Kbit/s

Audio Signal

Coded

128 Kbit/s
Adaptive time window:

- Time window relies on temporal masking
  - Single quantization level over 8-24 ms window
  - Time window relies on temporal masking

Nightmare scenario:

- Backward masking saves in most cases

Adaptive switching of time window:
The effects of MP3

- Quantization noise under signal
- Occasional other time-frequency 'holes'
- Chop off high frequency (above 16 KHz)
MP3 & Beyond

• MP3 is 'transparent' at ~ 128 Kbps for stereo
  (11x smaller than 1.4 Mbps CD rate)
  - only decoder is standardized:
    - 11x smaller than 1.4 Mbps CD rate
    - MP3 is transparent at ~ 128 Kbps for stereo
  - MP3 is, computer music language!
  - complete DSP/computer music language!
  - 'synthetic' component of MPEG-4 Audio
  - M4AOL

  - MPEG Audio, LSPs...
  - wide range of component encodings
  - multichannel etc.
  - 30% better (stereo at 96 Kbps?)
  - rebuild of MP3 without backwards compatibility

• MPEG-Audio

  - more flexible models → better encoders

• MPEG-AAC

  - only decoder is standardized:
Summary

• For coding, every bit counts
  - Detailed psychoacoustic models are key
  - Noise shaping, hides quantization noise
  - Wide-band coding
  - CELP residual modeling can go beyond speech
  - LPC modeling is old but good
  - Speech coding
  - Shannon limits
  - Take care over quantization domain & effects

For coding, every bit counts