EE E6820: Speech & Audio Processing & Recognition

Lecture 7: Audio compression and coding

Dan Ellis <dpwe@ee.columbia.edu>
Michael Mandel <mim@ee.columbia.edu>

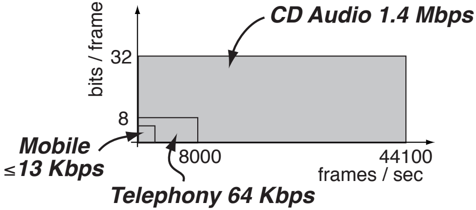
Columbia University Dept. of Electrical Engineering
http://www.ee.columbia.edu/∼dpwe/e6820

March 31, 2009

Outline

Compression & Quantization

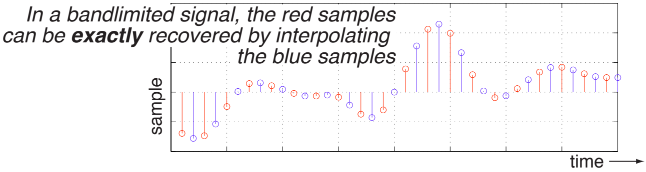
- How big is audio data? What is the bitrate?
  - F_s frames/second (e.g. 8000 or 44100)
    x C samples/frame (e.g. 1 or 2 channels)
    x B bits/sample (e.g. 8 or 16)
  - $\rightarrow$ $F_s \cdot C \cdot B$ bits/second (e.g. 64 Kbps or 1.4 Mbps)



- How to reduce?
  - lower sampling rate $\rightarrow$ less bandwidth (muffled)
  - lower channel count $\rightarrow$ no stereo image
  - lower sample size $\rightarrow$ quantization noise
- Or: use data compression

Data compression:
Redundancy vs. Irrelevance

- Two main principles in compression:
  - remove redundant information
  - remove irrelevant information
- Redundant information is implicit in remainder
  - e.g. signal bandlimited to 20kHz, but sample at 80kHz
  - $\longrightarrow$ can recover every other sample by interpolation:



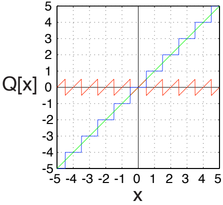
- Irrelevant info is unique but unnecessary
  - e.g. recording a microphone signal at 80 kHz sampling rate

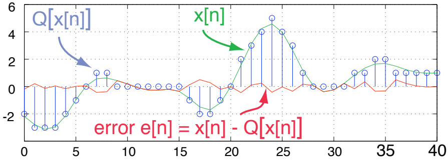
Irrelevant data in audio coding

- For coding of audio signals,
  irrelevant means perceptually insignificant
  - an empirical property
- Compact Disc standard is adequate:
  - 44 kHz sampling for 20 kHz bandwidth
  - 16 bit linear samples for $\sim$ 96 dB peak SNR
- Reflect limits of human sensitivity:
  - 20 kHz bandwidth, 100 dB intensity
  - sinusoid phase, detail of noise structure
  - dynamic properties - hard to characterize
- Problem: separating salient & irrelevant

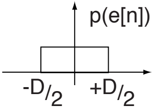
Quantization

- Represent waveform with discrete levels





- Equivalent to adding error $e[n]$:
  $x[n] = Q\left[x[n]\right] + e[n]$
- $e[n] \sim$ uncorrelated, uniform white noise



$\rightarrow$ variance $\sigma_e^2 = \frac{D^2}{12}$

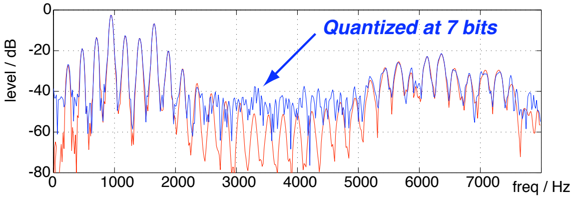
Quantization noise (Q-noise)

- Uncorrelated noise has flat spectrum
- With a B bit word and a quantization step D
  - max signal range $(x) = -(2^{B-1}) \cdot D \ldots (2^{B-1} - 1) \cdot D$
  - quantization noise $(e) = -D/2 \ldots D/2$

$\longrightarrow$ Best signal-to-noise ratio (power)

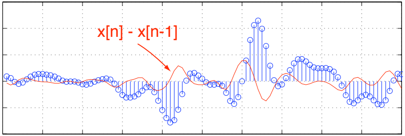
$$
\begin{aligned}
SNR &= E[x^2]/E[e^2] \\
&= (2^B)^2
\end{aligned}
$$

.. or, in dB, $20 \cdot \log_{10} 2 \cdot B \approx 6 \cdot B$ dB



Redundant information

- Redundancy removal is lossless
- Signal correlation implies redundant information
  - e.g. if $x[n] = x[n-1] + v[n]$
    $x[n]$ has a greater amplitude range
    $\rightarrow$ uses more bits than $v[n]$
  - sending $v[n] = x[n] - x[n-1]$ can reduce amplitude, hence bitrate



  - but: 'white noise' sequence has no redundancy ...
- Problem: separating unique & redundant

Optimal coding

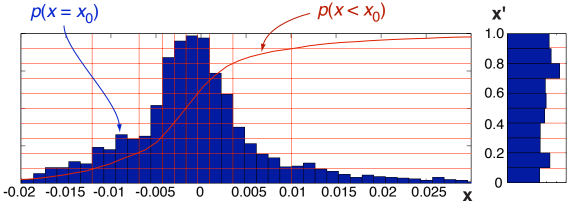
- Shannon information:
  An unlikely occurrence is more 'informative'

$p(\mathrm{A}) = 0.5$ $p(\mathrm{B}) = 0.5$	$p(\mathrm{A}) = 0.9$ $p(\mathrm{B}) = 0.1$
ABBBBAAABBABBABBABB	**AAAAABBAAAAAABAAAAB**
A, **B** equiprobable	**A** is expected; **B** is 'big news'

- Information in bits $I = -\log_2(\mathit{probability})$
  - clearly works when all possibilities equiprobable
- Optimal bitrate $\longrightarrow$ av.token length $=$ entropy $H = E[I]$
  - .. equal-length tokens are equally likely
- How to achieve this?
  - transform signal to have uniform pdf
  - nonuniform quantization for equiprobable tokens
  - variable-length tokens $\longrightarrow$ Huffman coding

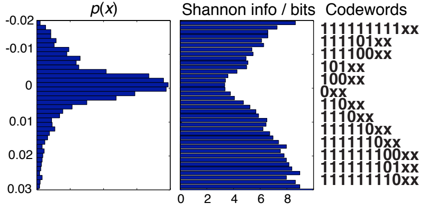
Quantization for optimum bitrate

- Quantization should reflect pdf of signal:



  - cumulative pdf $p(x < x_0)$ maps to uniform x'

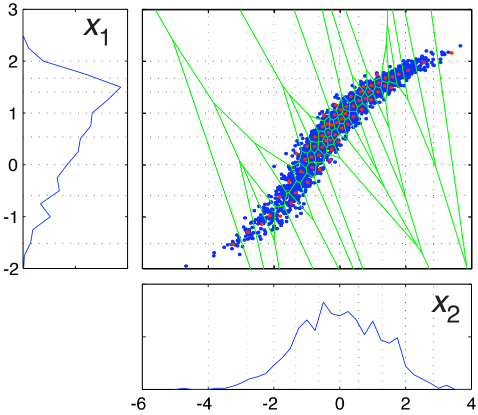
- Or, codeword length per Shannon $\log_2(p(x))$:



  - Huffman coding: tree-structured decoder

Vector Quantization

- Quantize mutually dependent values in joint space:



- May help even if values are largely independent
  - larger space x1,x2 is easier for Huffman

Compression & Representation

- As always, success depends on representation
- Appropriate domain may be 'naturally' bandlimited
  - e.g. vocal-tract-shape coefficients
  - can reduce sampling rate without data loss
- In right domain, irrelevance is easier to 'get at'
  - e.g. STFT to separate magnitude and phase

Aside: Coding standards

- Coding only useful if recipient knows the code!
- Standardization efforts are important
- Federal Standards: Low bit-rate secure voice:
  - FS1015e: LPC-10 2.4 Kbps
  - FS1016: 4.8 Kbps CELP
- ITU G.x series (also H.x for video)
  - G.726 ADPCM
  - G.729 Low delay CELP
- MPEG
  - MPEG-Audio layers 1,2,3 (mp3)
  - MPEG 2 Advanced Audio Codec (AAC)
  - MPEG 4 Synthetic-Natural Hybrid Codec
- More recent 'standards'
  - proprietary: WMA, Skype...
  - Speex ...

Outline

1. Information, Compression & Quantization
2. Speech coding
3. Wide-Bandwidth Audio Coding

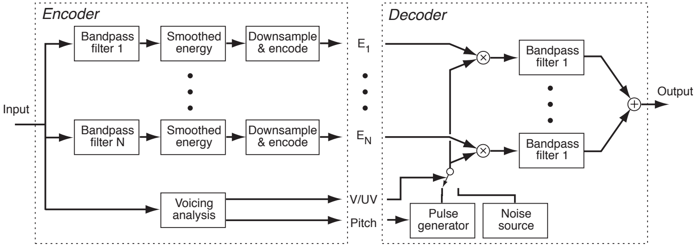
Speech coding

- Standard voice channel:
  - analog: 4 kHz slot ($\sim$ 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 & f_0 change slowly
- Applications:
  - live communications
  - offline storage

Channel Vocoder (1940s-1960s)

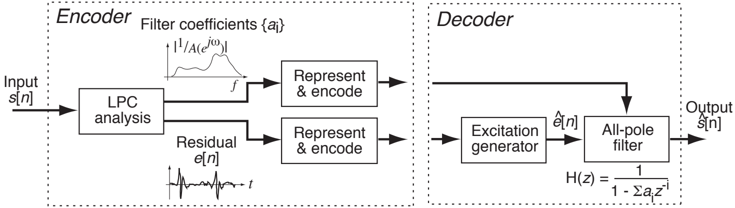
- Basic source-filter decomposition
  - filterbank breaks into spectral bands
  - transmit slowly-changing energy in each band



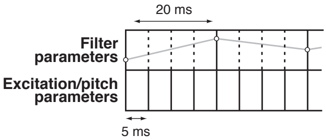
  - 10-20 bands, perceptually spaced
- Downsampling?
- Excitation?
  - pitch / noise model
  - or: baseband + 'flattening'...

LPC encoding

- The classic source-filter model

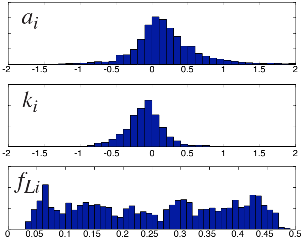


- Compression gains:
  - filter parameters are ∼slowly changing
  - excitation can be represented many ways



Encoding LPC **filter** parameters

- For 'communications quality':
  - 8 kHz sampling (4 kHz bandwidth)
  - $\sim$10th order LPC (up to 5 pole pairs)
  - update every 20-30 ms $\rightarrow$ 300 - 500 param/s
- Representation & quantization



- $\{a_i\}$ - poor distribution, can't interpolate
- reflection coefficients $\{k_i\}$: guaranteed stable
- LSPs - lovely!

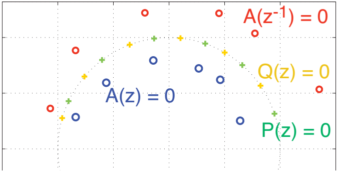
- Bit allocation (filter):
  - GSM (13 kbps): 8 LARs x 3-6 bits / 20 ms = 1.8 Kbps
  - FS1016 (4.8 kbps): 10 LSPs x 3-4 bits / 30 ms = 1.1 Kbps

Line Spectral Pairs (LSPs)

- LSPs encode LPC filter by a set of frequencies
- Excellent for quantization & interpolation
- Definition: zeros of

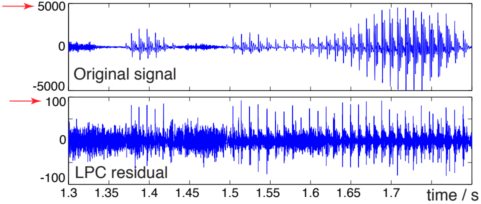
$$P(z) = A(z) + z^{-p-1} \cdot A(z^{-1})$$
$$Q(z) = A(z) - z^{-p-1} \cdot A(z^{-1})$$

  - $z = e^{j\omega} \to z^{-1} = e^{-j\omega} \to |A(z)| = |A(z^{-1})|$ on u.circ.
  - $P(z)$, $Q(z)$ have (interleaved) zeros when $\angle\{A(z)\} = \pm\angle\{z^{-p-1}A(z^{-1})\}$
  - reconstruct $P(z)$, $Q(z) = \prod_i (1 - \zeta_i z^{-1})$ etc.
  - $A(z) = [P(z) + Q(z)]/2$



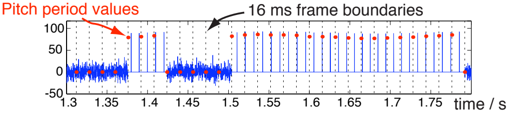
Encoding LPC **excitation**

- Excitation already better than raw signal:



  - save several bits/sample, but still > 32 Kbps

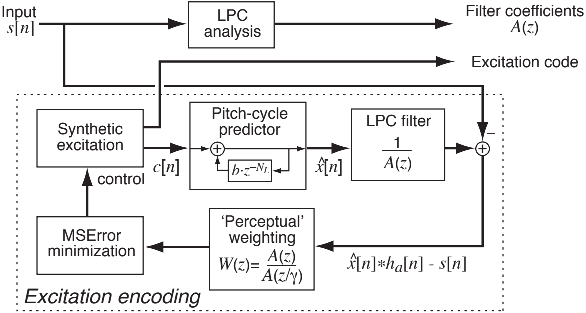
- Crude model: U/V flag + pitch period
  - $\sim$ 7 bits / 5 ms = 1.4 Kbps $\rightarrow$ LPC10 @ 2.4 Kbps



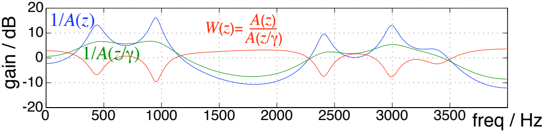
- Band-limit then re-extend (RELP)

Encoding excitation

- Something between full-quality residual (32 Kbps) and pitch parameters (1.4 kbps)?
- 'Analysis by synthesis' loop:

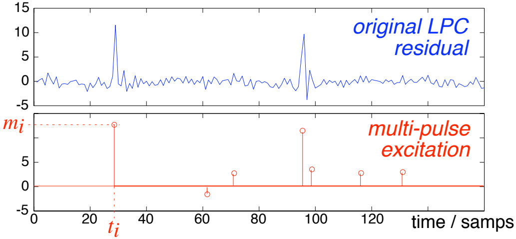


- 'Perceptual' weighting discounts peaks:



Multi-Pulse Excitation (MPE-LPC)

- Stylize excitation as N discrete pulses



  - encode as $N \times (t_i, m_i)$ pairs

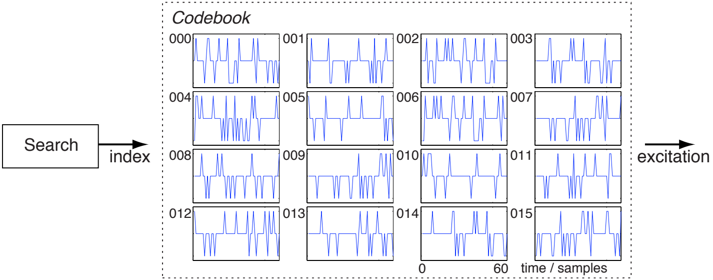
- Greedy algorithm places one pulse at a time:

$$E_{pcp} = \frac{A(z)}{A(z/\gamma)}\left[\frac{X(z)}{A(z)} - S(z)\right]$$
$$= \frac{X(z)}{A(z/\gamma)} - \frac{R(z)}{A(z/\gamma)}$$

  - $R(z)$ is residual of target waveform after inverse-filtering
  - cross-correlate h_γ and $r * h_\gamma$, iterate

CELP

- Represent excitation with codebook
  e.g. 512 sparse excitation vectors



  - linear search for minimum weighted error?

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

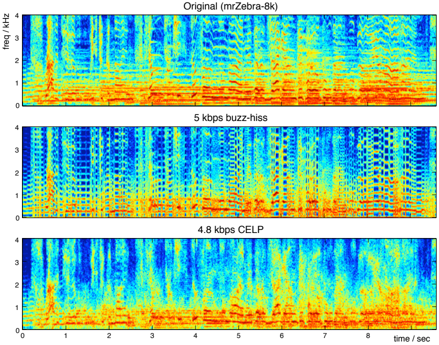
10 LSPs	4x4 + 6x3 bits =	34 bits
Pitch delay	4 x 7 bits =	28 bits
Pitch gain	4 x 5 bits =	20 bits
Codebk index	4 x 9 bits =	36 bits
Codebk gain	4 x 5 bits =	20 bits

  - 138 bits

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

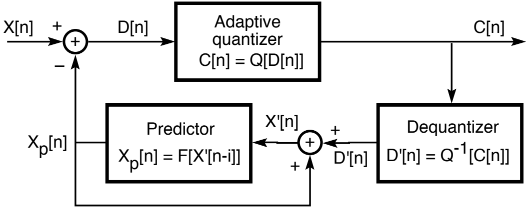


- LPC filter models vocal tract; also matches auditory system?
  - i.e. the 'source-filter' separation is good for relevance as well as redundancy?

Outline

Wide-Bandwidth Audio Coding

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

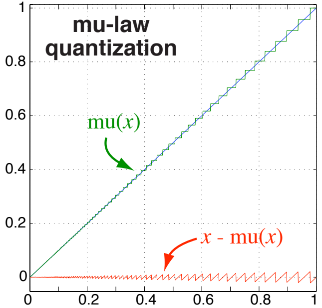


  - as prediction gets smarter, becomes LPC
    e.g. `shorten` - lossless LPC encoding
- Larger compression gains needs irrelevance
  - hide quantization noise with psychoacoustic masking

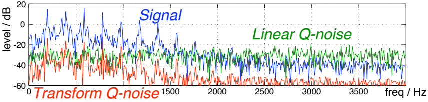
Noise shaping

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

- Have Q-noise scale with signal level
  - i.e. quantizer step gets larger with amplitude
  - minimum distortion for some center-heavy pdf

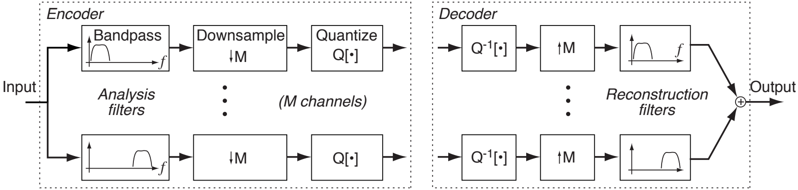


- Or: put Q-noise around peaks in spectrum
  - key to getting benefit of perceptual masking



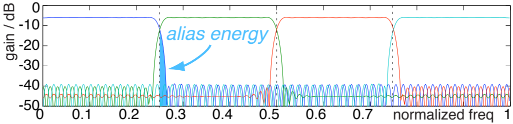
Subband coding

- Idea: Quantize separately in separate bands



  - Q-noise stays within band, gets masked

- 'Critical sampling' → $1/M$ of spectrum per band

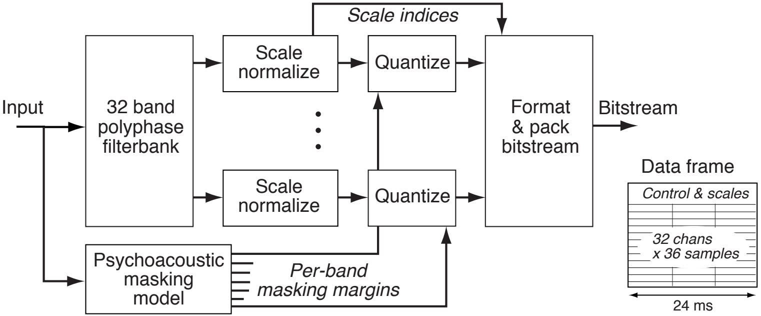


  - some aliasing inevitable

- Trick is to cancel with alias of adjacent band
  → 'quadrature-mirror' filters

MPEG-Audio (layer I, II)

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

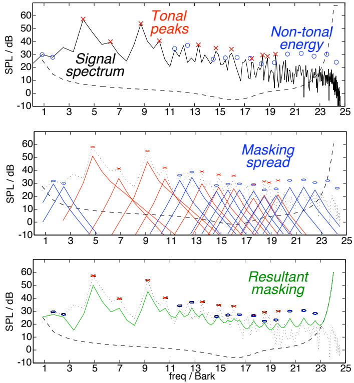


  - 22 kHz $\div$ 32 equal bands $=$ 690 Hz bandwidth
  - 8 / 24 ms frames $=$ 12 / 36 subband samples
  - fixed bitrates 32 - 256 Kbps/chan (1-6 bits/samp)
  - scale factors are like LPC envelope?

MPEG Psychoacoustic model

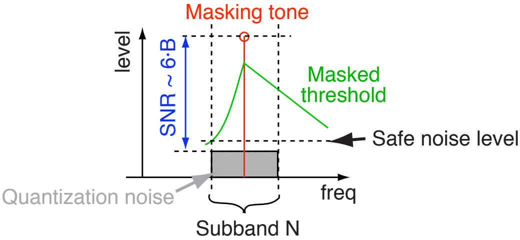
- Based on simultaneous masking experiments
- Difficulties:
  - noise energy masks ∼10 dB better than tones
  - masking level nonlinear in frequency & intensity
  - complex, dynamic sounds not well understood

- Procedure
  - pick 'tonal peaks' in NB FFT spectrum
  - remaining energy → 'noisy' peaks
  - apply nonlinear 'spreading function'
  - sum all masking & threshold in power domain



MPEG Bit allocation

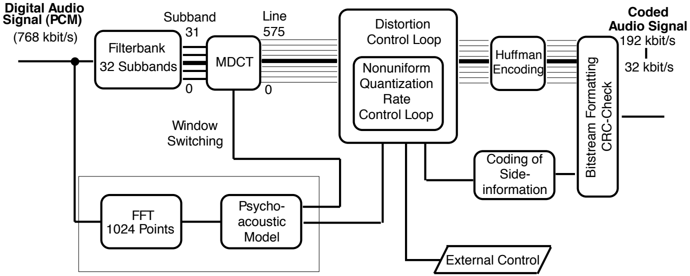
- Result of psychoacoustic model is maximum tolerable noise per subband



  - safe noise level → required SNR → bits B
- 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
- Bands with no signal above masking curve can be skipped

MPEG Audio Layer III

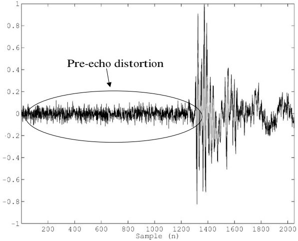
- 'Transform coder' on top of 'subband coder'



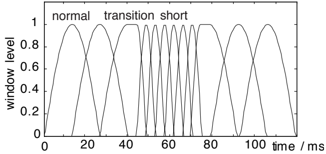
- Blocks of 36 subband time-domain samples become 18 pairs of frequency-domain samples
  - more redundancy in spectral domain
  - finer control e.g. of aliasing, masking
  - scale factors now in band-blocks
- Fixed Huffman tables optimized for audio data
- Power-law quantizer

Adaptive time window

- Time window relies on temporal masking
  - single quantization level over 8-24 ms window
- 'Nightmare' scenario:

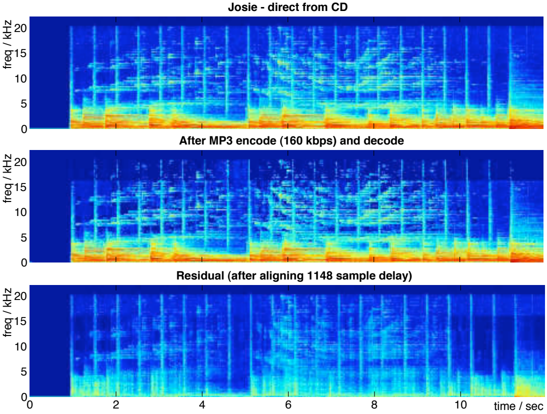


  - 'backward masking' saves in most cases
- Adaptive switching of time window:



The effects of MP3

- Before & after:



  - chop off high frequency (above 16 kHz)
  - occasional other time-frequency 'holes'
  - quantization noise under signal

MP3 & Beyond

- MP3 is 'transparent' at $\sim$ 128 Kbps for stereo (11x smaller than 1.4 Mbps CD rate)
  - only decoder is standardized: better psychological models $\rightarrow$ better encoders
- MPEG2 AAC
  - rebuild of MP3 without backwards compatibility
  - 30% better (stereo at 96 Kbps?)
  - multichannel etc.
- MPEG4-Audio
  - wide range of component encodings
  - MPEG Audio, LSPs, ...
- SAOL
  - 'synthetic' component of MPEG-4 Audio
  - complete DSP/computer music language!
  - how to encode into it?

Summary

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

References

Ben Gold and Nelson Morgan. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. Wiley, July 1999. ISBN 0471351547.

D. Pan. A tutorial on MPEG/audio compression. *IEEE Multimedia*, 2(2):60–74, 1995.

Karlheinz Brandenburg. MP3 and AAC explained. In *Proc. AES 17th International Conference on High Quality Audio Coding*, pages 1–12, 1999.

T. Painter and A. Spanias. Perceptual coding of digital audio. *Proc. IEEE*, 80(4): 451–513, Apr. 2000.