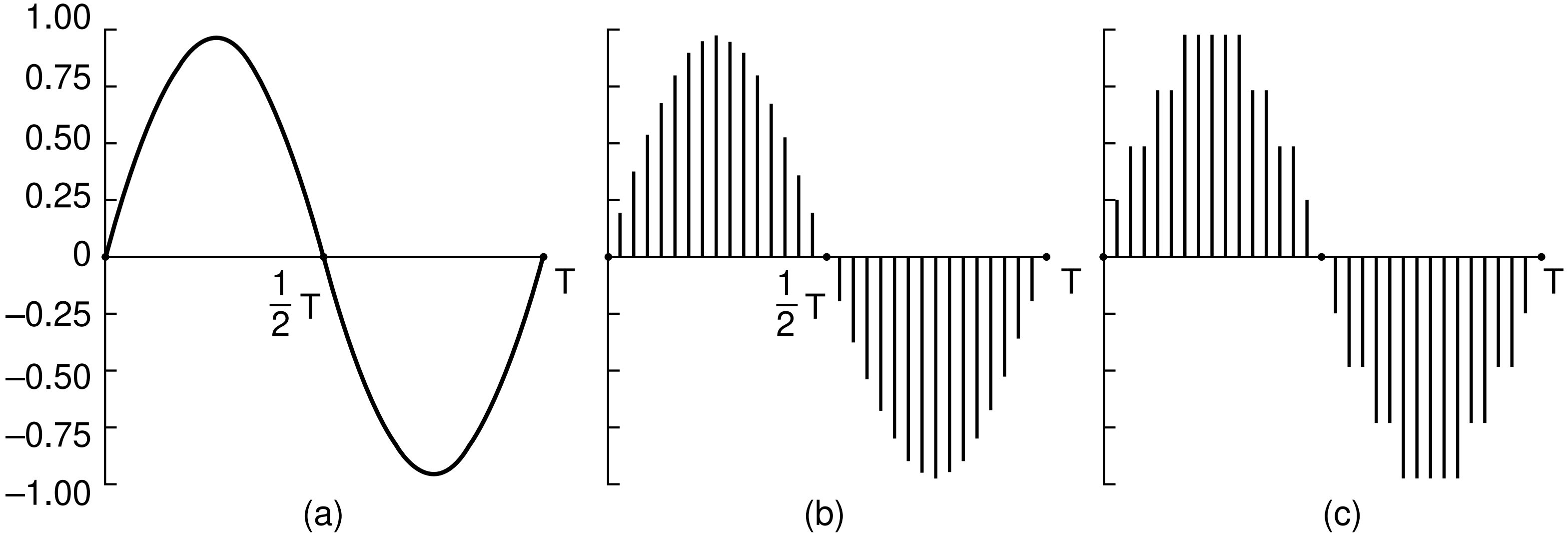
Multimedia Review

- digital audio and video
- audio encodings
- quality evaluation
- video encodings
- JPEG
- MPEG
- H.263

Digital Audio

- sample each audio channel and quantize ➠ pulse-code modulation (PCM)
- Nyquist bound: need to sample at twice (+ ϵ) the maximum signal frequency
- analog telephony: 300 Hz – 3400 Hz ➠ 8 kHz sampling
- audio CD: 44,100 Hz sampling (based on video equipment used for early recordings)
- signal-to-(quantization) noise ratio:
- increase dynamic range
- highly redundant ➠ compression

Digital Audio: Sampling



Digital Audio: Compression

Alternatives for compression:

- companding: non-linear quantization ➠ μ-law (G.711)
- waveform: exploit statistical correlation
- model: model voice, extract parameters (e.g.: pitch)
- subband: split signal into bands (e.g., 32) and code individually ➠ MPEG audio coding

Newer codings: make use of masking properties of human ear

Digital Audio: Compression

coding	kb/s	use
LPC-10	2.4	robotic, secure telephone
GSM	13.0	European mobile phone
G.729	8.0	mobile telephony
G.723	5.3/6.3	videophones
G.711	64.0	telephone
MPEG L3	128.0	CD stereo
16 bit/44.1 kHz	1411	compact disc

See also `http://www.cs.columbia.edu/~hgs/audio`

Distortion: Signal-to-Noise Ratio

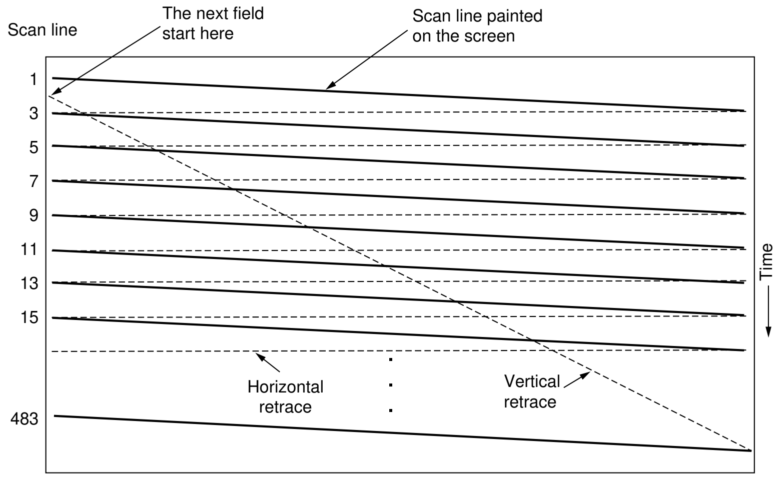
- error $r(n) = x(n) - y(n)$
- variances $\sigma_x^2, \sigma_y^2, \sigma_r^2$
- $\sigma_u^2 = \frac{1}{M} \sum_{n=1}^{M} u^2(n)$
- SNR (decibel, dB) $= 10 \log_{10}(\sigma_x^2/\sigma_r^2)$
- for amplitudes: dB $= 20 \log_{10} x/y$
- also: dynamic range
- hearing range: 0 dB $\ldots$ 120 dB (pain); conversation: 50 dB
- e.g.: FM radio $\approx$ 40 dB, TV $<$ 40 dB, audio CD: 96 dB

Distortion: Other Measures

- SNR *not* a good measure of perceptual quality
- ➠ segmental SNR: time-averaged blocks (say, 16 ms)
- frequency weighting
- subjective measures:
  - A-B preference
  - subjective SNR: comparison with additive noise
  - mean opinion score (MOS): 5 (excellent) to 1 (bad)
  - intelligibility: diagnostic rhyme test ➠ low estimate of sentence intelligibility

Analog Video

system	where	lines (disp.)	size	f/s
NTSC	US, Japan	525 (483)	4:3	30
PAL	Europe	625 (576)	4:3	25



reduce flicker ➠ interlace scanning: 2 fields/frame

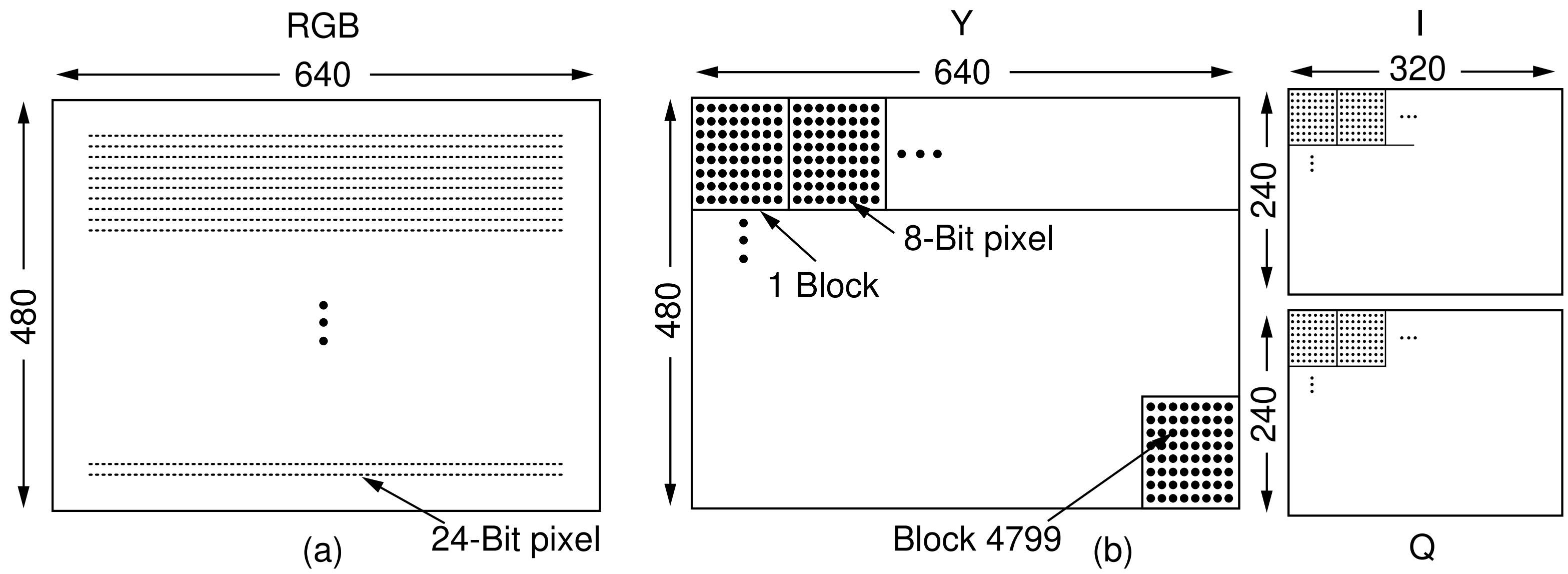
Digital Video

format	lines	pixels
SCIF	480	704
CIF	288	352
QCIF	144	176

- camera ➠ Red, Green, Blue ➠ chrominance, luminance
- YUV: Y = chrominance, UV: color difference (red, blue − Y)

Digital Video

eye more sensitive to luminance ➠ subsample chrominance: 2:1:1, 4:1:1, 4:2:2



treat chrominance and luminance differently

Video Coding

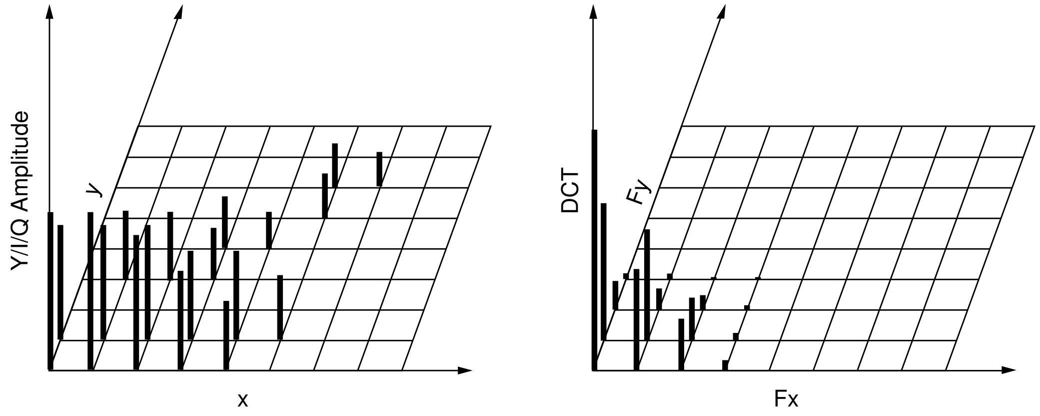
lossless (entropy): (X ray images!) run-length, statistical encoding, Huffman Coding, Lempel-Ziv Coding (GIF).

lossy: ➠ exploit spatial and temporal redundancy

- **transform:** $\rightarrow$ frequency domain, higher quantization steps for higher frequencies, energy compaction
- **vector quantization:** map NxN block into N^2-dimensional space and find closest in *codebook*
- **model-based:** geometric description ➠ very low bit-rate

Video Coding: JPEG (Joint Photographic Experts Group)

- individual (still) pictures
- lossless or lossy
- typically about 2 Mb/s for video stream
- discrete cosine transform: samples ➠ blocks (16x16 Y, 8x8 UV) ➠ frequency domain
- code frequency samples with different resolution



Video coding: JPEG (Joint Photographic Experts Group)

- DC or average (0,0) value: difference to previous block
- quantization:

DCT Coefficients

150	80	40	14	4	2	1	0
92	75	36	10	6	1	0	0
52	38	26	8	7	4	0	0
12	8	6	4	2	1	0	0
4	3	2	0	0	0	0	0
2	2	1	1	0	0	0	0
1	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Quantized coefficients

150	80	20	4	1	0	0	0
92	75	18	3	1	0	0	0
26	19	13	2	1	0	0	0
3	2	2	1	0	0	0	0
1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Quantization table

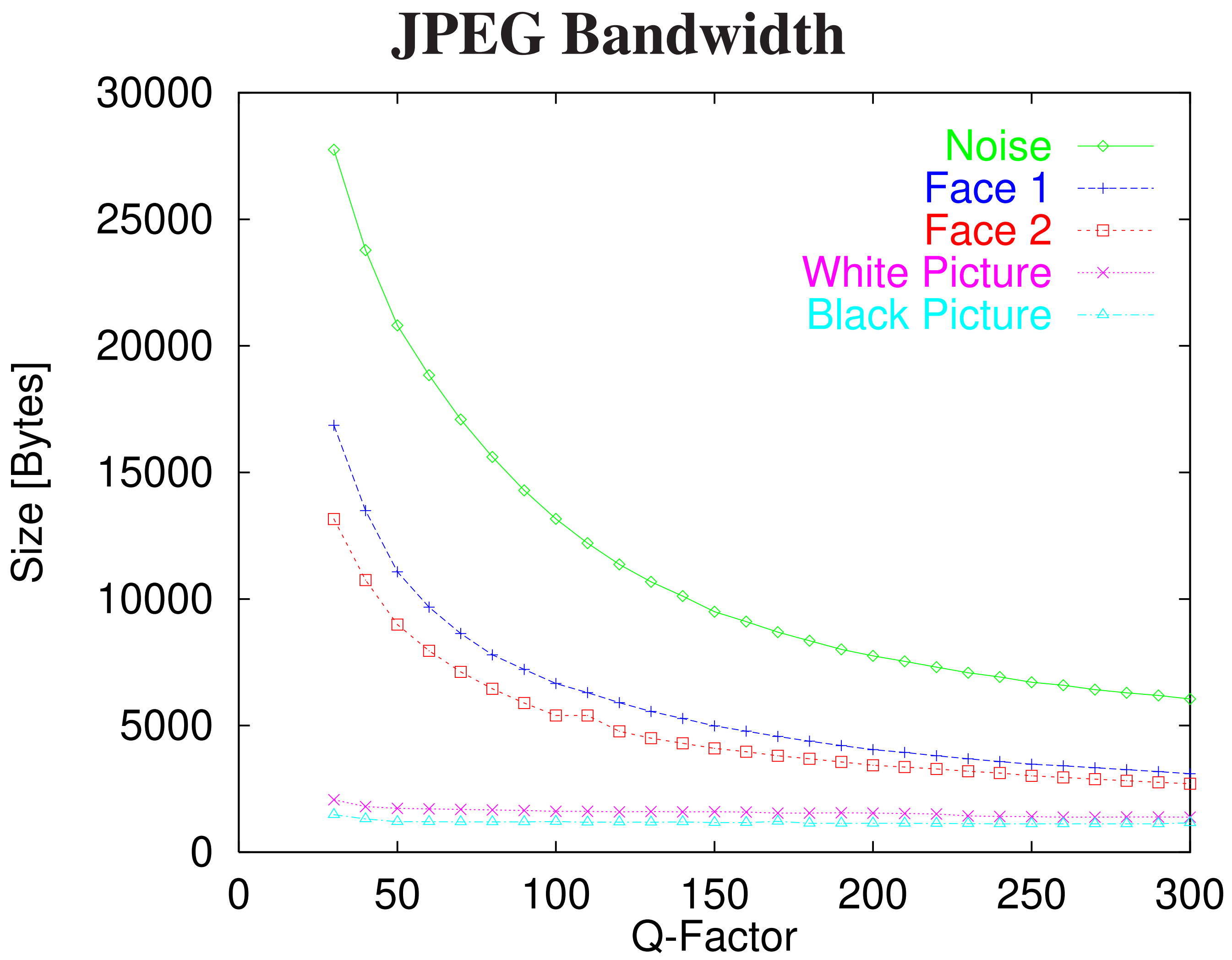
1	1	2	4	8	16	32	64
1	1	2	4	8	16	32	64
2	2	2	4	8	16	32	64
4	4	4	4	8	16	32	64
8	8	8	8	8	16	32	64
16	16	16	16	16	16	32	64
32	32	32	32	32	32	32	64
64	64	64	64	64	64	64	64

- zig-zag scan and run-length coding:

150	80	20	4	1	0	0	0
92	75	18	3	1	0	0	0
26	19	13	2	1	0	0	0
3	2	2	1	0	0	0	0
1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

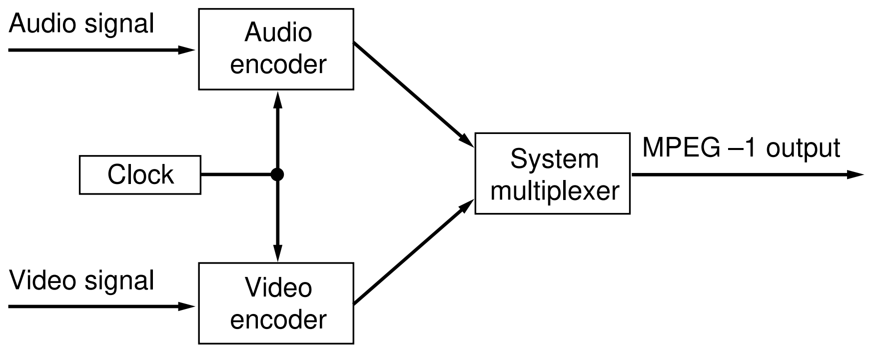
- Huffmann coding

See also `http://www.cs.columbia.edu/~hgs/video`


JPEG Bandwidth
Size [Bytes]
30000
25000
20000
15000
10000
5000
0
0
50
100
150
200
250
300
Q-Factor
Noise
Face 1
Face 2
White Picture
Black Picture

MPEG (Motion Picture Experts Group)

Audio and video:



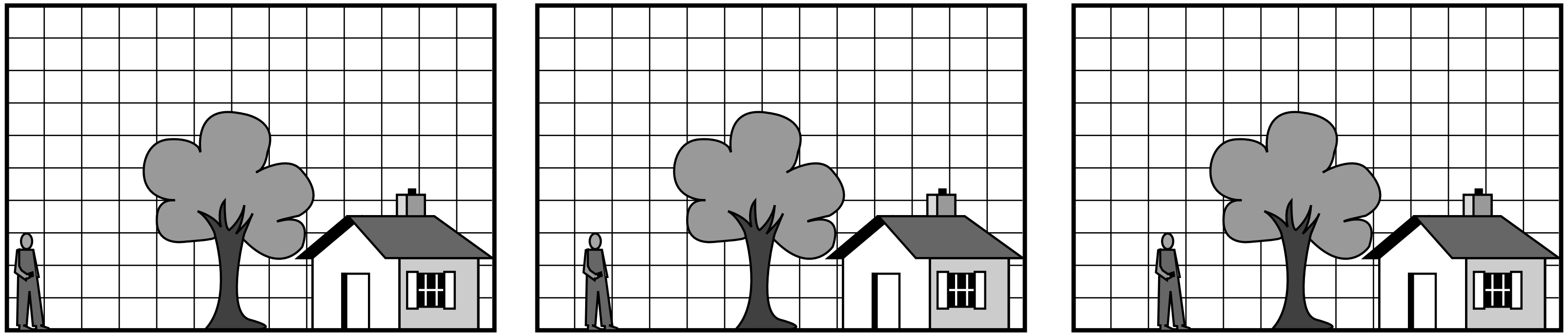
- ➠ JPEG + motion compensation ≈ H.261, MPEG video
- MPEG-1: 1.2 Mb/s fixed rate (CD ROM)
- MPEG-2: higher resolutions (HDTV), scaling
- image prediction: intra (I), forward (P), bidirectional (B) ➠ IBBPBBPBBPBBIBBPBBPB
- need I frames for error resiliency, joining movie in the middle

H.263

Video component of H.323 conferencing standard

- Similar to MPEG ➠ DCT, motion compensation, quantization
- Designed for conferencing, rates from 16 kbps and up
- Adds special motion compensation modes (overlapping and 8x8)
- Adds algebraic codes instead of Huffman codes
- I,B and P frames used

MPEG: Motion Compensation



- transmit “motion vectors” to account for panning and translational motion ➠ hard to find, easy to decode
- conditional replenishment: only transmit blocks that are different

Characteristics of digital audio and video

	audio	video
rate (kb/s)	$5 \ldots 64 \ldots 1500$	$20 \ldots 1500 \ldots 6000$
loss tolerance	$\leq 5\%$	$10^{-5} \ldots 10\%$
packet size	small	large
traffic	constant + silences	variable bit rate

One-way delay tolerance:

conference audio without echo cancellation: 40 ms

conference audio with echo cancellation: 150 ms (hand-off delay!)

playback audio/video: $\geq$ 500 ms (“VCR response”)

test for those faking deafness

Audio silence detection

- needed in conferences to avoid drowning in fan noise
- also reduces data rate
- in use in transoceanic telephony since 1950's (TASI: time-assigned speech interpolation)
- use energy estimate (μ-law already close) or spectral properties (difficult)
- difficulty: background noise, levels vary
- ➠ vary noise threshold: threshold = running average + hysteresis
- if above threshold, increase running average by one for each block
- if below threshold, update running average
- speech has soft (unvoiced) beginnings and endings ➠ *hang-over*, pre-talkspurt burst

Audio traffic models

talkspurt: constant bit rate: one packet every 20...100 ms ➠ mean: 1.67 s

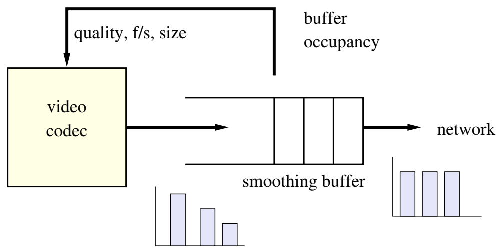
silence period: usually none (maybe transmit background noise value) ➠ 1.34 s

➠ for telephone conversation, both roughly exponentially distributed

- double talk for "hand-off"
- may vary between conversations... ➠ only in aggregate
- Common model for a single user is two state on-off Markov Process
- For N users (PBX, telephone switch), N+1 state birth-death Markov Process

Video traffic models

- easy case: fit into constant bit rate



- alternative: variable rate ➠ mux gain of $\approx 2 \ldots 6$
- short time scales: packets within slice or frame
- medium time scales: I, B, P packet pattern
- longer time scales: scene changes (every few seconds) ➠ higher rate
- looks similar at all time scales, long-term correlation, *heavy-tailed distribution* ➠ *self-similar*
- but: for short queues, long-term correlations don't matter