CS 598KN

Advanced Multimedia Systems Design

Lecture 4 – H.264 + Digital Audio + MP3

Klara Nahrstedt

Fall 2017

Overview

- H.264 and H.265
- Human Auditory System
- Digital Audio
- MP3 Encoding

H.264 OVERVIEW

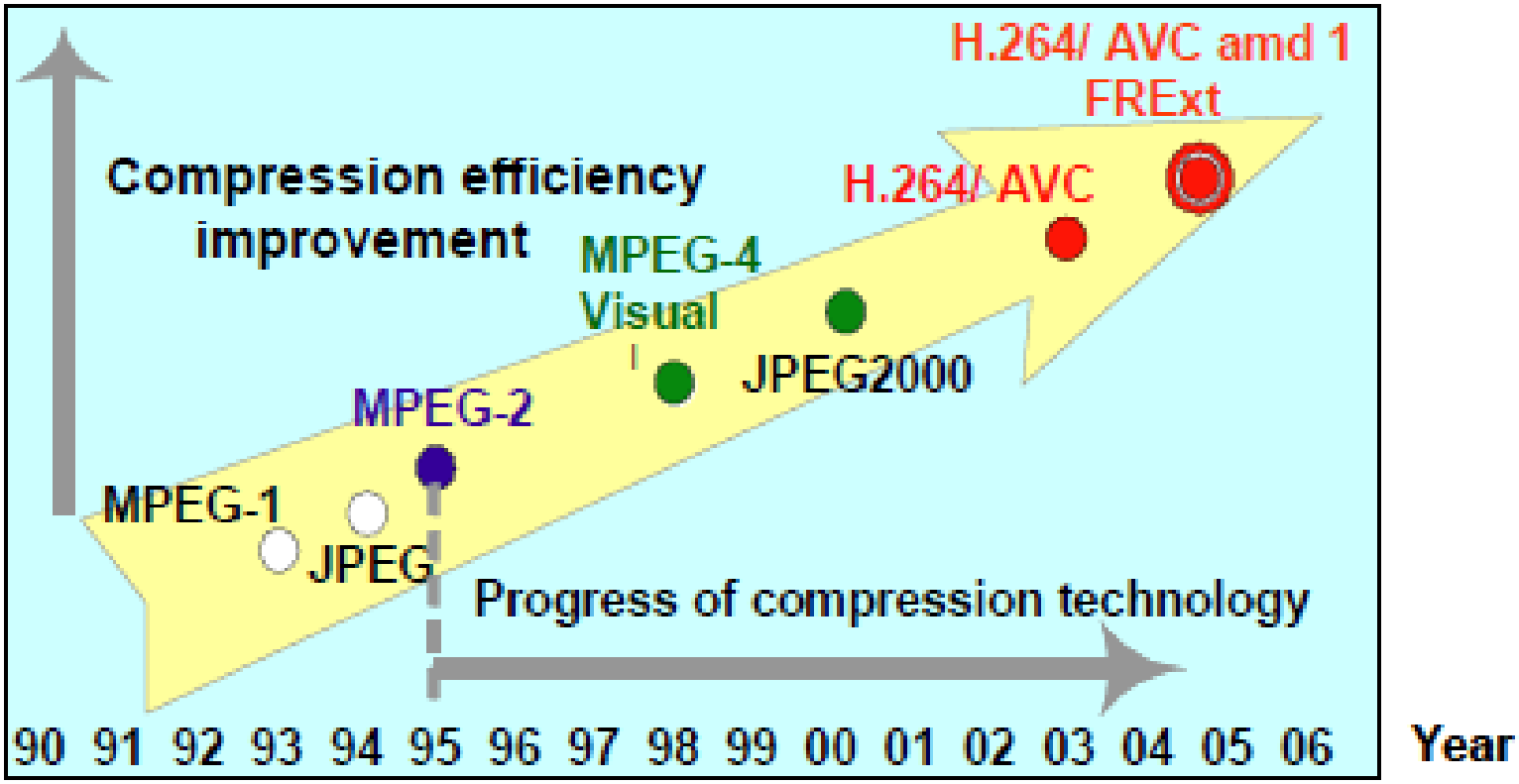
H.264 /AVC
(Advanced Video Coding)

- Developed by Joint Video Team
  - ITU-T's Video Coding Experts Group (VCEG)
  - ISO/IEC's Moving Picture Experts Group (MPEG)
- Reference model developed in 2003
- Amendment added in 2004
  - Fidelity Range Extensions (FRExt, Amendment 1)
    - Demonstrates coding efficiency against MPEG-2 with potentially 3:1 compression ratio

H.264/AVC History

- In 1998 – ITU VCEG finished its video coding standard, H.26L
  - First Call for Proposals to improve coding
    - In 1999 - First proposed design of new standard
- In 2001 – ISO MPEG finished its own video coding standard, MPEG-4, Part 2
  - Call For Proposals (CFP) to improve coding efficiency
  - VCEG provided its own design to MPEG's CFP
- Joint Forces VCEG+MPEG => New standard
  - **MPEG-4/Part 10 = MPEG-4 AVC = H.264 AVC=JVT = H.26L = H.264**

Evolution of Compression Technology



Source: Technology Overview, “AVC-Intra (H.264 Intra) Compression,” Panasonic Broadcast, 2007

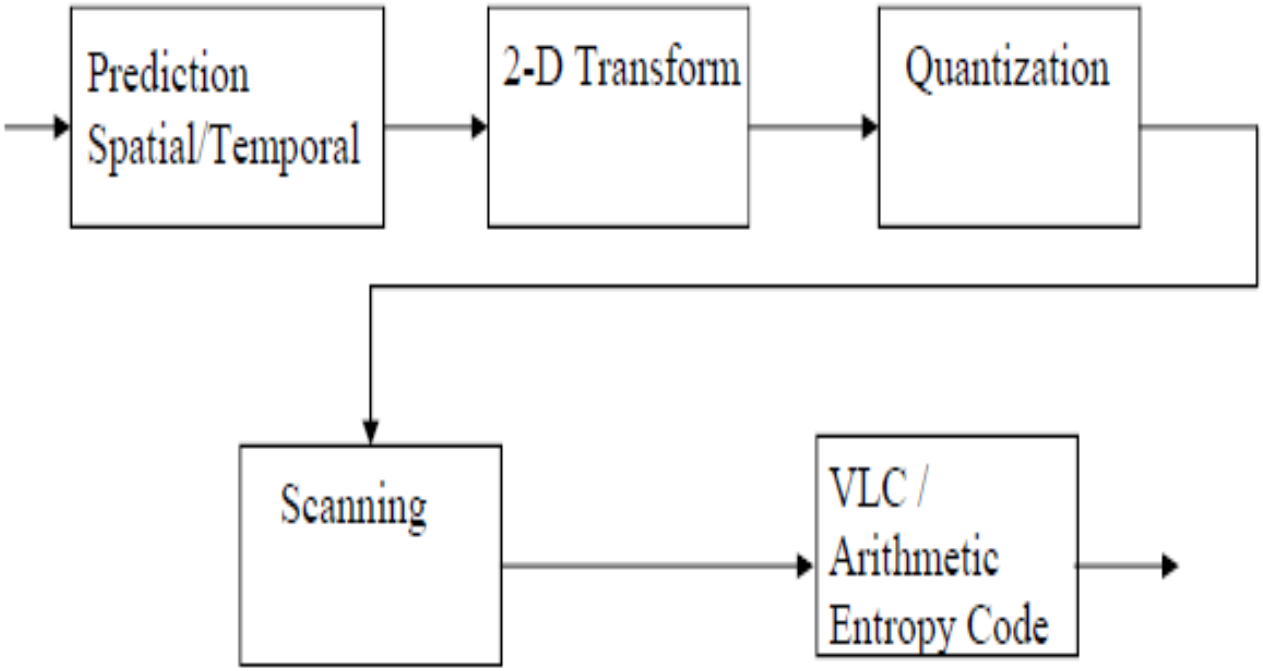
Applications of H.264

- Video conferencing
- Entertainment
  - Broadcasting over cable, satellite, cable modem, DSL,
  - Storage on DVD, hard disks
  - Video-on-demand
- Streaming video
- Surveillance/military apps
- Digital cinema

H.264 Coding Structure

- At basic level, coding structure of H.264 similar to MPEG-1, MPEG-2, MPEG-4/Part 2.
- Hierarchy of video sequence
  - Sequence (pictures (slides (macroblocks partitions (sub-macroblock partitions (block (samples))))))
  - Bits associated with slide layer and below - Video Coding Layer (VCL)
  - Bits associated with higher layers – Network Abstraction Layer (NAL)

H.264 Encoding Block Diagram



VLC – Variable Length Coding (e.g., Huffman Coding)

H.264 Image Preparation

- First version had
  - 4:2:0 luma chroma relations
  - RGB-to-YCbCr color space translation
  - Subsampling chroma components by 2:1
  - 8 bit sample precision
- FRExt version has
  - 4:2:2, 4:4:4 formats
  - Higher than 8 bits precision

H.264 Image Preparation

- Basic unit of coding – macroblock
  - In 4:2:0 chroma format
    - macroblock 16x16 pixels in luma
    - Macroblock 8x8 pixels in chroma
  - In 4:2:2 chroma format
    - Macroblock in chroma 8x16 pixels
  - In 4:4:4 chroma format
    - Macroblock in chroma 16x16 pixels
- Major processing unit becomes **Slice** (group of macroblocks)

H.264 Image Coding Tools in Slices

(New in comparison to previous MPEG-1,2/H.26X standards)

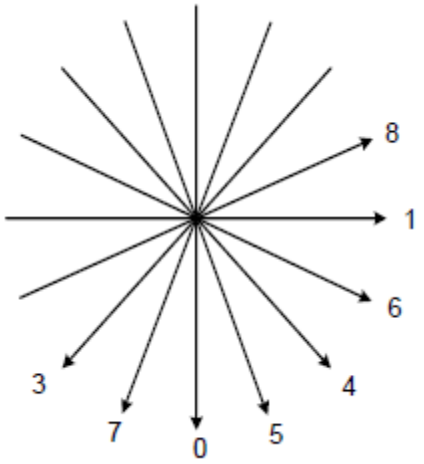
- Spatial and temporal prediction
- Lossless entropy coding
- Deblocking filter
- Residual color transform for efficient RGB coding
- Different slices use different coding tools
  - I-slice; P-slice, B-slice

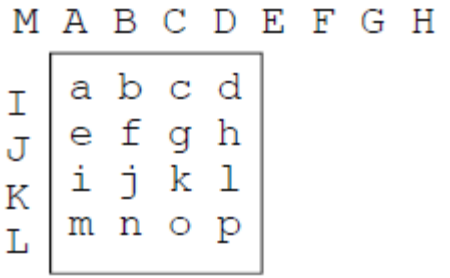
I-Slice Coding Algorithm

1. Spatially predict pixels from neighboring pixel values
2. Transform 4x4 or 8x8 blocks from spatial domain to other differentiable domain
3. Perform perceptual-based quantization
4. Scan quantized coefficients
    1. Zig-zag or field-based
5. Compress with Entropy Coding
    1. CAVLC (Context-Adaptive Variable Length Coding)
    2. CABAC (Context-Adaptive Binary Arithmetic Coding)

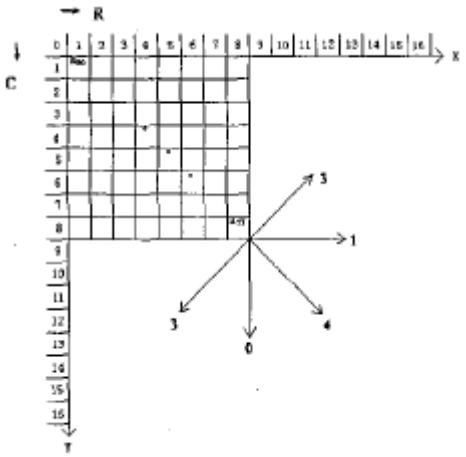
H.264 Intra-Spatial Prediction

- Due to Spatial correlation among pixels we can compress data by intra-spatial prediction
  - H.264 supports three basic types of **intra spatial predictions**





Spatial Prediction for 4x4 Block



Intra prediction directions

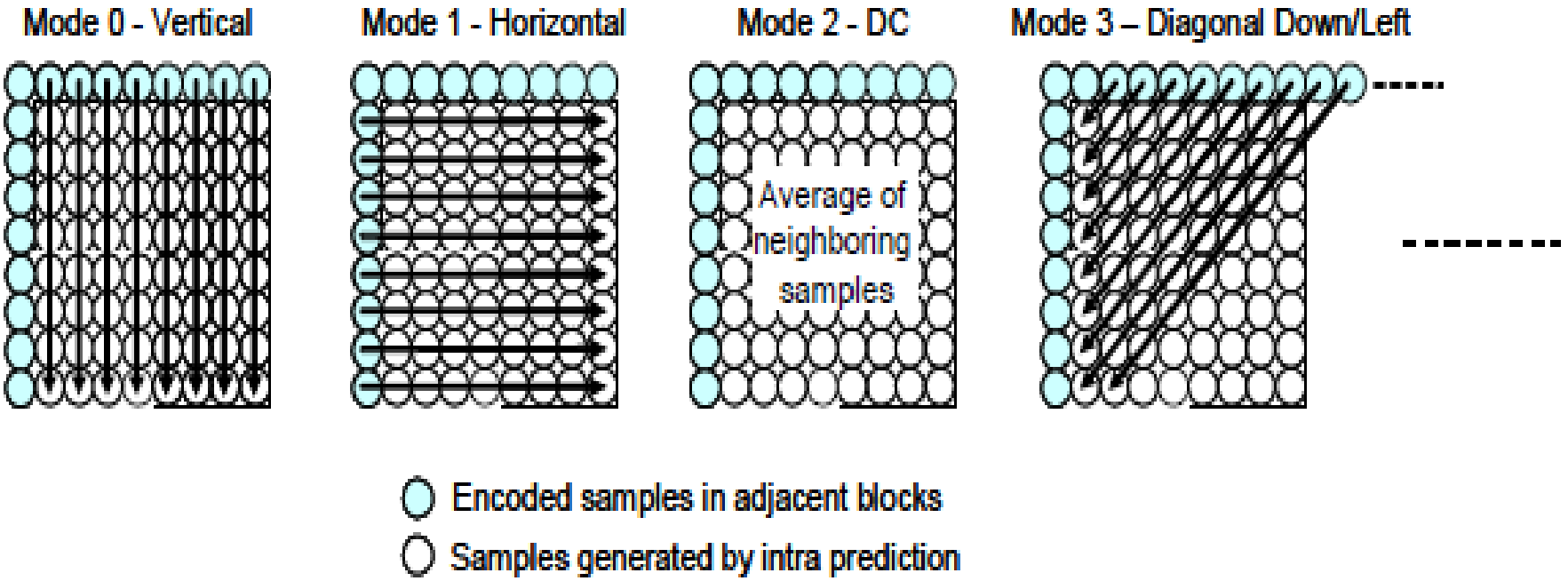
Comparison of Original and Intra Predicted Images

(a) Original input image

(b) Intra prediction image

Source: Technology Overview, “AVC-Intra Compression” Panasonic Broadcast, September 7, 2007

Examples of Spatial Intra Prediction Modes



Source: Technology Overview, “AVC-Intra Compression” Panasonic Broadcast, September 7, 2007

H.264 Transformation

- Goal of transformation – de-correlate data spatially
  - New approach
    - use integer inverse transform design, rather than IDCT
    - Use integer transform for forward transform
      - No mismatch between encoder and decoder

$$T_{4x4} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix}, \quad T_{8x8} = \begin{bmatrix} 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 \\ 12 & 10 & 6 & 3 & -3 & -6 & -10 & -12 \\ 8 & 4 & -4 & -8 & -8 & -4 & 4 & 8 \\ 10 & -3 & -12 & -6 & 6 & 12 & 3 & -10 \\ 8 & -8 & -8 & 8 & 8 & -8 & -8 & 8 \\ 6 & -12 & 3 & 10 & -10 & -3 & 12 & -6 \\ 4 & -8 & 8 & -4 & -4 & 8 & -8 & 4 \\ 3 & -6 & 10 & -12 & 12 & -10 & 6 & -3 \end{bmatrix}$$

Transformation matrix

H.264 Transform

- If *16x16 intra-prediction mode* with 4x4 transform, DC coefficients of the sixteen 4x4 luma blocks in the macro-block are further transformed by Hadamard transform

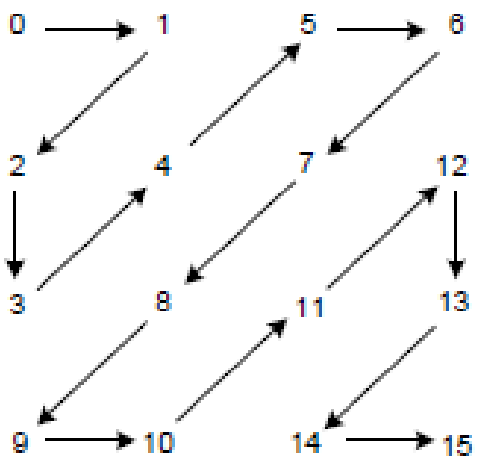
$$H_{2x2} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad H_{4x4} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix}.$$

Hadamard Transform for DC Coefficients in 16x16 macroblock

H.264 Perceptual-based quantization scaling

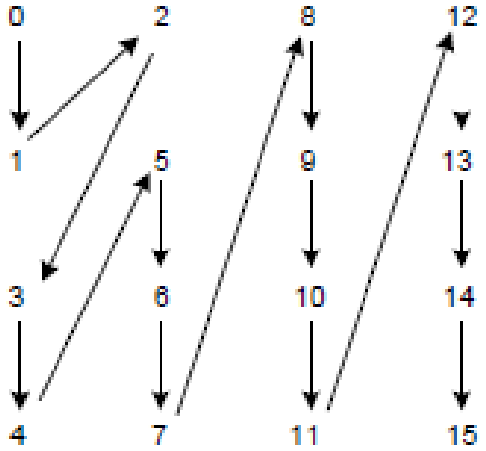
- Encoder specifies
  - For each transform block size
  - For intra and inter prediction
  - A customized scaling factor
- Decoder gets to tune quantization fidelity based on human visual system
- We don’t improve objective fidelity, but we improve subjective fidelity

Scanning



a

Frame-mode zig-zag scanning



b

Field-mode

- Frame-mode scan ordering designed to order the highest variance coefficients first and maximize number of consecutive zero-value coefficients
- Field-mode reflects decreasing correlation of the source data in the vertical dimension

Entropy Coding

- Lossless coding techniques – replace data with coded representation
- Variable Length Coding (VLC)
- Binary Arithmetic Coding (BAC)
- H.264/AVC supports context adaptation
  - ACVLC
  - ACBAC

Adaptive Encoding (Adaptive Huffman)

- Huffman code change according to usage of new words and new probabilities can be assigned to individual letters
- If Huffman tables adapt, they must be transmitted to receiver side

Adaptive Huffman Coding Example

Symbol	Code	Original probabilities
A	001	P(A) = 0.16
B	1	P(B) = 0.51
C	011	P(C) = 0.09
D	000	P(D) = 0.13
E	010	P(E) = 0.11

Symbol	Code	New Probabilities (based on new word BACAAB)
A	1	P(A) = 0.5
B	01	P(B) = 1/3
C	001	P(C) = 1/6
D	0000	P(D) = 0
E	0001	P(E) = 0

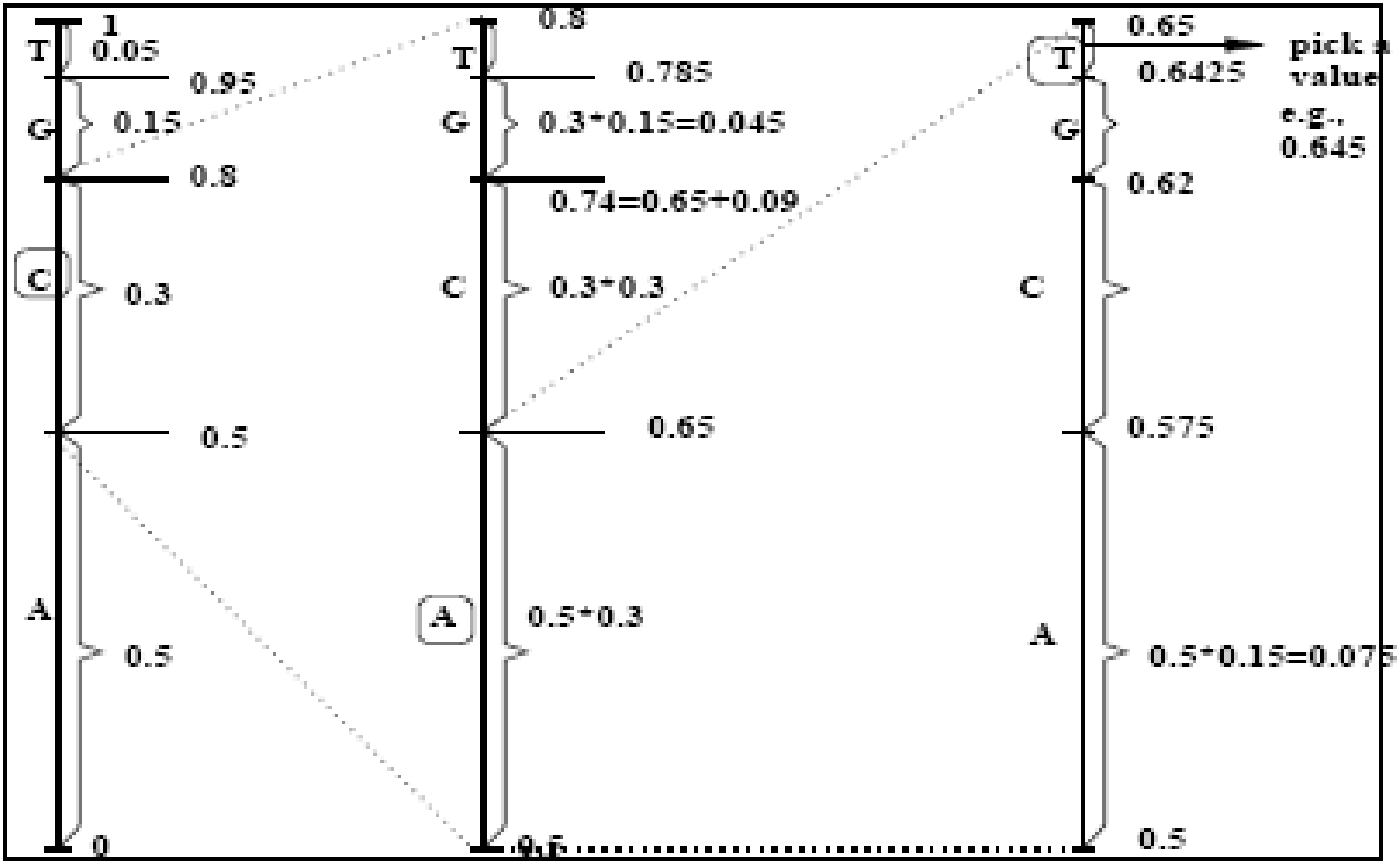
Arithmetic Coding

- Optimal algorithm as Huffman coding wrt compression ratio
- Better algorithm than Huffman wrt transmitted amount of information
  - Huffman – needs to transmit Huffman tables with compressed data
  - Arithmetic – needs to transmit length of encoded string with compressed data

Binary Arithmetic Coding

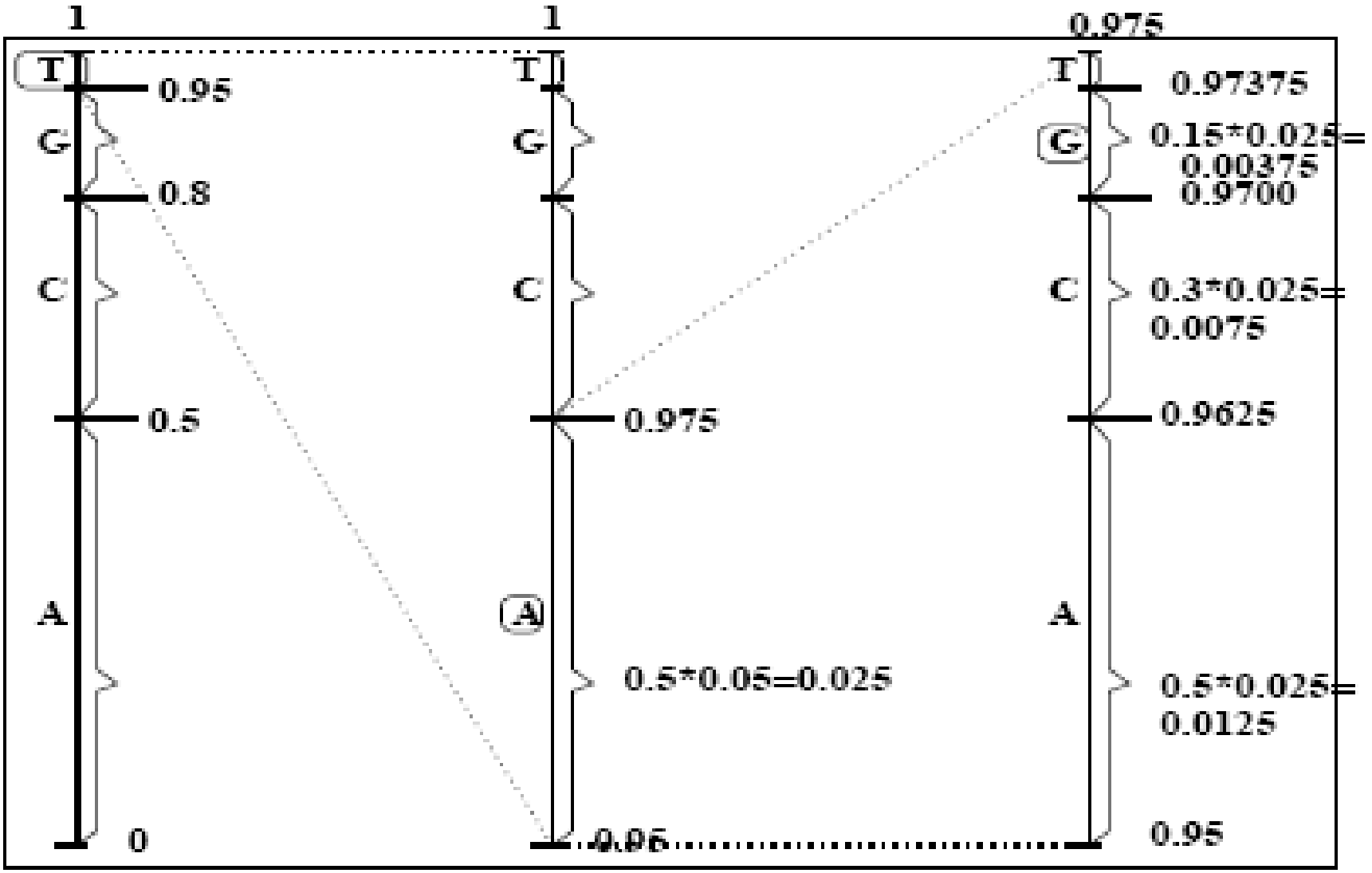
- Each symbol is coded by considering the prior data
- Encoded data must be read from the beginning, there is no random access possible
- Each real number (< 1) is represented as binary fraction
  - $0.5 = 2^{-1}$ (binary fraction = 0.1); $0.25 = 2^{-2}$ (binary fraction = 0.01), 0.625 = 0.5+0.125 (binary fraction = 0.101) ….

P(A)=0.5, P(C) = 0.3, P(G) = 0.15, P(T) = 0.05 => Encode CAT



Send Value: 0.645

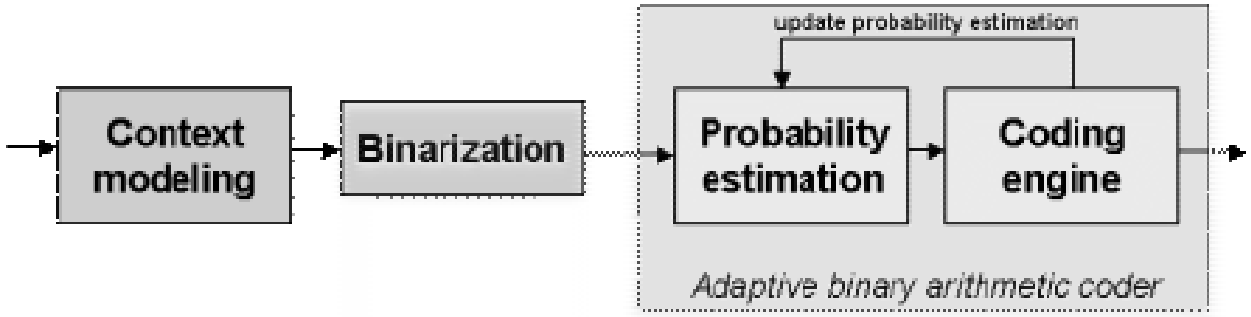
Given: P(A)=0.5, P(C) = 0.3, P(G) = 0.15, P(T) = 0.05 => Decode: 0.9715



Decoded Word: TAG

CABAC (Context-Adaptive Binary Arithmetic Coding)

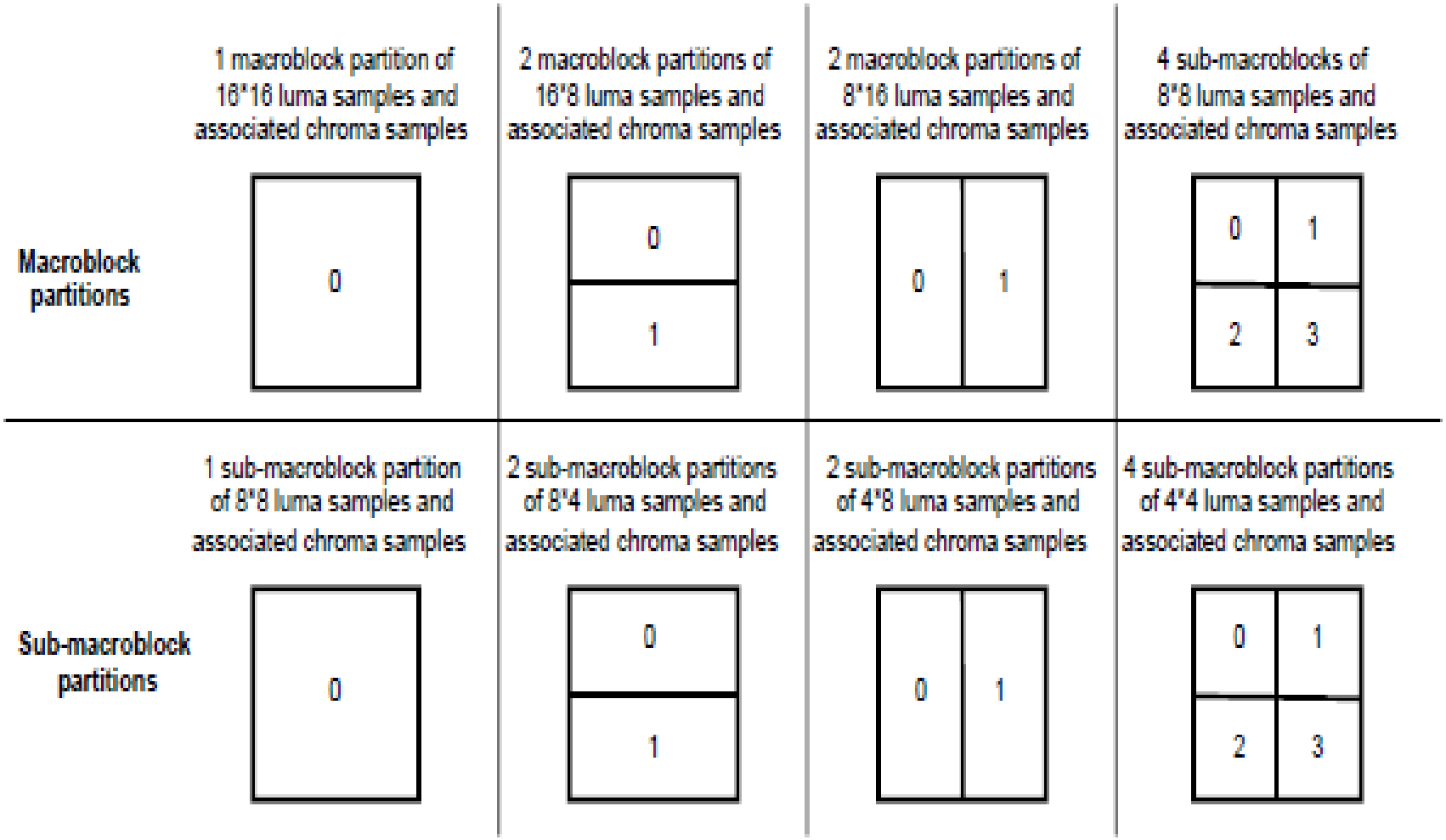
- CABAC mode improves compression efficiency by ~10% relative to CAVLC
- CABAC much more computationally complex



H.264 P-Slices

- Temporal prediction is used with estimation of motion between pictures
- Motion is estimated at 16x16 macro-block level or by partitioning macro-block into smaller regions 16x8, 8x16, 8x8, 8x4, 4x8, 4x4
- **One motion vector** can be sent for **each sub-macro-block partition**
- Motion is estimated from multiple pictures that lie **either in past or in future** in display order

Macro-block Partitions for Motion Estimation



H.264 B-Slices

- Temporal prediction with two motion vectors representing two estimates of motion per macro-block or sub-macro-block
- Consider any reference picture in future or past in display order
- **Weighted prediction concept** is further extended in usage of **weighted average** between two predictions

Deblocking Filter

- Video filter
  - Is applied to blocks in decoded video to improve visual quality and prediction performance , smoothing sharp edges which can form between macro-blocks
  - Operates on edges of 4x4 and 8x8 transform blocks
  - Is adaptive filter that adjusts its strength depending upon compression model of macro-block

Residual Color Transform Support

- Video captured and displayed in RGB
- Human visual system is better matched to luma (brightness) and chroma (hue & saturation) representations
- In MPEG-2, we have color transformation
  - RGB to YCbCr domain

$$Y = K_R * R + (1 - K_R - K_B) * G + K_B * B; \quad Cb = \frac{1}{2}\left(\frac{B - Y}{1 - K_B}\right); \quad Cr = \frac{1}{2}\left(\frac{R - Y}{1 - K_R}\right);$$

  - Problem: rounding errors, complex
- H.264 FRExt uses RGB to YCgCo

$$Y = \frac{1}{2}\left(G + \frac{(R + B)}{2}\right); \quad Cg = \frac{1}{2}\left(G - \frac{(R + B)}{2}\right); \quad Co = \frac{(R - B)}{2}.$$

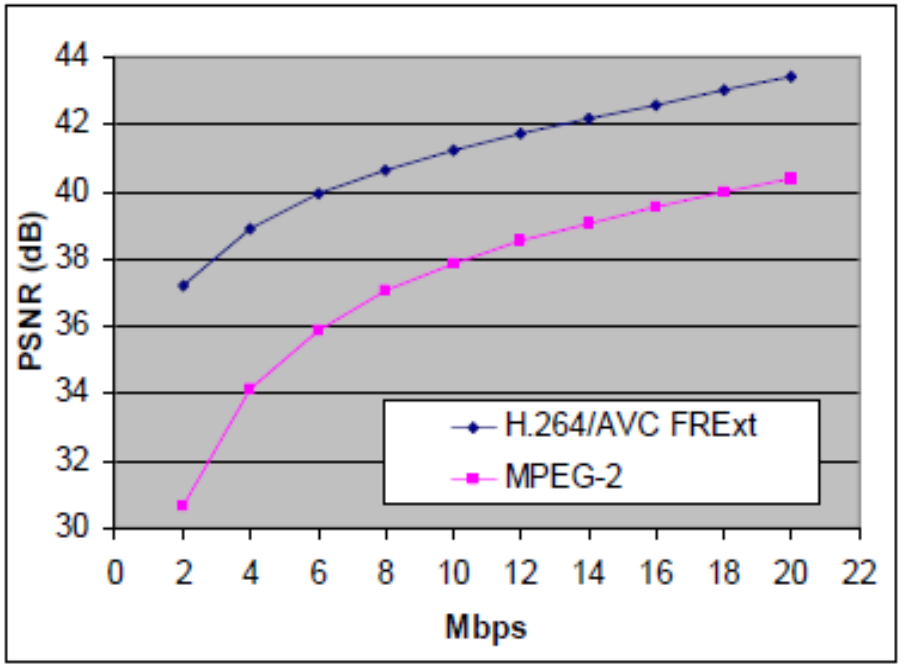
  - Reduces complexity, but not rounding errors

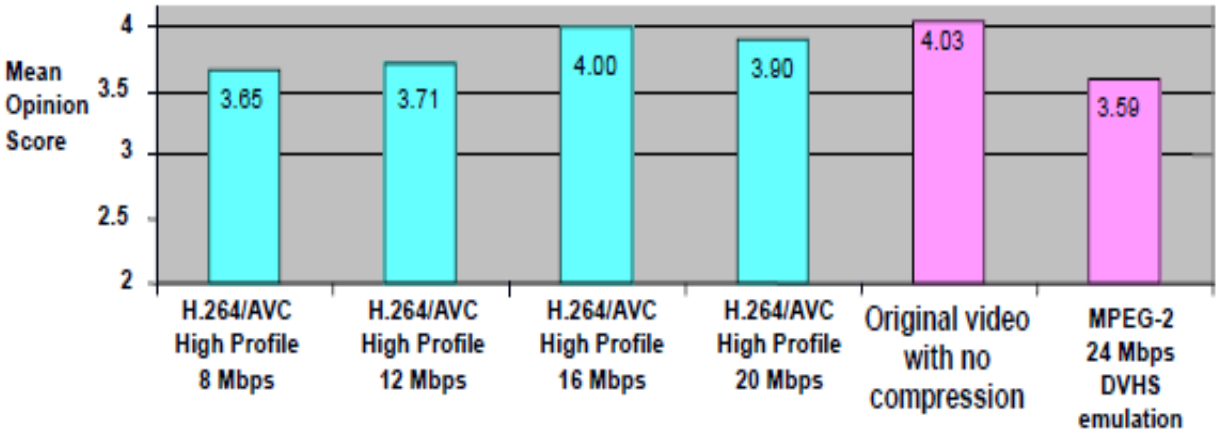
$$Co = R - B; \quad t = B + (Co \gg 1); \quad Cg = G - t; \quad Y = t + (Cg \gg 1);$$

  - Reduces complexity and rounding errors

Summary

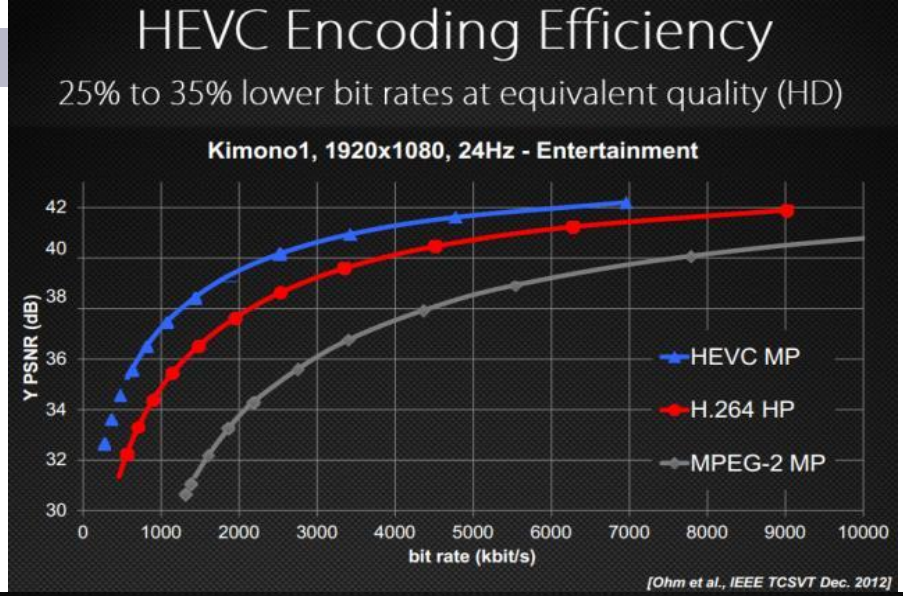
- H.264 AVC is a standard that has the potential to improve efficiency as well as subjective perception

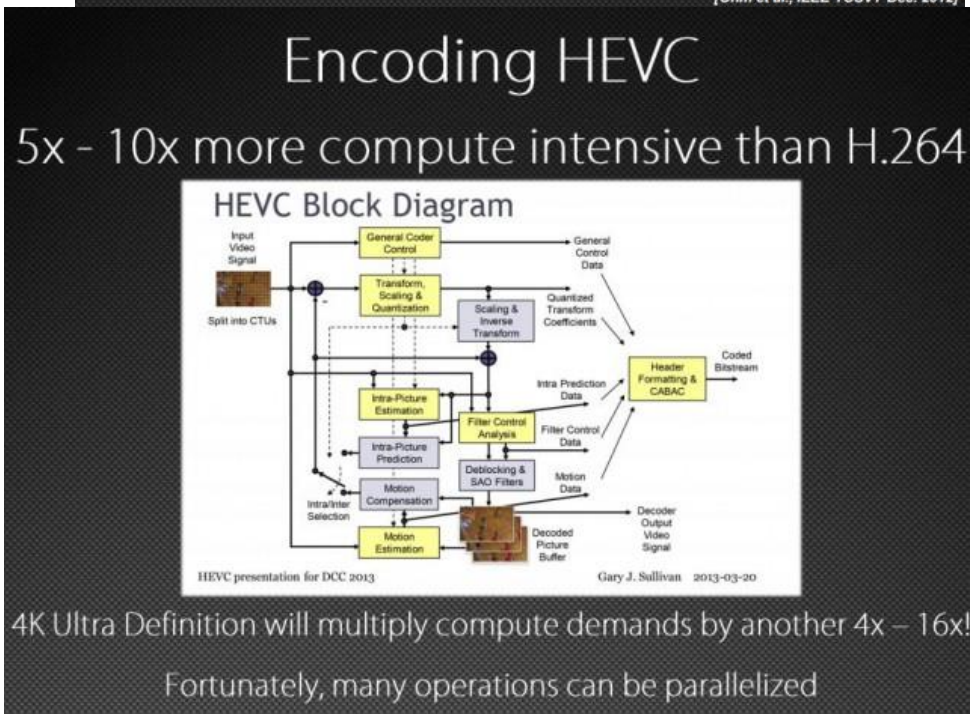




H.265 - HEVC

- HEVC (High Efficient Video Coding)
- HEVC is also h.265 and MPEG-H, Part 2
  - Much stronger storage efficiency than H.264 (50% reduction)
  - Compression of 8192×4320 images
  - CABAC preserved
  - H.265 codec is parallelized, 10bit color, higher frame rate (50fps)
- MKV – container that holds codec and format of video (e.g., H.264 or HEVC)





Source: https://www.extremetech.com/computing/162027-h-265-benchmarked-does-the-next-generation-video-codec-live-up-to-expectations

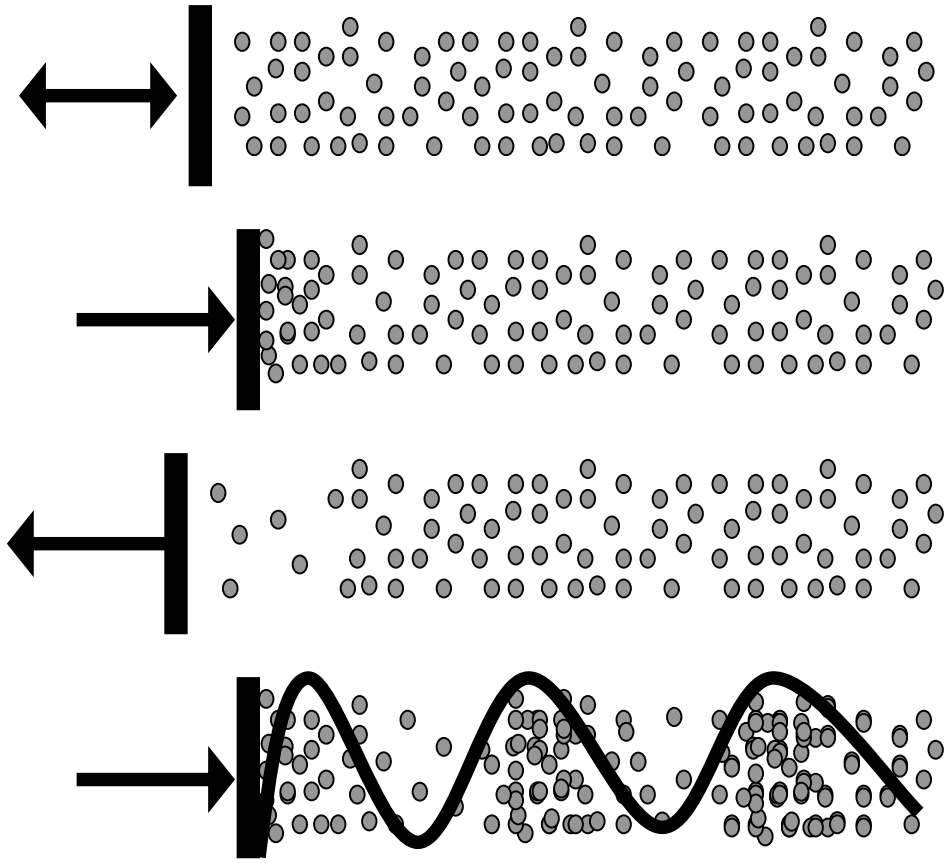
References

- Slides based on Literature Reference
  - Gary Sullivan et al, “The H.264/AVC Advanced Video Coding Standard: Overview and Introduction to the Fidelity Range Extensions”, SPIE Conf. on Applications of Digital Image Processing, 2004.
  - Zhang Nan et al., “Spatial Prediction Based Intra Coding”, IEEE ICME 2004
  - Technology Overview, “AVC-Intra (H.264 Intra) Compression”, Panasonic Broadcast, 2007

DIGITAL AUDIO CHARACTERISTICS

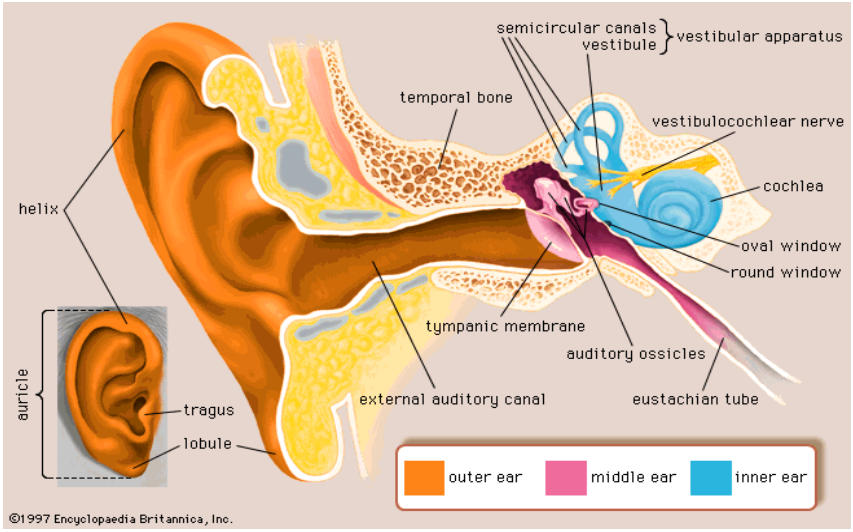
Auditory Perception

- Sound – physical phenomenon caused by vibration of material
- These vibrations trigger pressure wave fluctuations in the air
- Wave forms

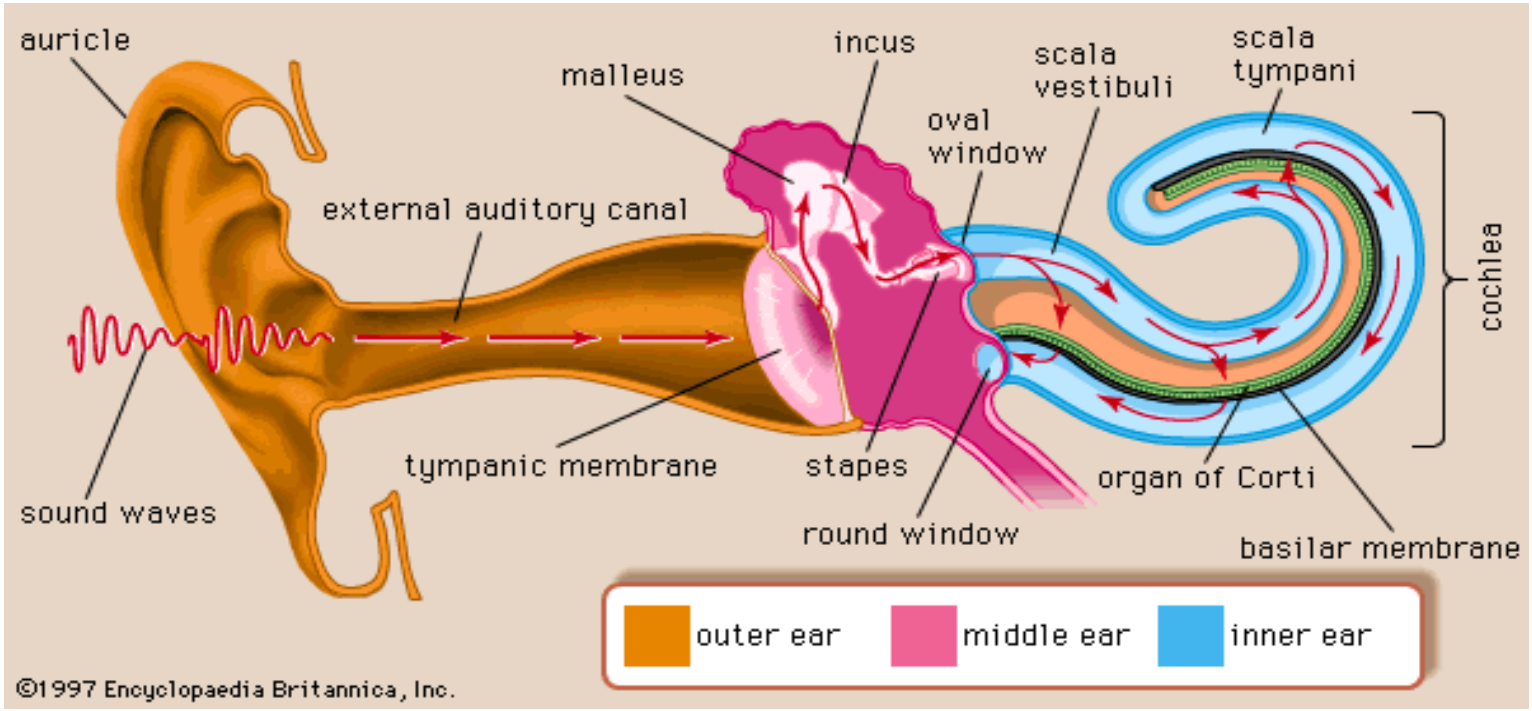


Auditory System

- Ears, parts of brain, and neural pathways
- Changes in pressure move hair-like fibers within the inner ear
- Movements result in electrical impulses sent to the brain

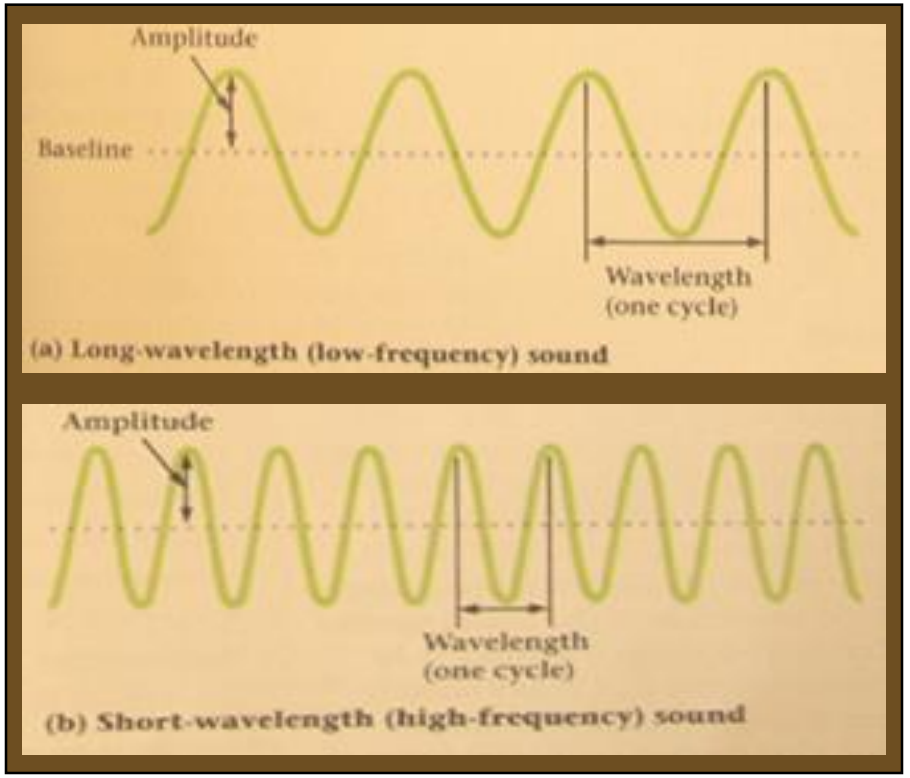


Process of Hearing (Transduction)



Physical Dimensions

- Amplitude
  - height of a cycle
  - relates to loudness
- Wavelength (*w*)
  - distance between peaks
- Frequency (λ)
  - cycles per second
  - relates to pitch
  - λw = velocity
- Most sounds mix many frequencies & amplitudes



Sound is repetitive changes in air pressure over time

Sound Perception and Psychoacoustics

- **Psychoacoustics**
  - Study the correlation between the physics of acoustical stimuli and hearing sensations
  - Experiments data and models are useful for audio codec
- **Modeling human hearing mechanisms**
  - Allows to reduce the data rate while keeping distortion from being audible

Psychological Dimensions

- **Loudness**
  - higher amplitude results in louder sounds
  - measured in decibels (db), 0 db represents hearing threshold
- **Pitch**
  - higher frequencies perceived as higher pitch
  - hear sounds in 20 Hz to 20,000 Hz range

Decibel Scale

- Describes intensity relative to threshold of hearing based on multiples of 10

$$dB = 10 \log \frac{I}{I_0}$$

I_0 is reference level = 10^{-12} W/m^2

Decibels of Everyday Sounds

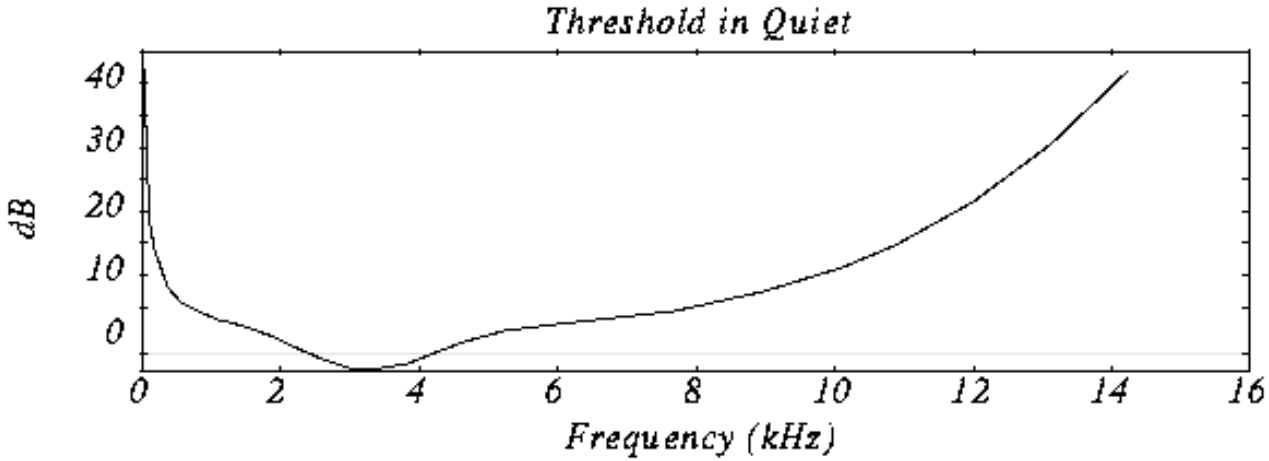
Sound	Decibels
Rustling leaves	**10**
Whisper	**30**
Ambient office noise	**45**
Conversation	**60**
Auto traffic	**80**
Concert	**120**
Jet motor	***140***
Spacecraft launch	***180***

Masking

- Perception of one sound interferes with another
- Frequency masking
- Temporal masking

Frequency Masking

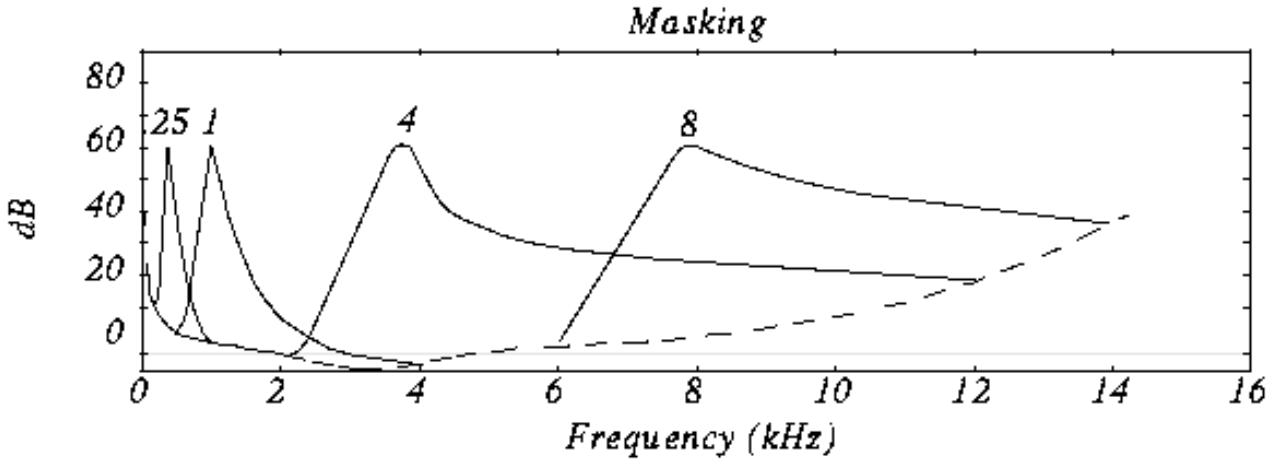
- Louder, lower frequency sounds tend to mask weaker, higher frequency sounds



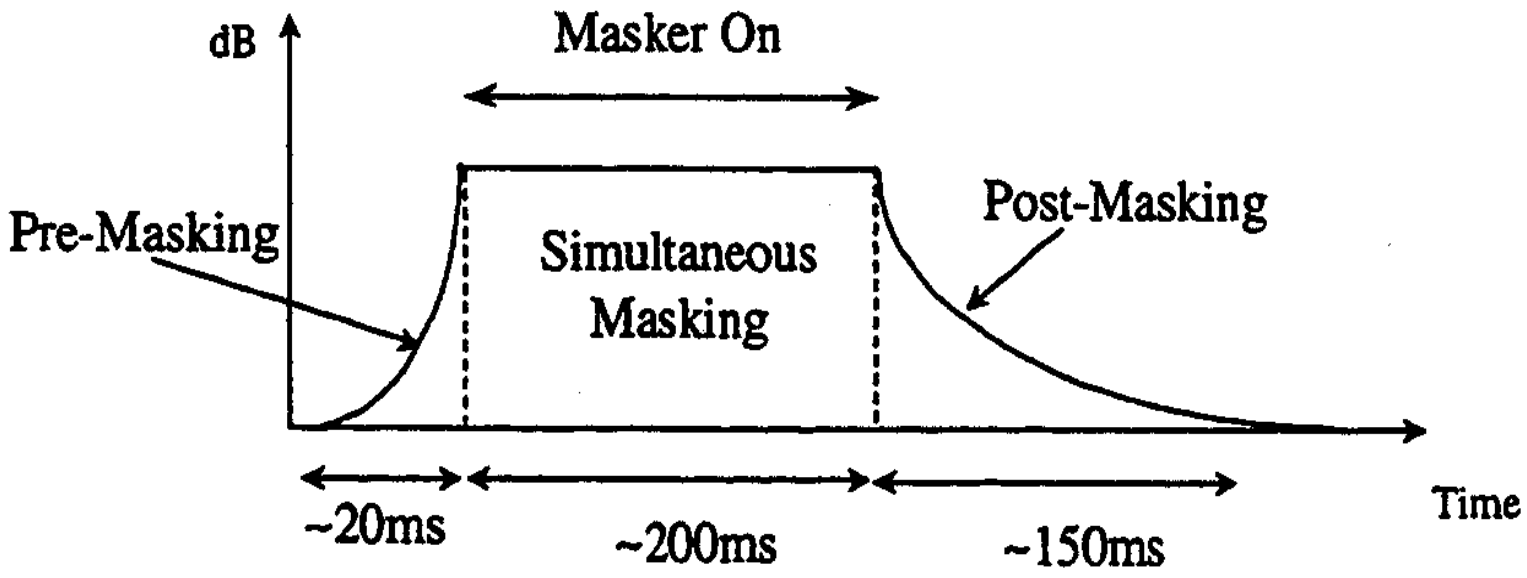
From http://www.cs.sfu.ca/CourseCentral/365/

Frequency Masking

- Louder, lower frequency sounds tend to mask weaker, higher frequency sounds

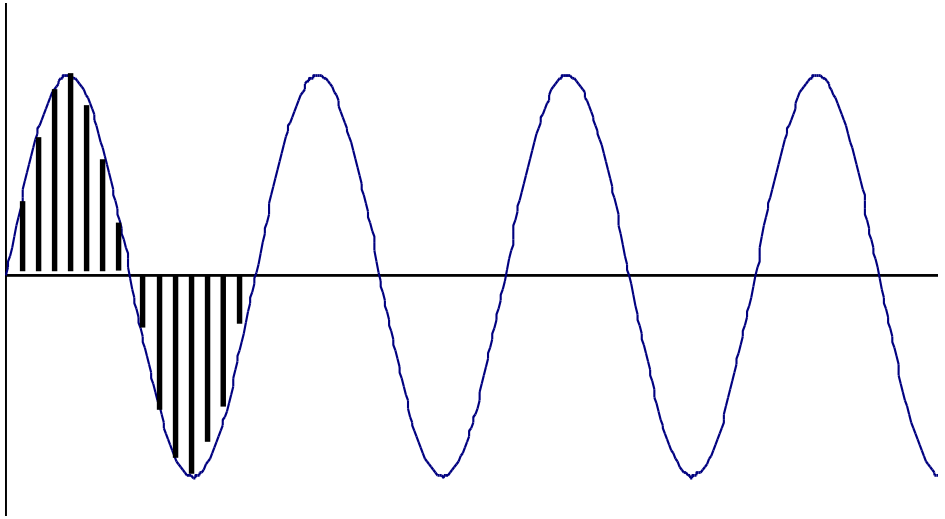


Temporal Masking



Digital Representation of Audio

- Must convert wave form to digital
  - sample
  - quantize
  - compress



Nyquist Theorem

For lossless digitization, the sampling rate should be at least twice the maximum frequency response.

- In mathematical terms:

$$f_s > 2*f_m$$

- where f_s is sampling frequency and f_m is the maximum frequency in the signal

Nyquist Limit

- max data rate = 2 H $\log_2$V *bits/second,* where
  H = bandwidth (in Hz)
  V = discrete levels (bits per signal change)

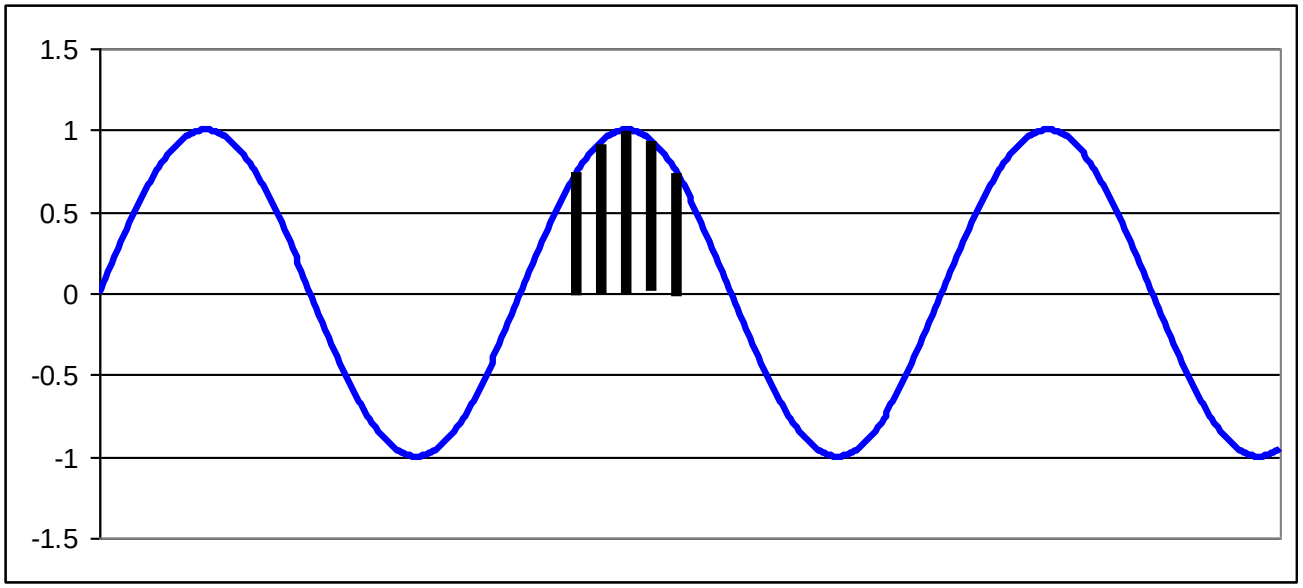
- Shows the maximum number of bits that can be sent per second on a *noiseless* channel with a bandwidth of H, if V bits are sent per signal
  - Example: what is the maximum data rate for a 3kHz channel that transmits data using 2 levels (binary) ?
  - (2x3,000xln2=6,000bits/second)

Sampling Ranges

- Auditory range 20Hz to 22.05 kHz
  - must sample up to to 44.1kHz
  - common examples are 8.000 kHz, 11.025 kHz, 16.000 kHz, 22.05 kHz, and 44.1 KHz
- Speech frequency [200 Hz, 8 kHz]
  - sample up to 16 kHz
  - but typically 4 kHz to 11 kHz is used

Sampling Rates	Used As...
8000	Telephony Standard, Popular in UNIX Workstations
11000	Quarter of CD rate, Popular on Macintosh
16000	G.722 Standard (Federal Standard)
18900	CD-ROM XA Rate
22000	Half CD rate, Macintosh rate
32000	Japanese HDTV, British TV audio, Long play DAT
37800	CD XA Standard
44056	Professional audio industry
44100	CD Rate
48000	DAT Rate

Quantization



Quantization

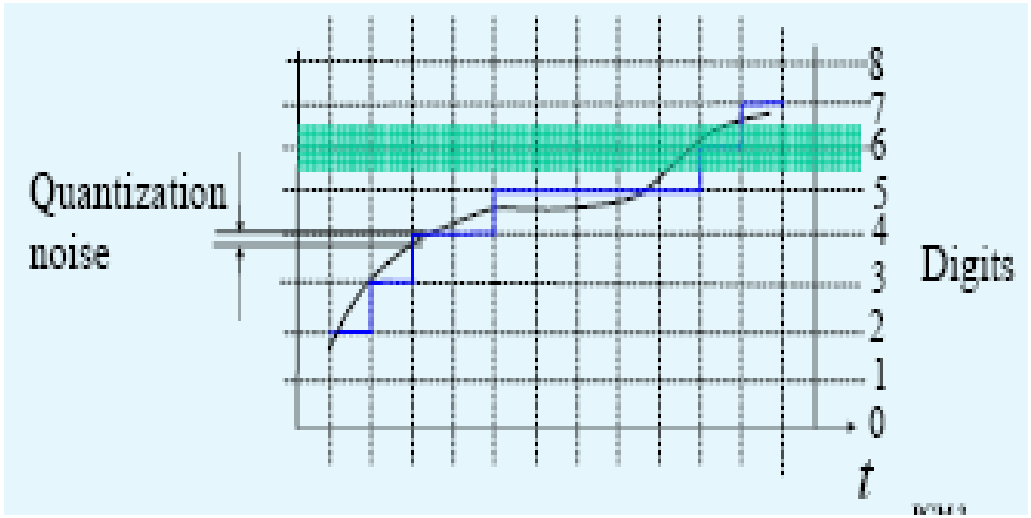
- Typically use
  - 8 bits = 256 levels
  - 16 bits = 65,536 levels
- How should the levels be distributed?
  - Linearly? (PCM)
  - Perceptually? (u-Law)
  - Differential? (DPCM)
  - Adaptively? (ADPCM)

Pulse Code Modulation

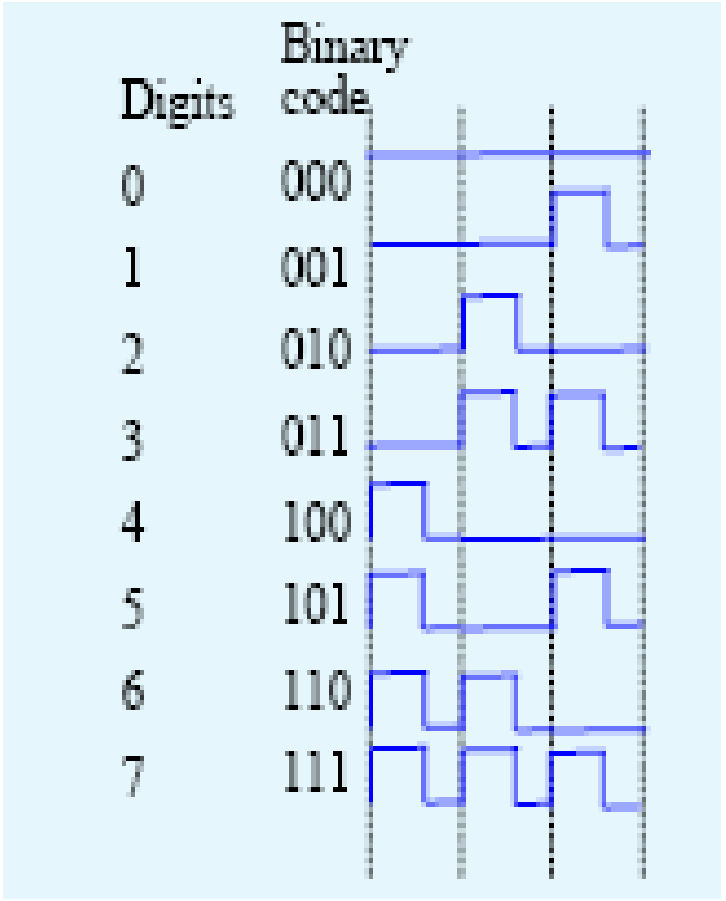
- Pulse modulation
  - Use discrete time samples of analog signals
  - Transmission is composed of analog information sent at different times
  - Variation of pulse amplitude or pulse timing allowed to vary continuously over all values
- PCM
  - Analog signal is quantized into a number of discrete levels

PCM Quantization and Digitization

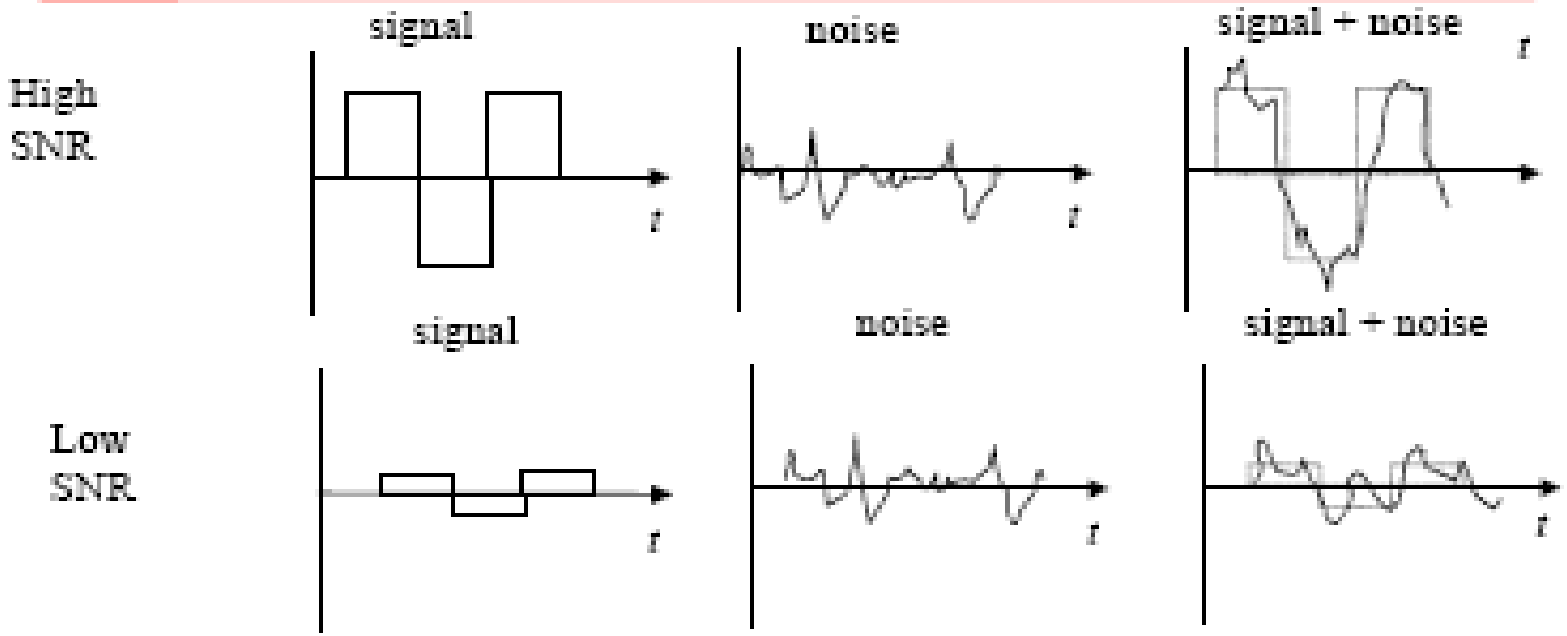
Quantization



Digitization



Signal-to-Noise Ratio



$$\text{SNR} = \frac{\text{Average Signal Power}}{\text{Average Noise Power}}$$

$$\text{SNR (dB)} = 10 \log_{10} \text{SNR}$$

Data Rates

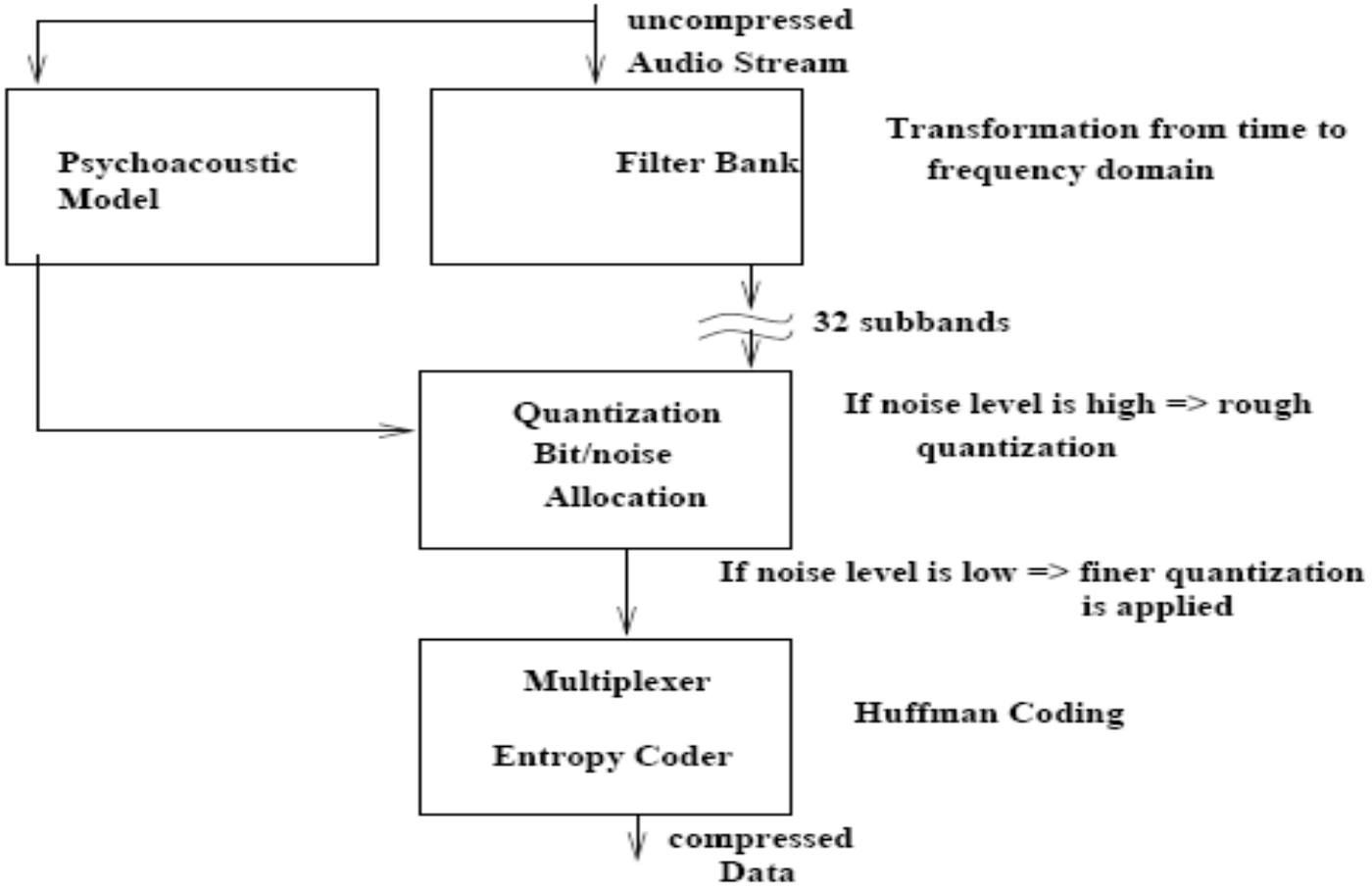
- Data rate = sample rate * quantization * channel
- Compare rates for CD vs. mono audio
  - 8000 samples/second * 8 bits/sample * 1 channel = 8 kBytes / second
  - 44,100 samples/second * 16 bits/sample * 2 channel = 176 kBytes / second ~= 10MB / minute

MPEG AUDIO CODING

MPEG Audio Encoding

- Characteristics
  - Precision 16 bits
  - Sampling frequency: 32KHz, 44.1 KHz, 48 KHz
  - 3 compression layers: Layer 1, Layer 2, Layer 3 (MP3)
    - Layer 3: 32-320 kbps, target 64 kbps
    - Layer 2: 32-384 kbps, target 128 kbps
    - Layer 1: 32-448 kbps, target 192 kbps

MPEG Audio Encoding Steps



MPEG Audio Filter Bank

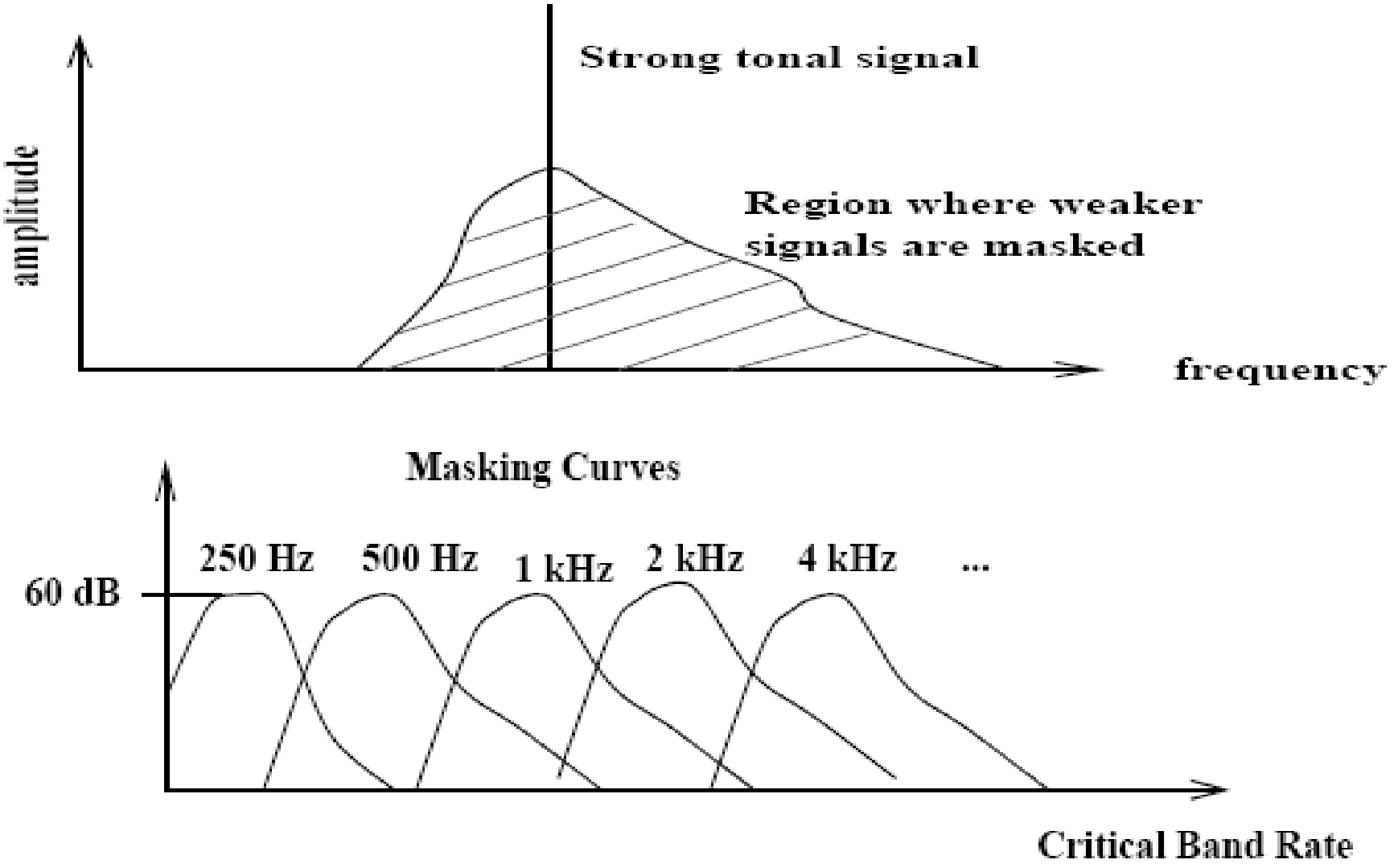
- Filter bank divides input into multiple sub-bands (32 equal frequency sub-bands)
- Sub-band i defined

$$S_t[i] = \sum_{k=0}^{7} 3 \sum_{j=0}^{7} \cos\left(\frac{(2i+1)(k-16)\pi}{64} * (C[k+64j] * x[k+64j] \right)$$

- $i \in [0,31], S_t[i]$ - filter output sample for sub-band I at time t, C[n] – one of 512 coefficients, x[n] – audio input sample from 512 sample buffer

MPEG Audio Psycho-acoustic Model

- MPEG audio compresses by removing acoustically irrelevant parts of audio signals
- Takes advantage of human auditory systems inability to hear quantization noise under auditory masking
- Auditory masking: occurs when ever the presence of a strong audio signal makes a temporal or spectral neighborhood of weaker audio signals imperceptible.

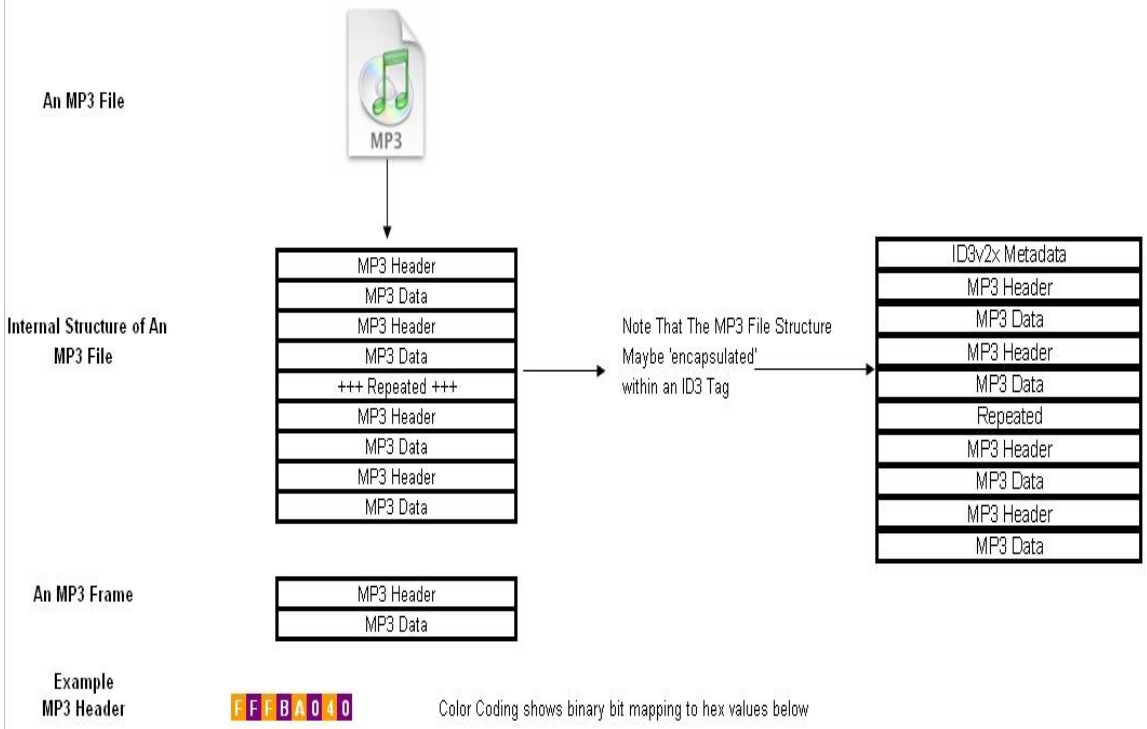


MPEG/audio divides audio signal into frequency sub-bands that approximate critical bands. Then we quantize each sub-band according to the audibility of quantization noise within the band

MPEG Audio Bit Allocation

- This process determines number of code bits allocated to each sub-band based on information from the psycho-acoustic model
- Algorithm:
  1. Compute mask-to-noise ratio: MNR=SNR-SMR
     - Standard provides tables that give estimates for SNR resulting from quantizing to a given number of quantizer levels
  2. Get MNR for each sub-band
  3. Search for sub-band with the lowest MNR
  4. Allocate code bits to this sub-band.
     - If sub-band gets allocated more code bits than appropriate, look up new estimate of SNR and repeat step 1

MP3 Audio Format



Example MP3 Header

FFFBA040 Color Coding shows binary bit mapping to hex values below

Detail Of An MP3 Header

Bits	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Binary	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
Hex	F			F			F							B			A				0						4				0	
Meaning	MP3 Sync Word												Version	Layer		Error Protection	Bit Rate				Frequency		Pad. Bit	Priv. Bit	Mode		Mode Extension (Used With Joint Stereo)		Copy	Original	Emphasis	
Value	Sync Word												1 = MPEG	01 = Layer 3		1=No	1010 = 160				00 = 44100 Hz		0 = Frame is not padded	Unknown	01=Joint Stereo		0 = Intensity Stereo Off	0 = MS Stereo Off	0=Not Copyrighted	0=Copy Of Original Media	00=None	

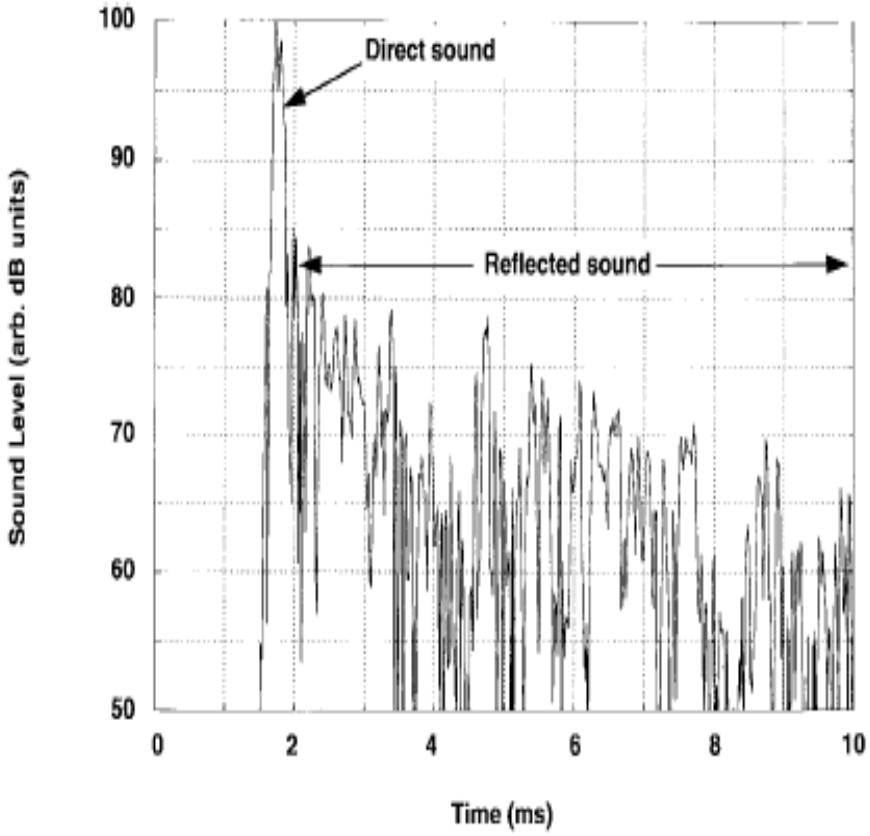
Source: http://wiki.hydrogenaudio.org/images/e/ee/Mp3filestructure.jpg

MPEG Audio Comments

- Precision of 16 bits per sample is needed to get good SNR ratio
- Noise we are getting is quantization noise from the digitization process
- For each added bit, we get 6dB better SNR ratio
- Masking effect means that we can raise the noise floor around a strong sound because the noise will be masked away
- Raising noise floor is the same as using less bits and using less bits is the same as compression

Acoustic Limitations for Spatial Audio in Desktop Environments

- Desktop Environment
  - Two loudspeakers on sides of video monitor
  - Small room (office room, home living room)
- Audible distortion of reproduced sound
  - Effect of **discrete early reflections** - Dominant source of monitoring non-uniformities
    - These non-uniformities appear in the form of frequency-response anomalies in rooms where difference between direct and reflected sound level for firs t15 ms is less than 15 dB



Conclusions

- Audio is very important part of multimedia with its own characteristics and challenges
  - Audio quality is more important than video quality!!
- Systems and networks need to carefully consider the compression techniques audio uses to deliver high quality audio
  - Most well known MP3 and now coming MP4
- Systems and networks need to carefully consider psycho-acoustic features, head-related transfer functions, to enable immersive audio
  - Immersive audio is possible and available (methods exist) but non-trivial to implement with major challenges due to physical and technological challenges

References

- Slides based on Literature Reference
  - D. Pan, “A Tutorial on MPEG/Audio Compression”, 2002, Readings in Multimedia Computing and Networking, Editors Kevin Jeffay and HongJiang Zhang