

EE3414

Multimedia Communication Systems - I

Audio Coding

Yao Wang

Polytechnic University, Brooklyn, NY11201

<http://eeweb.poly.edu/~yao>

Outline

- Psychoacoustic model of human hearing
 - Threshold in quiet
 - Frequency masking
 - Temporal masking
- Basic steps in perceptual audio coding
 - Quantization basics
 - Subband analysis
 - Bit allocation based on masking threshold
- MPEG audio coding
 - MPEG1 audio layers (including MP3) and technical differences
 - MPEG-2 audio coding (BC and AAC)
 - MPEG-4 audio coding

Speech vs. Audio Coding

- Speech coding
 - Targeted for telephony applications
 - High rate waveform-based speech coder: for comfortable, natural sound, use simple predictive coding techniques
 - Low rate model-based speech coders: for intelligible speech, sufficient for communication purposes, use speech-production models (a filter driven by an excitation signal)
- Audio coding
 - For high quality production of music (including speech) in multiple channels
 - Music has a much wider bandwidth and multichannels
 - Waveform-based to retain the natural sound quality
 - Make extensive use of human hearing properties in determining the quantization levels in different frequency bands
 - Each frequency component is quantized with a step-size that depends on the hearing threshold
 - Don't code if the ear cannot hear it!

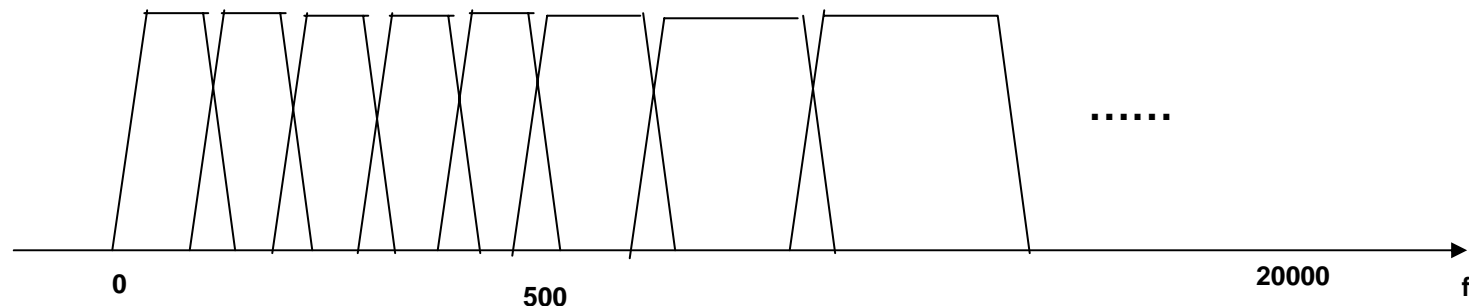
Psychoacoustic Model of Human Hearing

- Ear as a filter bank
- Three masking effects:
 - Threshold in quiet
 - Frequency masking
 - Temporal masking

Ear as a Filterbank

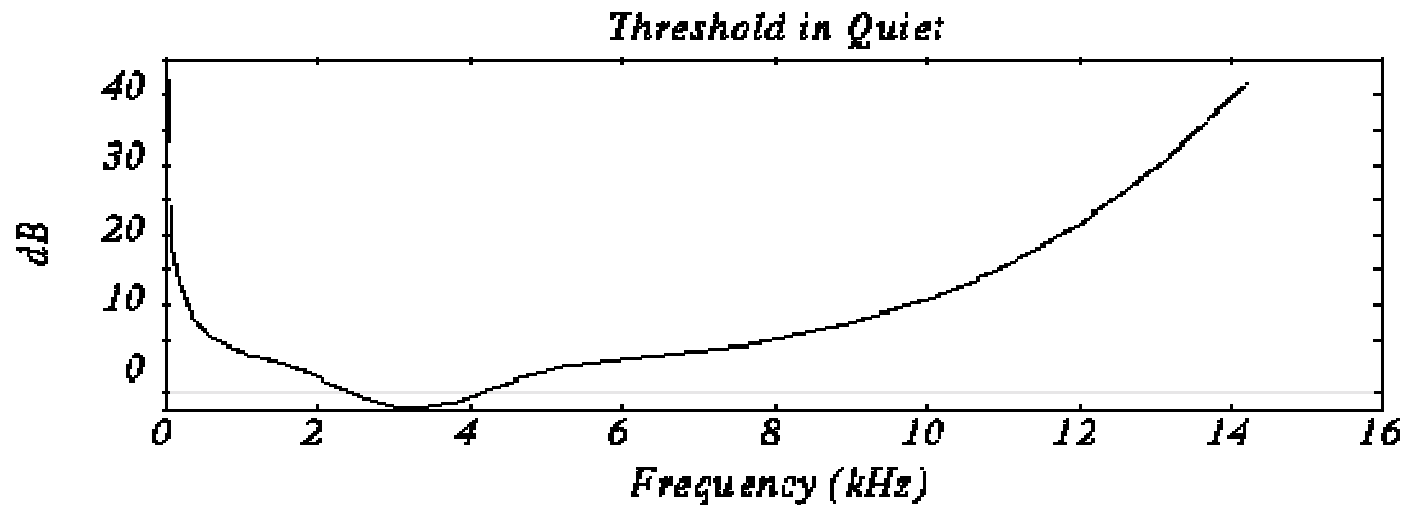
- The auditory system can be roughly modeled as a filterbank, consisting of 25 overlapping bandpass filters, from 0 to 20 KHz
 - The ear cannot distinguish sounds within the same band that occur simultaneously.
 - Each band is called a critical band
 - The bandwidth of each critical band is about 100 Hz for signals below 500 Hz, and increases linearly after 500 Hz up to 5000 Hz
 - 1 bark = width of 1 critical band

$$\text{Bark} = \begin{cases} f/100, & f \leq 500\text{Hz} \\ 9 + 4\log_2(f/1000), & f > 500\text{Hz} \end{cases}$$



Threshold in Quiet

Put a person in a quiet room. Raise level of 1 kHz tone until just barely audible. Vary the frequency and plot

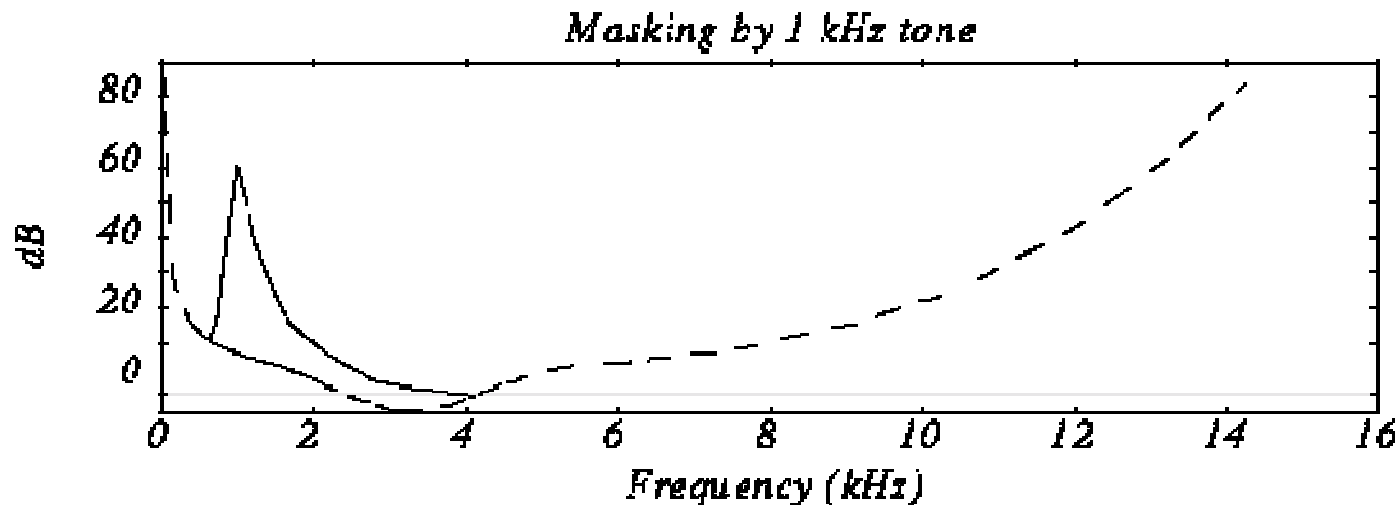


The threshold levels are frequency dependent. The human ear is most sensitive to 2-4 KHz.

From <http://www.cs.sfu.ca/fas-info/cs/CC/365/li/material/notes/Chap4/Chap4.4/Chap4.4.html>

Frequency Masking

Play 1 kHz tone (*masking tone*) at fixed level (60 dB). Play *test tone* at a different level (e.g., 1.1kHz), and raise level until just distinguishable. Vary the frequency of the test tone and plot the threshold when it becomes audible

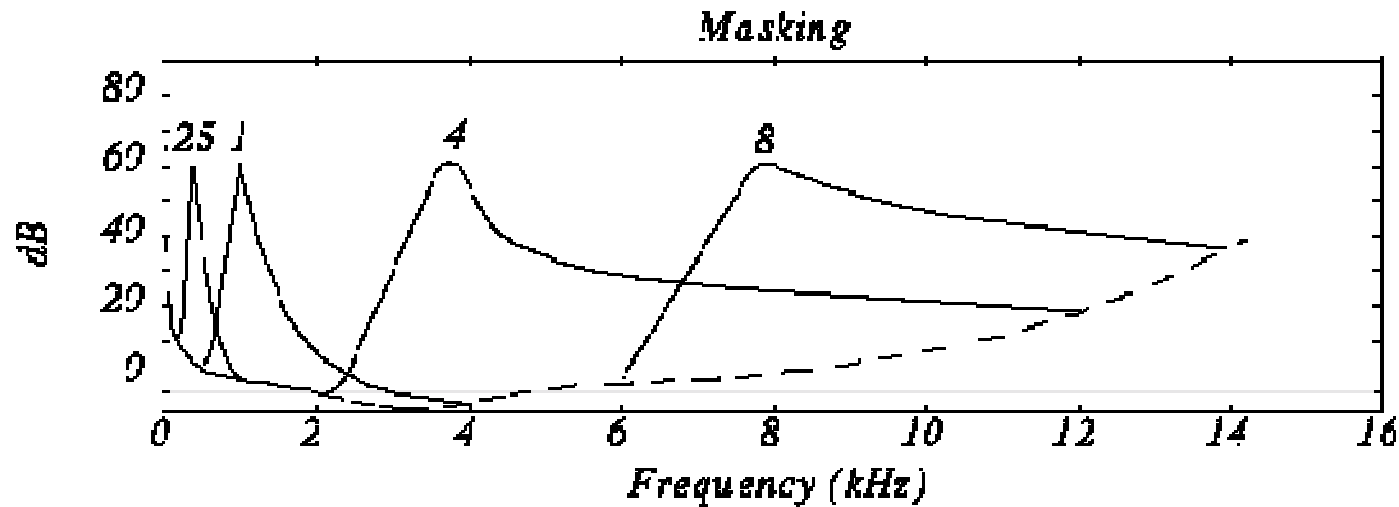


The threshold for the test tone is much larger than the threshold in quiet, near the masking frequency

From <http://www.cs.sfu.ca/fas-info/cs/CC/365/li/material/notes/Chap4/Chap4.4/Chap4.4.html>

Frequency Masking

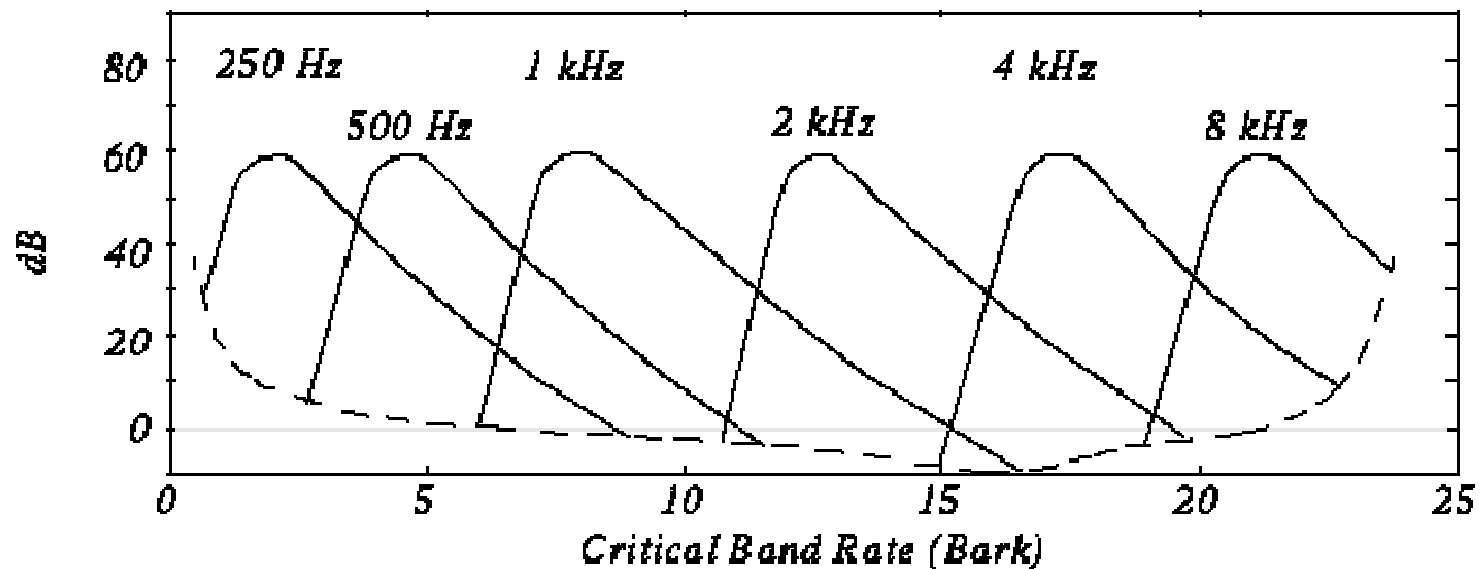
Repeat the previous experiment for various frequencies of masking tones yields



From <http://www.cs.sfu.ca/fas-info/cs/CC/365/li/material/notes/Chap4/Chap4.4/Chap4.4.html>

Frequency Masking on Critical Band Scale

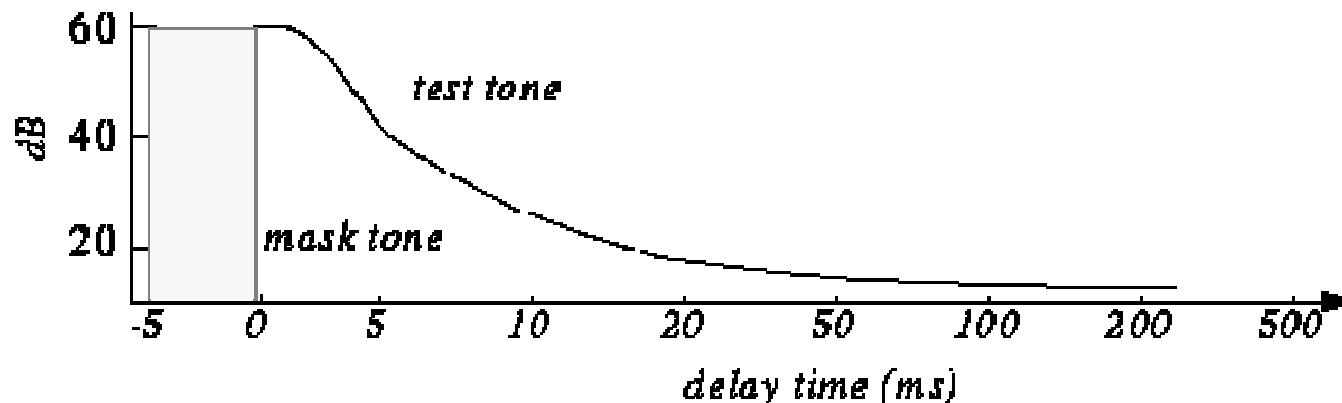
- Critical bands: The widths of the masking bands for different masking tones are different, increasing with the frequency of the masking tone.



From <http://www.cs.sfu.ca/fas-info/cs/CC/365/li/material/notes/Chap4/Chap4.4/Chap4.4.html>

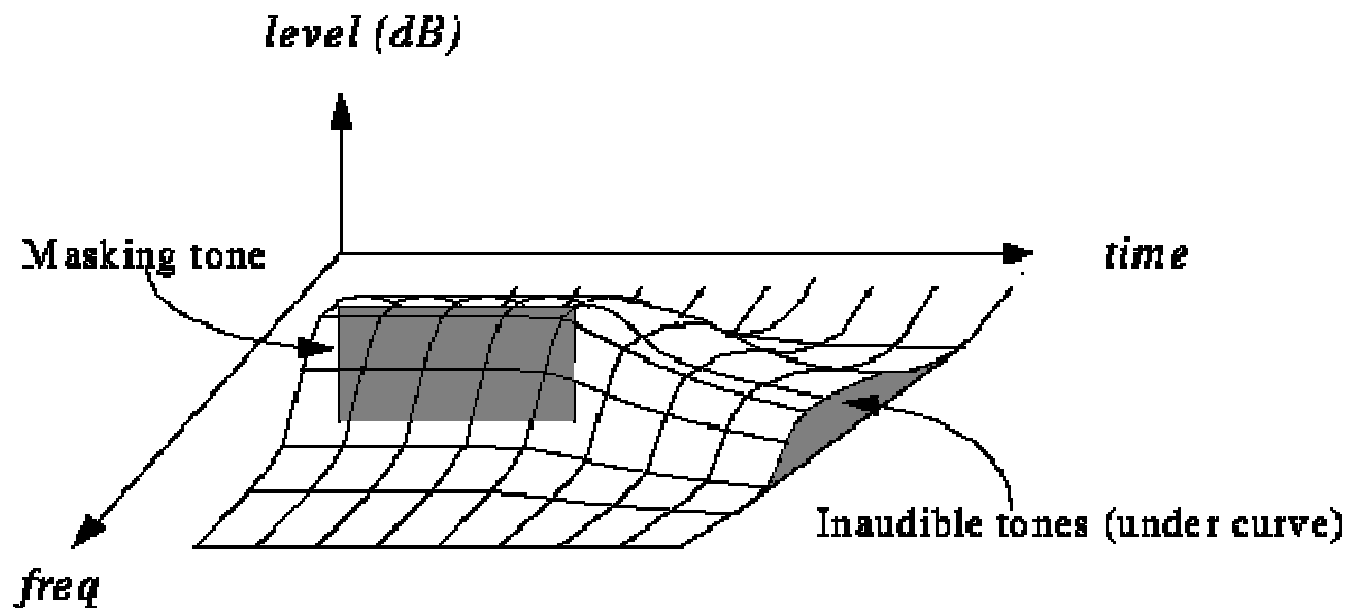
Temporal Masking

- If we hear a loud sound, then it stops, it takes a little while until we can hear a soft tone nearby
- Play 1 kHz *masking tone* at 60 dB, plus a *test tone* at 1.1 kHz at 40 dB. Test tone can't be heard (it's masked). Stop masking tone, and measure the shortest delay time after which the test tone can be heard (e.g., 5 ms). Repeat with different level of the test tone and plot. The weaker is the test tone, the longer it takes to hear it.



From <http://www.cs.sfu.ca/fas-info/cs/CC/365/li/material/notes/Chap4/Chap4.4/Chap4.4.html>

Total Effect of Frequency and Temporal Masking

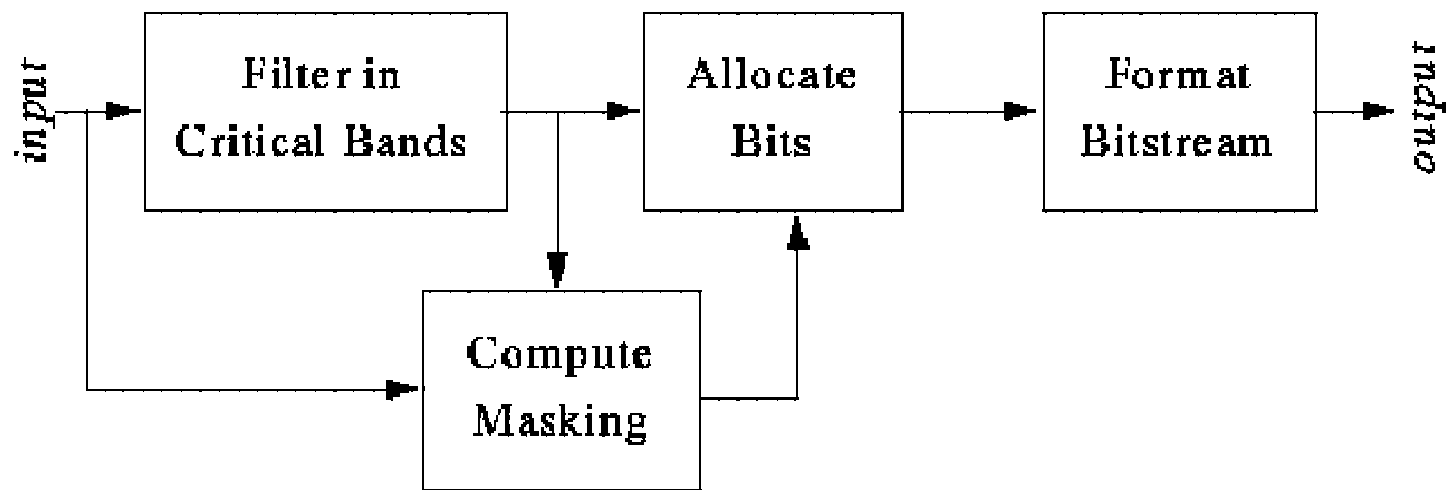


From <http://www.cs.sfu.ca/fas-info/cs/CC/365/li/material/notes/Chap4/Chap4.4/Chap4.4.html>

Perceptual Audio Coding: Basic Ideas

- Decompose a signal into separate frequency bands by using a filter bank
- Analyze signal energy in different bands and determine the total masking threshold of each band because of signals in other band/time
- Quantize samples in different bands with accuracy proportional to the masking level
 - Any signal below the masking level does not need to be coded
 - Signal above the masking level are quantized with a quantization step size according to masking level and bits are assigned across bands so that each additional bit provides maximum reduction in perceived distortion.

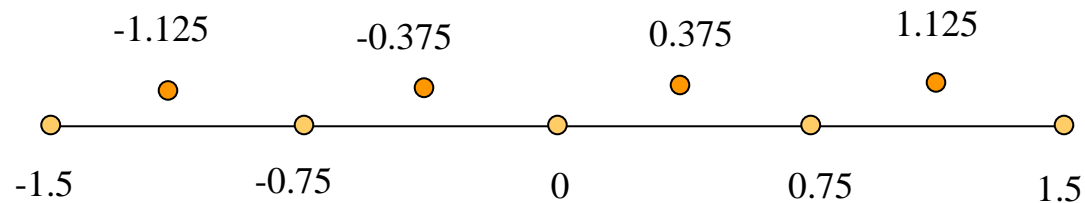
Perceptual Audio Coding Block Diagram



From <http://www.cs.sfu.ca/fas-info/cs/CC/365/li/material/notes/Chap4/Chap4.4/Chap4.4.html>

Quantization Basics: Review (1)

- The quantization error for a uniform quantizer with stepsize Q is approximately uniformly distributed in $(-Q/2, Q/2)$, or with a variance of $Q^2/12$ (this is the quantization noise).



Any number (f) between 0 and 0.75 (Q) is quantized to 0.375 ($Q/2$). Maximum error $q=f-Q(f)$ is $Q/2$ (if $f=0.75$) or $-Q/2$ (if $f=0$). If the source is uniform, then the error q will be uniformly distributed in $(Q/2, -Q/2)$.

$$p(q) = \begin{cases} 1/Q & \text{if } q \in (-Q/2, Q/2); \\ 0 & \text{otherwise} \end{cases}; \quad \sigma_q^2 = \text{Var}(q) = E\{q^2\} = \int_{-Q/2}^{Q/2} q^2 p(q) dq = \frac{Q^2}{12}$$

Quantization Basics: Review (2)

- Assuming the original signal is uniformly distributed over a range of B . With R bits/sample, we can use 2^R levels. The stepsize Q is related to the bit rate R by $Q=B/(2^R)$
- The quantization noise is reduced by 6 dB for every additional bit ($Q \rightarrow Q/2$)

$$SNR = 10 \log_{10} \frac{\sigma_f^2}{\sigma_q^2}$$

$$\sigma_q^2 = \frac{Q^2}{12} = (B/2^R)^2 / 12 = B^2 / 12 * 2^{-2R}$$

$$SNR(R+1) - SNR(R) = 10 \log_{10} \frac{\sigma_q^2(R)}{\sigma_q^2(R+1)}$$

$$= 10 \log_{10} \frac{2^{-2(R)}}{2^{-2(R+1)}} = 10 \log_{10} 2^2 \approx 6 \text{dB}$$

Example

- Assume that the levels of 16 of the 32 bands are :

Band	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Level (db)	0	8	12	10	6	2	10	60	35	20	15	2	3	5	3	1

- Assume that If the level of the 8th band is 60dB, it gives a masking of 12 dB in the 7th band, 15dB in the 9th.
- Level in 7th band is 10 dB (< 12 dB), so ignore it.
- Level in 9th band is 35 dB (> 15 dB), so send it.
- > Can encode with up to 2 bits (= 12 dB) of quantization error. If the original sample is represented with 8 bits, then we can reduce it to 6 bits.

From <http://www.cs.sfu.ca/fas-info/cs/CC/365/li/material/notes/Chap4/Chap4.4/Chap4.4.html>

MPEG Standards Overview

- MPEG: motion picture expert group of the International Standards Organization (ISO)
- MPEG-1: Defines coding standards for both audio and video, and how to packetize the coded audio and video bits to provide time synchronization
 - Total rate: 1.5 Mbps
 - Video (352x240 pels/frame, 30 frame/s): 30 Mbps -> 1.2 Mbps
 - Audio (2 channels, 48 K samples/s, 16 bits/sample): 2*768 kbps -> <=0.3 Mbps
 - Applications: web movies, MP3 audio, video CD
- MPEG-2: for better quality audio and video
 - Video: 720x480 pels/frame, 30 frames/s: 216 Mbps - > 3-5 Mbps
 - Audio (5.1 channels), Advanced audio coding (AAC)
- MPEG-4: targeted for a variety of applications, with wide range of quality and bit rate, but improved quality mainly at low bit rate
 - For internet audio video streaming

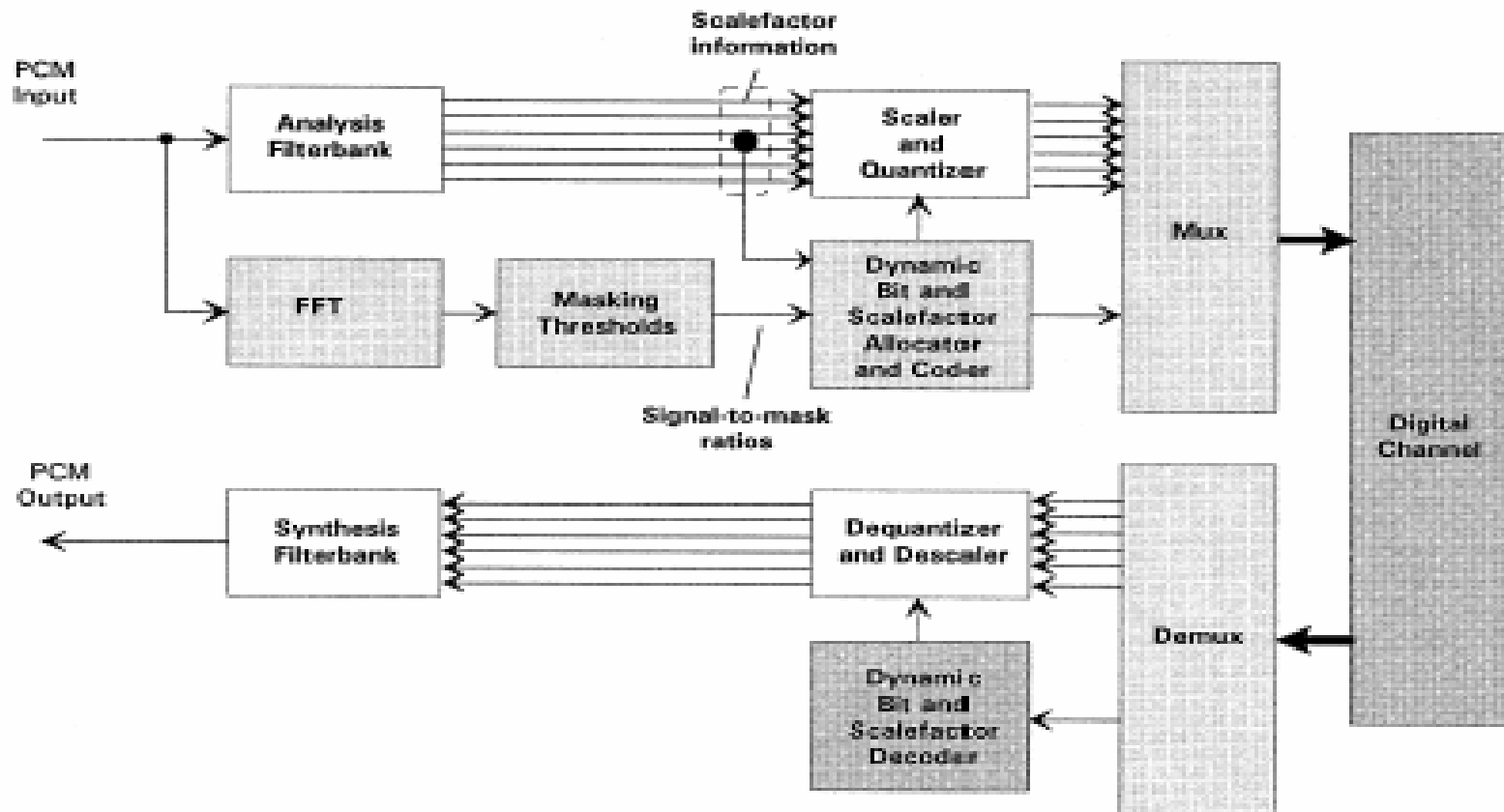
Basic Steps in MPEG-1 Audio Coding

1. Use convolution filters to divide the audio signal into 32 frequency subbands --> *sub-band filtering*.
2. Determine amount of masking for each band based on its frequency (*threshold-in-quiet*), and the energy of its neighboring band in frequency and time (*frequency and temporal masking*) (this is called the *psychoacoustic model*).
3. If the energy in a band is below the masking threshold, don't encode it.
4. Otherwise, determine number of bits needed to represent the coefficient in this band such that the noise introduced by quantization is below the masking effect (Recall that 1 additional bit reduces the quantization noise by 6 dB).
5. Format bitstream: insert proper headers, code the side information, e.g., quantization scale factors for different bands, and finally code the quantized coefficient indices, generally using variable length encoding, e.g. Huffman coding.

MPEG-1 Audio Layers

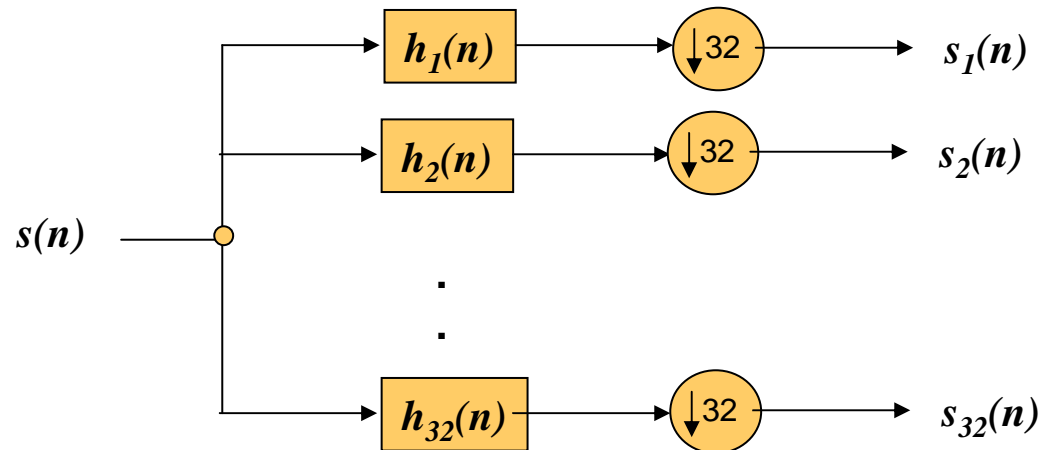
- Layer 1: DCT type filter with equal frequency spread per band. Psychoacoustic model only uses frequency masking.
- Layer 2: Same filter bank as layer 1. Psychoacoustic model uses a little bit of the temporal masking.
- Layer 3 (**MP3**): Layer 1 filterbank followed by MDCT per band to obtain non-uniform frequency division similar to critical bands. Psychoacoustic model includes temporal masking effects, takes into account stereo redundancy, and uses Huffman coder.
- At the time of MPEG1 audio development (finalized 1992), Layer 3 was considered too complex to be practically useful. But today, layer 3 is the most widely deployed audio coding method (known as MP3), because it provides good quality at an acceptable bit rate. It is also because the code for layer 3 is distributed freely.

MPEG Layer I/II Block Diagram



from Peter Noll *MPEG Digital Audio Coding Standards*

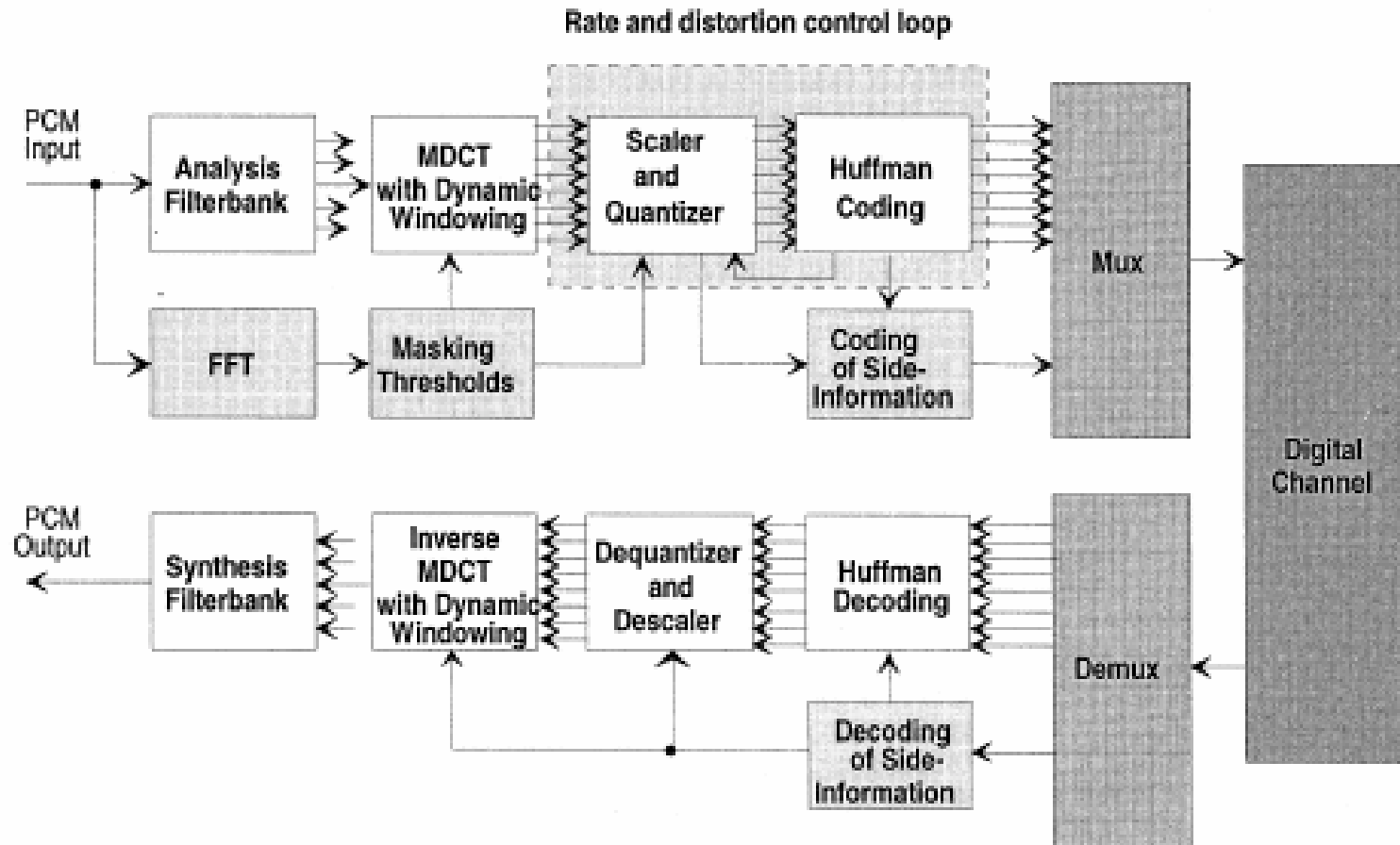
Generating Frequency Bands Using a Filterbank



$$h_{\text{sub}(k)}(n) = h(n) \cos \left[\frac{(2k+1)\pi n}{2M} + \varphi(k) \right];$$
$$M = 32; k = 0, 1, \dots, 31; n = 0, 1, \dots, 511$$

**$h(n)$ is a low-pass prototype filter $h(n)$, 512 samples long
All the filters are obtained by shifting $h(n)$ by modulating with Cosine**

MPEG Layer III Block Diagram



from Peter Noll *MPEG Digital Audio Coding Standards*

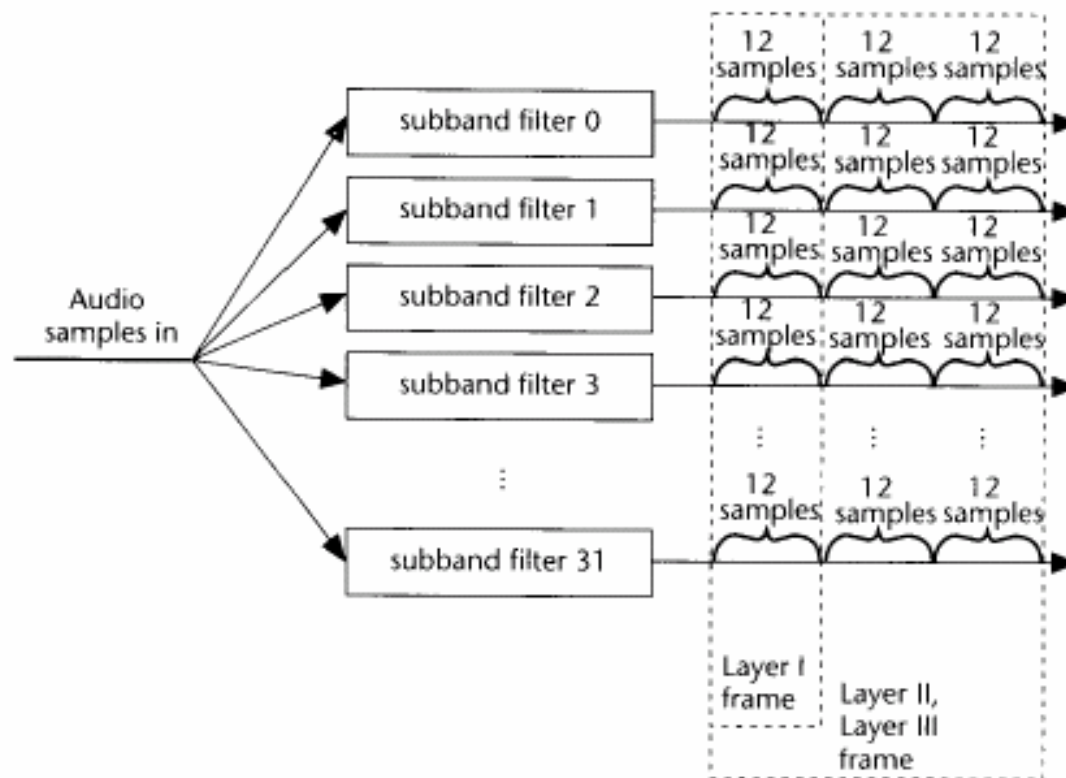
Subband Filtering for Layer 3

- In order to achieve a higher frequency resolution closer to critical band partitions, the 32 subband signals are subdivided further in frequency content by applying, to each of the subbands, a 6- or 18-point modified DCT (MDCT) block transform, with 50% overlap; yielding $32*6=192$ or $32*18=576$ bands.

Subband Filtering and Framing

- Input sequence is separated into 32 frequency bands. Each subband filter produces 1 sample out for every 32 samples in.
- Layer 1 processes 12 samples at a time in each subband. All 12 samples in the same band are scaled by the maximum value and quantized with the same bit allocation.
- Layer 2 and Layer 3 process 36 samples at a time. The 36 samples in the same band are quantized with the same bit allocation, but with 3 separate scale factors, one for each group of 12 samples.

Subband Filtering and Framing



From <http://www.cs.sfu.ca/fas-info/cs/CC/365/li/material/notes/Chap4/Chap4.4/Chap4.4.html>

MPEG-1 Audio Layers: Performance Comparison

Layer	Target bitrate	Ratio	Quality @ 64 kbits	Quality @ 128 kbits
Layer 1	192 kbit	4:1	---	---
Layer 2	128 kbit	6:1	2.1 to 2.6	4+
Layer 3	64 kbit	12:1	3.6 to 3.8	4+

(=MP3)

5 = perfect, 4 = just noticeable, 3 = slightly annoying, 2 = annoying, 1 = very annoying

raw data rate per audio channel: 48 kHz sample/s * 16 bits/sample = 768 kbps

From <http://www.cs.sfu.ca/fas-info/cs/CC/365/li/material/notes/Chap4/Chap4.4/Chap4.4.html>

Performance Comparison

TABLE 40.3 Approximate MPEG-1 Bit Rates for Transparent Representations of Audio Signals and Corresponding Compression Factors (Compared to CD Bit Rate)

MPEG-1 audio coding	Approximate stereo bit rates for transparent quality	Compression factor
Layer I	384 kb/s	4
Layer II	192 kb/s	8
Layer III	128 kb/s ^a	12

^a Average bit rate; variable bit rate coding assumed.

CD bit rate: 44.1 KHz, 16 bits/sample, stereo: $44.1K \cdot 16 \cdot 2 = 1.41Mbps$

From P. Noll, "MPEG digital audio coding standards".

MPEG2 Audio: Overview

- Audio format: 5 Channel (3/2 stereo)
- Two modes:
 - Backward compatible to MPEG1 (BC)
 - Advanced audio coding (AAC)

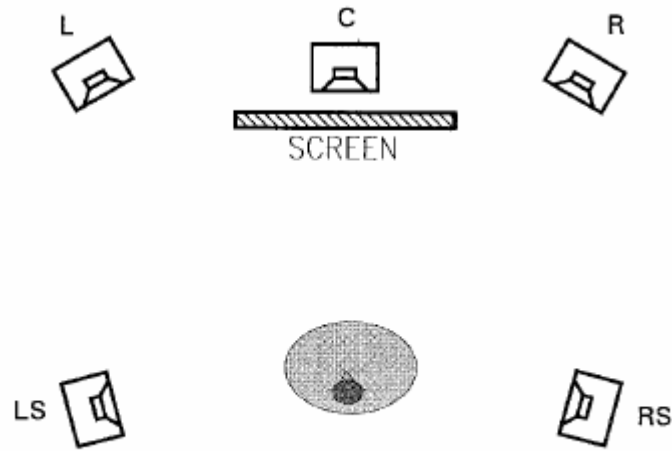


FIGURE 40.15: 3/2 Multichannel loudspeaker configuration.

from Peter Noll *MPEG Digital Audio Coding Standards*

Backward Compatible Mode

- Down-mix 5 channels to left and right signals and code as in MPEG1, and send additional signals for reconstituting the 5 channel as extension signals.

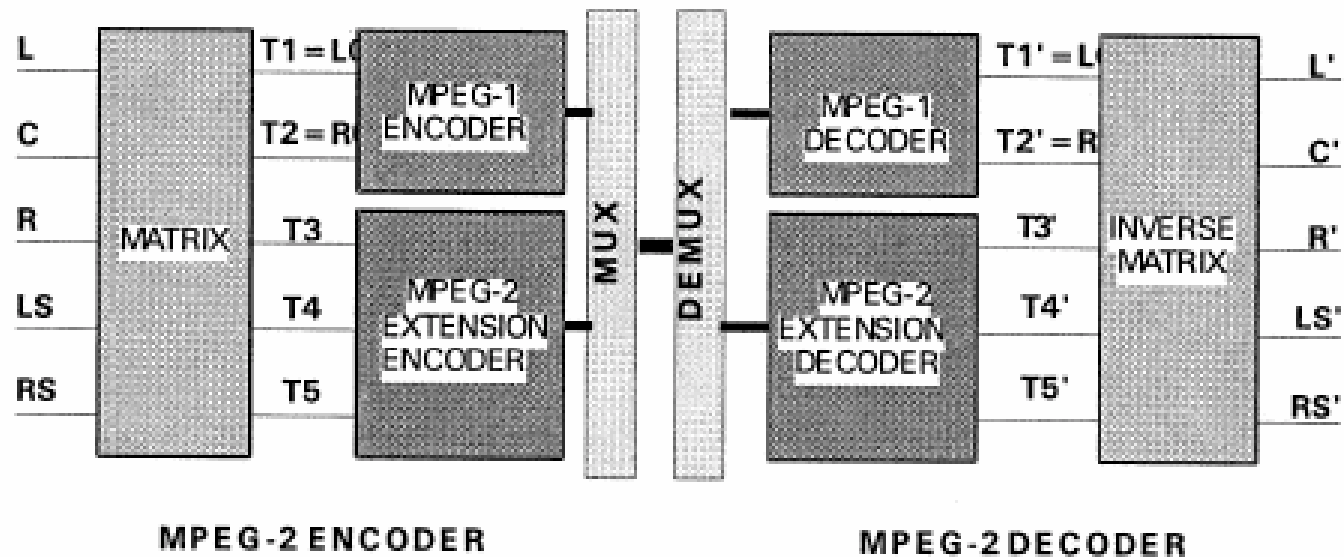


FIGURE 40.16: Compatibility of MPEG-2 multichannel audio bit streams.

from Peter Noll *MPEG Digital Audio Coding Standards*

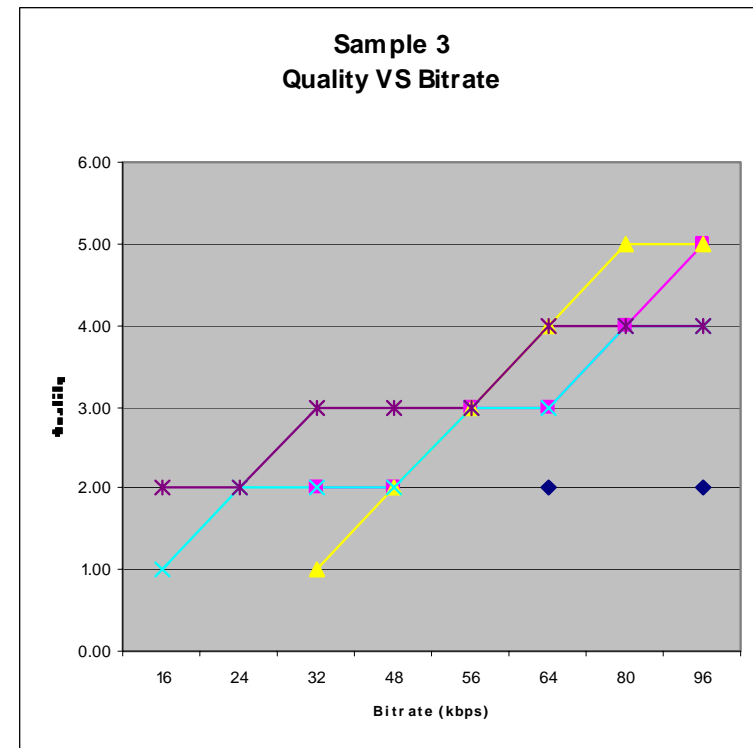
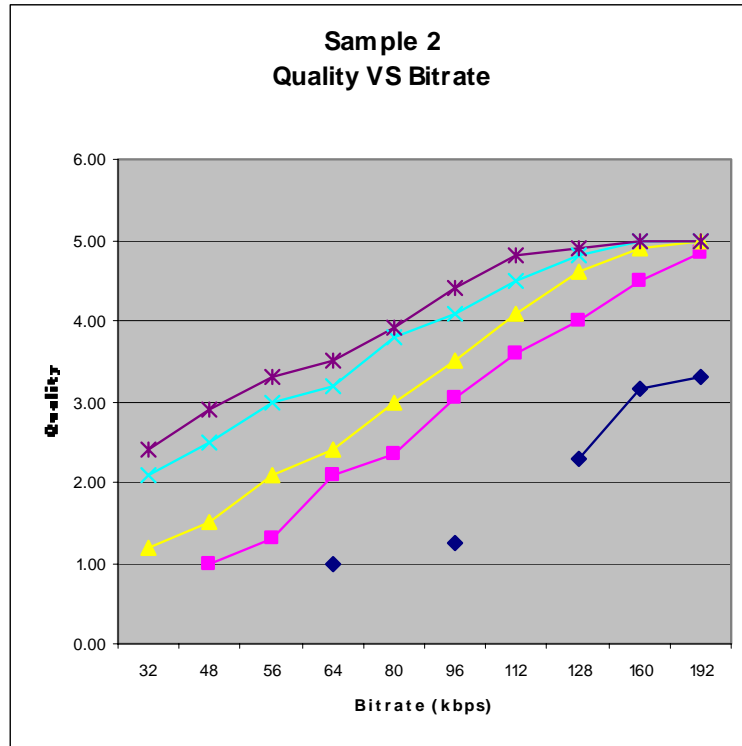
MPEG2 AAC

- Main components:
 - Time to frequency mapping by using filterbank (generating 2048 or 256 bands using MDCT)
 - Temporal noise shaping on the MDCT coefficients
 - Psychoacoustic modeling
 - Quantization and coding
 - Optional Preprocessing
 - Optional temporal prediction
- 3 Profiles
 - Main – Variable length DCT, noiseless coding, etc.
 - Low Complexity – No temporal noise shaping & time domain prediction
 - Sampling Rate Scalability – preprocessor allows for sampling rates of 6, 12, 18, & 24 KHz
- Performance:
 - AAC at 320 kbps and BC at 640 kbps are indistinguishable from original 5 channel audio (3.5 Mbps)
 - AAC can deliver high quality stereo at 128 kbps

MPEG4 Audio: Overview

- Integrates different applications within one framework:
 - Speech, audio, text-to-speech (synthetic audio), MIDI
- Uses 3 Core Coders
 - Parametric coding for low bit rate speech
 - Analysis-by-synthesis for medium bit rates
 - Sub-band/Transform coding for high bit rates (MPEG4 AAC)
- Low Delay (LD) Encoding / Decoding
- Quality Scalability

Quality vs. Bit Rate Testing



- ◆— MPEG 1- Layer I
- MPEG 1- Layer II
- ▲— MPEG 1- Layer III
- ×— MPEG 4 - LC
- *— MPEG 4 - LTP

Test results by Anthony Caliendo & Sherida Subrati, EE3414 S03

Audio Comparison

- Original sample (44.1 KHz, 16 bit/sample, stereo, 1.44mbps)
- Coded at 64 kbps
 - M1L1
 - M1L2
 - M1L3
 - M4LC
 - M4LTD (long term prediction)
- Coded at 128 kbps
 - M1L1
 - M1L2
 - M1L3
 - M4LC
 - M4LTD

Sound created by Anthony Caliendo & Sherida Subrati, EE3414 S03



MP3 Audio PlayList

What should you know?

- The properties of the auditory system
 - Ear as a filterbank
 - Masking effects: threshold-in-quiet, frequency/temporal masking
- Basic components in perceptual audio coding
 - Subband decomposition, bit allocation based on psychoacoustic model, quantization and coding
- MPEG1 audio
 - What are the three layers? What are their differences in techniques and performances
- MPEG2 audio
 - What are the two modes (BC and AAC)
 - How does MPEG2 achieve backward compatibility with MPEG1?
 - How does AAC improves upon MP3?
- MPEG4 audio
 - What are the applications covered?

References

- Peter Noll, MPEG Digital Audio Coding Standards, Chapter in: IEEE Press/CRC Press "The Digital Signal Processing Handbook" (ed.: V.K. Madisetti and D. B. Williams), pp. 40-1 - 40-28, 1998. Available at (<http://www.ff.vu.lt/studentams/tekstai/vizualizavimas/mpeg%20audio%20coding.pdf>) (copies provided)
- Z. N. Li and M. Drew, Fundamentals of multimedia, Prentice Hall, 2004. Chapter 14: MPEG audio compression.
- "Audio compression", <http://www.cs.sfu.ca/fas-info/cs/CC/365/li/material/notes/Chap4/Chap4.4/Chap4.4.html>
- D. Pan, "A Tutorial on MPEG/Audio Compression", IEEE Multimedia, pp. 60-74, summer issue, 1995
- K. Brandenburg, "MP3 and AAC Explained", AES 17th Intl conf. on high quality audio coding, 1999. Available at: <http://mpeg.telecomitalia.com/tutorials.htm>