4.4. Audio Compression

Simple Audio Compression Methods
Psychoacoustics
MPEG Audio Compression

- *Reference:* Chapter 6 of Steinmetz and Nahrstedt

- *Reference:* Davis Pan, "A Tutorial on MPEG/Audio Compression", IEEE Multimedia, Vol. 2, No. 2, pp. 60-74, 1995.

4.4.1 Simple Audio Compression Methods

- Traditional lossless compression methods (Huffman, LZW, etc.) usually don't work well on audio compression (the same reason as in image compression).

The following are some of the Lossy methods:

- Silence Compression - detect the "silence", similar to run-length coding

- Adaptive Differential Pulse Code Modulation (ADPCM)

  e.g., in CCITT G.721 -- 16 or 32 Kbits/sec.

  (a) Encodes the difference between two or more consecutive signals; the difference is then quantized --> hence the loss
  (b) Adapts at quantization so fewer bits are used when the value is smaller.

  - It is necessary to predict where the waveform is headed --> difficult

  - Apple has proprietary scheme called ACE/MACE. Lossy scheme that tries to predict where wave will go in next sample. About 2:1 compression.

- Linear Predictive Coding (LPC) fits signal to speech model and then transmits parameters of model --> sounds like a computer talking, 2.4 kbits/sec.

- Code Excited Linear Predictor (CELP) does LPC, but also transmits error term --> audio conferencing quality at 4.8 kbits/sec.

4.4.2 Psychoacoustics

Human hearing and voice

- Frequency range is about 20 Hz to 20 kHz, most sensitive at 2 to 4 KHz.
- Dynamic range (quietest to loudest) is about 96 dB
- Normal voice range is about 500 Hz to 2 kHz
  - Low frequencies are vowels and bass
  - High frequencies are consonants

Critical Bands

- Human auditory system has a limited, frequency-dependent resolution. The perceptually uniform measure of frequency can be expressed in terms of the width of the *Critical Bands.*
  It is less than 100 Hz at the lowest audible frequencies, and more than 4 kHz at the high end. Altogether, the audio frequency range can be partitioned into 25 critical bands.
- A new unit for frequency *bark* (after Barkhausen) is introduced:

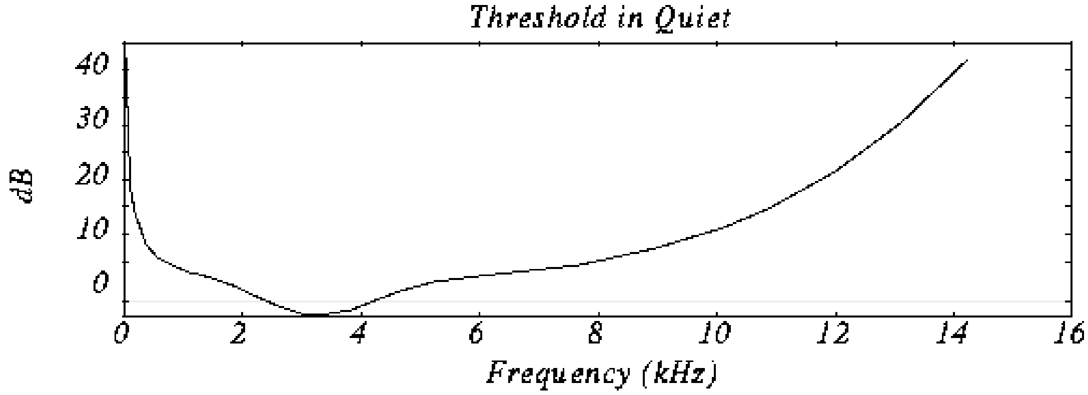
  1 Bark = width of one critical band

  For frequency < 500 Hz, it converts to $freq / 100$ Bark,

  For frequency > 500 Hz, it is $9 + 4 \cdot \log_2(freq/1000)$ Bark.

Sensitivity of human hearing in relation to frequency

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

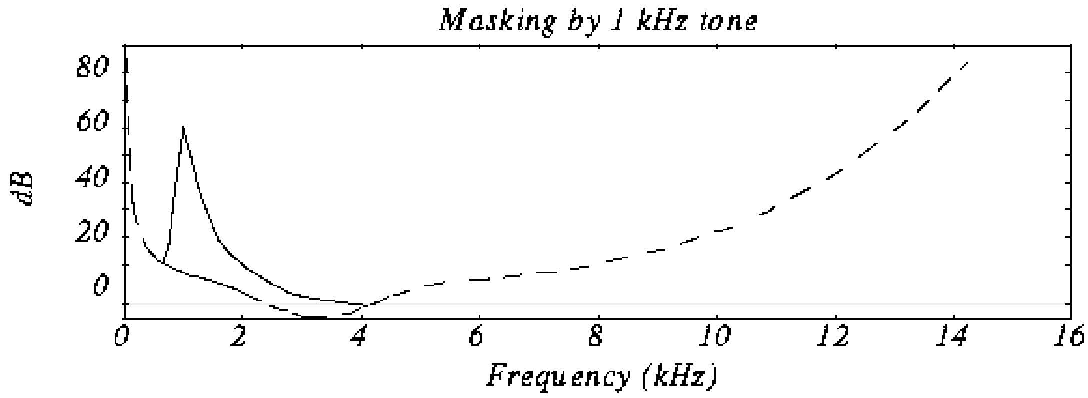


Frequency Masking

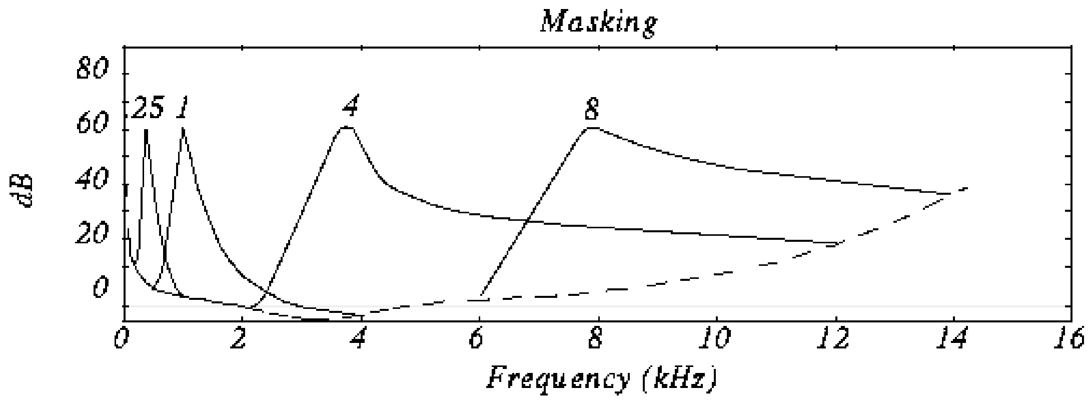
Question: Do receptors interfere with each other?

- Experiment: Play 1 kHz tone (*masking tone*) at fixed level (60 dB). Play *test tone* at a different level (e.g., 1.1 kHz), and raise level until just distinguishable.

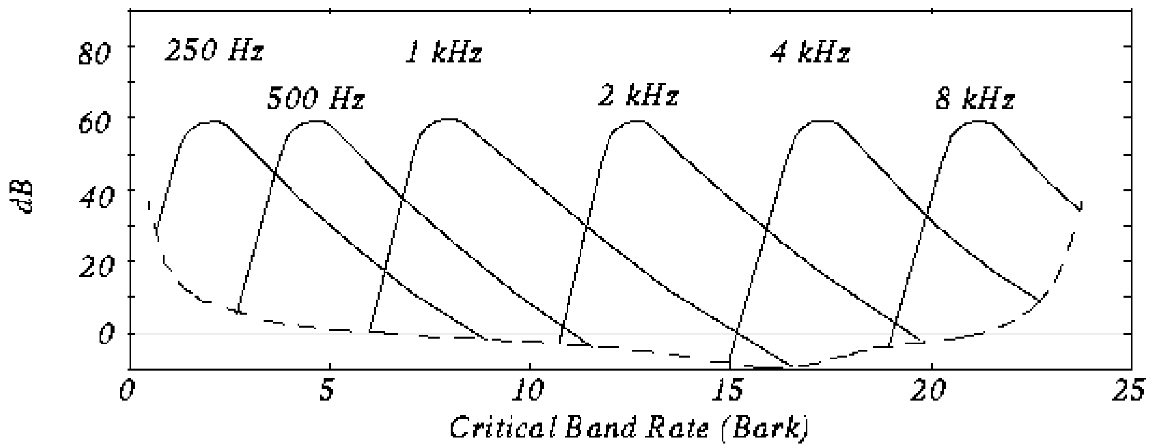
- Vary the frequency of the test tone and plot the threshold when it becomes audible:



- Repeat for various frequencies of masking tones



- Frequency Masking on critical band scale:



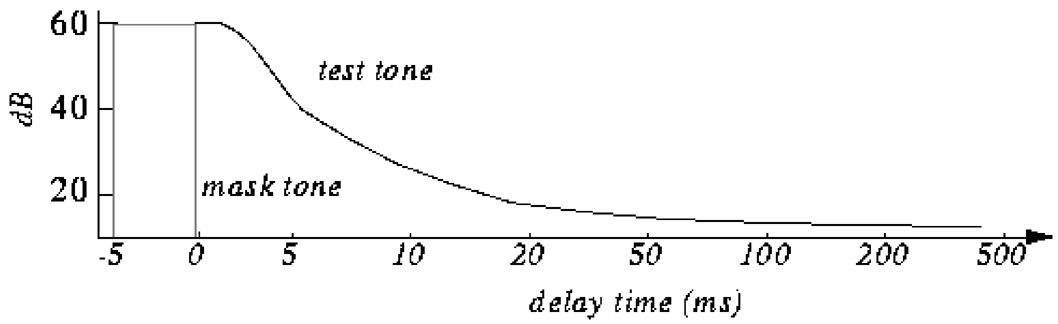
Temporal masking

- If we hear a loud sound, then it stops, it takes a little while until we can hear a soft tone nearby.
- Experiment: 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).

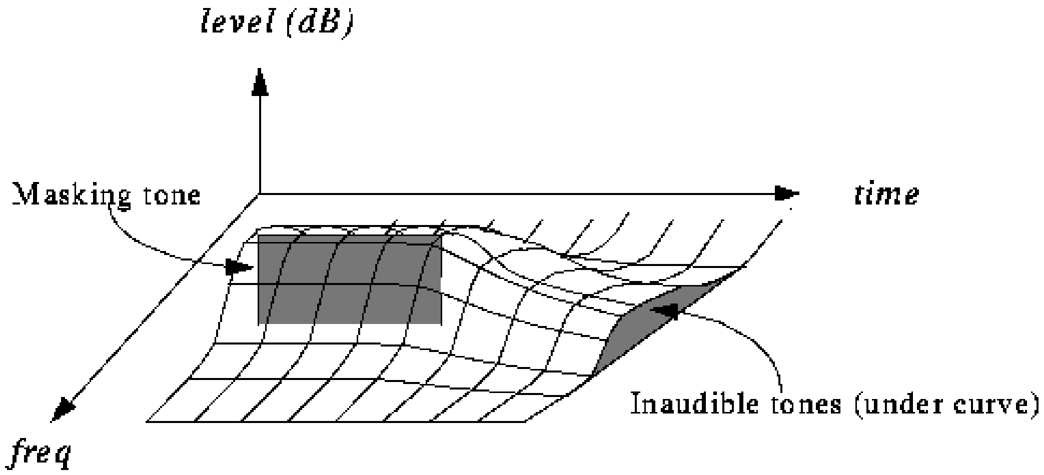
  Stop masking tone, then stop test tone after a short delay.

  Adjust delay time to the shortest time when test tone can be heard (e.g., 5 ms).

  Repeat with different level of the test tone and plot:



- Total effect of both frequency and temporal maskings:



4.4.3 MPEG Audio Compression

Some facts

- MPEG-1: 1.5 Mbits/sec for audio and video

  About 1.2 Mbits/sec for video, 0.3 Mbits/sec for audio

  (Uncompressed CD audio is 44,100 samples/sec * 16 bits/sample * 2 channels > 1.4 Mbits/sec)

- Compression factor ranging from 2.7 to 24.
- With Compression rate 6:1 (16 bits stereo sampled at 48 KHz is reduced to 256 kbits/sec) and optimal listening conditions, expert listeners could not distinguish between coded and original audio clips.
- MPEG audio supports sampling frequencies of 32, 44.1 and 48 KHz.
- Supports one or two audio channels in one of the four modes:

  1. Monophonic -- single audio channel
  2. Dual-monophonic -- two independent channels, e.g., English and French
  3. Stereo -- for stereo channels that share bits, but not using Joint-stereo coding
  4. Joint-stereo -- takes advantage of the correlations between stereo channels

Steps in algorithm:

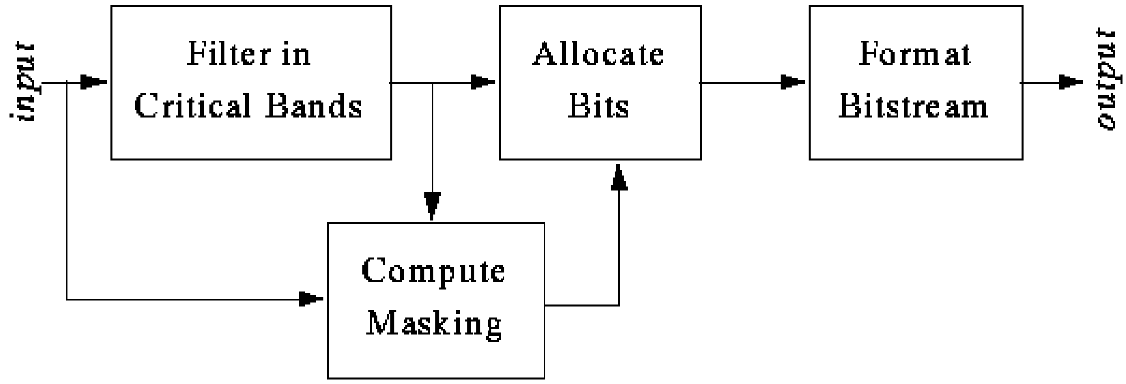
1. Use convolution filters to divide the audio signal (e.g., 48 kHz sound) into 32 frequency subbands --> *subband filtering*.

2. Determine amount of masking for each band caused by nearby band using the *psychoacoustic model* shown above.

3. If the power in a band is below the masking threshold, don't encode it.

4. Otherwise, determine number of bits needed to represent the coefficient such that noise introduced by quantization is below the masking effect (Recall that one fewer bit of quantization introduces about 6 dB of noise).

5. Format bitstream



Example:

- After analysis, the first levels of 16 of the 32 bands are these:

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

- 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.
  [Only the amount above the masking level needs to be sent, so instead of using 6 bits to encode it, we can use 4 bits -- a saving of 2 bits (= 12 dB).]

MPEG Layers

- MPEG defines 3 layers for audio. Basic model is same, but codec complexity increases with each layer.

- Divides data into frames, each of them contains 384 samples, 12 samples from each of the 32 filtered subbands as shown below.

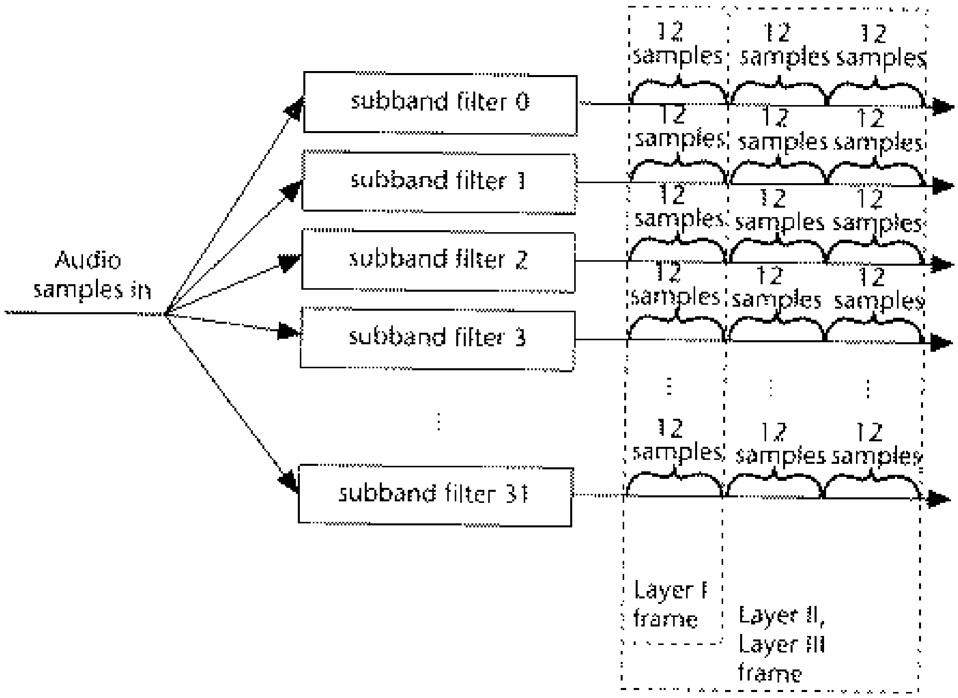


Figure: Grouping of Subband Samples for Layer 1, 2, and 3

- Layer 1: DCT type filter with one frame and equal frequency spread per band. Psychoacoustic model only uses frequency masking.

- Layer 2: Use three frames in filter (before, current, next, a total of 1152 samples). This models a little bit of the temporal masking.

- Layer 3: Better critical band filter is used (non-equal frequencies), psychoacoustic model includes temporal masking effects, takes into account stereo redundancy, and uses Huffman coder.

  Stereo Redundancy Coding:

  - Intensity stereo coding -- at upper-frequency subbands, encode summed signals instead of independent signals from left and right channels.

  - Middle/Side (MS) stereo coding -- encode middle (sum of left and right) and side (difference of left and right) channels.

Effectiveness of MPEG audio

Layer	Target Bit-rate	Ratio	Quality at 64 kb/s	Quality at 128 kb/s	Theoretical Min. Delay
Layer 1	192 kb/s	4:1	---	---	19 ms
Layer 2	128 kb/s	6:1	2.1 to 2.6	4+	35 ms
Layer 3	64 kb/s	12:1	3.6 to 3.8	4+	59 ms

- Quality factor: 5 - perfect, 4 - just noticeable, 3 - slightly annoying, 2 - annoying, 1 - very annoying
- Real delay is about 3 times of the theoretical delay

Further Exploration

MPEG Resources on the Web.

Last Updated: ***3/23/00***