5: Music Compression

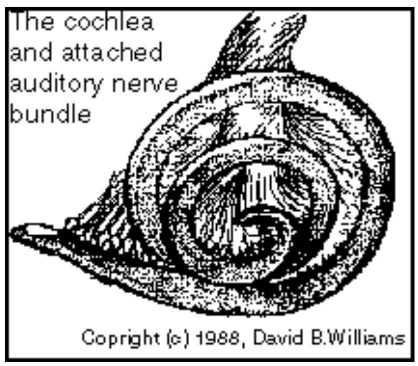
Mark Handley

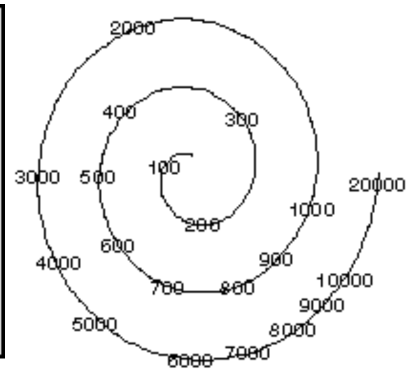
Music Coding

- LPC-based codecs model the sound source to achieve good compression.
  - Works well for voice.
  - Terrible for music.
- What if you can't model the source?
  - Model the limitations of the human ear.
  - Not all sounds in the sampled audio can actually be heard.
  - Analyze the audio and send only the sounds that can be heard.
  - Quantize more coarsely where noise will be less audible.

Human Ear

- Cochlea does a mechanical fourier transform





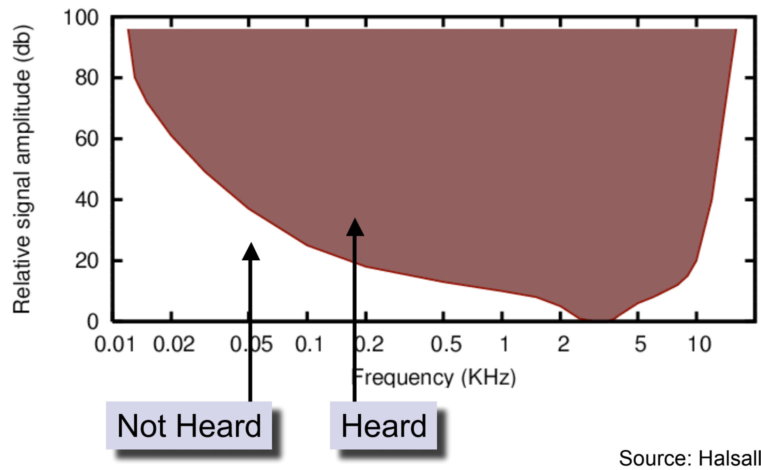
From http://www.hitl.washington.edu/publications/hollander/2.html

Amplitude Sensitivity

- Dynamic range is ratio of maximum signal amplitude to minimum signal amplitide (measured in decibels).
  - $D = 20 \log (A_{max}/A_{min})$ dB
  - Human hearing has dynamic range of ~ 96dB
- Sensitivity of the ear is dependent on frequency.
  - Most sensitive in range of 2-5KHz.

Amplitude Sensitivity

Frequencies only heard if they exceed a sensitivity threshold:

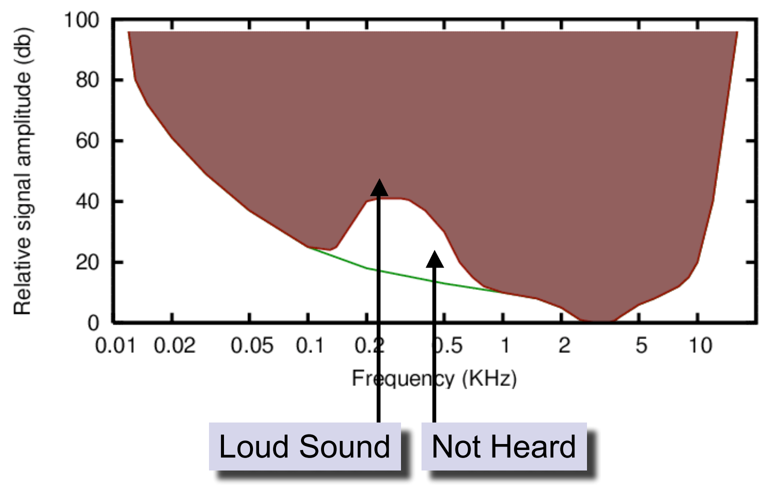


Source: Halsall, p184

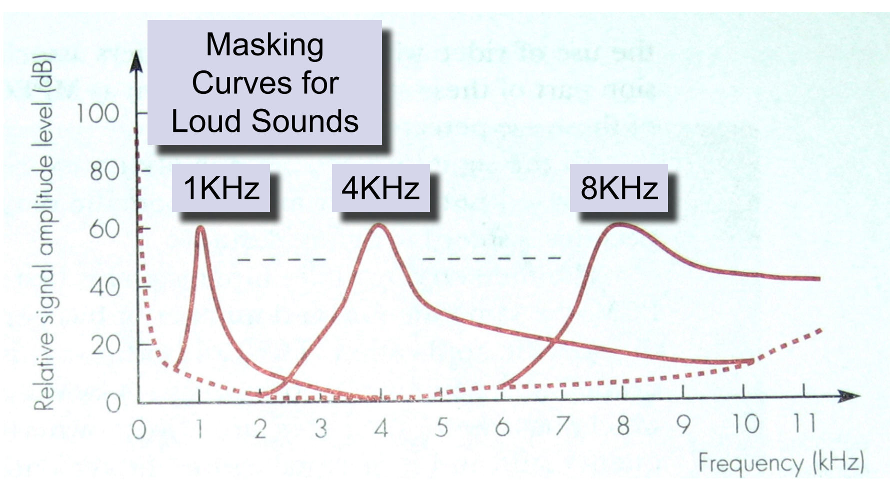
Frequency Masking

- The sensitivity threshold curve is distorted by the presence of loud sounds.
  - Frequencies just above and below the frequency of a loud sound need to be louder than the normal minimum amplitude before they can be heard.

Frequency Masking

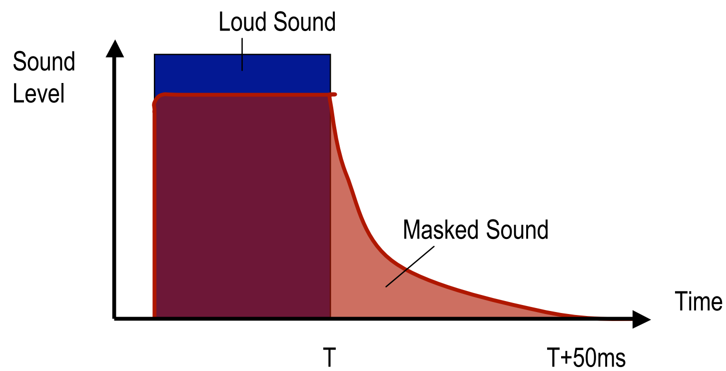


Frequency Masking



Temporal Masking

- After hearing a loud sound, the ear is deaf to quieter sounds in the same frequency range for a short time.



MPEG Audio

Layer	Application	Quality
1	Digital Audio Cassette	Hi-fi quality at 192Kb/s per channel
2	Digital Audio and Digital Video Broadcasting	Near CD-quality at 128Kb/s per channel
3	CD-quality audio over low bitrate channels	CD-quality at 64Kb/s per channel

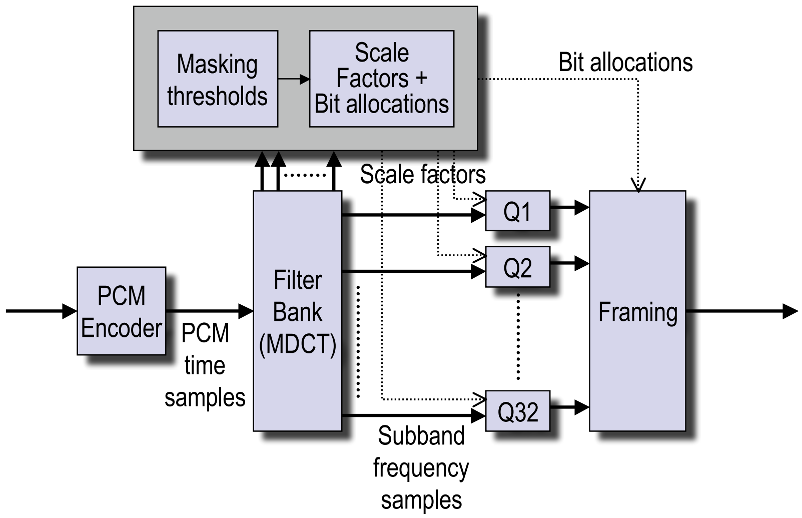
MPEG Audio Encoding

- Sample audio as PCM (typically 16 bit linear). 12 sets of 32 samples.
- Use filter bank to divide signal into 32 frequency bands
  - Maps time-domain samples into 12 values for each of 32 frequency subbands.
- Determine power in each subband.
- Use psychoacoustic model to predict masking for each subband.
  - If power in a subband is below masking threshold, don't code it.
  - Otherwise determine the number of bits needed to code the subband such that the quantization noise is below the masking threshhold.
    - [One fewer bit of quantization introduces ~6dB of noise]

MPEG Audio Layers

- Layer 1:
  - DCT type filter with one frame and equal frequency spread per band.
  - Psychoacoustic model only uses frequency masking.
- Layer 2:
  - Uses three frames in filter (1152 samples).
  - More compact encoding of scale factors and samples.
- Layer 3:
  - Better critical band filter is used (non-equal frequencies)
  - Psychoacoustic model includes temporal masking effects.
  - Takes into account stereo redundancy
  - Huffman encoding of quantized samples.

MPEG layer 1, 2 Encoder



MPEG Audio Layers

- Layer 1:
  - DCT type filter with one frame and equal frequency spread per band.
  - Psychoacoustic model only uses frequency masking.
- Layer 2:
  - Uses three frames in filter (1152 samples).
  - More compact encoding of scale factors and samples.
- Layer 3:
  - Better critical band filter is used (non-equal frequencies)
  - Psychoacoustic model includes temporal masking effects.
  - Takes into account stereo redundancy
  - Huffman encoding of quantized samples.

Huffman Coding

- Variable length coding, with most frequent codes using fewest bits and less frequent codes using more bits.
  - Provably optimal.

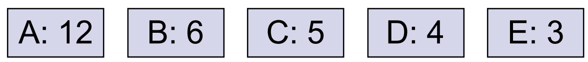
- Encoding done by building an encoding tree.
- Tree is built bottom up, based on frequencies of characters from the “text”.

Huffman Coding: Example

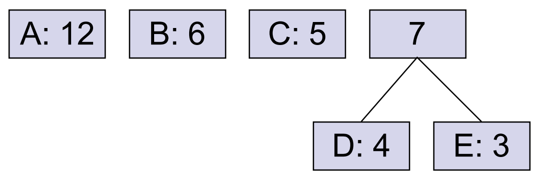
- Eg: five letter alphabet, A,B,C,D,E with frequencies in the text of A: 12, B: 6, C: 5, D: 4, E: 3
- Start with five separate subtrees:

A: 12	B: 6	C: 5	D: 4	E: 3

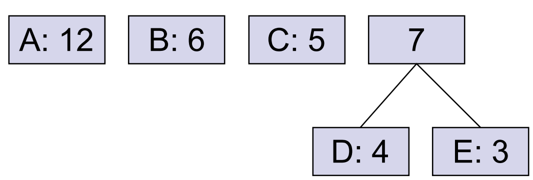
Huffman Coding: Example



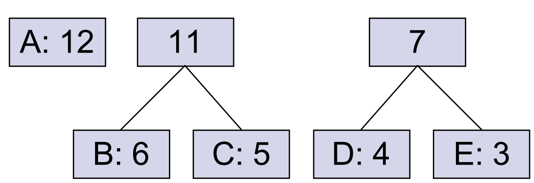
Subtrees for D and E have lowest frequencies, so merge them:



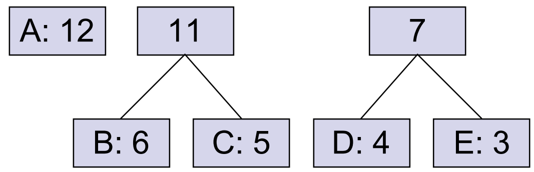
Huffman Coding: Example



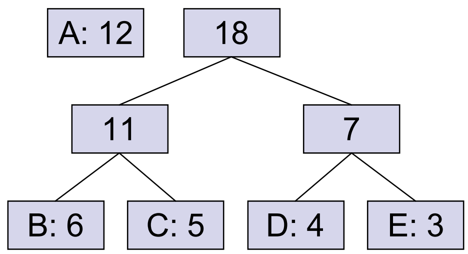
Subtrees for B and C have lowest frequencies, so merge them:



Huffman Coding: Example

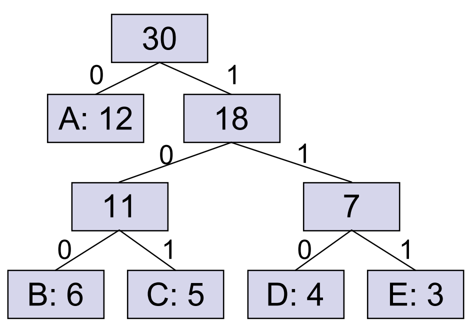


Subtrees for {B,C} and {D,E} have lowest frequencies; merge:



Huffman Coding: Example

Finally:

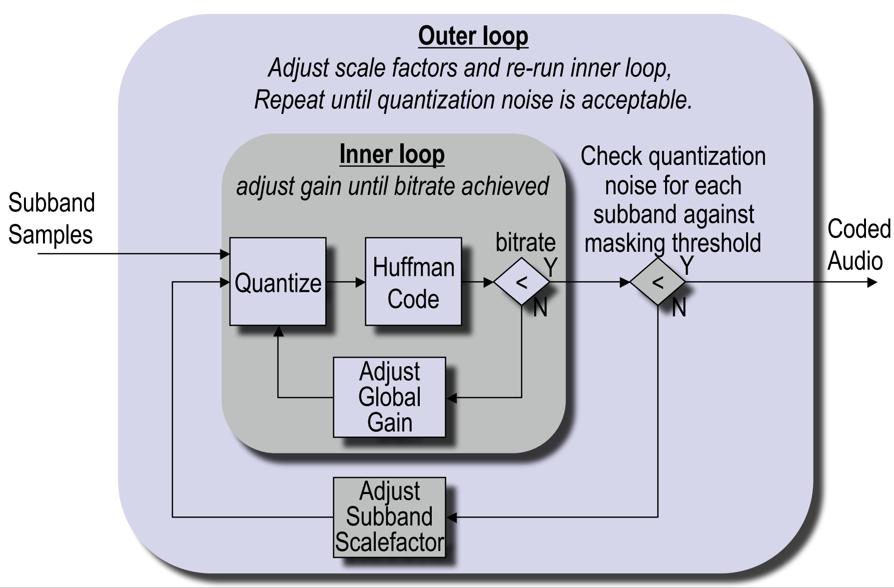


- To encode a character, walk down the tree. Each left branch adds a 0, each right branch adds a 1.
- Eg: A codes as 0, B codes as 100, E codes as 111

Fixed Bitrate Encoding in MP3

- Goal is to encode at a fixed bitrate.
  - Eg: 128Kb/s.
- Can't directly allocate bits to subbands because of Huffman encoding (don't know how many bits will result).
  - Use an iterative approach to changing the scale factors used in quantizing each subband.

MP3 Iterative Encoding

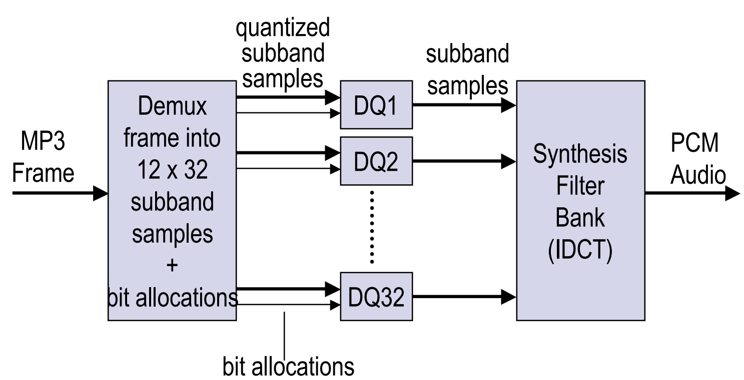


MP3 Stereo

- Multiple stereo modes:
  - Mono
  - Code each channel separately.
  - Joint stereo:
    - Code mean + difference.
    - For low frequencies, only code mean (you can't hear stereo at low frequencies)

MP3 Decoder

- No need for psychoacoustic model at decoder.
- Improved encoder can improve quality for any decoder.



Beyond MP3

- MP3 is no longer state of the art.
- Most newer codecs follow same general principles though.
  - MPEG 2 Advanced Audio Codec (AAC)
  - Ogg Vorbis
  - Windows Media Audio (WMA)

MPEG-2 AAC

AAC is derived from MP3. Main differences:

- **5.1 Surround Sound**
- **Better Filter bank**:
  - MP3 used a hybrid filter bank for backward compatibility reasons.
  - MPEG-2 AAC uses a plain Modified Discrete Cosine Transform
- **Temporal Noise Shaping** (TNS):
  - Shapes the distribution of quantization noise in time by prediction in the frequency domain.
  - Helps with voice signals.
- **Finer control of quantization resolution** - the given bit rate can be used more efficiently.
- **Bit-stream format**: the information to be transmitted undergoes entropy coding in order to keep redundancy as low as possible.

Ogg Vorbis

- Patent free, similar quality to AAC.
- Like AAC, MDCT used to transform to frequency domain.
- Psychoacoustic model used to determine the noise floor (envelope of masking effects) across frequency bands.
  - Includes simultaneous noise masking.
- Noise floor subtracted from MDCT components.
- Noise floor and MDCT residue coded separately using a codebook-based vector quantization algorithm.

Window Media Audio (WMA)

- MDCT-based codec, pretty similar to AAC and Ogg.
  - Frequency and temporal masking, then requantize.
- Main differences:
  - More block sizes to choose from (can trade off temporal vs frequency precision better).
  - Different use of Huffman coding:
    - Independently Huffman-code mantissa and exponent of floating-point quantized frequency values.
  - Mid/side coding of stereo.
    - Code L+R (mid) and L-R (side) separately.