EBU5303

Multimedia Fundamentals

Perceptual Encoding

Agenda

- Compression of audio is rarely lossless
- A-law encoding is a nonlinear companding method
- Psychoacoustics is the study of how the human ears and brain perceive sound
- MP3 and AAC compression methods use perceptual encoding

Reading

http://digitalsoundandmusic.com/chapters/ch5/

5.3.8 Algorithms for Audio Companding and Compression

Agenda

- Compression of audio is rarely lossless
- A-law encoding is a nonlinear companding method
- Psychoacoustics is the study of how the human ears and brain perceive sound
- MP3 and AAC compression methods use perceptual encoding

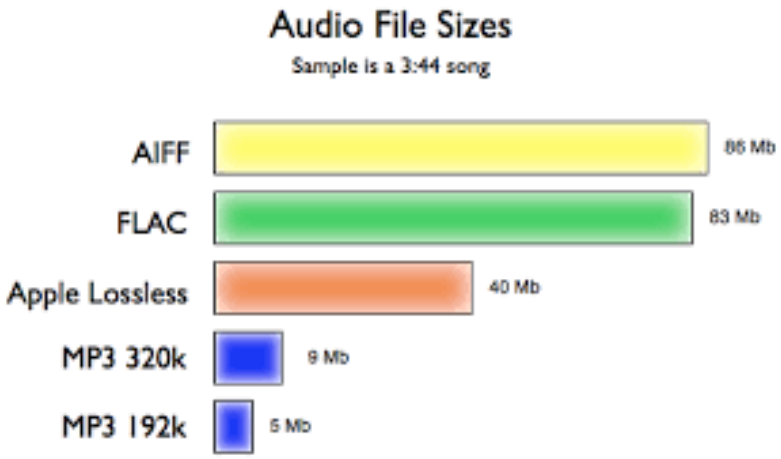
Audio Compression

- Uncompressed 3 minutes song in stereo = 25 MB
- Lossless compression does not work well (complex and unpredictable nature of sound waveforms)
- Obvious compression technique: **silence compression**
  - Detect silence = samples falling below a threshold
  - Treat them as zero and compress using RLE (Run Length Encoding)
  - Silence is rarely absolute $\Rightarrow$ not strictly lossless

Lossless Audio Compression

- Used for editing or further compression, for archival storage, or as master copies.
- Compression ratios are typically around 50–60% of original size.
- Lossless audio compression formats: Free Lossless Audio Codec (FLAC), Apple's Apple Lossless (ALAC), MPEG-4 ALS, Microsoft's Windows Media Audio 9 Lossless (WMA Lossless), Monkey's Audio, TTA, and WavPack

Audio File Sizes (example)



- FLAC uses linear prediction to convert the audio samples.
- There are two steps: the predictor and the error coding.
- The difference between the predictor and the actual sample data is calculated and is known as the residual.
- The residual is stored efficiently using Golomb-Rice coding (a type of entropy encoding)
- FLAC also uses run-length encoding for blocks of identical samples, such as silent passages.

Agenda

- Compression of audio is rarely lossless
- A-law encoding is a nonlinear companding method
- Psychoacoustics is the study of how the human ears and brain perceive sound
- MP3 and AAC compression methods use perceptual encoding

Companding

- Companding = compressing and expanding
- Companding = reducing the bit depth
- When reducing the bit depth with a linear method, rounding down has a greater impact on low amplitude samples than on high amplitude ones.

  E.g.: when reducing the bit depth from 16 to 8 bits, all amplitudes between 0 and 255 become 0 (i.e. they are lost!).

Exercise

Let's assume that converting from 16 bits to 8 bits is done by dividing the sample value by 256 and rounding down.

Consider the following two16 bits samples' amplitudes: 32767 and 255.

- Convert the two samples to 8 bits.
- Compare the errors from rounding these two samples.

A-law

- A-law: a nonlinear companding method.
- With A-law encoding, not all samples are encoded in the same number of bits.
- The human auditory system is believed to be a logarithmic process in which high amplitude sounds do not require the same resolution as low amplitude sounds: the human ear is more sensitive to quantisation noise in small signals than large signals.
- A-law coding apply a logarithmic quantisation function to adjust the data resolution in proportion to the level of the input signal: smaller signals are represented with greater precision – more data bits – than larger signals.

A-law

For a given input *x*, the equation for A-law encoding is as follows,

$$F(x) = \operatorname{sgn}(x) \begin{cases} \frac{A|x|}{1+\ln(A)}, & |x| < \frac{1}{A} \\ \frac{1+\ln(A|x|)}{1+\ln(A)}, & \frac{1}{A} \leq |x| \leq 1, \end{cases}$$

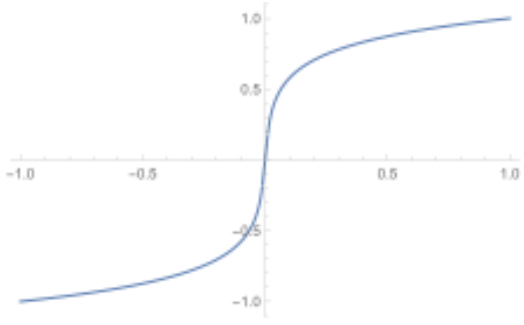
where *A* is the compression parameter. In Europe, $A = 87.6$.

A-law expansion is given by the inverse function,

$$F^{-1}(y) = \operatorname{sgn}(y) \begin{cases} \frac{|y|(1+\ln(A))}{A}, & |y| < \frac{1}{1+\ln(A)} \\ \frac{\exp(|y|(1+\ln(A))-1)}{A}, & \frac{1}{1+\ln(A)} \leq |y| < 1. \end{cases}$$

A-law

The A-law function has the effect of “spreading out” the quantization intervals more at lower amplitudes.



Plot of F(x) for A = 87.6

Agenda

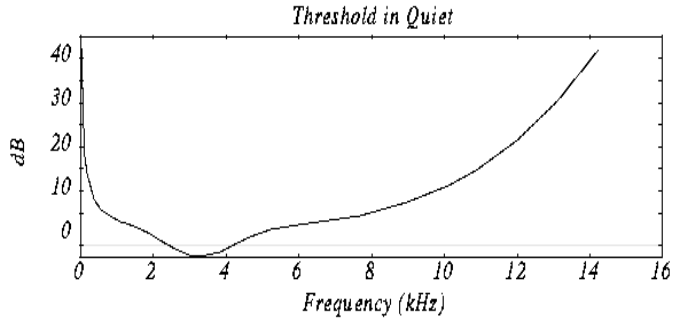
- Compression of audio is rarely lossless
- A-law encoding is a nonlinear companding method
- Psychoacoustics is the study of how the human ears and brain perceive sound
- MP3 and AAC compression methods use perceptual encoding

Psychoacoustics

- A branch of science studying the psychological and physiological responses associated with sound (including noise, speech and music), i.e. the study of how the human ears and brain perceive sound.
- Psychoacoustical experiments have shown that human hearing is nonlinear in a number of ways, including perception of octaves, perception of loudness, and frequency resolution.

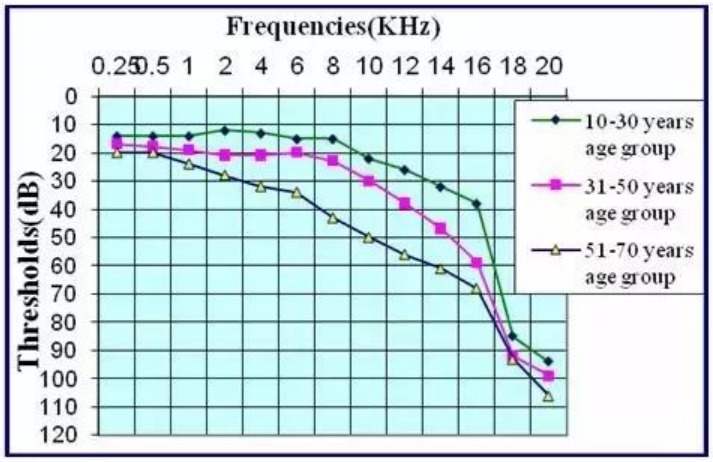
Threshold of Hearing

- Humans hear best (i.e., have the most sensitivity to amplitude) in the range of about 1000 to 5000 Hz, which is close to the range of the human voice.
- Threshold of hearing = minimal level at which sound can be heard



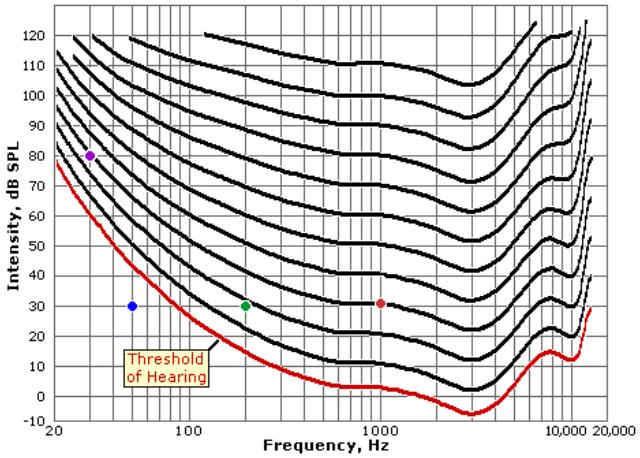
Threshold of Hearing

The threshold of hearing changes with age.



Threshold of Hearing

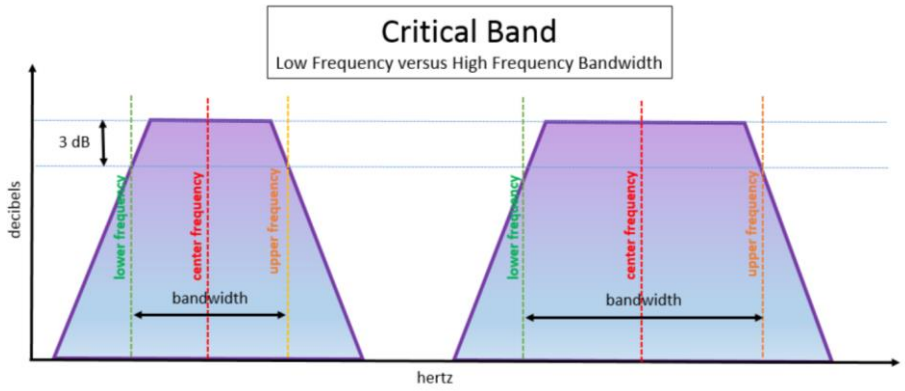
The graph shows how loud a sound at a certain frequency needs to be, for it to be perceived as a certain loudness.



Critical Bands

- Human ability to distinguish between frequencies decreases nonlinearly from low to high frequencies: at the very lowest audible frequencies, we can tell the difference between pitches that are only a few Hz apart, while at high frequencies the pitches must be separated by more than 100 Hz before we notice a difference.
- This difference in frequency sensitivity arises from the fact that the inner ear is divided into **critical bands**.
- Each band is tuned to a range of frequencies (in a similar manner to bandpass filters), and there are 24 critical bands in the human hearing range.
- Critical bands for low frequencies are narrower than those for high ones.

Critical Bands



Critical Bands

Critical Band (Bark)	Center Frequency (Hz)	Bandwidth (Hz)
1	50	100
2	150	100
3	250	100
4	350	100
5	450	110
6	570	120
7	700	140
8	840	150
9	1000	160
10	1170	190
11	1370	210
12	1600	240
13	1850	280
14	2150	320
15	2500	380
16	2900	450
17	3400	550
18	4000	700
19	4800	900
20	5800	1100
21	7000	1300
22	8500	1800
23	10500	2500
24	13500	3500

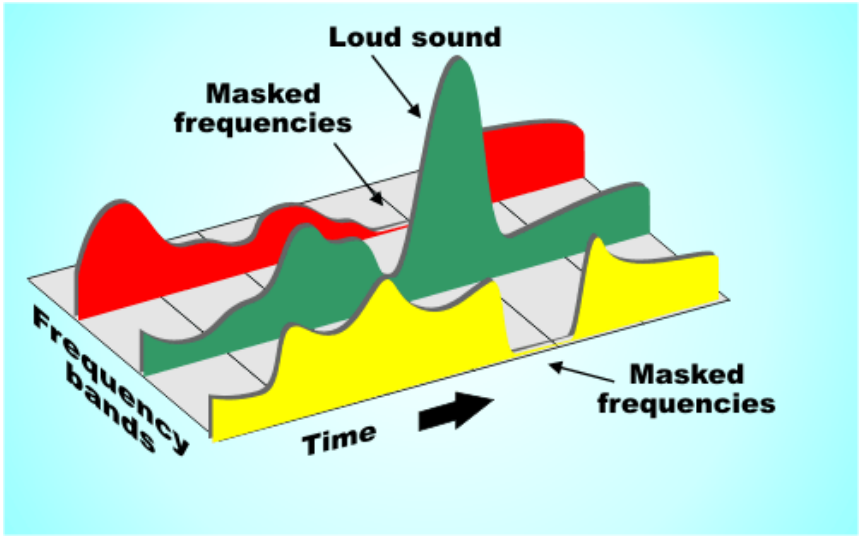
Critical Bands

- If two tones are in the same critical band, they are not easily distinguishable as separate, distinct tones.
- When the tones are about 4 Hertz apart, the ear hears a single tone with a low frequency modulation or beating.
- When the tones are about 70 Hertz apart, the ear hears a rapid modulation or beating.
- With a separation of 350 Hertz, the two tones are in different critical bands, and the ear can distinguish them from each other.

Frequency Masking

- **Frequency masking**: A loud tone may mask a softer tone of similar or higher frequency.
- Masking occurs when two frequencies are received by a critical band at about the same moment in time, one of the frequencies being significantly louder than the first such that it makes it inaudible.
- The loud frequency is called the **masking tone**, and the quiet one is the **masked frequency**.

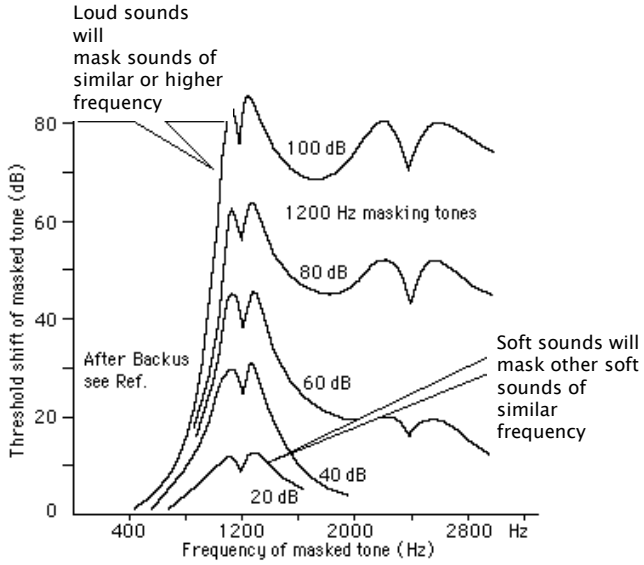
Frequency Masking



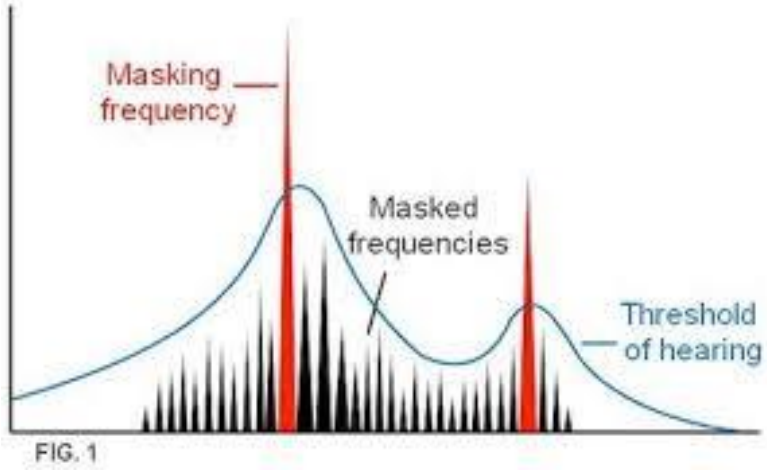
Question

How does frequency masking affect the threshold of hearing?

Frequency Masking

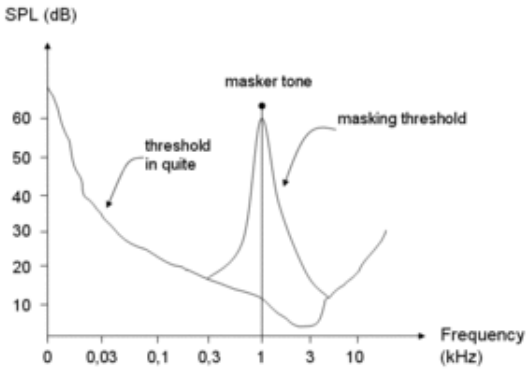


Frequency Masking

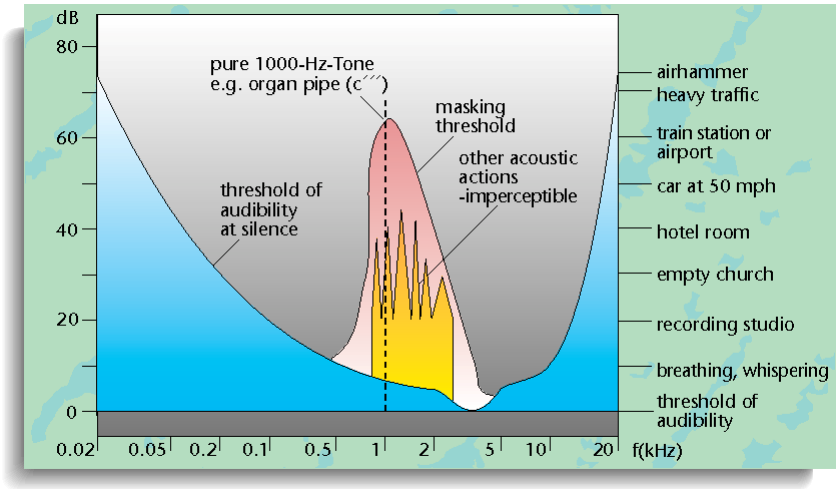


Frequency masking

- Masking causes the threshold of hearing to be raised within a critical band in the presence of a masking tone.
- The new threshold of hearing is called the **masking threshold**.

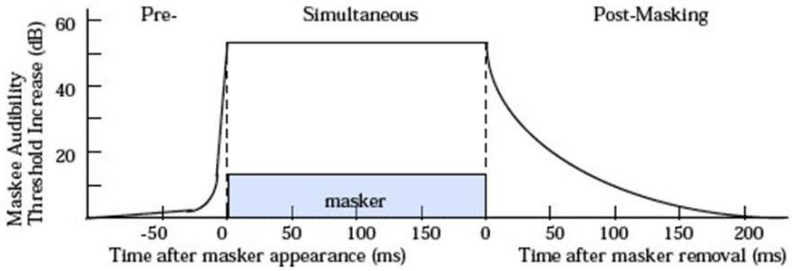


Frequency masking

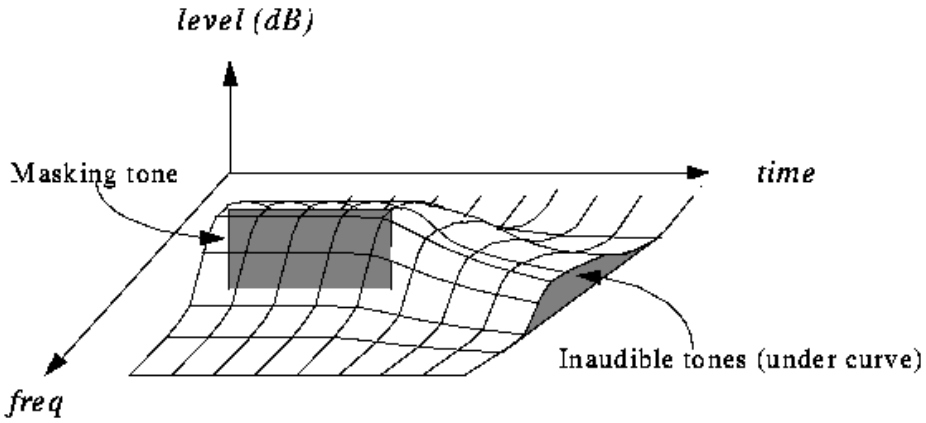


Temporal Masking

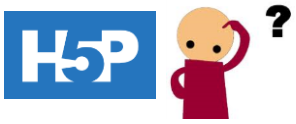
- **Temporal masking**: After a loud sound stops, there is a small delay before we can hear a softer tone.
- The duration of masking depends on the duration of the masker, its amplitude and its frequency.



Frequency and Temporal Masking

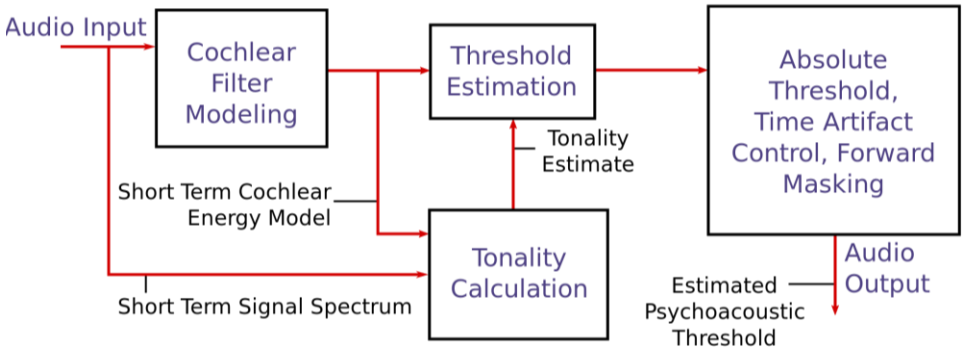


Question



How can psychoacoustics allow for effective lossy compression of audio signals?

Psychoacoustic Model



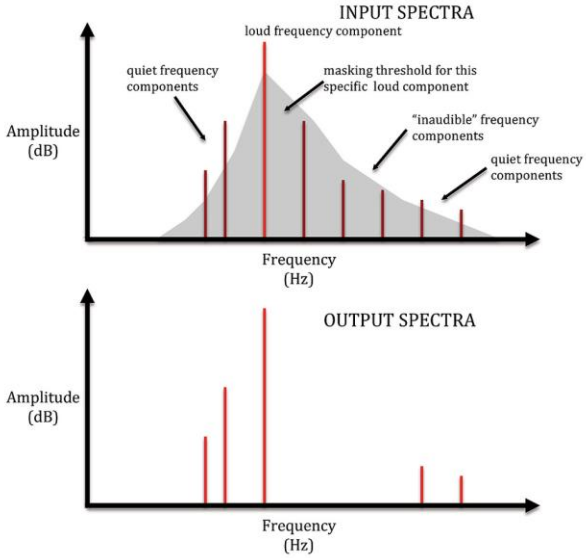
Agenda

- Compression of audio is rarely lossless
- A-law encoding is a nonlinear companding method
- Psychoacoustics is the study of how the human ears and brain perceive sound
- MP3 and AAC compression methods use perceptual encoding

Perceptual Encoding

- Based on psychoacoustics
- The goal of applying psychoacoustics to compression methods is to determine the components of sounds that human ears don't perceive very well, if at all.
- These are the parts that can be discarded, thereby decreasing the amount of data that must be stored in digitised sound.

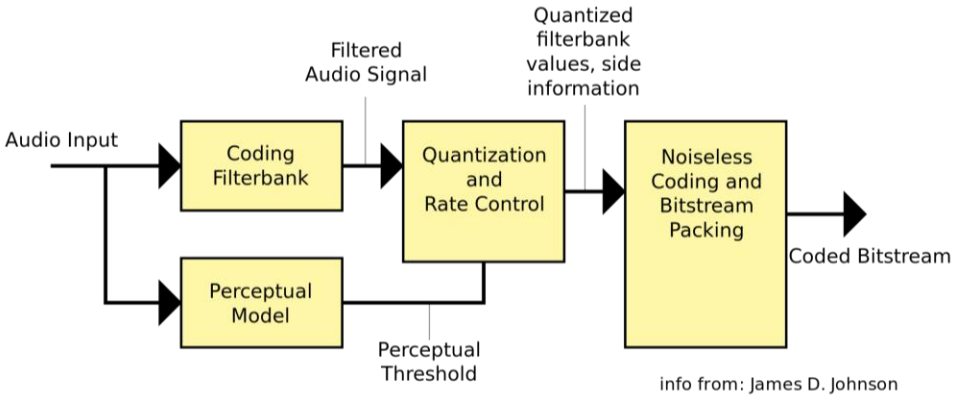
Frequency Masking and Encoding



Perceptual Audio Coder

1. Use convolution filters to divide the audio signal into frequency subbands --> *subband filtering*.
2. Determine amount of masking for each band caused by nearby band using a *psychoacoustic model*
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.
5. Format bitstream

Perceptual Audio Coder



Illustrative example

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

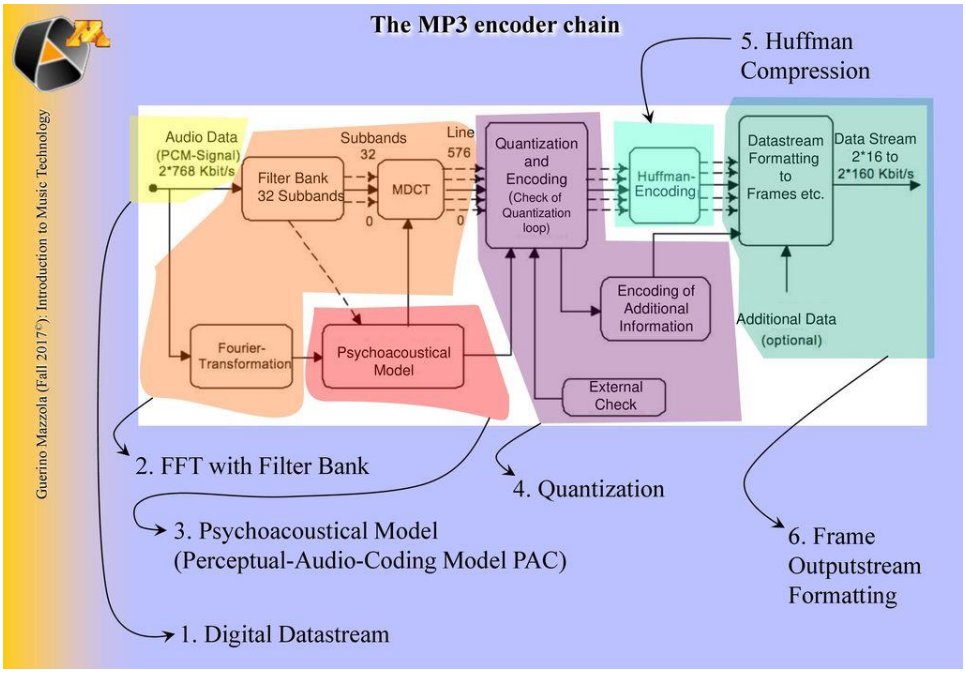
- If the level of the 8th band is 60dB, it gives a masking of 12 dB in the 7th band, 15dB in the 9th (perceptual model).
- Level in 7th band is 10 dB (< 12 dB), so ignore it.
- Level in 9th band is 35 dB (> 15 dB), only the amount above the masking level needs to be sent.

MPEG Audio

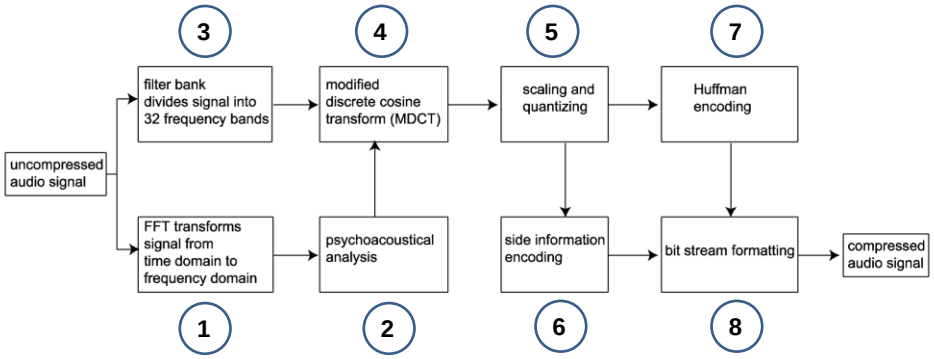
- **MP3**: formally MPEG-1 (and MPEG-2) Audio Layer III
  - 16 bits
  - Sampling rate: 32, 44.1, or 48 kHz
  - Bitrate: 32 to 320 kbps
- **AAC:** Advanced Audio Coding; part of the MPEG-2 and MPEG-4 specifications
  - More sample frequencies (8 kHz to 96 kHz)
  - Higher coding efficiency and simpler filter bank
  - 96 kbps AAC sounds better than 128 kbps MP3

MP3

- Compared to CD-quality digital audio (i.e. 2 channel signed 16-bit sampled at 44,100 Hz), MP3 compression can commonly achieve a 75 to 95% reduction in size.
- The MPEG standard does not include a precise specification for an MP3 encoder, but does provide example psychoacoustic models: implementers devise their own algorithms suitable for removing parts of the information from the audio input.
- The quality of MP3 encoded sound depends on the quality of the encoder algorithm as well as the complexity of the signal being encoded.

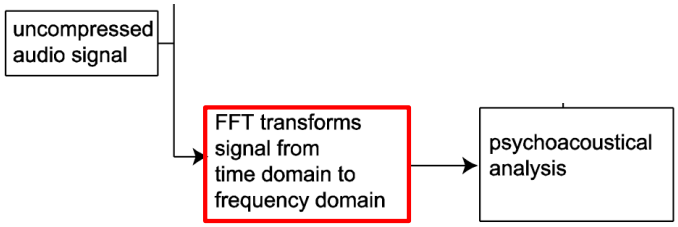


MP3 algorithm



1 FFT

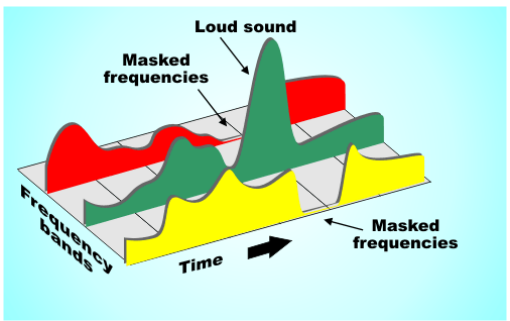
Divide the audio signal in frames of 1152 samples, and use the Fourier transform to transform the time domain data to the frequency domain, sending the results to the psychoacoustical analyser.



2 Psychoacoustic analyser

The psychoacoustic analyser identifies masking tones and masked frequencies in a local neighborhood of frequencies over a small window of time.

- It outputs a set of **signal-to-mask ratios** (**SMRs**).
- The SMR is the ratio between the amplitude of a masking tone and the amplitude of the minimum masked frequency in the chosen vicinity.

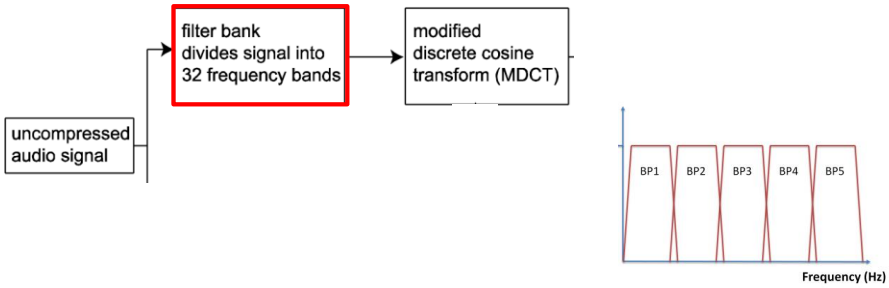


3

Filter bank

Divide each frame into 32 frequency bands between 0 and 22.05 kHz, using filter banks (bandpass filters).

- The 32 resulting bands are still in the time domain.
- That is, there are 32 sets of 1152 time-domain samples, each holding just the frequencies in its band.



4

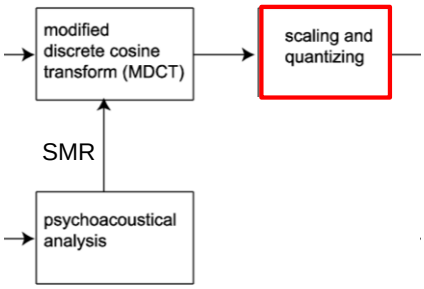
MDCT

Use the MDCT (Modified Discrete Cosine Transform) to divide each of the 32 frequency bands into 18 subbands for a total of 576 frequency subbands.

5 Scaling and Quantising

Sort the subbands into 22 groups, called scale factor bands.

- Based on the SMR, the scale factor bands cover several MDCT coefficients and more closely match the critical bands of the human ear.
- Use nonuniform quantisation combined with scaling factors: bands that have a lower SMR are multiplied by larger scaling factors because the quantisation error for these bands has less impact, falling below the masking threshold.



Scaling and Quantising example

Say that an uncompressed band value is 20,000 and values from all bands are quantised by dividing by 128 and rounding down. Thus the quantised value would be 156. When the value is restored by multiplying by 128, it is 19,968, for an error of 32/20,000 = 0.0016.

Now suppose the psychoacoustical analyser reveals that this band requires less precision because of a strong masking tone. Thus, it determines that the band should be scaled by a factor of 0.1. Now we have 20,000*0.1/128= 15. Restoring the original value we get 15*128/0.1 = 19200, for an error of 800/20,000 = 0.04.

Scaling and Quantising

An appropriate psychoacoustical analysis provides scaling factors that increase the quantisation error where it doesn't matter, in the presence of masking tones.

Scale factor bands effectively allow less precision (i.e., fewer bits) to store values if the resulting quantisation error falls below the audible level.

This is one way to reduce the amount of data in the compressed signal.

Exercise H5P

Say that an uncompressed band value is 5,000 and values from all bands are quantised by dividing by 128 and rounding down.

- What is the quantised value?
- What is the quantisation error?

Exercise

Now suppose the psychoacoustical analyser reveals that this band requires less precision because of a strong masking tone. Thus, it determines that the band should be scaled by a factor of 0.2.

- What is the new quantised value?
- What is the new quantisation error?

Side information and Huffman

6

Encode side information: side information is the information needed to decode the rest of the data, including where the main data begins, where scale factors and Huffman encodings begin, the Huffman table to use, the quantisation step, and so forth.

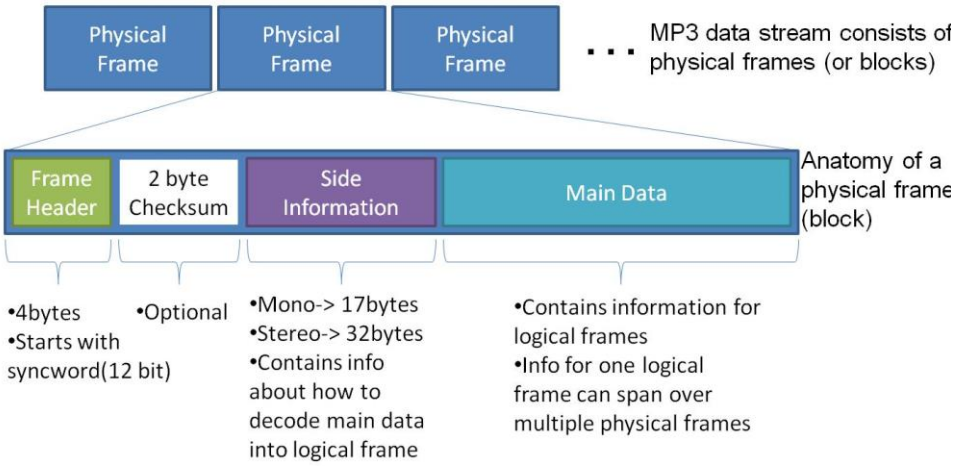
7

Use Huffman encoding on the resulting 576 quantised MDCT coefficients.

8

Bit stream

Put the encoded data into a properly formatted frame in the bit stream.



MP3 algorithm

/*Input: An audio signal in the time domain
Output: The same audio signal, compressed
/*
Process the audio signal in frames of 1152 samples

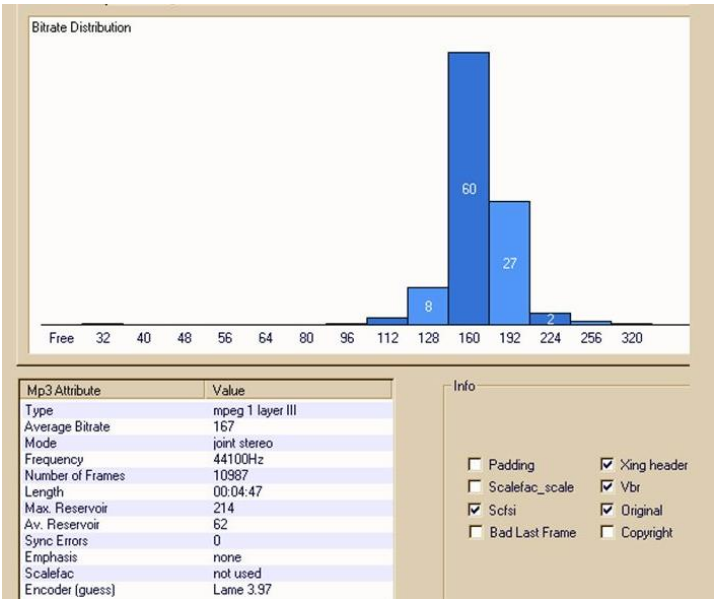
For each frame {
Use the Fourier transform to transform the time domain data to the frequency domain, sending the results to the
Psychoacoustical analyser {
Based on masking tones and masked frequencies, determine the signal-to-masking noise ratios (SMR) in areas across the frequency spectrum
} // end of Psychoacoustical analyser
Divide the frame into 32 frequency bands

For each frequency band {
Use the modified discrete cosine transform (MDCT) to divide each of the 32 frequency bands into 18 subbands, for a total of 576 frequency subbands
Sort the subbands into 22 groups, called scale factor bands, and based on the SMR, determine a scaling factor for each scale factor band
Use nonuniform quantisation combined with scaling factors to quantise
Encode side information
Use Huffman encoding on the resulting 576 quantised MDCT coefficients
Put the encoded data into a properly formatted frame in the bit stream
} // end of For each frequency band
} // end of For each frame

MP3 bitrates

- CD audio is 44100 samples per second, stereo and 16 bits per channel: the bitrate of uncompressed CD digital audio is 44100x32 = 1411 kbit/s.
- MP3 was designed to encode data at 320 kbit/s or less.
- A bit rate of 128 kbit/s is commonly used.
- Bitrates 128, 160 and 192 kbit/s represent compression ratios of approximately 11:1, 9:1 and 7:1 respectively.
- With some advanced MP3 encoders, it is possible to specify a given quality, and the encoder will adjust the bit rate accordingly (i.e. Variable Bit Rate or VBR).
- Average Bit Rate (ABR) is a type of VBR where the bitrate is allowed to vary for more consistent quality, but is controlled to remain near an average value chosen by the user, for predictable file sizes.

ABR



AAC

- AAC compression, the successor to MP3, uses similar encoding techniques but improves on MP3 by offering more sampling rates (8 to 96 kHz), more channels (up to 48), and arbitrary bit rates.
- Filtering is done solely with the MDCT, with improved frequency resolution for signals without transients and improved time resolution for signals with transients.
- Frequencies over 16 kHz are better preserved.
- The overall result is that many listeners find AAC files to have better sound quality than MP3 for files compressed at the same bit rate.

Summary

- Compression of audio is rarely lossless
- A-law encoding is a nonlinear companding method
- Psychoacoustics is the study of how the human ears and brain perceive sound
- MP3 and AAC compression methods use perceptual encoding