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

Audio Compression

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 -> not strictly lossless

Lossless Audio Compression

Used for editing or further compression, for archival storage, or as master copies
Compression ratios = 50–60%
FLAC is one of the lossless audio compression formats:

- uses linear prediction to convert the audio samples
- 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.

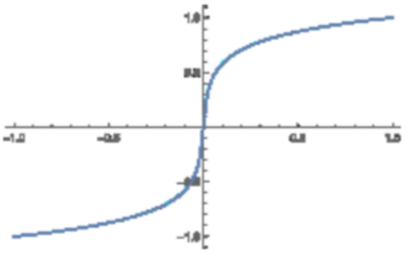
Companding

Companding = compressing and expanding, 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.
Eample: 16-bit samples(-32768 ~ 32767), 8-bit(-128~127)

1. The amplitude is 32767 -> then convert to 127 … 255, with the error is 255/32768 < 1%
2. The amplitude is 0~255, then convert to 0, with the error 100%

A-law: a nonlinear companding method

not all samples are encoded in the same number of bits
fundation: the human ear is more sensitive to quantisation noise in small signals than large signals. -> (high amplitude sounds do not require the same resolution as low amplitude sounds)
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
logarithmic quantisation function: **smaller signals** are represented with **greater precision** – more data bits – than larger signals.



Plot of F(x) for A = 87.6

A-Law function

Advantage of A-Law: presume some of the dynamic range, that would be lost if the lower method of reducing the bit depth

Psychoacoustics

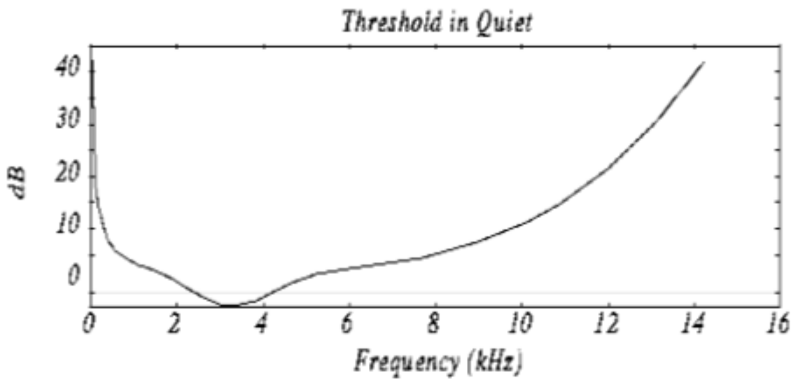
- the study of how the human ears and brain perceive sound
- Human hearing is non-linear

Threshold of Hearing

Threshold of Hearing = minimal level at which sound can be heard

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.

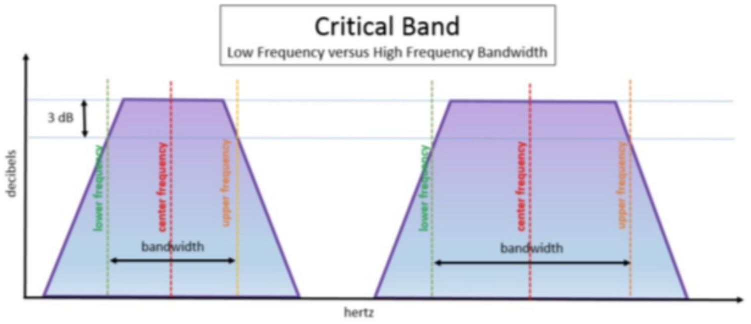
The threshold of hearing changes with age.



Critical Bands

The inner ear is divided into **24 critical bands** in the human hearing range

Fundation: Human ability to distinguish between frequencies decreases nonlinearly from low to high frequencies

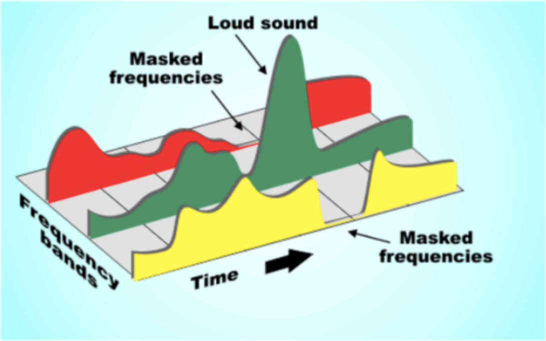


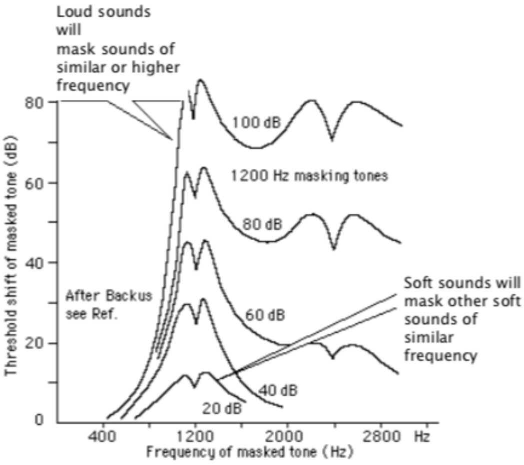
Critical bands for low frequencies are narrower than those for high ones
If two tones are in the same critical band, they are not easily distinguishable as separate, distinct tones.

- 4 Hertz apart, single tone with a low frequency modulation or beating
- 70 Hertz apart, a rapid modulation or beating
- 350 Hertz apart, the two tones are in different critical bands, can be distinguished

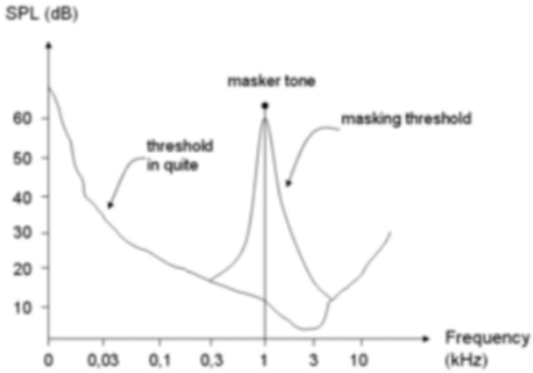
Frequency Masking

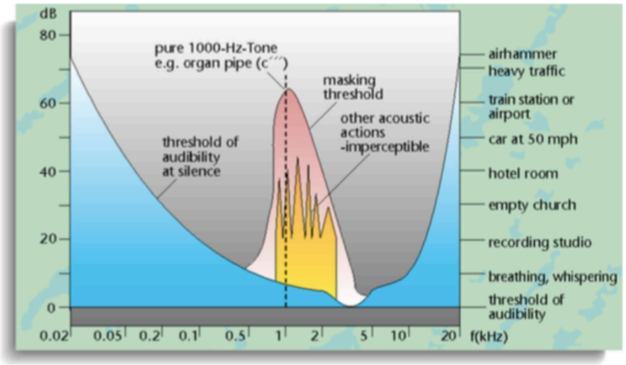
Frequency masking: A loud tone may mask a softer tone of similar or higher frequency.
Masking tone: the loud frequency
Masked tone: the quiet one





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**.

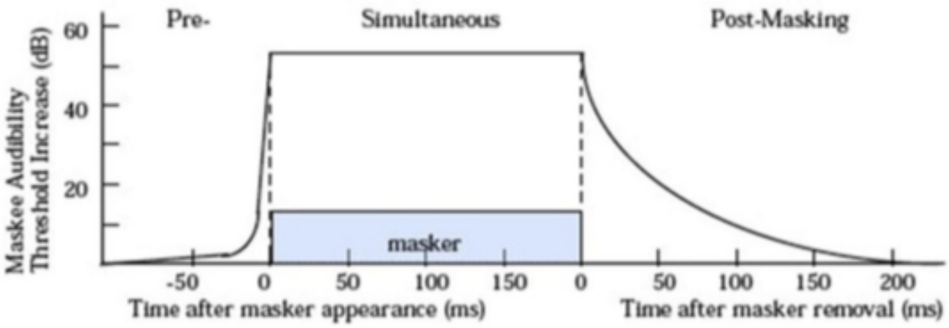


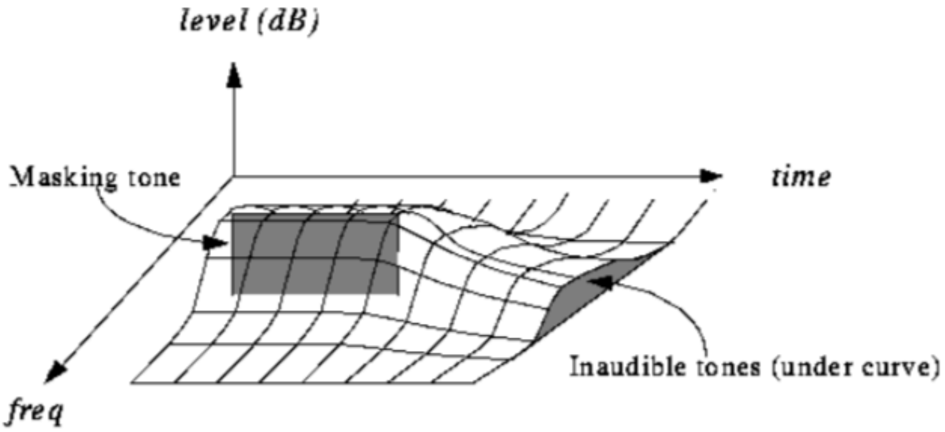


Question:
How does frequency masking affect the threshold of hearing?
Solution: The threshold of hearing will be raised within a critical band in the presence of masking tone (masking frequency)

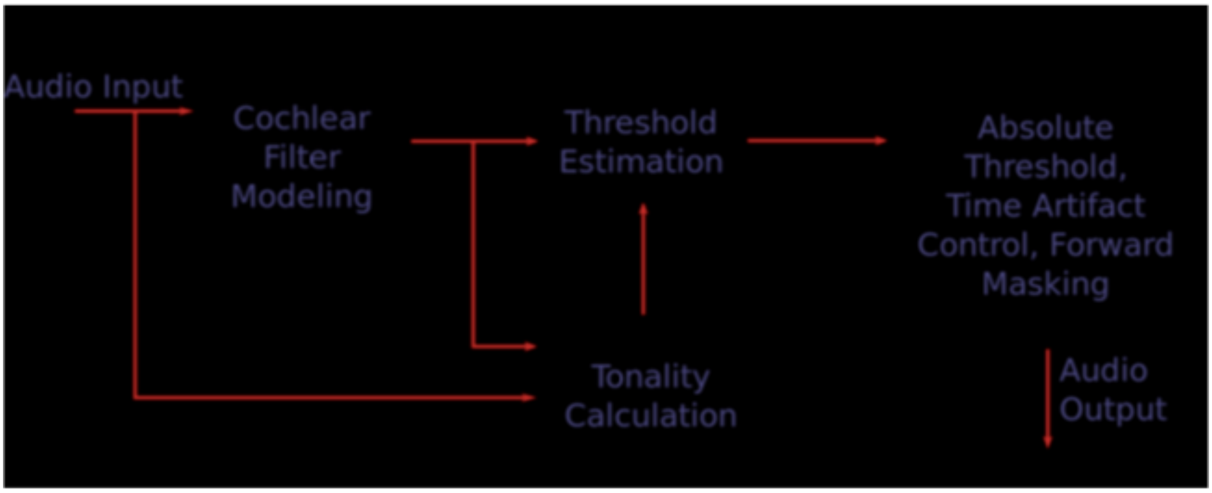
Temporal Masking

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

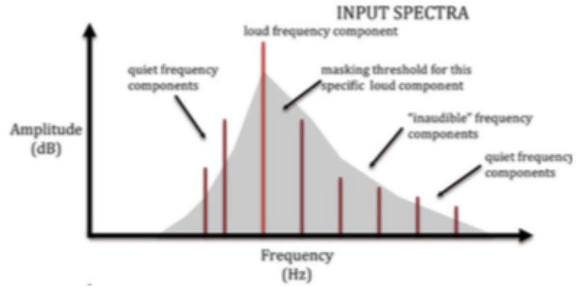


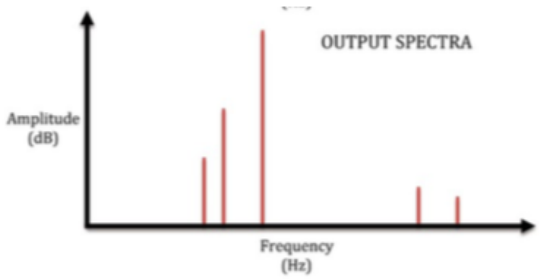
Pyschoacoustic model

Question:
How can psychoacoustics allow for effective lossy compression of audio signals?
Solution: According to the psychoacoustics knowledge, the new masking threshold can be estimated. With this new threshold, the tones below it are inaudible which can be removed.

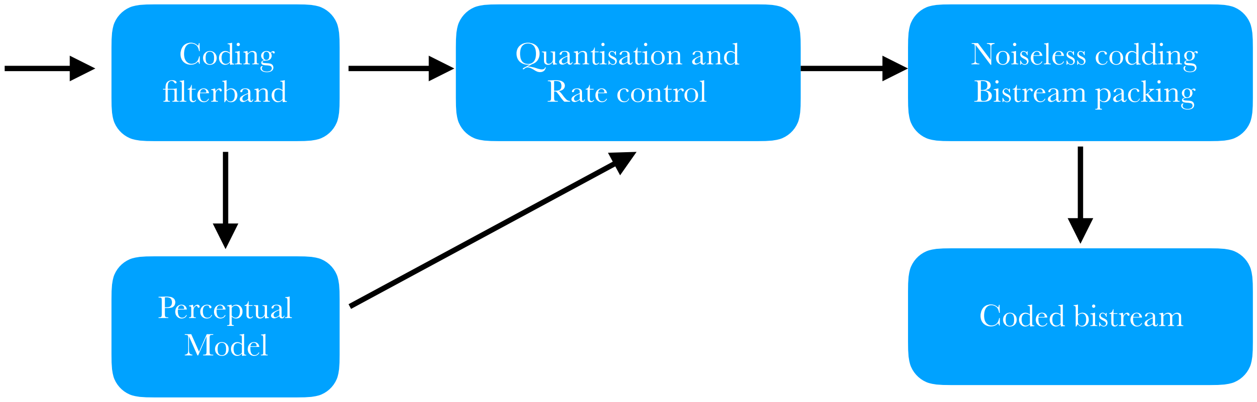
MP3 and AAC compression methods use perceptual encoding

Perceptual Encoding: 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. — decreasing the amount of data that must be stored in digitised sound
Frequency Masking and Encoding:





Perceptual Audio Coder


Coding filterband
Quantisation and Rate control
Noiseless codding Bistream packing
Perceptual Model
Coded bistream

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).
7: 10 < 12, discarded, 9: 35 > 15, perceived

MPEG Audio

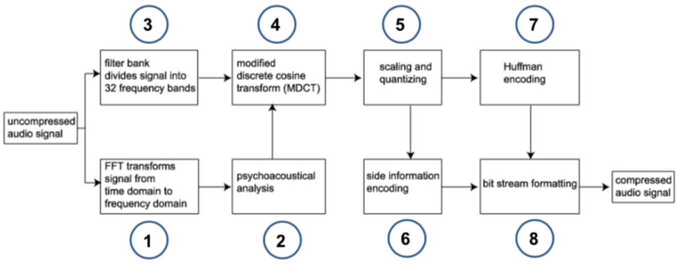
MP3: formally MPEG-1 (and MPEG-2) Audio Layer III
– **16 bits**
– Sampling rate: 32, 44.1, or 48 kHz
– Bitrate: 32 to 320 kbps
AAC: Advanced Audio Coding; part of the MPEG- 2 and MPEG-4 specifications
– More sample frequencies (8 kHz to 96 kHz)
– Higher coding efficiency and simpler filter bank

– 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 quality of MP3 encoded sound depends on the quality of the encoder algorithm as well as the complexity of the signal being encoded


3
filter bank divides signal into 32 frequency bands
4
modified discrete cosine transform (MDCT)
5
scaling and quantizing
7
Huffman encoding
uncompressed audio signal
FFT transforms signal from time domain to frequency domain
psychoacoustical analysis
side information encoding
bit stream formatting
compressed audio signal
1
2
6
8

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.

outputs: outputs a set of signal-to-mask ratios (SMRs)

SMR: the ratio between the amplitude of a masking tone and the amplitude of the minimum masked frequency in the chosen vicinity.

The SMR at a given frequency is expressed as the difference (in dB) between **the SPL of the masker** and **the masking threshold at that frequency**

3. Filter bank

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

- 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**

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

Lower SMR — larger masking threshold, the quantisation error is more likely to fall below the threshold. In this case, less precision is allowable (scale factor | a larger scaling factor)

Example:

uncompressed band value is 20,000 and values from all bands are quantised by dividing by 128 and rounding down

20000/128=156, 156*128=19968, error = 32/20,000 = 0.0016

suppose the psychoacoustical analyser reveals that this band requires less precision because of a **strong masking tone,** the band should be scaled by a factor of 0.1

20,000*0.1/128= 15, 15*128/0.1 = 19200, 800/20,000 = 0.04

6. Side information

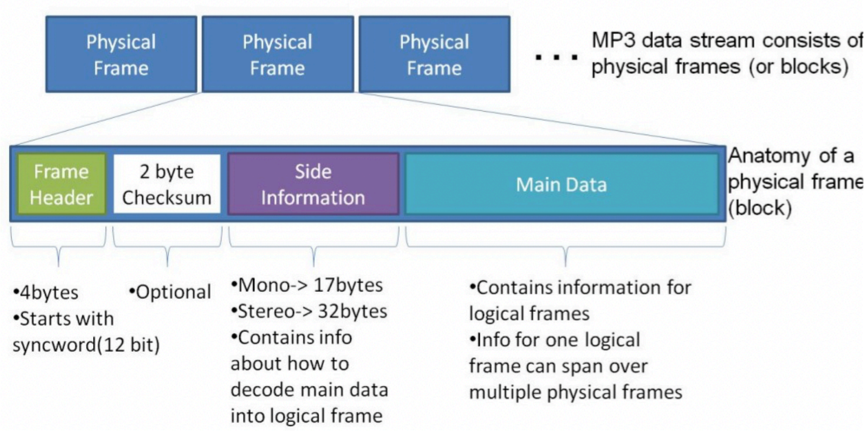
side information is the information needed to decode the rest of the data, including where the main data begins, where scale factors and Huffman encodings begin, the Huffman table to use, the quantisation step, and so forth.

7. Huffman encoding

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


Physical Frame
Physical Frame
Physical Frame
. . .
MP3 data stream consists of physical frames (or blocks)
Frame Header
2 byte Checksum
Side Information
Main Data
Anatomy of a physical frame (block)
•4bytes
•Starts with syncword(12 bit)
•Optional
•Mono-> 17bytes
•Stereo-> 32bytes
•Contains info about how to decode main data into logical frame
•Contains information for logical frames
•Info for one logical frame can span over multiple physical frames

Questions:

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?

Solution: 5000 / 128 = 39, quantised value = 39

Quantisation error: $5000 - 39 \cdot 128 = 8$, Error rate $= \dfrac{8}{5000} = 0.0016$

With the scaling factor of 0.2, $\dfrac{5000 \cdot 0.2}{128} = 7$, Quantisation error: $\dfrac{7 \cdot 128}{0.2} = 4480$,

Quantisation error: 5000 - 4480 = 520, Error rate $= \dfrac{520}{5000} = 0.104$

MP3 bitrates

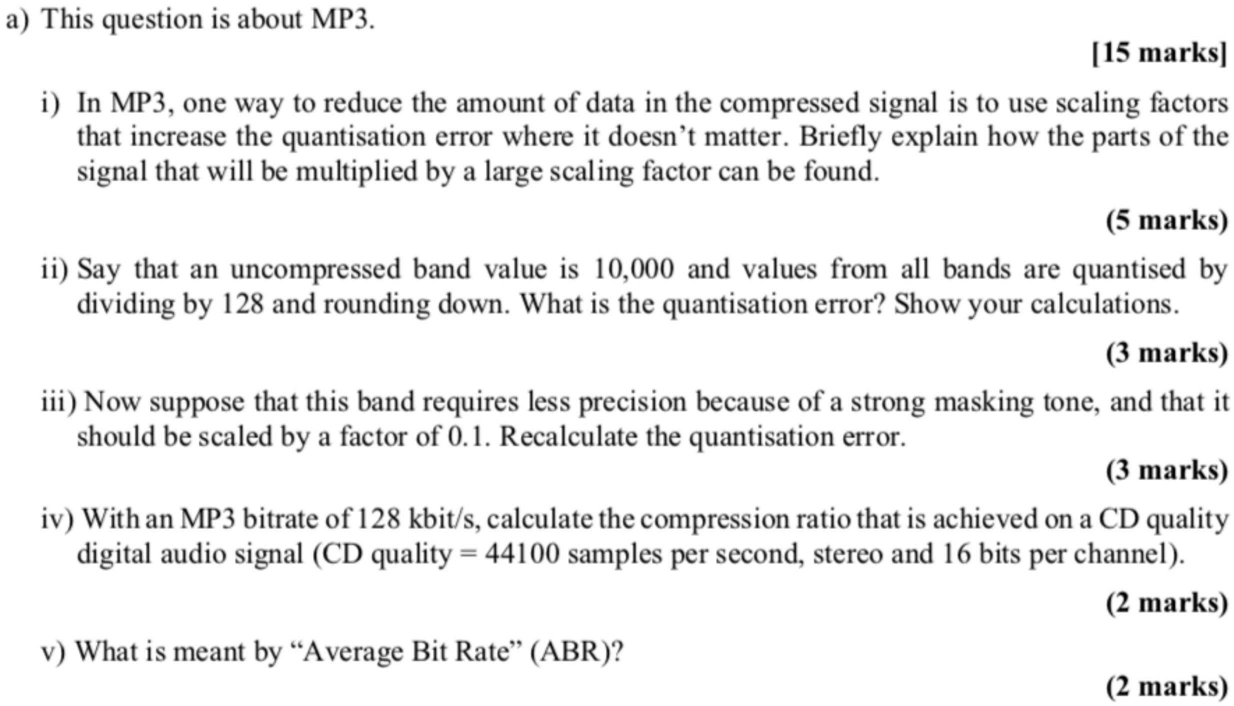
- MP3 was designed to encode data at 320 kbit/s or less
- **A bit rate of 128 kbit/s is commonly used (11:1 compression rate)**
- It is possible to **specify a given quality**, and the encoder will adjust the bit rate accordingly (i.e. Variable Bit Rate or VBR)
- **Average Bit Rate (ABR)** is a type 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

AAC

- AAC compression, the successor to MP3, use 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

Supplement:
MPEG-1/MPEG=3 for CD
MPEG-2 DVD, digital TV (support high resolution)
MPEG-4 animation, graphics and text (audiovisual objects)
MPEG-7 (adding info. Not compression), descriptors of various type of multimedia
MPEG-21 data protection

Questions is test



a) This question is about MP3.

[15 marks]

i) In MP3, one way to reduce the amount of data in the compressed signal is to use scaling factors that increase the quantisation error where it doesn't matter. Briefly explain how the parts of the signal that will be multiplied by a large scaling factor can be found.

(5 marks)

ii) Say that an uncompressed band value is 10,000 and values from all bands are quantised by dividing by 128 and rounding down. What is the quantisation error? Show your calculations.

(3 marks)

iii) Now suppose that this band requires less precision because of a strong masking tone, and that it should be scaled by a factor of 0.1. Recalculate the quantisation error.

(3 marks)

iv) With an MP3 bitrate of 128 kbit/s, calculate the compression ratio that is achieved on a CD quality digital audio signal (CD quality = 44100 samples per second, stereo and 16 bits per channel).

(2 marks)

v) What is meant by "Average Bit Rate" (ABR)?

(2 marks)

Solution:

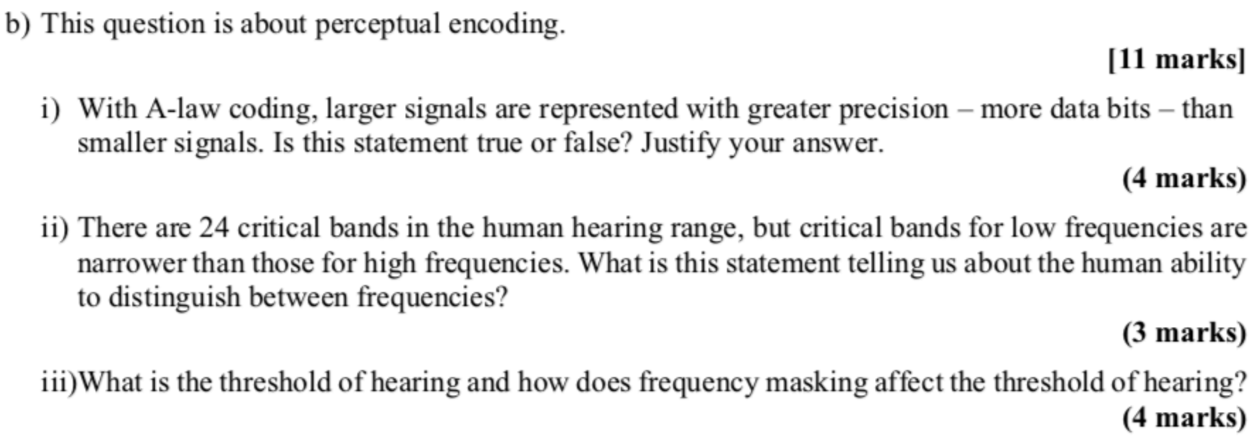
(1) In MP3, before the scaling and quantisation step, we need to figure out SMR by using psychoacoustic model. SMR is the difference between the Masking tone and the Masking threshold. If the psychoacoustic model outputs a small SMR value, we can use a large scaling factor, since in this case, the quantisation error is below the masking threshold and has less impact fow less quality.

(2) $\frac{10000}{128} = 78$, $10000 - 78 \cdot 128 = 16$, quantisation error $= \frac{16}{10000} = 0.0016$

(3) $\frac{10000 \cdot 0.1}{128} = 7$, $10000 - \frac{7 \cdot 128}{0.1} = 1040$, quantisation error $= \frac{1040}{10000} = 0.104$

(4) Compression rate $= \frac{44100 \cdot 16 \cdot 2}{128000} = 441 : 40$

(5) Average Bit Rate is a type of Variable Bit Rate where the bit rate is allowed to varied for more consistent quality, but is controlled to remain near to an average value for predictable file size.



b) This question is about perceptual encoding.

[11 marks]

i) With A-law coding, larger signals are represented with greater precision – more data bits – than smaller signals. Is this statement true or false? Justify your answer.

(4 marks)

ii) There are 24 critical bands in the human hearing range, but critical bands for low frequencies are narrower than those for high frequencies. What is this statement telling us about the human ability to distinguish between frequencies?

(3 marks)

iii)What is the threshold of hearing and how does frequency masking affect the threshold of hearing?

(4 marks)

Solution:

(1) False. Since human's ear is more sensitive to the quantisation noise in small signals than in large signals. Smaller signals should be represented with greater precision in order to presume the dynamic range.

(2) This statement tells us that human's ear can distinguish lower frequencies better than higher frequencies (Human ability to distinguish between frequencies decreases nonlinearly from low to high frequencies)

(3) The threshold of hearing is the minimum level at which sound can be heard. Frequency masking will raise the threshold of hearing within the critical band where frequency masking occurs (in the presence of the masking tone)