Psychoacoustic audio encoder

Audio Signal Processing

Jose Maria Perez-Macias

*October 17th 2016*

In this project we implement a psychoacoustic lossy audio encoder. These encoders use information about our perception of sound to remove information —or in other words, to introduce noise— without being noticed while optimizing bitrate. There are plenty of standards for psychoacoustic encoders. ATRAC, MPEG-1 L1–3, MPEG-2/4 AAC (Advanced Audio Coding), Dolby’s AC3, MPEG-4 HE AAC, MPEG-4 AAC LD, G722.1–2, apt-X, Fraunhofer ULD, OGC Vorbis, etc[[1]](#footnote-1). The main differences of these encoders are in the human audio perceptual model and the type of sub-band coding. These encoders are benchmarked based on quality versus bitrate.

These encoders are modelling the relationship between acoustic events and our auditory perception with the purpose to optimize the bitrate at a certain quality, which will stay “unnoticed”. One of the most important keywords in audio coding is “masking”, i.e. when weaker audible signals becomes inaudible in the presence of a louder signal. This happens in time and frequency. In this project, we will take advantage of the masking phenomena in the frequency domain.

The main idea is to hide quantization noise below the signal-dependent threshold of hearing (masked threshold), in other words, we are going to use a signal dependent number of bits for quantization. To estimate the minimum number of bits, we estimate the masking pattern of the source signal at each frequency sub-band. An extensive review on perceptual coding is done by Painter *et al* {Painter:2000fw}*.*

The quantization noise is the difference between the analog signal and its digital representation, and it is a result of the error in the quantization of the analog signal. The more bits we use, we increase the fidelity, and the noise becomes smaller. For example, if we use 3 bits we can distinguish 8 levels, with 16 bits, 65536 levels. In CD-audio quality, this means 2 channels \* 44.1Khz \* 16 bits + overheads = 4.32Mb/s (without overhead is 1.41Mb/s). The quantization noise is estimated to be SQNR = 20 log10(2Q) =~6.02 \*Q dB. This assumes that the input signal completely fills the band, and this is one of the reasons to sue compressors and expanders (to maximise the use of the dynamic range, and also, reduce the quantization noise). Assuming a nosie normally distributed we can obtain a maximum SNR = 0.62\*16 = 96.3dB. This is of course a simplification.

Threshold in quiet: this is the sound pressure level (SPL in dB) below which the human hearing of most people is unable to perceive a sine-tone. In most psychoacoustic models the threshold in quiet is normalized by the way a sine wave with amplitude of +/-0.5 in 16bit format reaches 0dB sound pressure level. <https://hydrogenaud.io/musepack/klemm/www.personal.uni-jena.de/~pfk/mpp/audiocoder_english.html#worumgehts>

Simultaneous masking: during an acoustic excitation

The psychoacoustic model analyses the signal input to estimate a signal-dependent masked threshold based on psychoacoustics.

We split the input data into uniform bands,

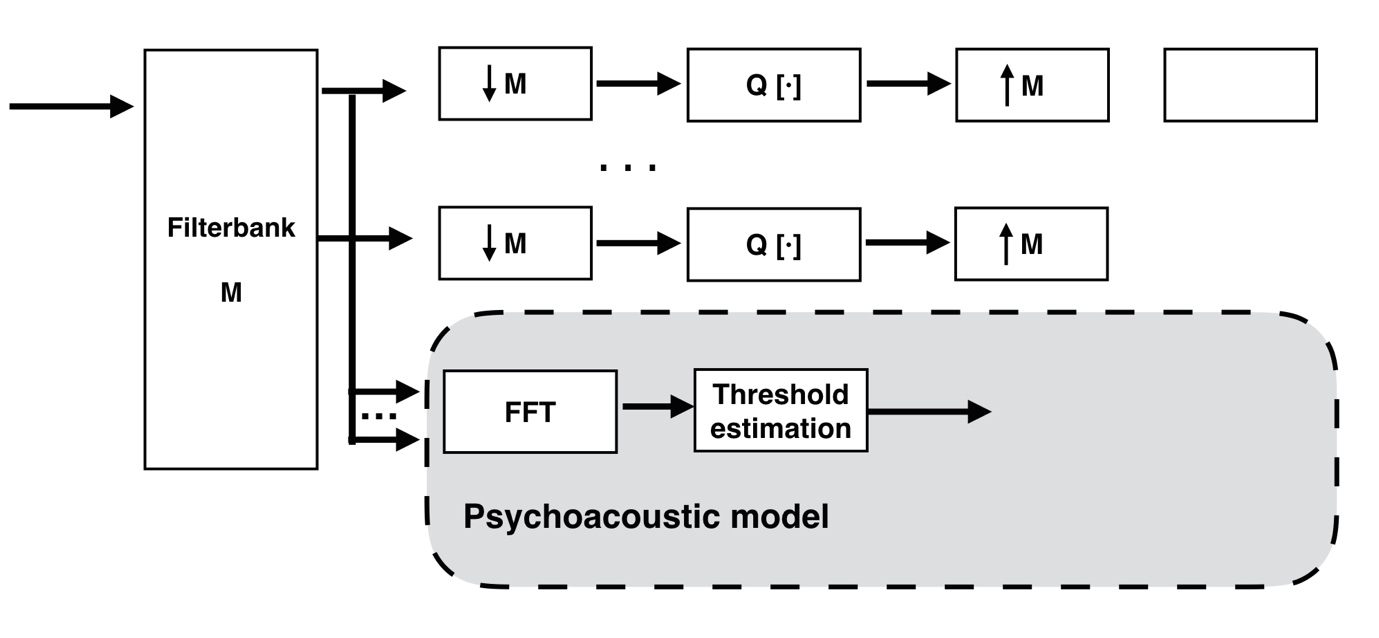
Quantizer adapted to the masking threshold.

A psychoacoustic lossy audio encoder is implemented by exploiting a uniform filter bank with critical sampling and perfect reconstruction along with a simplified psychoacoustic model to produce the masking threshold for each time-frequency point.

The analysis-synthesis filter band is an M-band MDCT-based uniform filter-bank to obtain a critically sampled, perfect reconstruction analysis-synthesis fitler bank.

The psychoacoustic model is implemented separately. It uses a DFT to evaluate the magnitude spectrum of the input signal in short time frames. The length of the frame is approximately 20 ms for fs.

To get good reconstruction without relying on the aliasing cancellation, which cannot be rely on the presence of quantization noise.



**Figure 1.** Schema of the encoder.

MDCT filter

Absolute threshold of hearing. We do not hear sounds that are too weak.

In this exercise we explore and study the MDCT filterbank with critical sampling, to then reconstruct the original signal (perfect reconstruction). The testing signal is a violin playing an F with a vibrato. Critical sampling “is a property where the combined sampling rate of the transformed frequency bands is the same as that of the original band” (Pulkki & Karjalainen, 2015). The MDCT is “a perfect reconstruction, real-valued, and critically sampled transform”(Pulkki & Karjalainen, 2015).

First we obtain the MDCT filterbank with M=10 bands using the supplied code. Results of the frequency response is illustrated in Figure 1, each filter ( is overlaid onto the same figure.

The filter is then applied to the input signal (our violin). Figure 2 depics the spectrogram for each of the bands of the filter. We can observe how the intensity is higher (more red), in the matching band.

We down sampled the signals at each band by a factor of M. Results are shown in Figure 3.

Now we could process each of the band, as we wish (not the purpose of the exercise). The next step after any processing is to reconstruct the signal. To do so we need to up-sample or interpolate it by a factor of M. Results are shown in Figure 4.

Finally, we filter the interpolated signal with the inverse filter. Results of the spectrograms (original vs. reconstructed) is shown in Figure 5. Finally, in Figure 6 we plot the residuals as a result of the subtraction of the original signal and the reconstructed signal. The resulting sound was heard and it sounded almost the same.

**Figure 2.** Spectrum and SPL of a windowed frame.

**Figure 3.** Masking threshold and threshold in quiet for frame X.

**Figure 4.** Masking threshold of current frame only (blue), masking threshold of previous frame multiplied with a decay constant (red line), and joined mask (black dots, the maximum of these two). Note that there is slight temporal masking happening in this frame at bands.

**Figure 5.** Sample Signal-to-Mask ratio plot from a 3.5s input signal.

**Figure 6.** Bits allocated to each sub-band.

**Figure 7.** Spectrum of input signal. The signal contains drums and bass guitar.

References

Pulkki, V., & Karjalainen, M. (2015). Communication Acoustics. John Wiley & Sons.

* Zwicker E, Fastl H, Psychoacoustics, Springer-Verlag, Berlin, Germany, 1990
* [2]  Painter T, Spanias A, A Review of Algorithms for Perceptual Coding of Digital Audio Signals,   http://www.eas.asu.edu/~speech/ndtc/dsp97.ps

# Code

1. http://www.iis.fraunhofer.de/content/dam/iis/de/doc/ame/conference/AES-116-Convention\_guideline-to-audio-codec-delay\_AES116.pdf [↑](#footnote-ref-1)