Effects of Quantization Error in MP3

Tyler Hattori
University of California, Santa Barbara
thattori@ucsb.edu

Abstract

The MP3 hybrid filter bank utilizes the psychoacoustic model to allocate bits in the encoding process in such a way that the quantization error is imperceptible to the human ear. While the MP3 model is generally successful in this task and enjoys widespread support in industry, its most basic implementation allows for unwanted artifacts in the reconstructed process when subjected to specific input signals. In this presentation, I will characterize these artifacts for the inputs of castanets, wideband male and female speech, and modern indie music. I will show that harmful auditory effects like pre-echo, reverb, and flanging can all arise from quantization error, but their presence is dependent on the type of input signal being encoded. Moreover, I will show that the effects of quantization error can be reduced by adjusting the length of the MP3 MDCT windows or the magnitude of the psychoacoustic masking threshold. In these cases, I will show that bitrate is sacrificed for performance.

1. Introduction

The MP3 hybrid filter bank uses 32 maximally decimated subbands. The output of these subbands are each fed into an 18-pt MDCT filter, quantized according to the psychoacoustic model, and Huffman encoded. The resulting bitstream is then Huffman decoded and the signal is reconstructed using an inverse quantization and IMDCT process and applying a 32-band synthesis filter bank. In this presentation, I will examine how different reconstructed signals are affected by changes specifically relevant to the psychoacoustic model and the MDCT. I obtained all of the plots and graphs presented by modifying and running the code outlined in Gerald Schuller's book on perceptual audio encoding [1].

2. The Psychoacoustic Model

2.1. Motivation

The psychoacoustic model is meant to identify components of an input signal that the human ear is unable to detect. It achieves this by calculating a "masking threshold," which defines the minimum magnitude a tone would need to be in order to be detected by the human ear. In the quantization step, bits are only allocated to frequency components above this magnitude—thereby removing unnecessary components in the input signal and saving on bitrate. The human ear has a natural masking threshold, which is well defined across its 20-20kHz bandwidth and is shown in Figure 1.

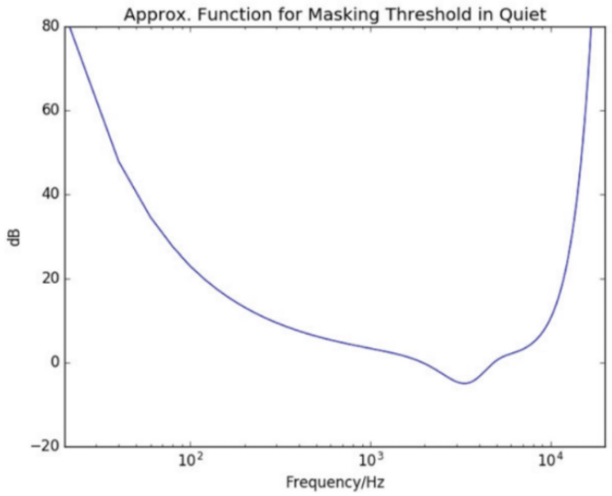

Figure 1. The natual masking threshold of the human ear

The human ear also has a certain property where an input tone will raise this natural masking threshold around the vicinity of its frequency. In other words, it is difficult for the human ear to hear a quiet tone when in the presence of a much stronger tone at a nearby frequency. The psychoacoustic model reflects this property by taking an FFT of the input signal and calculating a "spreading function"

for each peak in the input spectrum. These spreading functions model how the natural masking threshold of the human ear would be affected in the presence of the given tone. By adding each of these spreading functions to the natural masking threshold of the human ear, a new masking threshold is obtained and used as a reference for quantization.

2.2. Example for a Constant Tone Input

We can observe the psychoacoustic model in action by looking at its response to a 4.7kHz sine wave input. Figure 2a-d shows this process.

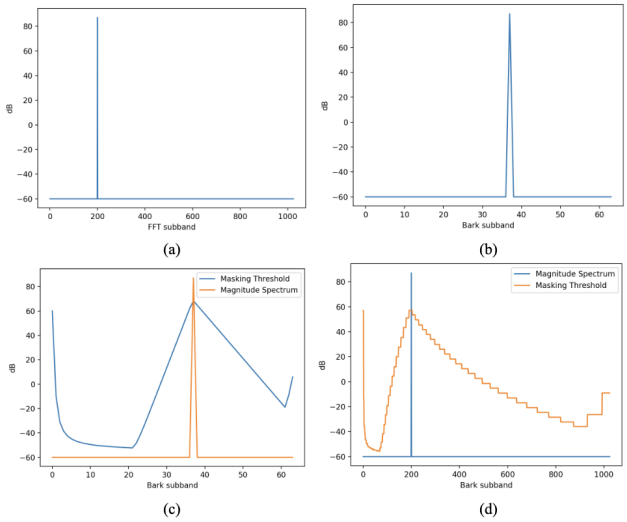

Figure 2. Steps to calculating the psychoacoustic masking threshold

Figure 2a shows the FFT of the input. As expected, there is only one frequency component for the constant tone input. Observe in Figure 1 that the human ear is not linearly sensitive in frequency. In fact, most of the information in a signal that the human ear cares about lies below the 4kHz range—even though it has a 20kHz bandwidth. Because of this trait, the spreding functions used in the psychoacoustic model are also nonlinear in frequency. To remedy this, the psychoacoustic model first converts the uniform FFT subbands into Bark subbands, which effectively spread out the information contained in lower frequencies. Figure 2b illustrates this conversion and Figure 2c shows the resultant masking threshold in the Bark domain. Notice that the spreading function is linear in the Bark domain. Figure 2d shows the masking threshold converted back to the uniform FFT domain, where the input's effect on the natural masking threshold of the human ear is clearly seen.
In the quantization process, components of the input signal that are the highest in magnitude relative to the psychoacoustic masking threshold (PMT) are allocated the most bits, while components below the PMT are allocated none. Therefore, it is relevant to discuss "reconstruction quality," which operates on the percentage scale and refers to manually raising or lowering the PMT in order to decrease or increase the amount of encoded information. The magnitude of the PMT is scaled according to $(PMT) * (100/Q)$, where Q is the reconstruction quality. For example, this means if $Q = 100$ the PMT is unchanged, but if $Q = 50$ the PMT is raised by 6dB (doubled in magnitude). Because less information is encoded, decreasing Q introduces quantization error while it saves on bitrate. In this presentation, I will decrease Q manually in order to characterize quantization error for different input signals.
Finally, notice that in Figure 2d the coefficients of the PMT are quantized. For the sake of time, I did not focus on this aspect of the psychoacoustic model because it is not immediately relevant or necessary to my analysis, but if you would like to see plots showing the levels of these quantization indices I can easily obtain them for you.

3. Quantization Error

I will now show how the PMT varies for three types of input signals. Specifically, I will look at how MP3 handles signals with sharp transients, speech, and music. Moreover, for each of these signals I will show how the reconstructed audio and bitrate are affected by changing the reconstruction quality Q and the MDCT window length L.
We will see that quantization error (generally) sounds like pre-echo for sharp transients, reverb for speech, and distortion for music. We will also see that removing these artifacts is almost always at the cost of bitrate.

3.1. Castanets

I will first look at pre-echo, which arises when the onset of a sharp transient in an input signal is unideally placed at the overlap between two windows in the MDCT. Because a castanets input signal has several sharp transients, we can hear in the reconstruction that a slight echo arises just before the onset of each transient. Figure 3a shows the FFT of the castanets input, figure 3b shows the calculated PMT, figure 3c shows the FFT of the reconstructed signal, and figure 3d shows the quantization error. These plots were found for $Q = 100$ and $L = 1024$.

As expected, the reconstructed signal only has frequency components where the original signal surpassed the PMT. The resulting quantization error signal (the difference between the original and the reconstruction) lies below the PMT entirely.
However, we notice that if we decrease the reconstruction quality from 100 to 90, we hear pre-echo in the output. In fact, the pre-echo is slightly audible for 100 quality. We can remove this effect by reducing L from 1024 to 512. Since this shortens the MDCT windows, even if the onset of a

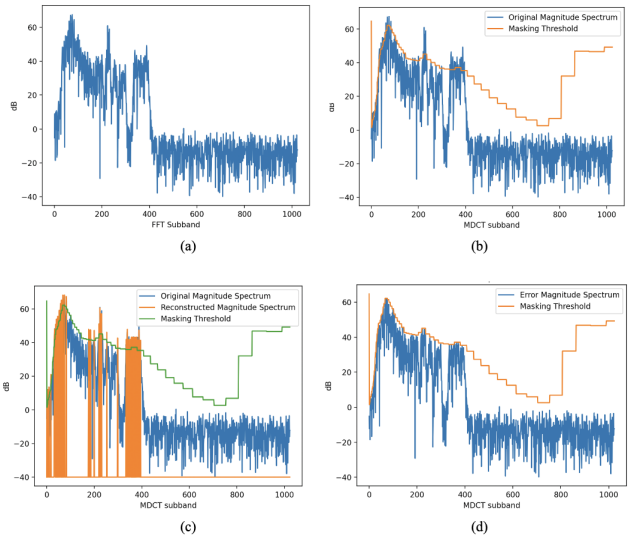

Figure 3. Calculated PMT for a castanets input

transient is placed in between two windows, the energy of the transient is not spread out as much in the reconstruction and the length of the pre-echo is minimized. This shows that simply decreasing the PMT is not always the most effective way to remove harmful artifacts. Figure 4 compares the spectrogram of the original signal (top) with the spectrogram of the reconstructed signal for different settings of Q and L.

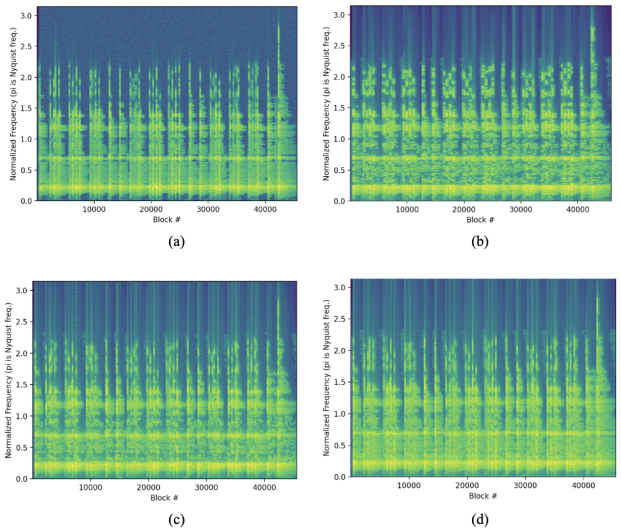

Figure 4. Spectrograms for the orignal castanets input (a), the reconstruction for $(Q, L) = (60, 1024)$ (b), $(Q, L) = (60, 256)$ (c), and $(Q, L) = (100, 256)$ (d)

Comparing Figure 4b with the spectrogram of the original signal in Figure 4a, we see that decreasing Q introduces holes in the spectrogram. This is because we are selecting which information should be encoded in the psychoacoustic model, and this loss of information is reflected in the spectrogram of the reconstruction. From Figure 4b, we also see that the quantization error gets spread out in moments immediately before the onset of a transient. From Figure 4c, we see that decreasing L reduces the length of these moments. From Figure 4d, we see that increasing Q to 100 fills in the holes in the spectrogram.

Listening to the reconstructed audio, we find that reducing L is very effective in removing pre-echo. However, I find that with $L = 256$ and $Q = 60$ a new effect arises which sounds a lot like flanging. This shows how quantization error does not always sound the same—even for a constant input signal. We will see that this flanging effect from decreasing L is more prominent for input signals with higher pitches.

Figure 5 shows how bitrate increases when we decrease L or increase Q.

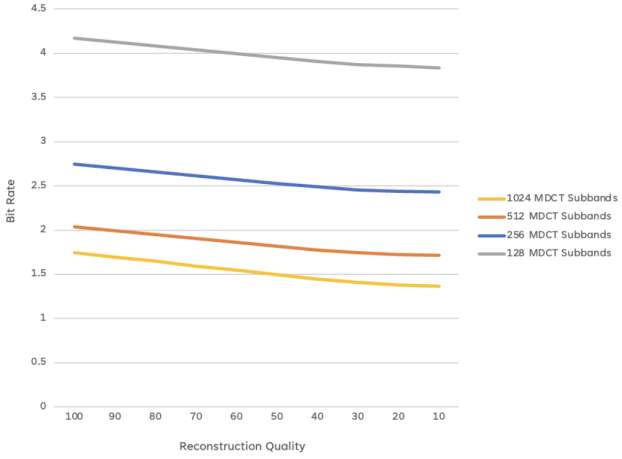

Figure 5. Bitrate vs Reconstruction Quality for a castanets input

This trend will stay consistent for the other input signals. It basically shows that encoding the input signal more accurately will always take more processing power. However, as we have seen in the case of the flanging effect, sometimes decreasing L (increasing bitrate) will simply make the quantization error sound different.

3.2. Male Speech

For wideband male speech, we are dealing with low pitches and longer harmonics. Figure 6a shows the FFT of the male speech input, figure 6b shows the calculated PMT, figure 6c shows the FFT of the reconstructed signal, and figure 6d shows the quantization error. These plots were found for $Q = 100$ and $L = 1024$.

We notice that the PMT for a male speech input is almost to 0dB at higher frequencies. This reflects how most of the information needed to reconstruct a male speech signal is

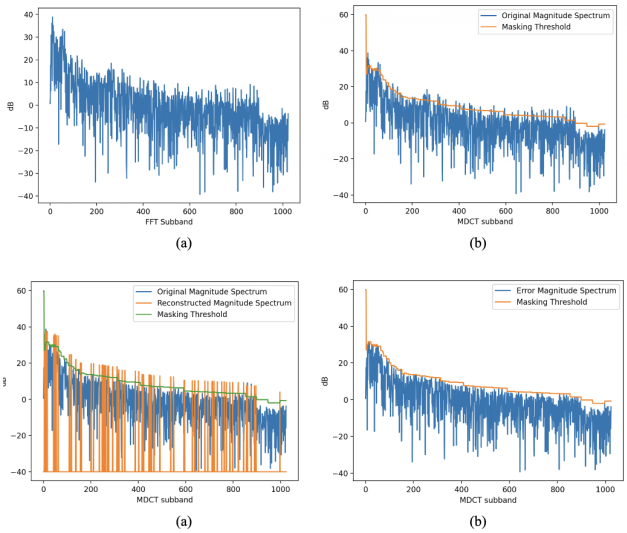

Figure 6. Calculated PMT for a male speech input

contained in the low frequencies. For this kind of input, the quantization error mostly takes the form of reverb, with some pre-echo arising at the start of certain words. This makes sense, as the energy of the quantization error is effectively being spread out constantly. In other words, it seems that this is the same phenomenon which led to pre-echo for a castanets input. The spectrograms of the original and the reconstructions are shown in Figure 7.

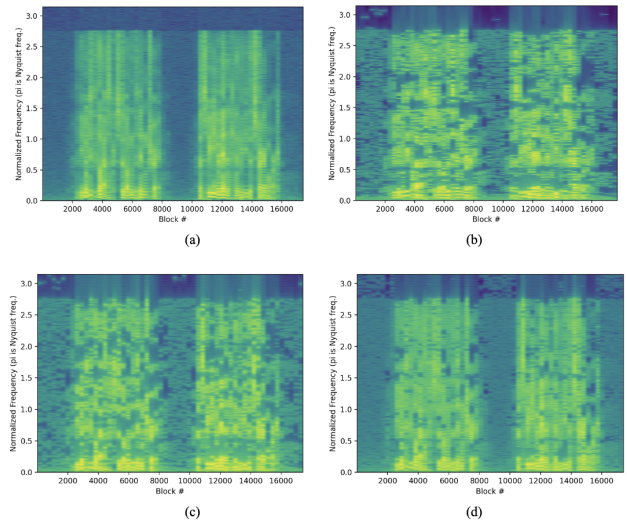

Figure 7. Spectrograms for the orignal male speech input (a), the reconstruction for $(Q, L) = (60, 1024)$ (b), $(Q, L) = (60, 256)$ (c), and $(Q, L) = (100, 256)$ (d)

Listening to the reconstructed audio, we find that reducing L is very effective in removing pre-echo. This is expected, as I hypothesized that the source of the reverb is the same source of the pre-echo in the castanets input (quantization error caused each to arise). However, I find that the reverb is more difficult to remove than the pre-echo, and with $L = 256$ and $Q = 60$ flanging is not heard nearly as much. This confirms that the flanging effect arises for input signals with higher frequency content, and the reverb effect is more prominent for signals with lower frequency content. I attribute this to the fact that lower frequencies have longer waveforms, which means they are more suseptible to the MDCT windowing issue. This is likely why decreasing L had a larger effect on the castanets signal than the male voice, as the quantization error arose due to the locations of the sharp onsets rather than the character of the entire signal.

Figure 8 shows how bitrate increases when we decrease L or increase Q.

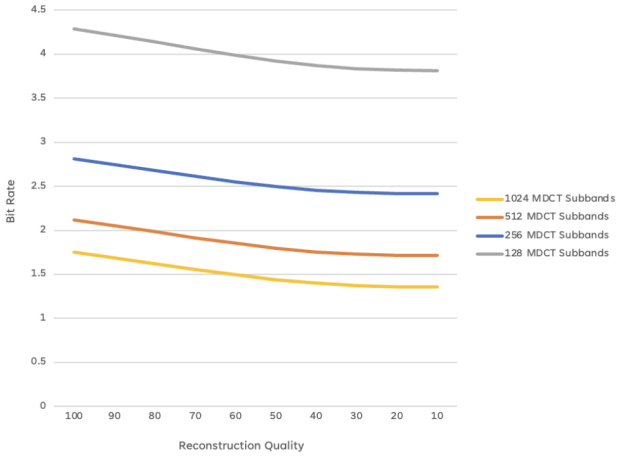

Figure 8. Bitrate vs Reconstruction Quality for a castanets input

We see that the trend is the same as before, as expected.

3.3. Female Speech

For wideband female speech, we are dealing with high pitches and shorter harmonics. Figure 9a shows the FFT of the female speech input, figure 9b shows the calculated PMT, figure 9c shows the FFT of the reconstructed signal, and figure 9d shows the quantization error. These plots were found for $Q = 100$ and $L = 1024$.

We notice that the PMT for a female speech input is much higher than 0dB at higher frequencies, which contrasts the case for male speech. We therefore expect less reverb and more flanging than the male speech reconstruction as we decrease L and Q. The spectrograms of the original and the reconstructions are shown in Figure 10.

Listening to the reconstructed audio, we confirm that the flanging effect is prominent when we use short windows and poor reconstruction quality. The case of female speech is interesting because I find that the best setting to use is

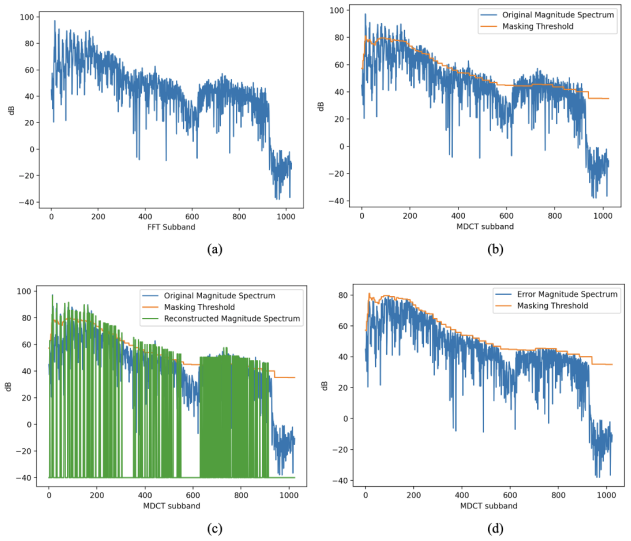

Figure 9. Calculated PMT for a female speech input

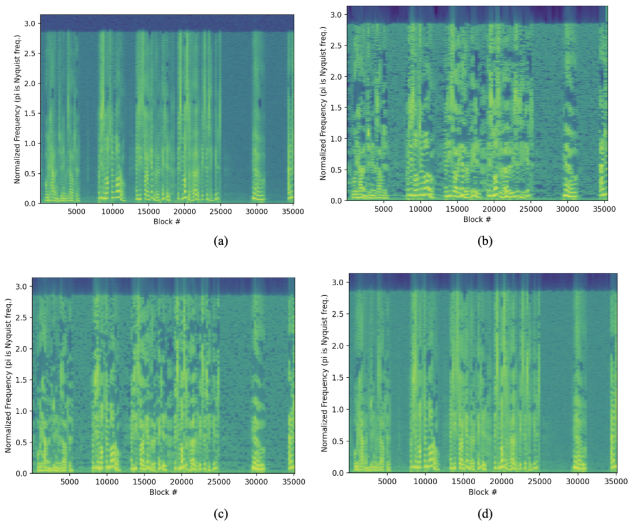

Figure 10. Spectrograms for the orignal female speech input (a), the reconstruction for $(Q, L) = (60, 1024)$ (b), $(Q, L) = (60, 256)$ (c), and $(Q, L) = (100, 256)$ (d)

$L = 512$. In the case of the castanets and male speech inputs, it was beneficial to reduce L in order to remove the pre-echo and reverb. However, in the case of female speech, the flanging effect is more of an issue than the reverb, so decreasing L as much as possible seems to not always be the best course of action. One could say that this is also true for the castanets input, as flanging will eventually become more of an issue than pre-echo as L is decreased, but the point is emphasized in the case of a female speec input.
Figure 11 shows how bitrate increases when we decrease L or increase Q.

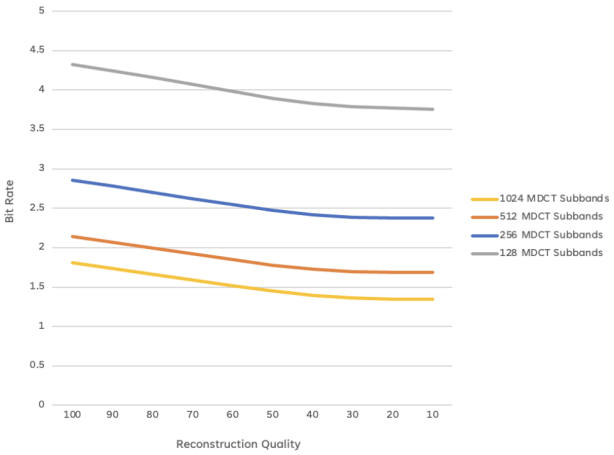

Figure 11. Bitrate vs Reconstruction Quality for a female speech input

We see that the trend is the same as before, as expected.

3.4. Indie Music

The music clip I have chosen contains a male voice, sharp transients from a snare drum, and high pitches from an electric guitar. With all of these ”basic” inputs meshed together, it becomes obvious that there is no simple fix to reducing the effect of quantization error—at least with the standard MP3 model. Figure 12a shows the FFT of the music input, figure 12b shows the calculated PMT, figure 12c shows the FFT of the reconstructed signal, and figure 12d shows the quantization error. These plots were found for $Q = 100$ and $L = 1024$.

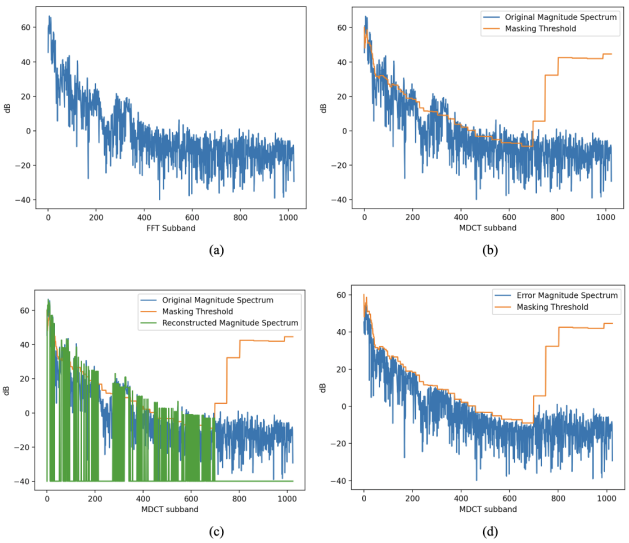

Figure 12. Calculated PMT for a music input

We notice that the PMT for a music input places emphasis

on both the low and high frequency components of the input signal. We therefore expect reverb, flanging, and pre-echo to arise as we decrease L and Q. The spectrograms of the original and the reconstructions are shown in Figure 13.

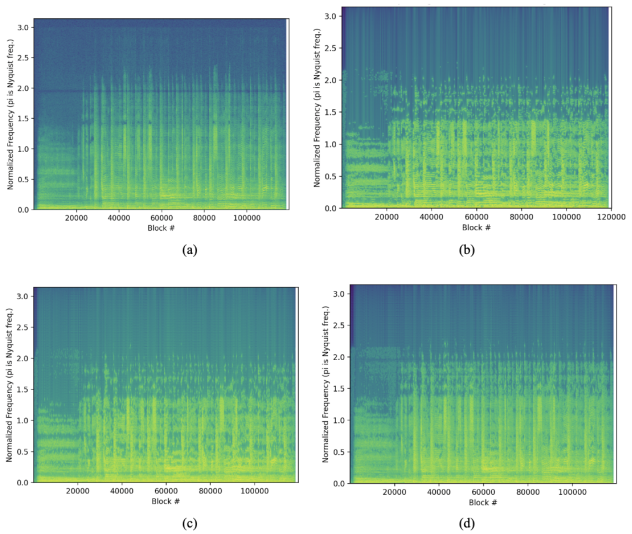

Figure 13. Spectrograms for the orignal music input (a), the reconstruction for $(Q, L) = (60, 1024)$ (b), $(Q, L) = (60, 256)$ (c), and $(Q, L) = (100, 256)$ (d)

Listening to the reconstructed audio, we confirm that general distortion is present for low Q and any L. Reducing this quantization error now becomes extremely case by case, as we must take into account the singer's voice, the locations of sharp transients, and the frequency content of each instrument. Most importantly, we see the need for more advanced developments to the MP3 model that can adapt to the content of specific inputs.
Figure 14 shows how bitrate increases when we decrease L or increase Q.

We see that the trend is the same as before, as expected.

4. Modern Developments to the MP3 Model

Two modern solutions to reducing quantization error I have found interest in are spectral band replication and adaptive filtering. I will briefly go over these developments.

4.1. Spectral Band Replication

Recall that the human ear can glean most of the information of a signal from its components below 4kHz. In spectral band replication, or bandwidth extension, only the low frequency content is directly encoded. Instead of encoding the higher frequency components directly, spectral band replication instead stores the shape of the power spectral envelope (PSE) for higher frequencies. In the reconstruction

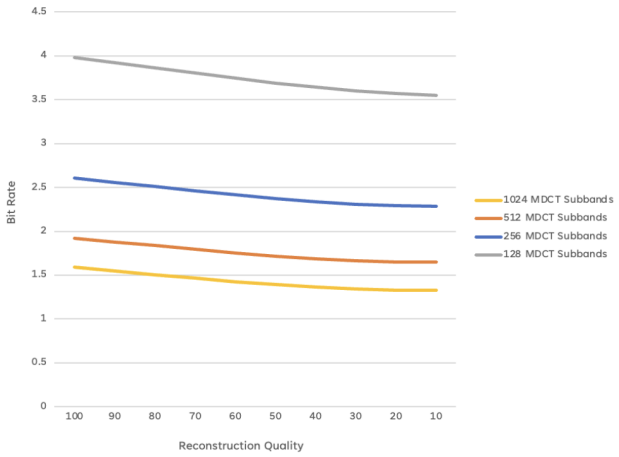

Figure 14. Bitrate vs Reconstruction Quality for a music input

process, the information of the lower band is simply replicated across the higher bands and scaled according to the PSE. The difference is almost indiscernible and the process saves on bitrate. Figure 15 illustrates this process [2].

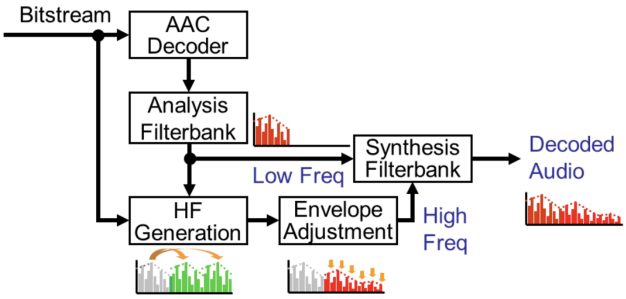

Figure 15. Spectral Band Replication (Bandwidth Extension)

Notice that applying the synthesis filterbank is still the last step in the reconstruction process.

4.2. Adaptive Filtering

Recall that pre-echo arises in signals with sharp transients and is reduced by shortening the length of the MDCT windows. Adaptive filtering uses transient detection to know when a sharp onset is coming and shorten the length of the MDCT windows for a brief period of time. This solution allows for the pre-echo to be minimized, but does not necessarily affect the rest of the reconstruction. Figure 16 illustrates the idea behind adaptive filtering [3][4].

Figure 16 top shows the typical MP3 process of using long 18-pt MDCT windows at each of the 32 outputs of the analysis filter bank. The set of 18 coefficients produced by each MDCT are calculated from 36-sample windows with 50 percent overlap between adjacent windows. In the reconstruction process, the inverse MDCT takes each of these 18-coefficient sets and calculates a corresponding 36-sample

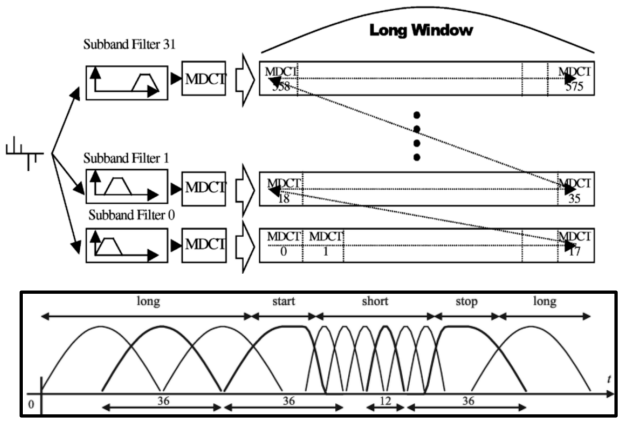

Figure 16. Adaptive filtering

window. These windows are then overlapped by 50 percent and added together to produce the reconstructed signal. With adaptive windowing, we can instead use 12-sample windows when a sharp onset is approaching by changing the length of the MDCT for a short period of time. This results in a greater number of encoded coefficients for this short interval, but it greatly reduces pre-echo.

5. Conclusion

I have shown that MP3 quantization error can produce various harmful effects in the reconstruction of different audio signals. In many cases, removing these harmful effects may be as simple as decreasing the masking threshold of the psychoacoustic model, but this method will lead to an increased bitrate and will not guarantee that every type of effect will be completely removed. For example, in the case of a castanets input, we saw that a better method for removing pre-echo came from decreasing the length of the MDCT windows. However, decreasing this length too far would introduce a flanging effect. Moreover, we saw that a music signal input can be characterized by a combination of several different basic input signals, so a simple process for removing any type of quantization error in music is not well defined. We conclude that the process of removing harmful artifacts in the reconstruction steps of the basic MP3 model should be regarded as a case by case issue dependent on the type of input signal. This illustrates the importance of more modern developments to perceptual audio encoding, such as subband replication and adaptive filtering. In many of these developments, the analysis and quantization steps of the MP3 model are tailored to specific characteristics of the input signal, which I have shown is crucial to a reliable model.

6. References

[1] Schuller, Gerald. (2020). Filter Banks and Audio Coding: Compressing Audio Signals Using Python. Springer International Publishing.

[2] Gaël Richard, Paris Smaragdis, Sharon Gannot, Patrick A Naylor, Shoji Makino, et al.. Audio Signal Processing in the 21st Century. IEEE Signal Processing Magazine, 2023, ff10.1109/MSP.2023.3276171ff. ffhal-04112575.

[3] Kiranyaz, Serkan and Qureshi, Ahmad and Gabbouj, Moncef. (2004). A Generic Audio Classification and Segmentation Approach for Multimedia Indexing and Retrieval. Audio, Speech, and Language Processing, IEEE Transactions on. 14. 1062 - 1081. 10.1109/TSA.2005.857573.

[4] Zieliński, T.P. (2021). Audio Compression. In: Starting Digital Signal Processing in Telecommunication Engineering. Textbooks in Telecommunication Engineering. Springer, Cham.