YouTube Audio Quality

You Tube is clearly used by a very large number of people. In the main they will be interested in watching vidoes of various types of content. But it also gets used to distribute, and make people aware of, audio recordings. A recent conversation on the "Pink Fish" webforum set me wondering about the technical quality of what is on offer from You Tube (YT). The specific comment was a claim that the 'opus' audio codec gives better results than the 'aac / mp4' alternative. So I decided to investigate...

Ideally to assess this requires a copy of what was uploaded to YT as a 'source' version which can then be compared with what YT then make available. By co-incidence I had also quite recently joined the Ralph Vaughan-Williams Society (RVWSoc). They have been putting videos up onto YT which provide excerpts of the recordings they sell on Audio CDs. These proved excellent 'tasters' for anyone who wants to know what they are recording and releasing. And when I asked, they kindly provided me with some examples to help me investigate this issue of YT audio quality and codec choice.

For the sake of simplicity I'll ignore the video aspect of this entirely and only discuss the audio side. The RVWSoc let me have 'source uploaded' copies of two examples. The choice of audio formats that they offerred of these videos are as follows:

```
Pan's Anniversary
Available audio formats for HZHVTr1w6L8:
ID        EXT   | ACODEC     ABR  ASR    MORE INFO
------------------------------------------------------------
139-dash m4a   | mp4a.40.5   49k 22050Hz DASH audio, m4a_dash
140-dash m4a   | mp4a.40.2 130k 44100Hz DASH audio, m4a_dash
251-dash webm  | opus      153k 48000Hz DASH audio, webm_dash
139       m4a   | mp4a.40.5   48k 22050Hz low, m4a_dash
140       m4a   | mp4a.40.2 129k 44100Hz medium, m4a_dash
251       webm  | opus      135k 48000Hz medium, webm_dash

Brass
Available audio formats for KsILRbZtTwc:
ID        EXT   | ACODEC     ABR  ASR    MORE INFO
------------------------------------------------------------
139-dash m4a   | mp4a.40.5   50k 22050Hz DASH audio, m4a_dash
140-dash m4a   | mp4a.40.2 130k 44100Hz DASH audio, m4a_dash
251-dash webm  | opus      149k 48000Hz DASH audio, webm_dash
139       m4a   | mp4a.40.5   48k 22050Hz low, m4a_dash
140       m4a   | mp4a.40.2 129k 44100Hz medium, m4a_dash
251       webm  | opus      136k 48000Hz medium, webm_dash
```

One aspect of this stands out immediately. This is the variety of audio sample rates (ASR) on offer. In each case only one version was uploaded, at a sample rate chosen by the RVWSoc. I had expected to see a choice of audio codecs (compression systems), but was quite surprised, in particuar, to see ASRs as low as 22·05k on offer. Given the main interest here is in determining what may give te highest audio quality I decided that analysis should focus on the higher, more conventional rates – 48k and 44k1. More generally, the above shows that – since in each case only one source file (and hence only one sample rate) was uploaded, some of the above offerred version at 48k or 44k1 also have been thorough a sample rate conversion as well as perhaps a codec conversion. Which introduces another factor that may degrade sound quality! In this case I had copies of what had been uploaded, so could determine which YT output versions had been

though such a rate conversion. However in general YT users won't know which version may have dodged that particular potential bullet!

Pan's Anniversary

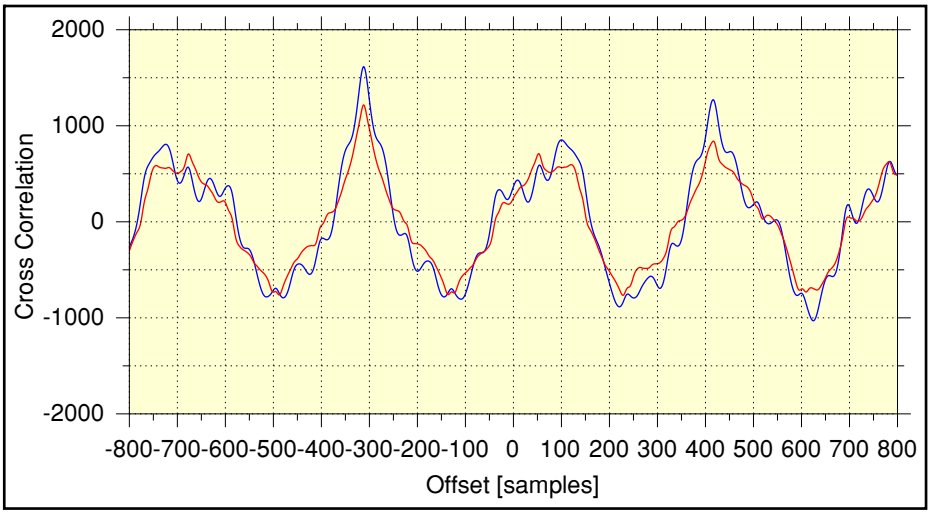
I'll begin the detailed comparisons with the video of an excerpt from the CD titled,"Pan's Anniversary". The version uploaded contains the audio in the form encoded in the aac(LC) codec at a bitrate of 194 kb/s and using a sample rate of 48k. The audio lasts 4 mins 54·88 sec. The table below compares this with the same aspects of the high ABR versions offerred by YT.

version	codec	ABR	bitrate (kb/s)	duration (m:s)
source	aac(LC)	48k	194	4:54·88
YT-140	aac(LC)	44k1	127 fltp	4:54·94
YT-251	opus	48k	135 fltp	4:54·90

We can see that YT-140 uses the same codec as the source, but alters the information bitrate. YT-521 transcodes the input aac(LC) to opus, but doesn't alter the sample rate. *Both* of the YT versions are of longer duration than the source uploaded. By loading the files into Audacity and examining the waveforms by eye it became clear that the YT versions were not time-aligned with each other, or with the source.

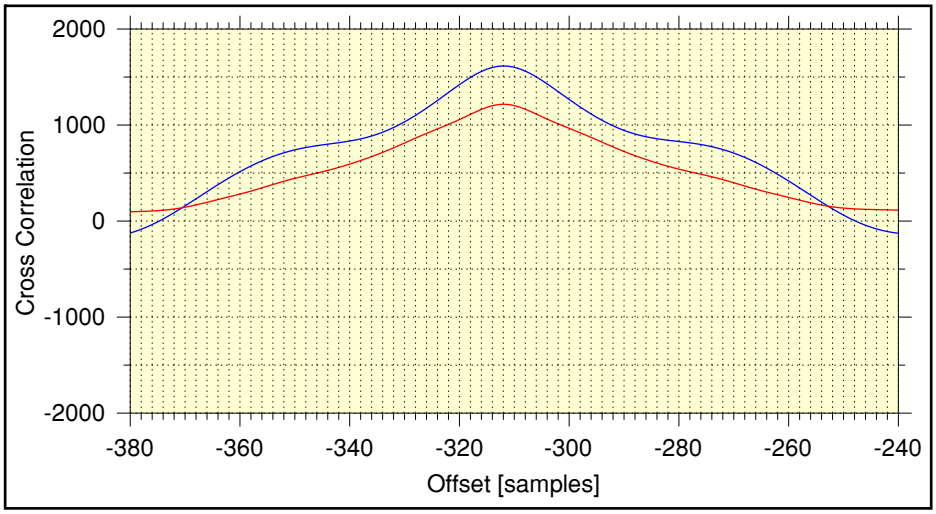
To avoid any changes caused by alteration of the ABR I decided to concentrate on comparing the source version with YT-251 – i.e. where the output uses the opus codec, not aac, but maintain's the source sample rate. Having chosen matching sample rates the simplest and easiest was to check how similar two versions are is to time-align them and then subtract, sample by sample, each sample in a sequence in one version from the nominally 'same instant' matching one in the other. If the patterns are the same, the result is a series of zero-valued samples. If they don't match we get a 'difference' pattern. However, first we have to determine the correct time-offset to align the sample sequences of the two versions.

In some cases that can be fairly obvious from looking at the sample patterns using a program like Audacity. But in other cases this is hard to see with enough clarity to determine with complete precision. Fortunately, we can use a mathematical method known as *cross-correlation* to show us the time alignment of similar waveform patterns. (See https://en.wikipedia.org/wiki/Cross-correlation if you want to know more about cross correlation.) This also can show us where the best alignment may occur in terms of any offset between the two patterns of samples being cross correlated.

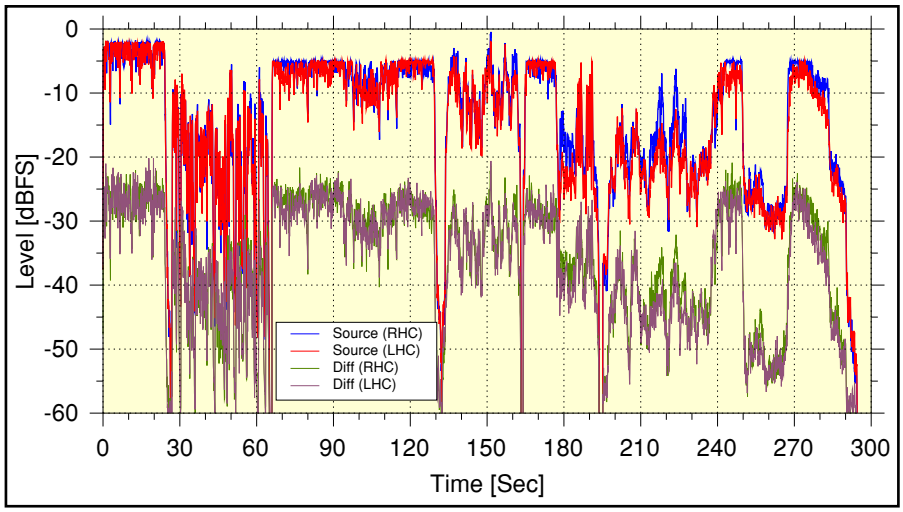


The above graph shows the resulf of cross correlating a section of the source and YT-251 versions of the audio. (The red and blue lines show the Left and Right channels of the stereo.) The results cover offsets over a range of +/- 800 samples. The process used 180,000 successive sample pairs from each set of samples. i.e. about 3·75 sec of audio from each.

The best alignment is indicated by the location of the largest peak. If the sample sequences were already aligned this would happen at an offset = 0. However he we can see that the YT-251 version is 'late' by just over 300 samples.



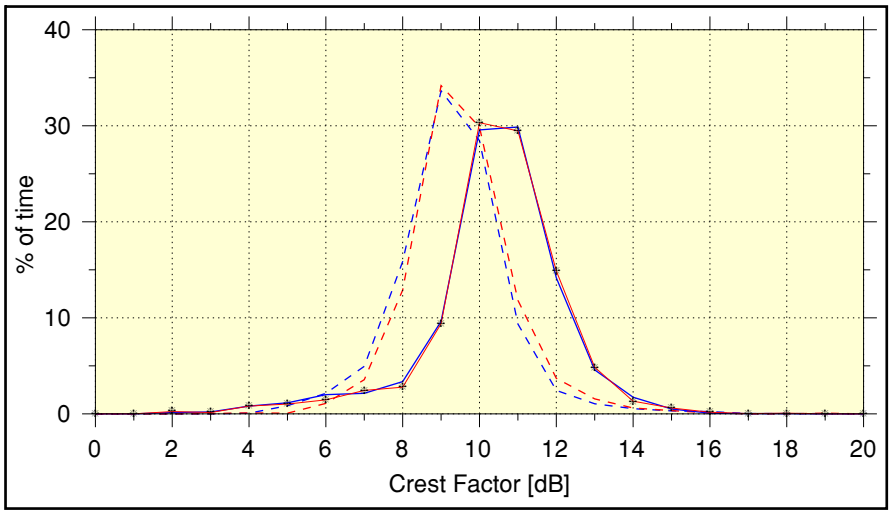
Zooming in, we can see that the peak is at an offset of -312 samples. Which at 48k sample rate corresponds to YT-251 being 6·5 milliseconds late. Having determined this I could trim 312 samples from the start of each channel of YT-251 and this aligned the two series of samples. (I also then had to trim the end to make them of equal length.) Once this was done it becomes possible to run though the samples and take a sample-by-sample difference between the source version and YT-251. This different set then shows the details of how the YT-251 output differs from the source version.



The graph above shows how the rms audio power level of the audio varies with time. The red and blue lines show the levels in the Pan's Anniversary source file. The green and magenta lines show the power levels versus time obtained from subtracting the source file sample values from the audio samples from YT-251. Ideally, we'd want a subtraction like this to produce a series of

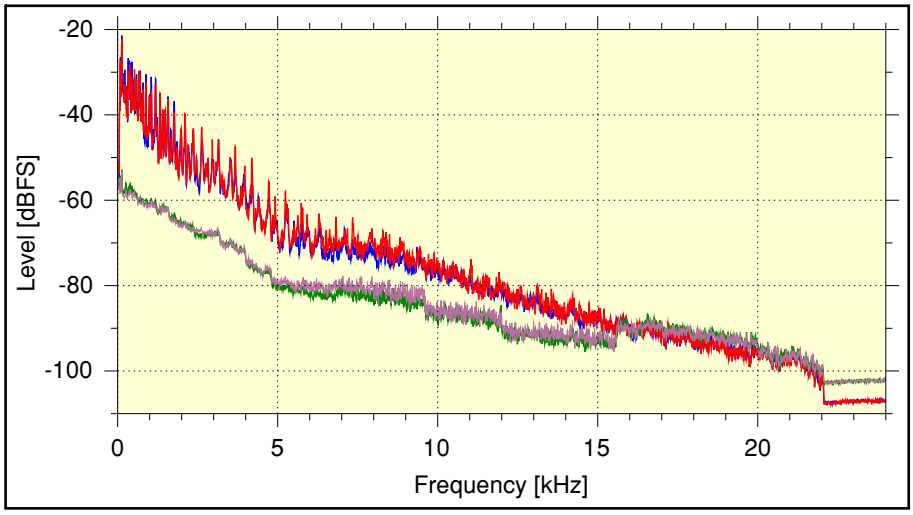
zeros as the difference samples because this would tell us that we got out from the YT processing exactly what had been submitted. But the above results show this clearly is ***not*** the case! There is a residual 'error' which is somewhere around 30 to 35dB below the input musical level.

In traditional terms for audio, 30dB would be regarded as a very poor signal/noise ratio. And if the change was considered as being equivalent to conventional distortion it would be assumed to indicate a level of around 3% distortion! So it represents a rather underwhelming result. However a more benign interpretation may be that it arises as a result of the process applied by YT slightly altering the overall amplitudes of the waveforms so they don't quite match – hence leave a non-zero difference when the input and output are substracted. With this in mind we can compare the input and output samples using other methods that aren't sensitive to an overall change in signal pattern levels.



The above graph compares the input and output files in terms of Crest Factor. This measures the peak/rms power levels of the waveform shapes defined by the series of samples. The broken lines show the Left and Right channel results for the source file sent to YT. The solid lines the equivalent results for YT-251. To obtain these results each set of samples was divided up into a series 0·1 Sec sections. The peak and rms power level of each was calculated, and the above shows how often a given value was obtained, grouped into 1dB wide statistical 'bins'. For a pure sinewave the peak/rms crest factor is 3dB. i.e. the peak levels are 3dB larger than the rms power. For well recorded music from acoustic instruments the Crest Factor tends to be in the range from a few dB up to well over 10dB for the most 'spiky' waveforms.

The result is interesting as we can see that the YT-251 output clearly exhibits a *different* Crest Factor distribution to the source file. It seems doubtful this could be produced by a simple change in the overall signal pattern level. (e.g. a simple volume control does change the overall level, but it should not change the *shape* of the audio waveform, and hence should leave the Crest Factor unaltered. If it *did* alter this, you'd be advised to replace the control with one that worked properly!) It is particularly curious that the Crest factor seems to be increased by having the audio pass though the YT processing. Although possibly this may arise due to an input which is aac(LC) coded being transcoded into 'opus' codec form. OTOH perhaps YT apply some form of 'tarting up' to make audio 'sound better'...



A more familiar way to show the character of an audio recording is to plot its spectrum. The above graph shows the spectrum of the Pan 'source' file (red and blue lines) and of the series of samples obtained by subtracting the YT-251 output sample series. (purple and green lines). We can then say that - at any given frequency - the bigger the gap between the red/blue lines and the green/purple ones, the closer the YT-251 output is to the source supplied to YT. Looking at the graph we can then see that the results indicate that the faithfulness of the YT-251 result to the input is at its highest at low frequencies where the gap is widests and the contributions to the overall signal level are greatest. However at higher frequencies the level of the error becomes a larger fraction of the input. And above about 16kHz the error level is actually *bigger* than the input signal power! (We can also see that the source, although at a 48k sample rate, has a sharp cutoff at just over 22 kHz. This indicates that that although what was submitted to YT was at 48k sample rate it was actually generated from a 44k1 (i.e. audio CD rate) version.)

The behaviour of the above spectra may will be another sign of changes that also produced an increase in the typical Crest Factor.

Having applied the above analysis to an example that produced a YT output using the opus codec we can now examine another example, this time using YT-140 output and a 44k1 source file that was uploaded to YT.