A Suggested Explanation For (Some Of) The Audible Differences Between High Sample Rate And Conventional Sample Rate Audio Material

Mike Story[1]

Background

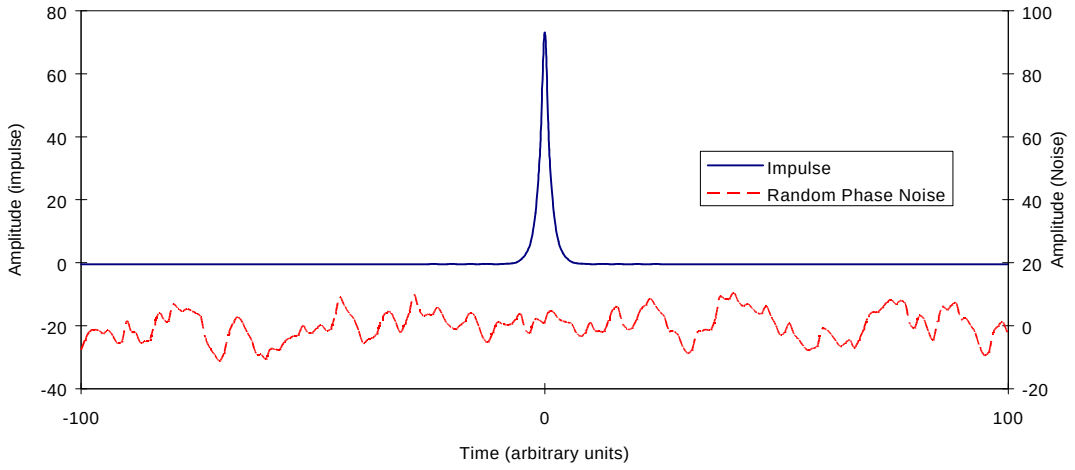
The bulk of the work on explaining how we hear has been concerned with frequency response. Frequency response[2] of the ear has been extensively investigated, and the frequency response of a recording or reproduction chain is by far the dominant factor in determining how realistic or intelligible the resulting signal is. Although there is some argument about exact numbers, about 20 kHz is generally accepted as the maximum frequency that humans can hear. We may not be able to tell one frequency from another above about 12 kHz - they all appear in the "max. frequency" bin of our discrimination mechanism - but we can tell they are there.

Digital audio systems have historically made use of this 20 kHz limit to set sampling rates. When CD formats were first established, the problem of storing the large amount of data needed for about 1 hours stereo playing time was substantial, so sample rates were set as low as reasonably possible, consistent with maintaining a 20 kHz bandwidth. 44.1 kS/s gave and still gives an unambiguous frequency range of 22.05 kHz (Nyquist principle).

In principle, if frequency response were the only issue, there would be no advantage in moving to formats with higher sampling rates. However, the evidence is otherwise. Direct comparisons of the same source material, recorded and reproduced at 44.1 kS/s, 96 kS/s and 192 kS/s show that there is an advantage in going to the higher rates - it sounds better! The descriptions of those used to making such comparisons tend to involve such terms as "less cluttered", "more air", "better hf detail" and in particular "better spatial resolution". We are left wondering - what mechanism can be at work? It seems unlikely that we have all suddenly developed ultrasonic hearing capabilities.

Actually, a little thought also suggests that frequency response cannot be the only factor at work in our hearing apparatus. Figure 1 shows two waveforms that have identical (power) spectra, and yet sound very different - a bandlimited impulse (a click) and a type of white noise. Other waveforms can easily be generated that have the same amplitude response, but sound (substantially) different still. Something else must be going on.

Figure 1 - Waveforms for two signals with identical Amplitude Spectra



[1] with *dCS* Ltd, Mull House, Great Chesterford Court, Great Chesterford, Saffron Walden, UK

[2] We refer to a recording or reproduction chain as having a frequency response, and we refer to a signal as having a frequency power spectrum - or sometimes, because frequency is so overwhelmingly important, just a spectrum.

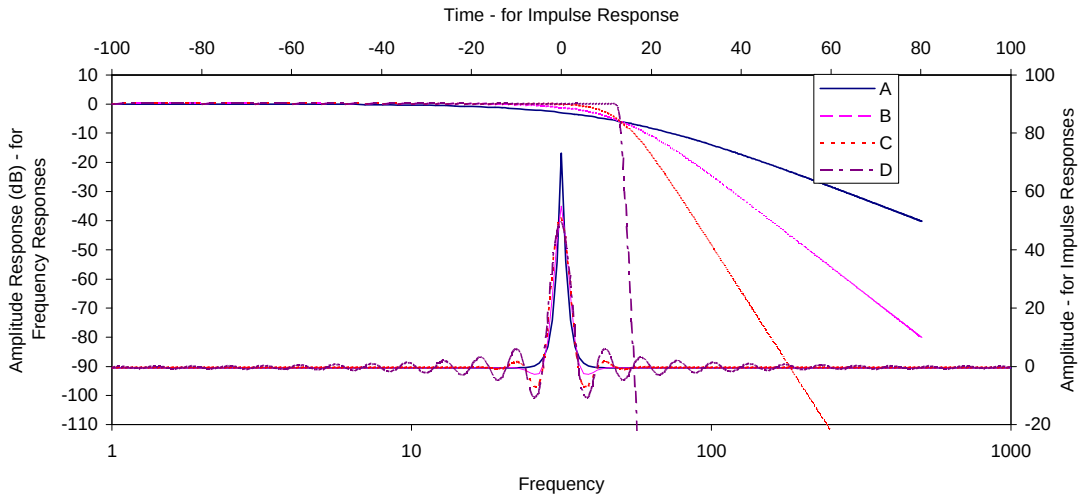
To generate the waveforms, we diddle with the phase of the individual frequency components. Mathematically, phase information is the only other variable needed to convert from a spectrum to a waveform - but phase is really a mathematicians play thing, that is not a natural in explaining real physical effects.

Sampling Effects

At this stage it is worth looking at some of the baggage that sampling a signal carries with it. One of the processes that needs to happen is anti-alias[3] filtering. On the recording side, anti-aliasing filtering must be carried out to prevent signals present above 20 kHz being aliased back on to the recording within the audio band. On the playback side, anti-alias filtering is usually carried out to prevent large amounts of high frequency energy being presented to subsequent items in the playback chain - such as amplifiers and speakers. In general, anti-aliasing filtering needs to be quite vigorous - typically from some fraction of a dB down at 20 kHz to -100dB or so at half sampling frequency (22.05 kHz for CDs).

Sharp filtering inevitably causes a ringing transient response. Filter designers and mathematicians are familiar with the problem - the effect is referred to as the Gibbs phenomenon. Figure 2 shows responses for four filters with increasingly sharp cutoffs, and the associated transient responses. It is worth noticing that the ringing sets in at quite a gradual slope (filter B) - it is not the extreme severity of the sharper filters that causes ringing, but the move away from a very gentle filter function (A).

Figure 2 - Roll Off and Impulse Response for Filters of Increasing Vigour



Anti-aliasing filtering causes this type of ringing transient response. The effect is well known and unavoidable, and tends to be dismissed as a mathematical irritation with no audible effect - because the frequency response is flat. The dismissal gains credibility, because if attempts are made to roll frequency response off a little earlier, recording engineers complain. Certainly, it seems that frequency response flat to 20 kHz to a dB or so is important.

Energy Dispersion

The ringing contains energy, and we can plot energy against time. For anti-aliasing filters we get the sort of shape shown in figure 3. This shows that although the energy in the input transient is concentrated at one time, the energy from the anti-alias filter is spread over a much longer time - the audio picture is “defocused”. We might be tempted to argue that the energy is ultrasonic, but this is certainly not the case at 44.1 or 48 kS/s - our bandwidth constraints mean that to get good anti-aliasing, we must filter as fast as we can, and only pass the audio bandwidth. Ergo - any energy in the output signal is in the audio band. At sample

[3] see, for example, “Digital Signal Processing”, Alan V.Oppenheim & Ronald W.Shafer, pub. Prentice-Hall 1975

rates above the standard, the energy in the ring still has the full bandwidth of the passband - maths tells us so. We can also note that the energy in the ringing is large - for a sharp filter it can be 12% (-9 dB) of the energy in the main lobe.

Figure 3 - Energy in Transient Ringing

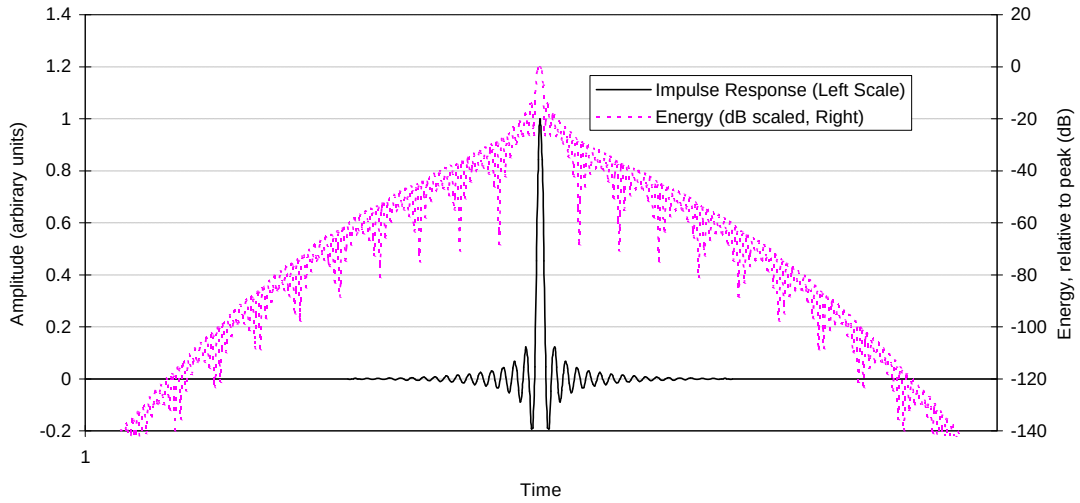
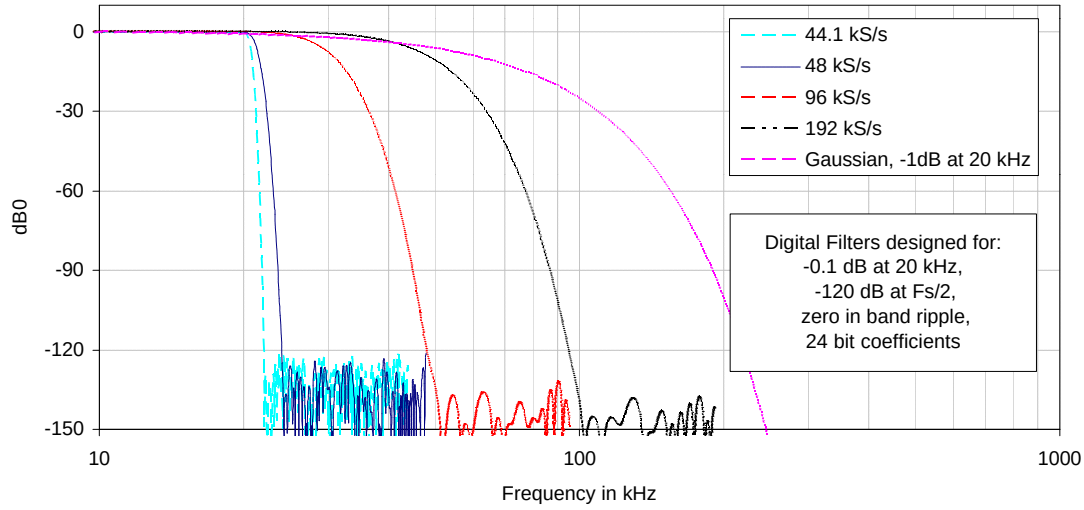


Figure 4 shows the frequency responses of some filters we might consider using, for audio purposes, at different sample rates. The digital filters do not represent any particular hardware, and are all designed to give the following performance:

- -0.1 dB at 20 kHz
- -120 dB at half sampling frequency
- no in-band ripple
- 24 bit coefficients

Figure 4 - Frequency Responses of Anti-Alias Filtering for Different Sample Rates

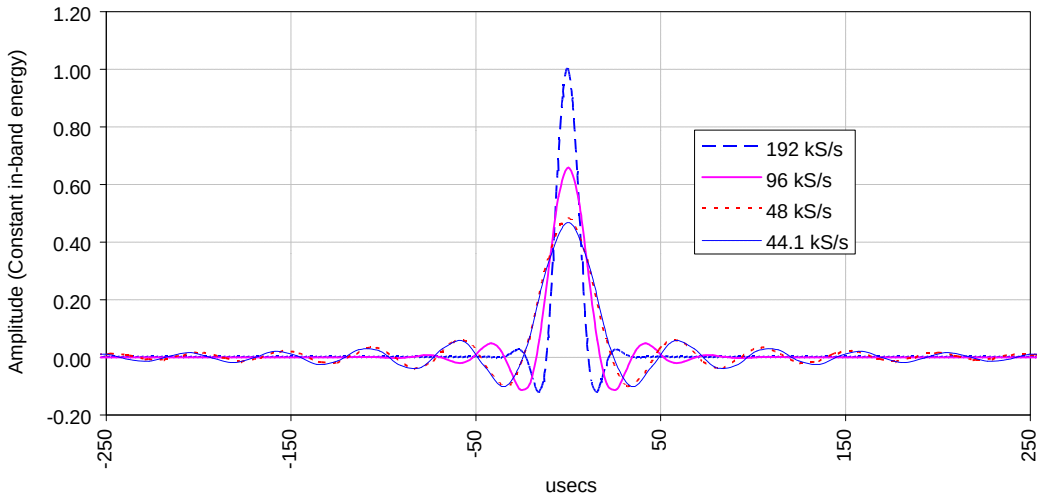


An analogue frequency response with a Gaussian filter set to -1dB at 20 kHz is shown for comparison. Gaussian filters have a non ringing transient response - if a filter rolls off faster than a Gaussian, it starts to ring. The ringing may be acceptable, of course. In practice, it is improbable one would use this filter, because -1dB at 20 kHz is unlikely to be acceptable for professional purposes. A filter that was flatter to 20 kHz, gentle to say 50 kHz and fast after that, with some ringing, might be more likely. The purpose of the comparison is to show just

how much extra bandwidth is needed to contain ringing (energy defocusing), and how much our ears may have taken for granted as the processing of the data they produce has evolved.

Figure 5 shows the transient responses of the different filters, operating at their different sample rates.

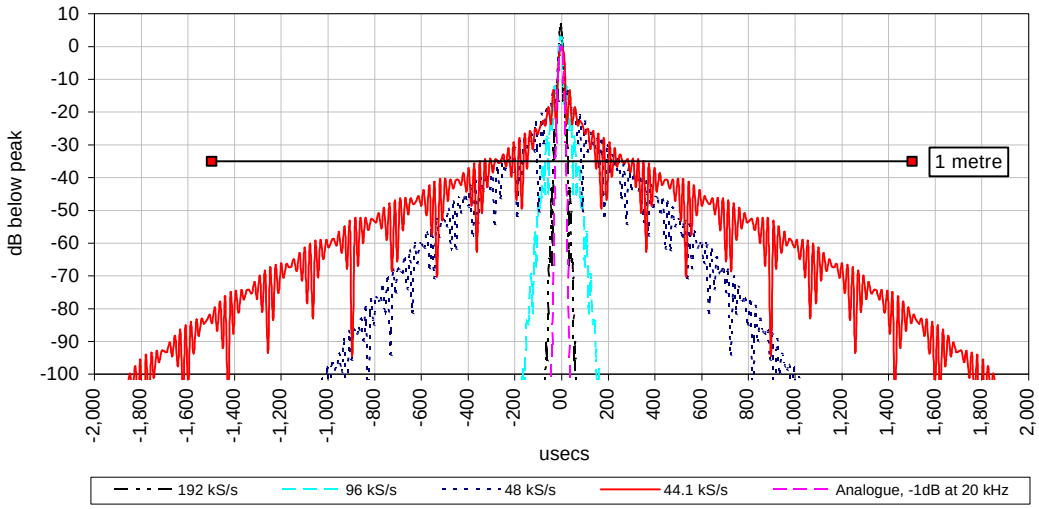
Figure 5 - Impulse Response vs Time for various sampling rates



Energy Dispersion at Different Sample Rates

Figure 6 shows the energy associated with the transient responses. 44.1 and 48 kS/s filters spread audible energy over 1 msec or more. The 96 kS/s filter is much better, keeping the vast bulk of the energy within 100 µsecs. The 192 kS/s filter can be very good indeed, keeping the energy within 50 µsecs. The analogue Gaussian filter is just a little better still, although the improvement is almost certainly academic because of energy dispersion from today's speakers and mics.

Figure 6 - Energy vs Time for various sampling rates



Taking into account the speed of sound, we can convert energy defocusing in the time domain to "smear" in distance estimation by the ears. Energy spread over ±500 μsecs is the same as a distance smear of ±15 cms. 96 kS/s keeps almost all the energy within about ±50 μsecs, or ±1.5 cms. One of the observations people make[4] about 96 kS/s material is that the spatial localisation of everything is very much better than 44.1 kS/s. 192 kS/s is better than this, although very dependent on amp and speaker performance to demonstrate it.

One can get oneself into a bit of a twist thinking about the energy in the ringing. After all, if it is in the audio band, allowing extra energy at higher frequencies through the system surely cannot cancel out some that is in the audio band? It does, though - so although we may not be able to hear energy above 20 kHz, its presence is mathematically necessary to localise the energy in signals below 20 kHz, and it is possible (and our contention) that we can hear its absence in signals with substantial high frequency content. A high sample rate system allows it through (fact) - and allows the high frequency signals to sound more natural (contention) but allowing better spatial energy localisation (fact).

It is our suggestion that some of the audible differences between conventional 44.1 kS/s and higher rates (88.2, 96, 176.4, 192 kS/s) may be related to this "energy smear" or defocusing caused by anti-alias filtering, and that the ear is sensitive to energy as well as spectrum. This is further backed up by our two original "same spectrum, sound diufferent" signals (figure 1). In the impulse, all the energy is concentrated at one time, whereas for the white noise the energy is uniformly spread over time. There is a precedent to this suggestion that the ear is sensitive to both spectrum and energy - the eye is as well. For sensitive vision or vision off the main beam, we use energy (luminance, or black and white information), whereas for detailed identification when we are looking at something, we use spectrum information (chrominance, or colour). In fact, most sensing processes are sensitive to energy. If the ear is sensitive to energy, it would almost certainly use the information for spatial localisation.

Multi Channel

In conclusion, it is worth noting that if this suggestion is correct, then it would be sensible for any multichannel audio formats to use one of the higher sampling rates. The purpose of multichannel is for better spatial localisation of sound sources - so it needs a sampling rate that can support this!

[4] even people with grey hair, who must have significant high frequency hearing loss