High-Resolution Audio

A perspective

Bob Stuart

Bob Stuart is a Fellow of the AES and lifelong student of the auditory sciences. Bob has advocated high-quality audio coding, contributing to ADA, JAS, as well as the DVD and Blu-ray standards.

What is High Resolution? The term is borrowed from optics. Resolution, or resolving power, is the ability of a device to produce separate images of closely spaced objects; a high-resolution image has clarity, depth, and no blur. In an image we can measure spatial or angular resolution, the impact of transmission, e.g., via a lens, or coding artifacts such as filtering or noise. The quest for better microscopes and astronomical telescopes has spurred research into "super-resolution" techniques to ameliorate fundamental resolution limits due to the wavelength of light, while in the digital domain, cameras and displays use increased pixel count, contrast, and color-depth to approach human visual acuity. We perceive resolution as definition.

In audio, high-resolution sound should be natural, resembling real life and many of the terms we use to qualify it, such as clarity, focus, transparency, and definition are borrowed from vision. If sound is natural, objects should have clear locations (position and distance) and separate readily into perceptual streams, particularly where environmental reverberation causes multiple arrivals closely separated in time—temporal resolution of microstructure in sound being analogous to spatial resolution in vision.

Resolution vanishes in a picture if the image is foggy or blurred no matter how many pixels are used. Similarly high-resolution sound only results from adequate operation of a complete chain—from performer to microphone to loudspeaker to listener. But what is adequate? High resolution in audio requires an absence of added noise, temporal blurring, and frequency—or content-correlated errors.

When we listen, it isn't the acoustic waveform or spectrum that we interpret but the spikes from around 30,000 afferent inner-hair-cell cochlear neurons.

As these signals travel through the brain stem to the auditory cortex, tonotopically organized neurons, initially coding for level, spectrum, modulations, onset, and offset, pass through nonlinear combining structures that exchange, encode, or extract a variety of temporal, spectral, envelope, and ethological features [2].

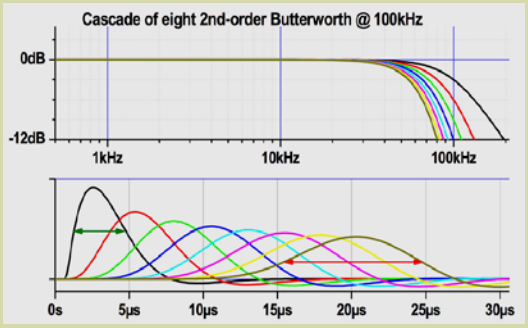

Fig. 1. The frequency and impulse responses of a cascade of low-pass filters.

By exploiting population coding, temporal resolution can approach 8 µs, and this precision reflects neural processing rather than being strictly proportional to our 18-kHz tonal bandwidth (an estimate of the upper "bin" of the cochlea and upper limit of pitch perception) [1].

Modern insights from psychophysics therefore imply we should preserve temporal structure at a finer scale for our ears to take advantage of the original sound; a distribution system that permits end-to-end resolution of 8 µs implies a Gaussian bandwidth of around 44 kHz.

A chain is as strong as its weakest link, and one limit on resolving events will be the minimum-phase transducers at each end and amplifiers in between, but as Fig. 1 shows, to achieve this target through a uniform cascade of eight stages, each needs a bandwidth closer to 100 kHz.

Notice that we just defined High Resolution in analog and not digital terms.

Sampling is the bridge between analog and digital, and for some time our understanding has been influenced by the Shannon theorem that shows that unambiguous reconstruction to analog can be possible if a signal of finite Fourier bandwidth is uniformly sampled at a (Nyquist) rate equal to at least twice its maximum frequency [3].

Sound is not inherently band limited, (indeed earlier analog recording systems showed relatively gentle high-frequency roll-off), but the theory implies (and current devices typically employ) sharp band-limiting and reconstruction kernels approximating sinc interpolation. Furthermore, the sounds that are important to us have properties that we can exploit and that this framework does not address, including self-similarity (manifesting as a finite rate of innovation) and temporal asymmetry (in natural sound cause precedes effect) [1][3][4].

A Fourier analyzer or a sinc sampling kernel use windows extending forward and backward in time, so that a frequency-domain description can be very unhelp-

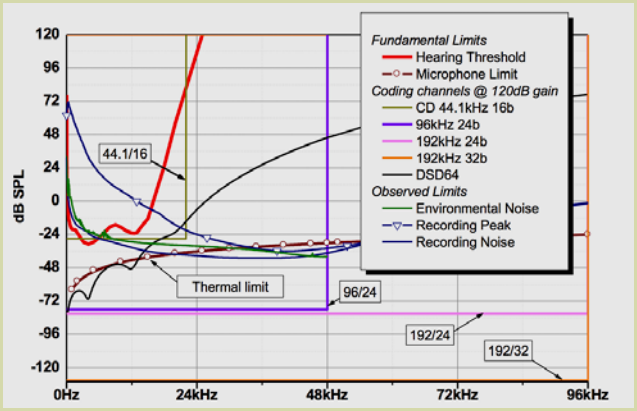

Fig. 2. Showing: fundamental limits of microphone noise [7], human hearing threshold (for uniformly-exciting noise) and an example of environmental noise. Also shown, for an acoustic gain of 110 dB SPL are the coding spaces for some channels and the background and peak levels for a low-noise piano recording (2L-082).

ful when the future of the signal is not known. By contrast, a neuron has to resolve whether or not to fire on the basis of what it sees "now" and pre-responses can give unnatural results.

Recent findings in neuroscience guide us to speculate that audio can be more efficiently transmitted if the channel coding is optimized for natural environmental sounds rather than specified with independent "rectangular" limits for frequency and amplitude ranges [5]. Direct evidence for the audibility of low-pass filters typically used in digital audio has been published [6].

Since everything we hear is subject to thermal noise, being lossless in the digital path is not strictly relevant. Modern A/D and D/A converters comprise modulators and chains of decimation or interpolation filters that add noise either side of the digital path.

At sample rates higher than 48 kHz we can use reconstruction filters with reduced pre- and post-rings. In fact sampling and reconstruction kernel-pairs can be selected that resemble neural kernels tuned to ensembles of natural sounds, and which provide less uncertainty of an event's duration [1][3][5].

In wideband music recordings, content-related signals normally fall below the inherent background noise above 48 kHz, suggesting that a sample-rate of 96 kHz could preserve most spectral content (see Fig. 2). Further increasing sampling rates while retaining sinc kernels yields diminished returns: the benefits coming from fewer decimation/interpolation steps in converters and shorter pre- and post-rings. Earlier we argued a preference for a natural, minimum-phase system response.

Increasing bit-depths from 16 to 24 or even 32 also gives diminishing returns. While more bits provide finer amplitude gradations, so long as every quantization is properly dithered, then bit depth just determines dynamic range, not amplitude resolution.

It's easy to overlook the extraordinary dynamic range that a 24-bit channel can encode. Fig. 2 shows coding spaces for 16, 24, and 32 bits. Also shown is the fundamental noise limit for a microphone [7]—we can't pick up quieter sounds, yet this spectrum can be described in an optimally shaped 17.5-bit 192-kHz LPCM channel. At lower sampling rates, the background noise in recordings does not normally justify using more than 18 bits.

Up to now, High Resolution in audio hasn't been usefully defined, which is a pity because without a secure bridge between auditory science and audio engineering, development can be haphazard.

The best so far [8], calls for components with 40-kHz analog bandwidth and sampling rates of 96 kHz or higher. Although a step in the right direction, this is barely adequate for a chain let alone one link.

If a reproducing system is to be flawless for the human listener, then its errors should be both natural and plausible. Although we tend to define errors in terms of our measuring instruments, we need to move away from the poor proxy of escalating sample-rates and bit-depths. Perhaps an even more fundamental definition can help?

If we stand right next to a well-played piano, the sound can be thrilling, vibrant, complex, and detailed. If we put our head under the lid (where microphones are sometimes placed) the sound has added brilliance that falls away if we move some meters away. There are a number of factors, but one is the way air itself changes sound.

We recently proposed that system errors should only resemble those introduced when sound travels a short distance through air, see Fig. 3. Within reasonable limits, air does not introduce distortion, but it does add thermal noise and temporal blur—progressively attenuating higher frequencies and slowing down transient edges. Such a system, placed between the listener and the performer, might not be noticed.

Systems designed from this viewpoint could then be quantified by the accumulation of "added distance."

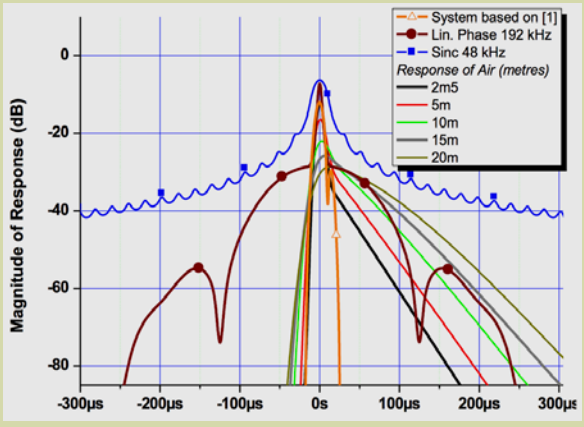

Fig. 3. End-to-end impulse magnitude response of a system based on [1] compared to typical 192- and 48-kHz systems and air at STP and 30% RH. Compared to typical 192-kHz sampling, temporal blur of the example is lowered by an order of magnitude.

REFERENCES

[1] Stuart, J. R. and Craven, P.G., "A Hierarchical Approach to Archiving and Distribution," 137th AES Convention, (2014), convention paper 9178.
[2] Oppenheim, J. M., et al., "Minimal Bounds on Nonlinearity in Auditory Processing," q-bio.NC (Jan 2013). arXiv:1301.0513
[3] Unser, M., "Sampling – 50 Years after Shannon," Proc. IEEE 88 No. 4, pp. 569–587 (Apr. 2000). http://dx.doi.org/10.1109/5.843002
[4] Vetterli, M., et al., "Sampling Signals wth Finite Rate of Innovation," IEEE Trans. Sig. Proc. 50, No. 6, pp. 1417–1428, (May 2002). http://dx.doi.org/10.1109/TSP.2002.1003065
[5] Lewicki, M.S., "Efficient Coding of Natural Sounds," Nature Neurosci. 5, 356–363 (2002). http://dx.doi.org/10.1038/nn831
[6] Jackson, H. M., Capp, M. D., and Stuart, J. R., "The Audibility of Typical Digital Audio Flters in a High-Fdelity Playback System," 137th AES Convention, (2014), convention paper 9174.
[7] Fellgett, P.B., 'Thermal Noise Limits of Microphones," J. IERE, 57 No. 4, pp. 161–166 (1987). http://dx.doi.org/10.1049/jiere.1987.0058
[8] Japan Audio Society, "Action Plan for High-Resolution Audio," (2014)