Audio Engineering Society

Convention Paper 9178

Presented at the 137th Convention
2014 October 9–12 Los Angeles, USA

A Hierarchical Approach to Archiving and Distribution

J. Robert Stuart [1], Peter G. Craven [2]

[1] *Meridian Audio Ltd, Huntingdon, PE29 6YE, UK*
jrs@meridian-audio.com

[2] *Algol Applications Ltd, London, SW19 3AR, UK*
peter@algol.co.uk

ABSTRACT

When recording, the ideal is to capture a performance so that the highest possible sound quality can be recovered from the archive. While an archive has no hard limit on the quantity of data assignable to that information, in distribution the data deliverable depends on application-specific factors such as storage, bandwidth or legacy compatibility. Recent interest in high-resolution digital audio has been accompanied by a trend to higher and higher sampling rates and bit depths, yet the sound quality improvements show diminishing returns and so fail to reconcile human auditory capability with the information capacity of the channel. By bringing together advances in sampling theory with recent findings in human auditory science, our approach aims to deliver extremely high sound quality through a hierarchical distribution chain where sample rate and bit depth can vary at each link but where the overall system is managed from end-to-end, including the converters. Our aim is an improved time/frequency balance in a high-performance chain whose errors, from the perspective of the human listener, are equivalent to no more than those introduced by sound travelling a short distance through air.

1. CONTEXT

This paper continues some of our earlier work on high-resolution audio and addresses a perceived misalignment between archival and distribution formats. Taking into account recent progress in auditory sciences, coding theory and listening tests, we introduce concepts enabling efficient high-quality sound distribution. Due to the wide scope many topics are introduced by reference.

In earlier times it was not always feasible to record with higher quality than the release format. These days the cost of preserving digital data continues to fall and so the form and sound quality of the archive deserve serious consideration. We need a variety of application-specific distribution formats, but if we are doing the best for the future, we should always store the best representation and it will rarely make sense to distribute these core assets in a format identical to that of the archive.

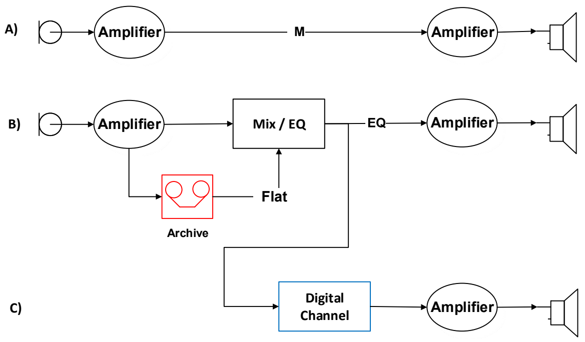

Figure 1. A) The 'perfect' replay chain; B) with added processes and storage and, in C) a digital channel.

1.1. System Model

The simplest possible sound reproducing system transfers a microphone feed to a remote loudspeaker, using the minimum of electronics: shown diagrammatically in Figure 1A.

Even this system does not convey the original sound perfectly. Although typically better than the loudspeaker, the microphone only samples the sound field and has intrinsic limitations in spectral and temporal response, directivity and dynamic range [27].

One aim is to distribute the sound of this microphone feed and yet the best it provides is a lossy representation, permanently imprinted by the microphone system and which we may only access with a loudspeaker. Although obvious, this point can mark a divergence between the archivist (who wants to preserve the original sound) and the producer (who sees the electrical signal, or a stored or modified version heard through a studio speaker as 'the master').

Normally a recording is stored and mastered, and these potentially benign steps are also shown in Figure 1B, where we see a 'flat master' and an 'EQ' or release mix.

1.2. Digital Coding

Recording preserves an analogy of the music waveform. Early recordings were mechanical and analogue magnetic tape followed. More recently the signal has been brought into a digital representation.

Analogue storage or transmission introduces distortion and noise that cannot be removed and also may disturb the time structure through wow, flutter, print-through etc. Analogue recordings tend to degrade with the passage of time, with each successive playback and cannot give repeatable results, whereas digital data can be maintained independent of storage formats and replay is at least perfectly repeatable.

Inside the digital domain, although careless coding or incautious signal processing may introduce distinctive problems, the prospect exists for transparent coding and processing; a topic tackled in some detail in [1].

Once the audio information is contained within digital data it can be transmitted through time or space losslessly and playback can be repeatable. However, the most critical steps remain at the analogue-digital (A/D) and digital-analogue (D/A) gateways and in the compromises and permanent limitations made at these points.

1.3. The Gateways

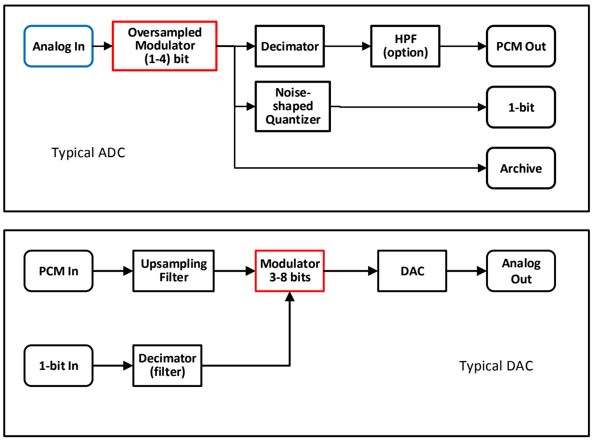

Figure 2. Internal blocks in typical A/D and D/A converters.

Early converters tended to be multibit and operate at the base sample rate. Figure 2 illustrates typical internal architectures of now widely used delta-sigma converters. Oversampling delta-sigma structures permit simplified analogue filtering and have the potential for the highest performance when using dither in a small-word-size hardware quantizer/modulator.[1] These concepts are explained in [15] and [14].

Even though it has significant problems as a release or distribution code, the output of the A/D modulator could be a more appropriate 'archive' than either the decimated

[1] In its most extreme form the modulator is 1 bit and the converter can sample or reconstruct at 64 or more times f_S. Although popular twenty years ago, this single-bit variant has significant problems because the modulator cannot be satisfactorily dithered and the shaped noise intrudes rapidly on the octave above 20 kHz [76].

multibit PCM output or the noise-shaped, quantized single-bit stream.

In fact an ideal system might connect the A/D modulator output directly to its counterpart in the D/A converter, using narrow PCM at a high-sample-rate; e.g. 384 kHz and 8 bits might be an ideal candidate [9], particularly if the sampling kernel were similar to a Gaussian.

We can see that the high-speed modulators are the critical stages, while for the PCM passed from one to the other, the sample rate may be chosen arbitrarily. The properties of the decimation and upsampling filters can significantly impact sound quality. A considerable part of research into high-resolution audio has centered on these filters and on varieties of dither [1][10][11][12][13][14][15] [16][72][54][55][59][62][65][66][67][76][77][78][79] [80][82].

1.4. High Resolution

The term 'high resolution' has a visual analogy; a high-resolution image has clarity, depth, absence of filtering or coding artefacts, little blur and is rapidly assimilated. In an image we can measure resolution of details, and the impact of coding or transmission, e.g. via a lens.

In audio, high resolution should also resemble real life: sounding natural, objects having clear locations (position and distance) and separate readily into streams (through absence of noise, distortion or modulation effects).

In the last decade it has become more common for recording professionals to self-select higher sample- or data-rate formats to improve sound quality. It's not uncommon to find recordings being laid down at 192 or even 384 kHz in 24- or 32-bit precision; data rate has become an unfortunate proxy for resolution.

Higher-than-CD data rate doesn't guarantee improved sound quality, but doubling or quadrupling sample rate from 44.1 or 48 kHz can show incremental improvements in exchange for a rapidly increasing file size [77] [32].

It is now widely accepted that one key benefit of higher sample rates isn't conveying spectral information beyond human hearing, but the opportunity to tackle the dispersive properties of brick-wall filtering. Wider-transition anti-alias and reconstruction filters directly shorten (proportionately) the impulse response and there is also more opportunity to apodize to remove extended pre- and post-rings [16].

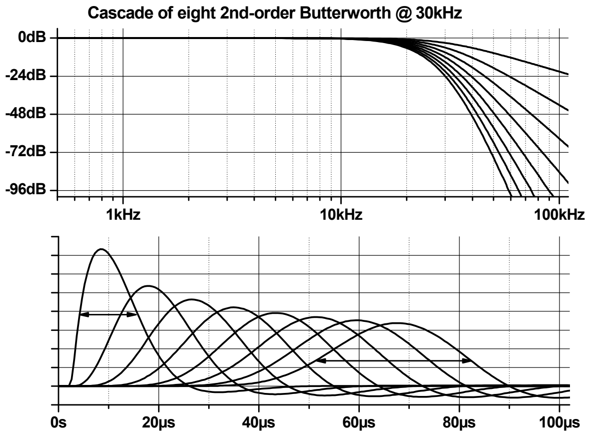

Figure 3. Showing the frequency and impulse responses of a cascade of eight 2nd-order Butterworth low-pass filters.

Providing the sampling kernel[2] is not too extended and that any subsequent quantization is properly dithered, then transient events can be accurately located in time [15]. However, higher sample rates do allow shorter details to be captured, improve dither convergence, and enable encoding kernels that provide much less uncertainty of an event's duration [14].

When considering the frequency and time responses of an end-to-end distribution channel, we must bear in mind that time dispersion or 'blur' can build up through a cascade of otherwise blameless components. Figure 3 illustrates the response of a cascade built up to eight stages, each with a 2nd-order roll-off at 30 kHz, possibly representing a microphone, preamplifier, mixer, converter pre- and post-filters, replay pre- and power amplifier and transducer. We imagine that such a chain might disguise aspects that a wider-band replay system would reveal and could confuse listening tests [28] [32].

A more severe viewpoint is to define a high-performance chain as one whose errors, from the perspective of the human listener, are equivalent to those introduced by sound travelling a short distance through air. Within reasonable limits, air does not introduce distortion, noise or modulation noise, but it does blur sound, progressively attenuating higher frequencies as shown in Figure 4 and slowing down transient edges. [92]

A system having similar properties, if placed between the listener and the performer, might not be noticed.

[2] As will be explained, sampling is modelled by convolution with a *kernel* such as a sinc function, followed by instantaneous sampling.

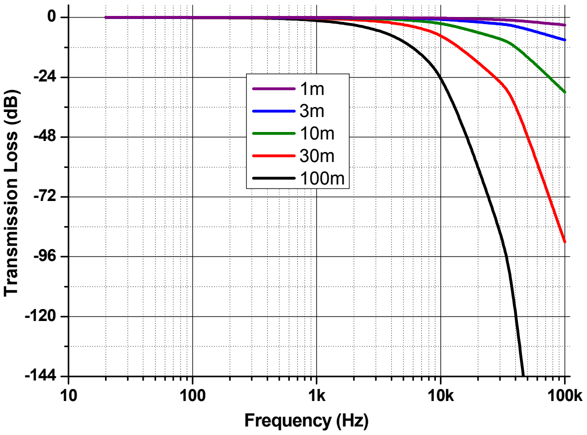

Figure 4. Attenuation of sound in air at STP and 30% relative humidity. Data from [92].

2. THE LISTENER

The quality of an audio channel can only be finally judged in its intended use: 'conveying meaningful content to human listeners'. The auditory sciences (psychoacoustics and neuroscience) help us to bridge listeners' impressions and engineering.

2.1. Psychoacoustics and Modelling

With care to context, psychoacoustics can help us estimate the audible consequence of imperfect 'conveying', allowing errors arising in the recording chain to be ranked. Such errors might be straightforward transmission failures, or take the form of noise, distortion, jitter, wow, flutter, etc. Essentially any change introduced can be isolated in measurement and modelled to estimate its impact in context. A special case is to estimate when channel errors might be inaudible.

Fundamental characteristics of the hearing system are complexity and non-linearity. To the listener, sounds have pitch and loudness rather than frequency and intensity, and the relationships between these measures are non-linear. Some non-linearities are extreme, such as: thresholds; detectability or loudness of a stimulus incorporating adjacent frequency elements; and masking by components slightly further away in time or frequency.

Perception refers most often to the 'low-level' behavior of the human auditory system where we are concerned with straightforwardly testable parameters like whether or not a simple stimulus is audible, or detectable in the presence of another (masker) sound, or distinguishable from a similar stimulus etc.

Psychoacousticians have designed auditory experiments which explore the limits of the human hearing system as a receiver – and which, in general, attempt to minimize the impact of cognition.

However, it is important also to consider the higher-level process of cognition – where sounds take on meaning. In cognition, higher-level processes modify the listener's ability to discriminate more, less or differently than indicated by the perceptual model.

In the cognitive process we hear 'objects' rather than 'stimuli' and we distinguish 'what' from 'where'. Mechanisms such as streaming exploit similarity, contrast and other cues to modify the basic percepts; so there is a risk that system errors which correlate to the signal, for example modulation noise, can attach to and modify 'perceived objects' [44].

2.2. Neuroscience and Modelling

Recently there has been considerable progress towards understanding how we hear, in particular in the related disciplines of neuroscience (helped in part by non-invasive imaging) and computational neuroscience; useful introductory texts are [41] [42] [43] and [7].

Neuroscience provides a second framework for enquiry and modelling and the approach tends to be different from traditional psychoacoustics. Rather than devise archetypical experiments to select between two alternatives [45], it is sometimes more useful to consider how neurons respond to the complexity of the natural world in which stimuli are not known in advance, but might instead be chosen from a large but representative set.

Regarding natural auditory stimuli, three important classes are the background sounds of the environment, animal vocalizations and speech. In ensembles, all three exhibit self-similarity and a general spectral tendency for amplitude to fall with frequency; environmental noise shows a very precise 1/f trend, see Figure 5.

Hearing is important for survival and we can't wait too long to make a decision. Steady-state signals are not normal; an averaging detector might take too long. So a better model is of a 'running commentary'; trying to make sense of the sounds as they arrive. To parse this running commentary we can't often 'rewind' into the short-term auditory memory and so strategies which robustly extract acoustic features in the presence of noise or interference have evolved.

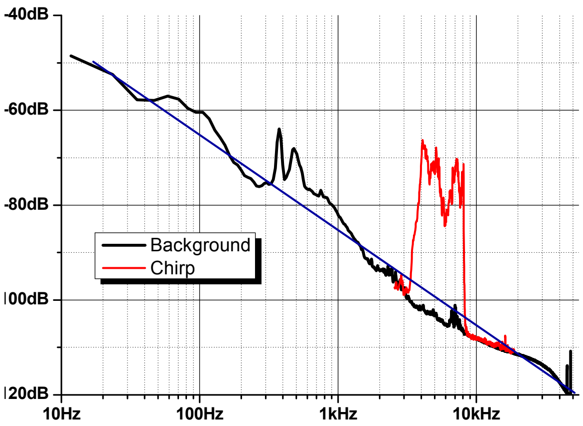

Figure 5. Environmental sound and bird call. Also included is a line at –20 dB/decade showing how precisely the spectrum follows the expected 1/f trend.

Our ability to externalize objects or to follow speech or a melody is amazingly robust and we can still understand an extensively modified or damaged stream of sound. However in current context, we want to ensure that we never stray into that prohibited area where meaning survives, but subtlety and ultimate realism do not.

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 – whose actions, in turn, are ultimately modified by a similar number of efferent (descending) neurons, many of which connect to the cochlear outer hair cells.

As the signals travel through the brain stem, the mid-brain and on to the auditory cortex (wherein finally, we 'hear'), tonotopically organized neurons, initially coding for level, spectrum, modulations, onset and offset, pass through combining structures which exchange, encode or extract a variety of temporal, spectral and ethological features [43][83].

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 bandwidth (an estimate of the upper 'bin' of the cochlea and upper limit of pitch perception) [56][58] [4][5][6][7][70][71].

The role of the descending neurons is not yet completely understood. At a simplified level they are implicated in gain control, in modifying feature extraction through attention, and perhaps most intriguingly in our context, conscious and unconscious control of the outer-hair-cell active process which is responsible for mechanical gain and 'filter width' implied in basilar-membrane motion.

This idea that this auditory-filter width can be responsive to attention and context has profound implications for detection and masking models [43].

In an important set of papers, Lewicki showed computational neural models proposing efficient auditory coding using kernels tuned to ensembles of natural sounds [2] [3]. His models evolved highly efficient, 'auditory filters' adapted to the three classes of natural sounds mentioned earlier and showed how each sound-class led to a different time-frequency balance and therefore filter bandwidth.

Filters adapted to animal vocalizations selected for fine frequency resolution. Speech drives fine frequency resolution in the region of 500Hz but selects for temporal resolution above 1.5 kHz, whereas environmental sounds preferred fine temporal discrimination – particularly at high frequencies. Although a model, these findings augment our understanding and reinforce that 'listening for' or 'attending to' 'objects' or 'streams' might indeed involve direct control of the cochlea [7][44].

It is also intriguing that environment-derived kernels bear resemblance to some we derive in Section 4.

Rieke *et al* [42] describe neurons that respond to higher moments of the stimulus; e.g. high-frequency auditory neurons which are not sensitive to phase, but instead encode the envelope of the sound-pressure waveform – a mechanism that we have long suspected in the context of perceiving the pre-ring of a sinc-kernel brick-wall filter at low and high sample rates [86][90].

These findings in neuroscience guide us to speculate that audio can be more efficiently transmitted if the channel coding is optimized for natural sounds rather than specified with independent 'rectangular' limits for frequency and amplitude ranges.

2.3. Temporal Limits

In certain circumstances the human hearing system is incredibly sensitive to temporal features. Since higher sample rates permit finer-grained details to be resolved it is important to understand where the limits for transparency may lie.

For the audio distribution channel we can consider temporal resolution in two aspects: its ability to maintain separation between closely spaced events (and not blur them together) and its ability to maintain a precise unquantized time-base within and between channels. Low-pass filtering may ultimately impact the separation of nearby events (hinted at in Figure 3) while filters in the digitizing process, that are sharper in the frequency domain, may also bring uncertainty to the start, stop and center of transient events.

Our ability to localize sounds swiftly and accurately is vital for survival. Sound intensity and arrival time provide important binaural cues and humans can discriminate inter-aural time differences as low as 10µs for frequencies below 1.5 kHz [33][35][37][38] (we are most sensitive in the region 0.8–1 kHz [90][85][88][89]) and as low as 6 µs for sounds with ongoing disparities, such as in reverberation [36] [84][86][34][64][63].

Other mechanisms have been investigated that hint to similar discrimination limits within a channel, i.e. monaurally, including: temporal fine structure in pitch perception, the comprehension of speech against a fluctuating background [91] and other cues [34].

It has been suggested by Kunchur that listeners can discriminate timing differences of the order of 7 µs [56] [57][58]. Woszczyk has also provided a convenient review of psychophysical and acoustic temporal factors [60].

In light of these psychophysical data, even though one limit on resolving events will always be the microphone system bandwidth, it would seem prudent to provide resolution for an archive that can resolve 3 µs. On the other hand, based on current recordings we have analyzed, and bearing in mind the response of microphones currently favored by recording engineers, a sensible target for today's distribution system would be of the order of 10 µs.

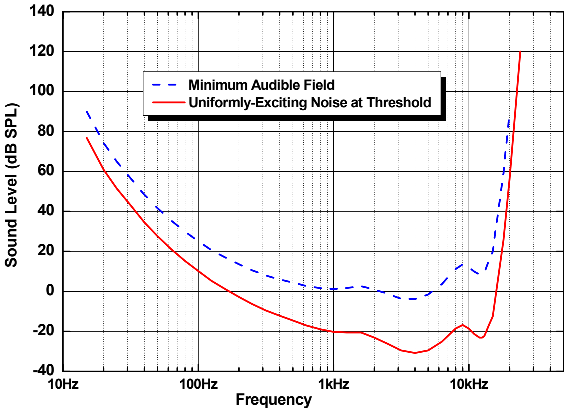

Figure 6. The upper curve is the minimum audible field threshold for pure tones. For evaluating noise spectra, the lower curve is uniformly exciting noise at threshold, from [28].

2.4. Spectral and Amplitude Limits

The standard hearing threshold for pure tones is shown in Figure 6 [46] [47] [48]. This minimum audible field has a standard deviation of around 10 dB and individuals are to be found whose thresholds are as low as –20 dB SPL at 4 kHz. Although the high-frequency response cut-off rate is always rapid, some can detect 24 kHz at high intensity. [68] [69] [81]

There are some fundamental physical limitations in analogue electronics (such as thermal and shot noise) and in the air itself. The human hearing system is extremely sensitive, in common with those of many mammals. It is thought that one limit of sensitivity derives from Brownian motion of molecules within cochlear fluid around the hair-cell receptors [49], and such is the efficiency of the outer ear that the mid-range limit for hearing is also close to that which would reveal the noise of Brownian motion in the air itself [7].

3. THE SIGNAL

3.1. Spectral Content of Music

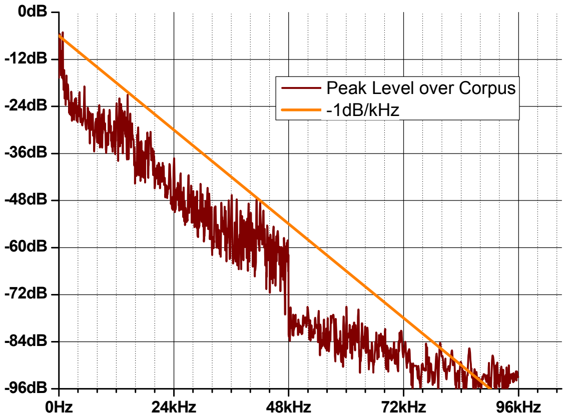

Figure 7. Peak spectral level gathered over a corpus of 96- and 192-kHz recordings.

There is significant content above 20 kHz in many types of music, as an analysis of high-rate recordings summarized in Figure 7 has revealed. One notable and common characteristic of musical instrument spectra is that the power declines, often significantly, with rising frequency.

Even though some musical instruments produce sounds above 20 kHz [53] it does not necessarily follow that a transparent system needs to reproduce them; what matters is whether or not the means used to reduce the bandwidth can be detected by the human listener.

3.2. Noise in Recordings

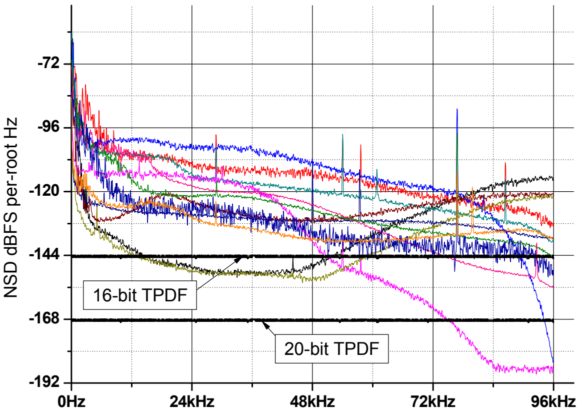

Figure 8. Examples of background noise in 192 kHz 24-bit commercial releases. Also shown is TPDF dither noise for 192-kHz 16- and 20-bit quantization. Curves plotted as noise-spectral-density in 1-Hz bandwidth.

Above we see measurements of noise in recordings, chosen to range from reissues from 60-year-old unprocessed analogue tape to modern digital recordings. Obviously these analyses embody the microphone and room noise of the original venue, but in some, analogue tape-recorder noise. Even the best recorder's noise floor is above that of an ideal 16-bit channel.

It is worth noticing that a 20-bit PCM channel is more than adequate to contain these recordings and that consequently 32-bit precision offers no clear benefit.

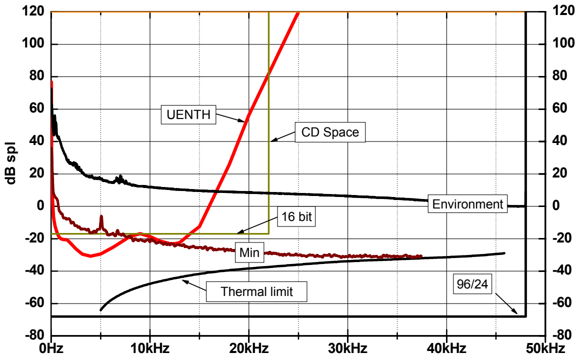

Figure 9. Showing the noise-spectral density of the lowest-background recording analyzed (Min) set to a replay gain of 0dBFS = 120dB SPL and in context of the uniformly exciting threshold noise from Figure 6. Also shown are the thermal-noise microphone limit, the environmental background noise from Figure 5 and coding spaces for CD and 96-kHz 24-bit PCM.

3.3. Environment and Microphones

Fellgett derived the fundamental limit for microphones, based on detection of thermal noise, shown for an omnidirectional microphone at 300°K in Figure 9 [52].

Cohen and Fielder included useful surveys of the self-noise for several microphones [51]. Inherent noise is less important if the microphone is close to the instrument and mixing techniques are used, but for recordings made from a normal listening position then the microphone is a limiting factor on dynamic range – more so if several microphones are mixed. Their data showed one microphone with a noise-floor 5 dB below the human hearing threshold, but other commonly used microphones show mid-band noise 10 dB higher in level than just-detectable noise. This further suggests that those recordings can be entirely distributed in channels using 18–20 bits.

3.4. Properties of Music

Content of interest to human listeners has temporal and frequency structure and never fills a coding space specified with independent 'rectangular' limits for frequency and amplitude ranges. As we noted in Section 2.2, environmental sounds show a 1/f spectral tendency. Ensembles of animal vocalizations and speech have self-similarity which leads to spectra that decline steadily with frequency. Music is similar but the levels decline at a progressively increasing rate, as seen in Figure 7.

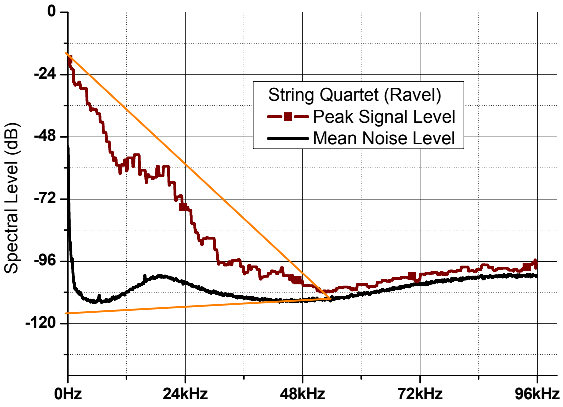

Figure 10. Showing the peak spectral level and background noise in a 192 kHz 24-bit recording of the Guarneri Quartet playing Ravel's String Quartet in F, 2nd movement.

There are several very significant points to be seen in Figure 10. Firstly the declining trend of peak level with frequency is classic, as is the background noise spectrum. We see that at around 52 kHz the curves converge and above that region we must assume that noise will obscure any higher-frequency details of the content.

This picture of the content occupying a 'triangular' space is common in all recordings we have analyzed and the converging point is usually below 48 kHz, with the highest so far being at 60 kHz.

From this spectral viewpoint, we could deduce that the information content relating to the original signal in the channel[3] occupies a space (within the 192-kHz 24-bit outer envelope) equivalent to that of a stream having a peak data-rate of 960 kb/s. The question is whether we can restrict the capture to just that signal-related information without disturbing the sound.

This insight has profound implications for the design of an efficient yet essentially lossless coding scheme as we will show in the next section.

4. ENCAPSULATION

When converting analogue audio to a digital representation, the waveform is sampled in time and amplitude. Amplitude quantization and dither have been well described in the literature [10–15], while system performance consequences are previously covered in [1], so here we concentrate on sampling and subsequent reconstruction to continuous time (analogue).

Sampling captures timing information present in the original continuous time signal, while reconstruction presents that information in a form that is accessible to the ear[4].

4.1. Sampling

In the several decades since both Shannon [17] and Nyquist [18] there has been considerable development in understanding of sampling theory. Shannon's sampling theorem shows how appropriate band-limiting allows repeated resampling of a signal without build-up of alias products. At a gross level, a communications system can then be characterized by a single number, its bandwidth, which is the narrowest bandwidth of any of the filters or subsystems that have been cascaded.

Overwhelming convenience has thus led to the notion that 'brickwall' bandlimiting is the ideal, the common specifications of passband, stopband and transition band measuring the deviation of anti-alias filters from that ideal. However, the tradition of brickwall filtering has not been enshrined in law and there are possibilities to balance time vs. frequency uncertainty to efficiently code sounds of consequence to the human listener. [5] [19–26]

The impulse response of an 'ideal' Shannon-sampled system is a 'sinc' function which has a fairly sharp central pulse but also a pre-ring and a post-ring, which build up and die away slowly.

Some may wonder how a time-domain analysis can tell us anything different from a more conventional frequency-domain analysis, since it is known that the frequency-domain and time-domain descriptions of a linear system are completely equivalent. If a human cannot hear above say 18 kHz, how can a pre-ring at a frequency of 20 kHz or 22 kHz be of any consequence?

One answer is to consider that a Fourier analyzer uses a window that extends both forwards and backwards in time. Thus although the two descriptions are equivalent if one considers the global signal, the frequency-domain description is very unhelpful in thinking about the situation at a particular point in time when the future of the signal is not known. A neuron has to make a decision on whether or not to fire on the basis of what it sees *now*.[6]

Can a sampled system convey time differences that are shorter than the periods between successive samples? An intuitive answer might be 'no' [60], but we note that even when convolved with a sinc function, an arbitrarily small displacement of an impulse can be detected on the basis

[3] Excluding information that merely allows one to accurately reconstruct noise and other processing artefacts.

[4] More precisely, reconstruction to continuous time is the first step in rendering to an acoustic signal, for we are not proposing to present samples to the brain via a neural implant! Were we to do so, it would be arguable that the sampling kernel should mimic the cochlear kernel, which has a finite width [3]. But for acoustic rendering, the requirement is that the total effect of sampling, reconstruction and rendering plus the cochlear kernel should not be significantly different from that of the cochlear kernel alone. This unfortunately places a tighter time constraint on the sampling and reconstruction process.

[5] Brickwall filtering is *idempotent*: once done, it can be cascaded arbitrarily without further loss. Here however we are considering the *total* end-to-end processing, which is *not cascaded* so different considerations apply.

[6] If we are expecting the nerve cell to ignore pre-responses, how does it know when it sees the 'real' peak that it is not merely a pre-response for something even bigger that is yet to come?

of waveform comparison, assuming one has sufficient signal-to-noise ratio.

Instantaneous sampling without any filtering is not recommended, for the sampling would then be vulnerable to high-frequency noise (even to the megahertz region).

Further, a Dirac impulse would not be registered at all if it happened to occur between the sampling instants. Intuitively one would at least integrate over one sample period, as illustrated in Figure 11 (upper). Here a transient falling entirely within the sample period corresponding to *Sample 0* will be integrated and the value of *Sample 0* will represent the area of the transient.

If the transient moves to the right, there will be no change in the sample values until the transient crosses into the adjacent territory of *Sample 1*. Positional information has been lost, indeed quantized, so the above 'intuitive' answer was correct for this case.

The information loss can be avoided by using an integration kernel in the form of a triangle or dual ramp that spans two sample periods, as shown in Figure 11 (lower). By comparing the values of *Sample 0* and *Sample 1,* both the area and the position of the transient can now be unambiguously determined.

These possibilities are extended in [23] wherein it is shown that by using a higher-order B-spline kernel,[7] it is possible to determine separately the intensities and positions of two or more pulses even if they lie within the same sampling period!

The equations are somewhat daunting and the process relies on signal samples having excellent signal-to-noise ratio, so the present authors are not suggesting that the ear is able to perform this feat. Nevertheless this paper is one of several that highlight possibilities for non-traditional sampling methods [19][20][21][22][24] [25][26].

4.2. Reconstruction

Reconstruction can be regarded as the dual of sampling and approached in a similar way. Thus, it is not recommended to present the samples as unfiltered Dirac spikes to subsequent equipment. Even convolving each spike with a rectangle of width one sample period (which is equivalent to a zero-order hold) still generates theoretically infinite slew rates at the transitions.

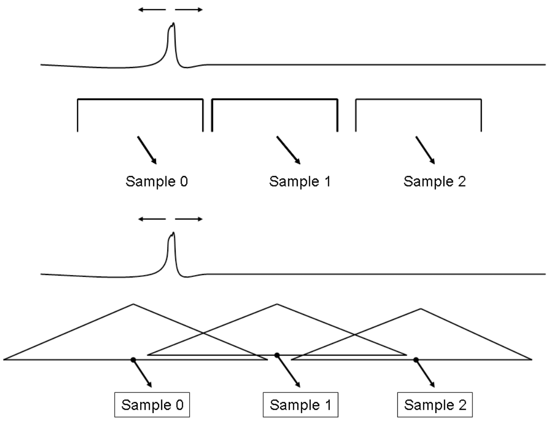

Figure 11. Illustrating the rectangular (upper) and triangular (lower) kernels described in the text.

It thus seems that convolution with a triangle function is the least that is needed to produce a signal that can be handled satisfactorily. This is equivalent to linear interpolation between sample values.

If sampling and reconstruction each use a triangular kernel, then simplistically[8] the total impulse response is a 3rd-order B-spline, of total width four sampling periods. That is a total width of 42 µs at a sample rate of 96 kHz and a time from 10% of peak to the peak of 13.2 µs.

Unfortunately, that is not the end of the story, for we also have to correct a frequency response droop from the 3rd-order B-spline which, for 96-kHz sampling, amounts to 2.5dB at 20 kHz (or 3 dB if sampling at 88.2 kHz).

To meet a criterion such as 0.1dB for the maximum acceptable 20-kHz droop we have generally used a maximally-flat minimum-phase 3rd-order FIR digital flattening-filter immediately prior to the triangle convolution in the reconstruction.

The flattening filter increases the total length of the end-to-end impulse response by three sample periods, giving a total length of seven sample periods. Inevitably, the impulse response is then no longer a single pulse, there being a negative downswing, a positive, and another negative following, as shown in Figure 12 below.

[7] The triangle being considered a B-spline of order 1.

[8] The complication is that because of the sampling, the total system is not time-translation invariant and so does not have a unique 'impulse response' – the response is slightly different according to the position of an original impulse relative to the sampling points.

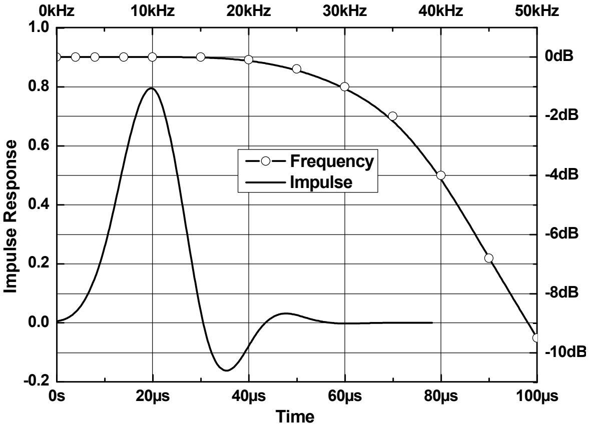

Figure 12. Overall end-to-end frequency and impulse responses of the example system described. The upper curve (open circles) shows dB vs frequency (axes right and top). The lower curve, shows amplitude vs time. The central 80% of energy is consistent with our 10 µs resolution target from Section 2.3, and is very short considering the 96-kHz intermediate transmission path.

4.3. Transparency

Continuing the argument from Section 3.4, we can infer from Figure 10 that the noise-floor of the recording is prolifically described by a 24-bit channel and, using a suitable dither, the word-size could be reduced for distribution with no impact. Psychoacoustic modelling and listening tests show us that, providing the noise from requantization stays 10dB below the original noise spectral density at frequencies below 15 kHz, there is no audible consequence [28][31].

Figure 13 shows spectra relating to the same 192-kHz recording as in Figure 10. The open squares are peak spectral density but after filtering (convolving) with a kernel that attenuates higher frequency components, especially in the range 48 kHz–96 kHz, somewhat more steeply than the triangular kernel discussed in Section 4.1.

When sampled at 96 kHz, frequencies that lie above 48 kHz in the filtered spectrum fold back to mirror-image positions below 48 kHz in the down-sampled spectrum, as shown by the filled squares.

Loss of signal information is minimal. From Figure 10 we deduce that nearly everything above 48 kHz is noise from the recording system.

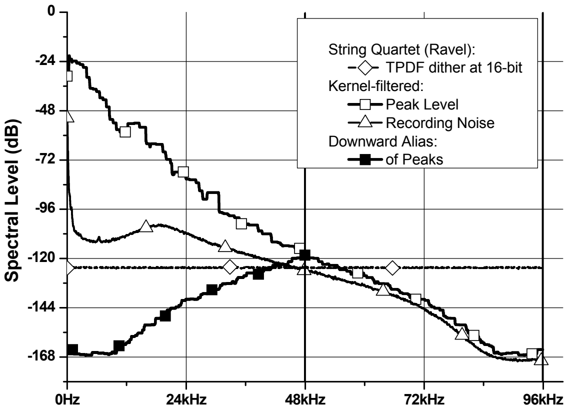

Figure 13. Showing the kernel-filtered noise and peak spectrum along with aliasing, as described in the text.

We cannot chop these frequencies without introducing pre-responses or increasing blur: the sampling process merely reproduces them at the 'wrong' frequency. At the frequency where they are reproduced, they are far below the kernel-filtered-noise from the original recording except very close to 48 kHz, and at least 40 dB below for image frequencies below 20 kHz.

Since, as Figure 13 also shows, the resampling could be benignly quantized to 16 bits, preferably selecting appropriate dither with possibly mild noise-shaping, these aliased components would be covered with a benign inaudible noise.[9] We therefore assert that the audible effect of these aliased images is miniscule.

Aliasing in the frequency domain is equivalent to the time-domain phenomenon of an impulse response that depends on where, relative to the sampling instants, the original stimulus was presented: see footnote 8. Since, according to the frequency-domain description, the downward-sampled components are concealed by original noise, we consider them to be innocuous. There is also upward aliasing introduced by the reconstruction process: here we rely on plausibility arguments, verified by listening to the final result that these alias products, lying above 48 kHz, are inaudible and low enough in level to avoid slew-rate or other problems.

Of course there is also blur caused by the kernel filter and it might be supposed that the sampling and reconstruction filter would inevitably degrade the sound to some small extent. However, this blur is less than conventional

[9] At any sensible acoustic gain that dither would be below the threshold of hearing.

methods and listening tests using commercial 192-kHz material consistently show very positive results.

In our work we have used very detailed listening with recording professionals to help us evolve this coding paradigm which seeks a wholly different time vs frequency balance in representing a musical work.

While these concepts might surprise some, the theory of sampling has evolved considerably since Shannon and Nyquist and, in several other disciplines, such as image processing or astronomy, undersampling can increase resolution with careful application-specific thinking [19][20][21][22][23][24] [25][26].

4.4. Hierarchical Aspects

The above discussion can be extended from sampling a continuous (e.g. analogue) signal to resampling a signal that has already been sampled at a higher rate. The authors have had some success with this process using a triangular kernel, but for much commercial source material something closer to a B-spline of order 4 or 5 has been found preferable.

Using the techniques described in Sections 4.1 and 4.3, a signal that has already been sampled (e.g. at 192 kHz) could conceptually be resampled to another rate (e.g. 96 kHz) by first reconstructing to a continuous-time signal and then sampling that signal at the new rate.

This procedure is recommended only when the two sample rates bear an integer relationship[10] and in practice one would not execute it directly but as a model for a digital filter to resample from the original to the final sample rate in one operation.

Alternatively, one may abandon the conceptual model and design a digital resampling filter directly, respecting the criteria that we have here identified as desirable. These would include minimal blur consistent with a frequency characteristic that ensures aliased products will not be objectionable and downward aliases preferably remaining below an original noise-floor as described in Section 4.3.

Using such filters, we have been able to take a 192-kHz sampled signal, resample to 96 kHz for more economical transmission to a listener, then resample again to 192 kHz in order to optimally feed a D/A converter in the listener's decoder. The downsampling filter has six taps at 192 kHz and the upsampling filter (which includes flattening for droop in the down-sampler) also has six taps at 192 kHz, giving a combined response of 11 taps.

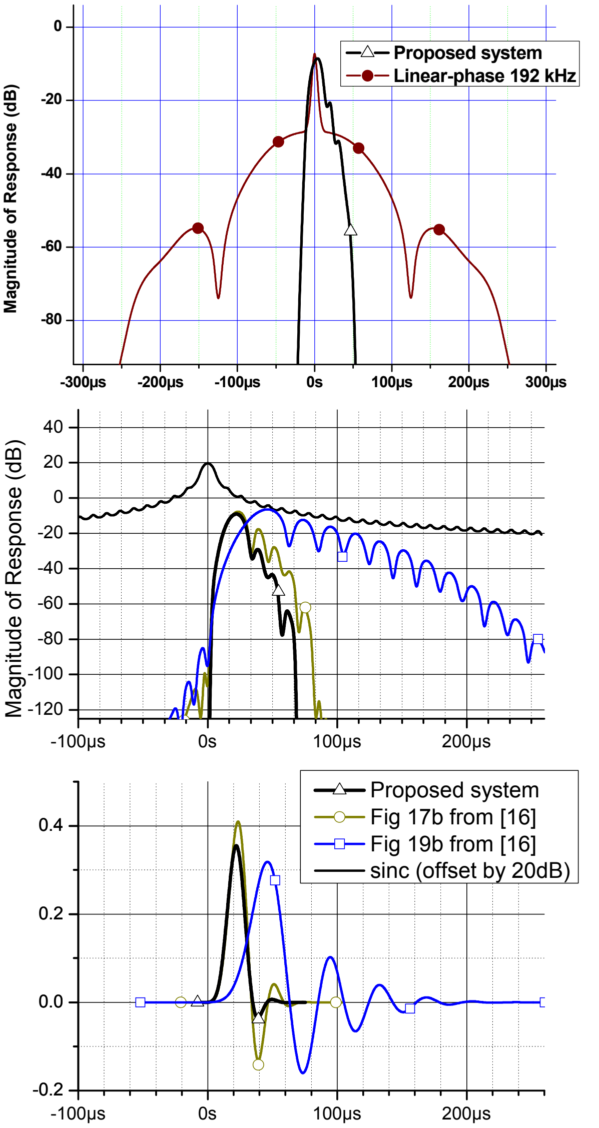

Figure 14. Impulse responses: upper and middle as dB magnitude and lower as amplitude. Comparing the discussed end-to-end system with: (upper) a typical linear-phase cascade at 192 kHz; (middle and lower) show examples from [16] of a 192-kHz apodized filter (Fig. 17b as open circles) and a 96-kHz apodized design (Fig. 19b as open squares). Even though it has been transmitted at 96 kHz, the encapsulation method shows substantially improved temporal fidelity over the earlier 192-kHz or 96-kHz designs. The middle panel also includes the more usual sinc response for 96-kHz (offset 20dB vertically for clarity).

[10] Otherwise there will be beats between the two sample rates.

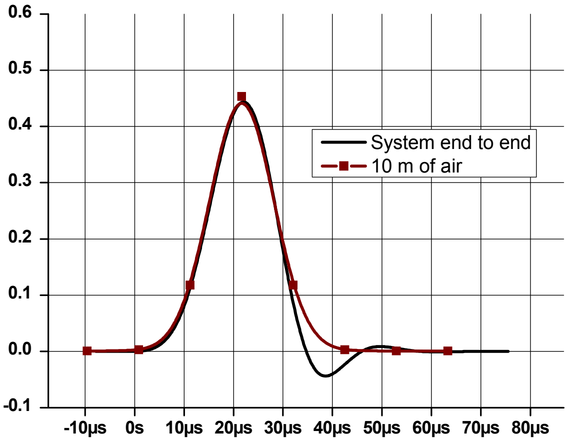

Figure 15. Showing the end-to-end impulse response of the process. Also, marked with rectangles, the impulse response of a Gaussian filter having the same attenuation at 40 kHz as 10 m of air at STP and 30% RH, corresponding to the middle curve in Figure 4.

Assuming that a preceding A/D samples using a triangular kernel and that a following D/A reconstructs also using a triangular kernel, the end-to-end response, shown in Figure 15, introduces considerably less blur than transmission at 96 kHz using conventional filters, as shown in Figure 14 below.

The end-to-end response can also be compared, as shown in Figure 15, with the impulse response of a Gaussian filter having the same attenuation at 40 kHz as 10 metres of air at 30% RH.

We thus have recipes for downward and upward conversion within a hierarchy of rates such as 44.1, 88.2, 176.4 and 352.8 kHz, however these methods do not provide satisfactory conversion from, for example, 96 kHz to 88.2 kHz. This is another reason why it is not recommended that the down-sampled signal be stored in the archive. Even if it sounds wonderful it is 'locked' into its own sample-rate family and cannot be transported to another without significant loss.

If a recording has been archived at 192 kHz and it is required to produce an 88.2-kHz version, a suitable procedure would be firstly to convert the sample rate to 176.4 kHz by conventional means, using severe filtering to suppress aliases, and then to convert to 88.2 kHz by the methods described here. This second conversion can be expected to provide substantial suppression of ringing and other artefacts near 88.2 kHz caused by the first sample rate converter, so one may hope that the audibly deleterious effects of conventional resampling will largely be avoided.

4.5. Distribution System

Using the coding concepts described above, it is possible to re-code a PCM signal so as to preserve both spectral and temporal features of the content in a smaller coding space. The encoding kernel should be chosen to best match each song (track) and should be kept constant for that segment, it may also take into account knowledge of the A/D converter or prior processing. To maximize potential sound quality and efficiency, both ends of the chain must be involved.

The receiver (decoder) should implement an appropriate up-sampling reconstruction, a flattening filter matching the chosen encoding kernel, and a platform-specific D/A manager. Ideally we should improve efficiency and ensure consistency by using end-to-end subtractive dither.

Conceptually, we are trying to connect the A/D and D/A modulators together with a signal that encapsulates the *entire* sound of the original but without artefacts that imply lack of resolution, and to package it for efficient distribution. The authors have used lossless buried-data signaling within the channel to carry instructions, metadata and authentication. We are illustrating a distribution method which, since the encapsulated version is monitored in the studio, is not only lossless in delivery, but also more dependable than if arbitrary D/A converters are used at playback.

This method is also efficient. For example, the Ravel segment illustrated in Figure 10 can be encapsulated into a distribution file containing all the relevant spectral and temporal information of the 192-kHz 24-bit original (9.2 Mbps) using an average data of 922 kbps.

5. CONCLUDING REMARKS

It is a fact of modern digital-audio life that some signals are not band-limited, anti-alias filters are not ideal and quantizations are not always dithered. However, in the context of distribution, we show that self-similarity in signals allow us to employ innovation-rate concepts while optimizing for temporal accuracy – appropriate for separating and locating environmental and music sounds.

Using insights from the auditory sciences, we review targets for dynamic range, frequency response and time response. We point out that for digital distribution, overall analogue-to-analogue temporal 'blur' makes a better performance metric than sample rate; an upper limit of 10 µs blur should ensure transparency.

We advocate distribution using lossless compression, lossless processing and hierarchical up/down-sampling;

we highlight the quality and efficiency gains possible if the encoder and decoder are mutually aware and each matched to their respective analogue converters.

We suggest that for the current music archive, an efficient distribution channel-coding may be based on non-sinc kernels that provide a music-appropriate coding, using ‘encapsulation filters’ at resampling paired with complementary reconstruction at playback, resulting in channels whose end-to-end degradations meet the target and are more comparable to those of sound passing a short distance through air.

This approach to re-coding results in superior sound and significantly lower data-rate when compared to unstructured encoding and playback, and has been enthusiastically supported in listening trials with a number of recording and mastering engineers, artists and producers.

To potentiate archives we recommend that modern digital recordings should employ a wideband coding system which places specific emphasis on time and frequency and sampling at no less than 384 kHz.

6. PATENT NOTICE

Some aspects of the technology described here are covered by patent applications.

7. ACKNOWLEDGEMENTS

This paper covers a sustained enquiry and the authors are particularly grateful to our co-workers in Meridian Audio and Algol; in particular to Richard Hollinshead, Trefor Roberts, Alan Wood, Michael Capp and Malcolm Law. Our enquiry involved many listening sessions and in-studio comparisons of sources and coding options which could not have been accomplished without particular help from Warner Music, especially Craig Kallman, George Lydecker, Spencer Chrislu, Scott Levitin, Craig Anderson, and Mike Jbara while Sony Music also enabled in-studio listening and we are very grateful to Brooke Ettlee and to Mark Wilder of Battery Studios. George Lydecker and Mark Wilder helped us to capture detailed measurements and characterizations of various analogue tape recorders, desks and AD/DA converters. We would also like to thank Tony Faulkner.

8. REFERENCES

[1] Stuart, J.R. ‘Coding for High-resolution Audio Systems’, *J. Audio Eng. Soc.*, Vol. 52, No. 3 (March 2004)

[2] Lewicki, M.S. ‘Efficient Coding of natural sounds’, *Nature Neurosci.* **5**, 356-363 (2002)

[3] Smith, E.C., Lewicki, M.S., ‘Efficient auditory coding’, *Nature* Vol. 439, pp. 978–982, (Feb. 2006)

[4] Ahveninem, J., Kopco, N., Jääskeläinen, I.P., ‘Psychophysics and neuronal bases of sound localization in humans’, *Hearing Research* **307,** pp. 86–97 (2014)

[5] King, A.J., Schnupp, J.W.H., Doubell, T.P., ‘The shape of ears to come: dynamic coding of auditory space’, *Trends in Cognitive Sciences* Vol.5 No.6, pp. 261–270, (June 2001)

[6] Brand, A., Behrend, O. et al., ‘Precise inhibition is essential for microsecond interaural time difference coding’, *Nature* Vol. 417 pp. 543–547, (May 2002)

[7] Plack, C.J. (ed.), *The Oxford Handbook of Auditory Science: Hearing*, **3** OUP (2010)

[8] Bluvas, E.C., Gentner, T.Q., ‘Attention to natural auditory signals’, *Hearing Research* **305,** pp. 10–18, (2013)

[9] ADA, ‘Proposal of Desirable Requirements for the Next Generation’s Digital Audio’, *Advanced Digital Audio Conference,* Japan Audio Society (April 1996)

[10] Vanderkooy, J., and Lipshitz, S.P., ‘Digital Dither: Signal Processing with Resolution Far Below the Least Significant Bit’, *AES 7th International Conference – Audio in Digital Times*, Toronto, 87–96 (1989)

[11] Craven, P.G., and Gerzon, M.A., ‘Compatible Improvement of 16-Bit Systems Using Subtractive Dither’, *AES 93rd Convention,* San Francisco, preprint 3356 (1992)

[12] Gerzon, M.A., and Craven, P.G., ‘Optimal Noise Shaping and Dither of Digital Signals’, *87th AES Convention*, New York, preprint 2822 (1989)

[13] Gerzon, M.A., Craven, P.G., Stuart, J.R., and Wilson, R.J., ‘Psychoacoustic Noise Shaped Improvements in CD and Other Linear Digital Media’, *AES 94th Convention*, Berlin, preprint 3501 (March 1993)

[14] Widrow, B., Kollár, I., *Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications*, CUP, Cambridge, UK, ISBN: 0521886716 (2008)

[15] Lipshitz, S.P., Vanderkooy, J., 'Pulse-Code Modulation – An Overview', *J. Audio Eng. Soc.*, Vol. 52, No. 3, pp. 200–214 (March 2004)

[16] Craven, P.G., 'Antialias Filters and System Transient Response at High Sample Rates', *J. Audio Eng. Soc.*, Vol. 52, No. 3, pp. 216–242, (March 2004)

[17] Shannon, C.E., 'Communication in the Presence of Noise', *Proc.IRE*, vol. 37 (1), pp. 10–21, (Jan. 1949)

[18] Nyquist, H., 'Certain topics in telegraph transmission theory,' *Trans. Amer. IEE*, vol. 47, pp. 617–644, (1928)

[19] Unser, M. 'Sampling – 50 Years after Shannon', *Proc. IEEE* vol. 88 No. 4, pp. 569–587 (Apr. 2000)

[20] Gensun, F., 'Whittaker-Kotelnikov-Shannon Sampling Theorem and Aliasing Error', *J. Approx. Theory* **85**, pp. 115–131, (1996).

[21] Butzer, P.L., Stens, R.L., 'Sampling Theory for not necessarily band-limited functions: A historical review', *SIAM Review* Vol. 34, No. 1, pp. 40–53, (Mar. 1992)

[22] Eldar, Y.C., Michaeli, T., 'Beyond Bandlimited Sampling: Nonlinearlities, Smoothness and Sparsity' *CCIT Report* #698, (Jun. 2008)

[23] Dragotti, P.L., Vetterli, M., Blu, T., 'Sampling Signals With Finite Rate of Innovation', *IEEE Trans.Sig. Proc*. Vol. 50, No. 6, pp. 1417–1428, (May 2007)

[24] Herley, C., Wong, P.W., 'Minimum Rate Sampling and Reconstruction of Signals with Arbitrary Frequency Support', *IEEE Trans. Information Theory* Vol. 45, no. 5, pp. 1555–1564 , (July 1999)

[25] Pohl, V., Yang, F., Boche, H., 'Causal Reconstruction Kernels for Consistent Signal Recovery', *EUSIPCO*, Bucharest, pp. 1174–1178, (2012)

[26] Unser, M., Aldroubi, A., 'A general sampling theory for non-ideal acquisition devices,' *IEEE Trans. Signal Processing*, vol. 42, pp. 2915–2925, Nov. (1994)

[27] Peus, S., 'Measurements on Studio Microphones', *AES 103rd Convention*, Preprint 4617 (Sep. 1997)

[28] Stuart, J.R., 'Noise: Methods for Estimating Detectability and Threshold', *J. Audio Eng. Soc.*, **42**, 124–140 (March 1994)

[29] Stuart, J.R. 'Predicting the audibility, detectability and loudness of errors in audio systems' *AES 91st convention*, New York, preprint 3209 (1991)

[30] Stuart, J.R. 'Estimating the significance of errors in audio systems' *AES 91st convention*, New York, preprint 3208 (1991)

[31] Stuart, J.R. 'Psychoacoustic models for evaluating errors in audio systems' *Proceedings of the Institute of Acoustics,* **13**, part 7, 11–33 (1991)

[32] Jackson, H.M., Capp, M.D., Stuart, J.R., 'The audibility of typical digital audio filters in a high-fidelity playback system', to be presented at *AES 137th Convention* (Oct. 2014)

[33] Henning, G.B., 'Detectability of interaural delay in high-frequency complex waveforms', *J. Acoust. Soc Am*., **55**, No. 1, 84–90, (1974)

[34] Krumbholz, K., Patterson, R.D., 'Microsecond temporal resolution in monaural hearing without spectral cues?' *J. Acoust. Soc Am*., **113**, No. 5, 2790–2800, (2003)

[35] Klump, R.G., Eady, H.R., 'Some Measurements of Interaural Time Difference Thresholds', *J. Acoust. Soc Am*., **28**, No. 5, 859–860, (1956)

[36] Hancock, K.E., Delgutte, B., 'A Physiologically Based Model of Interaural Time Difference Discrimination', *The Journal of Neuroscience*, 24(32):7110 –7117 (Aug. 2004)

[37] Wightman, F.L., Kistler, D.J. 'The dominant role of lowfrequency interaural time differences in sound localization', *J.Acoust. Soc. Am.* **91**, 1648-1661 (1992)

[38] Yost, W.A. 'Discrimination of interaural phase differences', *J. Acoust. Soc. Am.* **55**, 1299-1303 (1974)

[39] Nordmark, J.O., 'Binaural time discrimination', *J. Acoust. Soc Am*., **35**, No. 4, 870–880, (1976)

[40] Lagadec, R., 'New Frontiers in Digital Audio', *AES 89th Convention*, Los Angeles preprint #3002 (1990)

[41] Schnupp, J., et al., *Auditory Neuroscience: Making Sense of Sound*, ISBN 978-0-262-11318-2, MIT Press (2011)

[42] Rieke, F., et al, *Spikes: Exploring the Neural Code*, ISBN 978-0-262-18174-7, MIT Press (1997)

[43] Rees, A., Palmer, A.R., (eds.), *The Oxford Handbook of Auditory Science: The Auditory Brain*, **2** OUP (2010)

[44] Bregman, A.S., *Auditory Scene Analysis: The Perceptual Organization of Sound,* (The MIT Press, 1990)

[45] Green, D.M., Swets J.A. *Signal Detection Theory and Psychophysics*, ISBN 0-471-32420-5, New York: Wiley (1966)

[46] Robinson, D.W., and Dadson, R.S., 'Acoustics – Expression of physical and subjective magnitudes of sound or noise in air', ISO131 (1959)

[47] Robinson, D.W., and Dadson, R.S., 'A redetermination of the equal-loudness relations for pure tones', *Brit. J. Appl. Physics*, vol. 7, pp. 166–181 (May 1956)

[48] Dadson, R.S., and King, J.H., 'A determination of the normal threshold of hearing and its relation to the standardisation of audiometers', *J. Laryngol. Otol.,* **66**, 366–378 (1952)

[49] Harris, G.G., 'Brownian Motion in the Cochlear Partition', *J. Acoust. Soc Am.,* **44** No. 1, 176–186 (1968)

[50] Pumphrey RJ: 'Upper limit of frequency for human hearing', *Nature*, **166**, 571 (1950)

[51] Cohen, E.A., Fielder, L.D., 'Determining Noise Criteria for Recording Environments', *J. Audio Eng. Soc*., **40**, 384–402 (May 1992)

[52] Fellgett, P.B., 'Thermal noise limits of Microphones', *J. IERE,* **57** No. 4, 161–166 (1987)

[53] Boyk, J., 'There's life above 20 kilohertz! A survey of musical instrument spectra to 102.4 kHz', http://www.cco.caltech.edu/~boyk/spectra/spectra.htm (2000)

[54] Stuart, J.R., and Wilson, R.J., 'Dynamic Range Enhancement using Noise-Shaped Dither at 44.1, 48 and 96 kHz', *AES 100th Convention*, Copenhagen (1996)

[55] Acoustic Renaissance for Audio, 'DVD: Pre-emphasis for use at 96 kHz or 88.2 kHz', https://www.meridian-audio.com/ara/dvd_96k.pdf (Nov. 1996)

[56] Kunchur, M.N., 'Temporal resolution of hearing probed by bandwidth restriction', *Acta Acustica*, **94**, 594–603 (2008)

[57] Kunchur, M.N., 'Audibility of temporal smearing and time misalignment of acoustic signals', *Technical Acoustics*, http:/www.ejta.og, **17**, (2007)

[58] Kunchur, M.N., 'Auditory mechanisms that can resolve 'ultrasonic' timescales', *AES 128th Convention*, London, (May 2010)

[59] Plenge, G.H., Jakubowski, H., Schone, P., 'Which bandwidth is necessary for optimal sound transmission', *AES 62nd Convention*, Brussels, preprint 1449 (March 1979)

[60] Woszczyk, W., 'Physical and perceptual considerations for high-resolution audio', *AES 115th Convention*, New York, preprint 5931 (Oct 2003)

[61] Schulze, H., Langner, G., 'Auditory cortical responses to amplitude modulations with spectra above frequency receptive fields: evidence for wide spectral integration', *J. Comp. Physiol. A*, **185** 493–508 (1999)

[62] Dunn, J., 'Anti-alias and anti-image filtering: the benefits of 96 kHz sampling rate formats for those who cannot hear above 20 kHz', *AES 103rd Convention*, Amsterdam, preprint 4734 (May 1998)

[63] Gaskell, H., Henning, G.B., 'Forward and backward masking with brief impulsive stimuli', *Hearing Research*, **129** 92–100 (1999)

[64] Henning, G.B., 'Monaural phase sensitivity with Ronken's paradigm', *J. Acoust. Soc. Am.* **70(6)** 1669–1673 (Dec 1981)

[65] Stuart, J.R., and Wilson, R.J., 'A search for efficient dither for DSP applications', *AES 92nd Convention*, Vienna, preprint 3334 (1992)

[66] Stuart, J.R., and Wilson, R.J., 'Dynamic Range Enhancement Using Noise-shaped Dither Applied to Signals with and without Pre-emphasis', *AES 96th Convention*, Amsterdam, preprint 3871 (1994)

[67] Stuart, J.R., 'Auditory modelling related to the bit budget', *Proceedings of AES UK Conference 'Managing the Bit Budget'*, 167–178 (1994)

[68] Castro Silva, I.M.de, Feitosa, M.A.G., 'High-frequency audiometry in young and older adults when conventional audiometry is normal', *Brazilian Journal of Otorhinolaryngology* **72(5)** 665–672 (Sep/Oct 2006)

[69] Kurakata, K., Mizunami, T., Matsushita, K., Ashihara, K., 'Statistical distribution of normal hearing thresholds under free-field listening

conditions', *Acoust. Sci. & Tech.* **26,5** 440–446 (2005)

[70] Siveke, I., Ewert, S., Grothe, B., Wiegrebe, L., 'Psychophysical and Physiological Evidence for Fast Binaural Processing', *The Journal of Neuroscience*, **28(9)** 2043–2052 (Feb 2008)

[71] Heil, P., Neubauer, H., 'A unifying basis of auditory thresholds based on temporal summation', *PNAS* **100** 6151–6156 (May 2003)

[72] Akune, M., Heddle, R.M., and Akagiri, K., 'Super Bit Mapping: Psychoacoustically Optimized Digital Recording', *AES 93rd Convention,* San Francisco, preprint 3371 (1992)

[73] Buus, S. et al. 'Tuning curves at high-frequencies and their relation to the absolute threshold curve' in Moore, B.C.J. and Patterson, R.D. (eds.), *Auditory Frequency Selectivity*, (Plenum Press, 1986)

[74] Shailer, M.J., Moore, B.C.J., Glasberg, B.R., Watson, N., Harris, S. 'Auditory filter shapes at 8 and 10kHz' *J. Acoust. Soc. Amer.*, **88**, 141–148, (1990)

[75] Yoshikawa, S., Noge, S., Ohsu, M., Toyama, S., Yanagawa, H., Yamamoto, T., 'Sound Quality Evaluation of 96-kHz Sampling Digital Audio', *AES 99th Convention*, New York, Preprint 4112 (1995)

[76] Lipshitz, S.P., Vanderkooy, J, 'Why 1-Bit Sigma-Delta Conversion is Unsuitable for High-Quality Applications', *AES 110th Convention*, Amsterdam, preprint 5395 (May 2001)

[77] Pras, A., Guastavino, C., 'Sampling rate discrimination: 44.1 kHz vs. 88.2 kHz', *AES 128th Convention*, Amsterdam, preprint 8101 (May 2010)

[78] Leonard, B., 'The downsampling dilemma: perceptual issues in sample rate reduction', *AES 124th Convention*, Amsterdam, preprint 7398 (May 2008)

[79] Nishiguchi, T., Hamasaki, K., 'Differences of Hearing Impressions among Several High Sampling Digital Recording Formats', *AES 118th Convention*, Barcelona, preprint 6469 (May 2005)

[80] Nishiguchi, T., Hamasaki, K., Iwaki, M., Ando, A., 'Perceptual Discrimination between Musical Sounds with and without Very High Frequency Components', *AES 115th Convention*, New York, preprint 5676 (Oct 2003)

[81] Ashihara, K., Kurakata, K., Mizunami, T., Matcsushita, K., 'Hearing threshold for pure tones above 20kHz', *Acoust. Sci. & Tech* **27(1)** 12–19 (2006)

[82] Story, M., 'A suggested explanation for (some of) the audible differences between high sample rate and conventional sample rate audio material', http://www.cirlinca.com/include/aes97ny.pdf (Sep 1997)

[83] Talavage, T.M., et al., 'Tonotopic Organization in Human Auditory Cortex Revealed by Progressions of Frequency Sensitivity', *J. Neurophysiology*, Vol. 91 1282-1296, (2004)

[84] Hartmann, W.M., Macauley, E.J., 'Anatomical limits on interaural time differences: An ecological perspective', *Frontiers in Neuroscience*, (Feb 2014)

[85] Brughera, A., Dunai, L. and Hartmann, W.M. (2013) 'Human interaural time difference thresholds for sine tones: The high-frequency limit,' *J. Acoust. Soc. Am*. 133, 2839-2855.

[86] Joris, PX, 'Envelope coding in the lateral superior olive. II. Characteristic delays and comparison with responses in the medial superior olive', *J Neurophysiol*. 1996 Oct; 76(4):2137-56

[87] Bernstein, L.R., 'Auditory processing of interaural timing information: new insights.' *J Neurosci Res*. 2001 Dec 15;66 (6):1035-46.

[88] Hartmann, W.M., et al., 'Interaural time difference thresholds as a function of frequency', *Adv Exp Med Biol*. 2013;787:239-46

[89] Smith, R.C.G., Price, R.R., 'Modelling of Human Low Frequency Sound Localization Acuity Demonstrates Dominance of Spatial Variation of Interaural Time Difference and Suggests Uniform Just-Noticeable Differences in Interaural Time Difference', *PLoS One*. 2014; 9(2): e89033

[90] Takahashi, T.T., 'The neural coding of auditory space', *J Exp Biol*. 1989 Sep;146:307-22

[91] Moore, B.J.C., 'The role of temporal fine-structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people'. *Journal of the Association for Research in Otolaryngology*, 9:399-406, 2008.

[92] Kay, G.W.C., Laby, T.H., 'Tables of Physical and Chemical Constants', section 2.4.1, online at NPL, http://bit.ly/1rkaKGv