Coding for High-Resolution Audio Systems*

J. ROBERT STUART, *AES Fellow*

Meridian Audio Ltd., Huntingdon, PE29 6EX, UK

What do we mean by high resolution? The recording and replay chain is reviewed from the viewpoints of digital audio engineering and human psychoacoustics. An attempt is made to define high resolution and to identify the characteristics of a transparent digital audio channel. The theory and practice of selecting high sample rates such as 96 kHz and word lengths of up to 24 bit are examined. The relative importance of sampling rate and word size at various points in the recording, mastering, transmission, and replay chain is discussed. Encoding methods that can achieve high resolution are examined and compared, and the advantages of schemes such as lossless coding, noise shaping, oversampling, and matched preemphasis with noise shaping are described.

0 INTRODUCTION

Audio reproduction starts and ends with a vibration in air and we perceive it through a hearing mechanism that we experience as analog although it is not exclusively analog in operation. There has been debate over why an analog signal should be stored or transmitted digitally. Arguments in favor include that a digital representation may be transmitted without loss or interference and can be stored and manipulated in ways that avoid the distortions introduced by equivalent analog processes. Analog storage or transmission always introduces distortion and noise that cannot be removed, and may also threaten the time structure of the sounds through wow or flutter effects. Analog recordings also tend to degrade with the passage of time.

While badly designed digital coding or poorly executed digital processing can introduce quite distinctive problems, nowadays the prospect exists for transparent coding and processing—a topic tackled in this paper.

Every generation aims to capture great performances and to make them available to a wide audience. So the designer of a recording system should also consider the nature and quality of the archive. During the recording and playback stages, the audio properties of the capture and rendering processing are crucial. When it comes to distributing the recording, it is the audio properties of the channels in the distribution carrier that normally limit the delivered sound quality.

The Compact Disc (CD) was the first widely available digital audio carrier, and over a 20-year period it proved the effectiveness of an optical disc. A long-term audiophile criticism of the CD has been that it lacks the resolution to reproduce all the detail in a musical performance. The limitations of the CD's 44.1-kHz, 16-bit, linear-PCM coding are understood and covered in this paper.

As digital audio has evolved, the capabilities of the channels at both ends of the reproducing chain have come to be superior to those of the CD. High-quality audio practice now recognizes the CD channel as a "bottleneck," and recordings are routinely made and sometimes played back using equipment whose performance potential is considerably higher than that of the carrier. These concepts are illustrated in Figs. 1 and 2.

Techniques for maximizing the human auditory potential of the CD channel have also evolved, including psychoacoustic optimizations [1], subtractive dither [2], in-band noise shaping [3]–[7], buried-data techniques [8], [9] and dither [3], [10]. Some of these will be discussed in this paper.

Higher resolution audio promises better sound than the CD, and the potential for this has already been demonstrated in carriers that permit a wider frequency response, more channels, and greater dynamic range, such as DVD-Audio or SACD. The development of high-density formats based on DVD and its successors force the audio community to make choices on the best way to deliver improved sound, and open up the intriguing prospect of distribution channels that may be transparent to the human listener. During the development of any new distribution format the most important considerations are archive, integrity, resolution, dimensionality, and carrier channel coding.

*Manuscript received 2004 January 29.

In developing a recording chain we need to consider the cost–benefit at each step in the processes illustrated in Fig. 1. There is one, possibly unique, performance we wish to capture, and so the penalty for degrading the archive is high. The equipment and techniques used in the mixing and mastering stages have a weighting that reflects the number of successive processes that may be required to prepare the recording for release. Ideally all processing at this stage will both be transparent and have a sensible safety margin. The trickiest cost–benefit decision applies to the selection of coding on the release format itself. Here overspecification leads to amplified costs in playback equipment and/or loss of playing time. The replay section is the simplest because the user has choices about how to enjoy the recording. In this chain there may be one studio, (hundreds of) millions of players and discs, but only one, unique archive for each performance. These factors explain why higher resolution is now routinely used in studios and why audio coding may change throughout the chain, as illustrated in Fig. 2.

1 INFORMED CHOICES

1.1 Bit Budgets

In the Carrier block of Fig. 1 the signal is in the distribution format. The per-layer capacity of optical discs has steadily risen from 650 Mbyte on CD to 4.7 Gbyte on DVD, and upward of 30 Gbyte on blue-laser carriers. Despite the increasing capacity of modern optical discs, the choices still need to be made concerning the number of channels, the coding to use, or how it should be optimized.

Every distribution channel has a bit budget. In modern carriers such as DVD there is a tradeoff between data capacity given to audio, picture elements, video, ROM contents, and, of course, playing time. For the audio, therefore, the channel designer should avoid oversatisfying one of the requirements in an unbalanced way. Such as by providing excessive bandwidth at the expense of precision, playing time, or a reduction in the number of channels available for three-dimensional representation. For example, the ARA [11] suggested that it was necessary to deliver an audio bandwidth of 26 kHz, with a dynamic range equivalent to that of well implemented 20-bit linear PCM channels. Beyond that point it was felt that further benefits would not accrue until the sound delivered had, by whatever means, been rendered fully three-dimensional.

Informed decisions rely on some form of cost–benefit analysis, but how do we calculate the cost part of the equation if it involves a change in sound quality?

1.2 Models of Human Listening

The quality of an audio channel can only be finally judged in its intended use: “conveying meaningful program material to human listeners.” Psychoacoustics can provide a bridge between the listener’s impression and the engineer’s objective physical understanding. Psychoacoustics can help us to understand the potential consequence to the listener of imperfect “conveying,” applying a measure to any error arising in the channel.

These channel errors need not be transmission failures, but can take the form of noise, distortion, jitter, wow, flutter, and so on. Essentially any change introduced by an

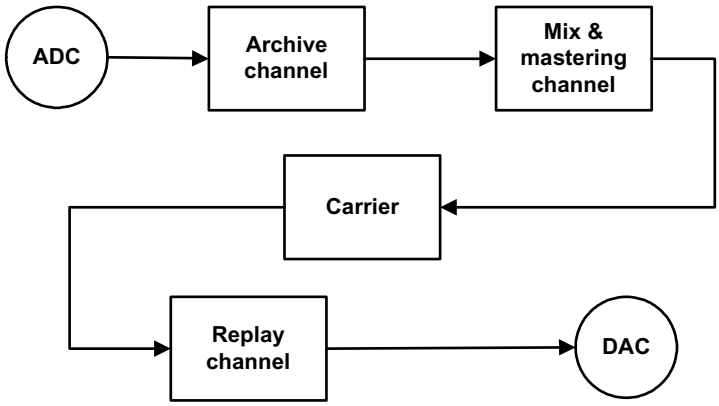

Fig. 1. Simplified block diagram of a reproducing chain.

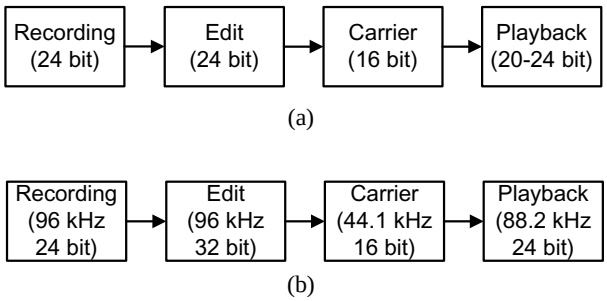

Fig. 2. Example CD chains. (a) Resolution (word size) is limited by carrier. (b) Higher sampling rates and precision are used in preparation while upsampling and/or resolution enhancement is employed on playback.

audio device can be viewed as an added error which may be isolated in measurement and examined by psychoacoustic modeling to estimate its impact. A special case is to try to estimate when channel errors are inaudible—inaudible errors imply transparency, and with some high-resolution coding schemes inaudible errors may be an attainable goal.

Psychoacoustic modeling to estimate the significance of errors can be extremely valuable, but every model or analogy has areas of validity, and the results cannot always be extrapolated. For example, we need to distinguish between perception and cognition. Perception refers to the "low-level" behavior of the human auditory system, where we are concerned with straightforwardly testable parameters such as whether or not a simple stimulus is audible, or detectable in the presence of another (masker) sound, or distinguishable from a similar stimulus. The psychoacoustic literature is full of 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. The study of auditory perception then is an inquiry to answer the question: "to what extent is the auditory system capable of detecting a stimulus, stimulus change, or error?"

Sometimes we also need to consider the higher level process of cognition—where sounds take on meaning. In the cognitive process we are looking for ways in which the higher-level process modifies the listener's ability to discriminate more or less than that indicated by the perceptual model. In the cognitive process we hear "objects" rather than "stimuli," and mechanisms such as streaming and grouping modify the significance of basic percepts.

Fundamental characteristics of the hearing system are complexity and nonlinearity. To the listener, sounds have pitch and loudness rather than frequency and intensity, and the relationships between these measures are nonlinear.

The existence of a threshold is an extreme example of nonlinearity, but so is the fact that the detectability or loudness of a stimulus includes elements that are nearby in frequency while components slightly further away can sometimes mask other sounds, making them seem quieter or inaudible.

A direct consequence of such nonlinearities is that to estimate the audible significance of any stimulus, we have to know its acoustic parameters, including sound intensity. Devices we are characterizing may precede the loudspeaker or volume control, and so we need to know the effective acoustic gain of the system, and this is defined as the SPL that could result from a full-scale sine wave signal, that is 0 dBFS.

The author uses auditory modeling to illuminate the discussion in this paper, the background for which is fully explained in [7] and [12].

2 ARCHIVE

It is important to maximize the archive potential of recordings. In previous generations of recording systems the archive pretty much took care of itself, in that it was not really possible to consider maintaining an archive containing data of significantly higher quality than the release format.

In the early days of the CD the performance potential of the originating equipment did not differ much from the CD standard. In fact it was a while before analog-to-digital converters that genuinely matched the channel potential of the carrier became widely available. Over the ensuing 20 years it has become customary to record and master at higher resolution, dropping to 44.1 kHz 16 bit for the release format, as illustrated in Fig. 2.

However, as time goes on, the cost of storing digital audio data for previously unthinkable periods has fallen rapidly and the format for archiving deserves serious consideration.

The Advanced Digital Audio (ADA) conference [13] encouraged identifying the archive "artifact" at an early stage and developing strategies to retain this for future generations, independent of the release-format recording. While maintaining an archive in more than one form is difficult and potentially prohibitively expensive, we should at least bear the possibility in mind. For this reason the block "Archive channel" has been isolated in Fig. 1.

As an example, Fig. 3 shows the internal architecture of the widely used delta–sigma analog-to-digital converter.

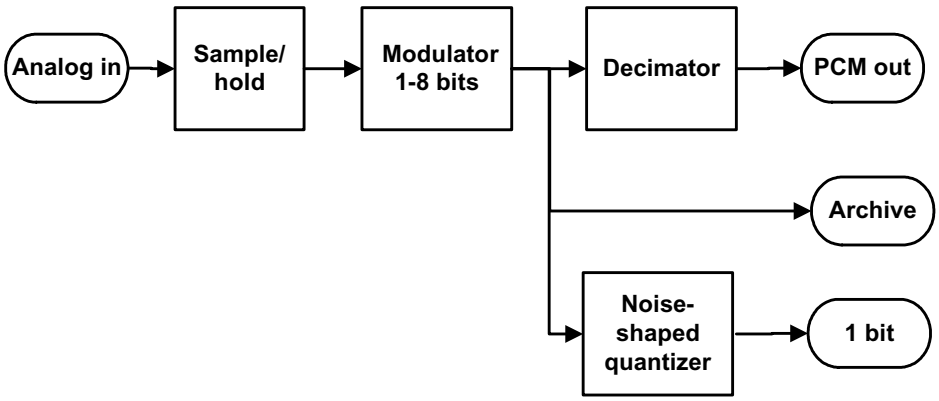

Fig. 3. Block diagram of delta–sigma analog-to-digital converter. A small word-size modulator (between 1 and 8 bit) operates at significant oversampling, maybe between 128 and 8 times base rate of 44.1 or 48 kHz. Wide (24-bit) linear PCM is developed in a dithered decimator, a 1-bit stream is developed in a high-order noise-shaped quantizer (assuming modulator uses more than 1 bit) and modulator output is labeled Archive.

The earliest analog-to-digital converters tended to be multibit and to operate at the base sample rate f_s. Oversampling delta–sigma structures permit simplified antialias filtering and have the potential for higher linearity through using a modest word size in the quantizer. In its most extreme form the modulator is 1 bit and the converter can sample at 64 or more times f_s. Although popular 10 years ago, the single-bit variant has substantial problems of jitter, "birdies," noise modulation, and instability that arise because the modulator cannot be dithered. Because of these problems with 1-bit coding, modern converters have tended to use narrow PCM (4–8 bit) in the modulator [14]. Although there is an engineering challenge to attaining perfect differential linearity in a hardware modulator, the fact that it can be dithered correctly means that overall linearity is improved and, importantly, errors tend not to be correlated with the signal. These concepts are explained well in Lipshitz and Vanderkooy's tutorial [15].

It could be argued that even though it has significant problems as a release or distribution code, the output of the modulator is a more appropriate "archive" than either the decimated multibit PCM output or the noise-shaped and quantized single-bit stream.

Of course it simply may not be possible to capture this signal, either because it is not made available or because there is no suitable recording apparatus on hand. In that case we should capture the processed output which has the minimum impact on transparency, for both this process and those that follow. This paper will argue that for these reasons the PCM output will be the most dependable alternative.

Although presented in the context of an analog-to-digital converter, this argument is valid in every circumstance in which, through processing, noise or errors are added to the digital data. In Fig. 3 the block labeled Decimator is a digital filter that may typically convert a highly oversampled 4–8-bit code into a 24-bit PCM stream at 96 or 192 kHz. While it has been thoroughly understood for some time that dither with a triangular probability distribution function (TPDF) and 2 least-significant bits (LSBs) amplitude can be used to eliminate all nonlinear distortion from DSP processes of this sort [15], it has not always been deployed in single-chip converters or even in some converters sold for professional use. As a result some PCM systems have been exposed to unnecessary criticism. When dither is used to maintain perfect linearity, the archive argument still applies because the multibit PCM signal has been filtered (removing any high-frequency information above half the output sample rate) and has a slightly degraded signal-to-noise ratio due to the dithered quantization. However, in contrast, the noise-shaped quantization process that produces the single-bit coding raises the high-frequency noise floor, introduces uncorrectable correlated errors, and the resulting format is less computationally convenient for either archive or processing for mixing, mastering, or playback.

3 DIGITAL AUDIO PROCESSING

Uniform linear multibit PCM is a very powerful method of encoding analog audio. Provided that both the correct level of TPDF dither is used in the quantizer, and the signal has no content above the Nyquist frequency (half the sampling rate), then the system has infinite resolution of both time and amplitude (see the worked examples in [15]).

The resolution offered in multibit linear PCM is arbitrarily extendable by selecting higher sampling rates and/or quantizer word size. Very quiet sounds may be masked either by nearby noise in the signal or by the signal-independent and (optionally) white-spectrum noise introduced by the dithered quantizer. This uncorrelated additive noise is benign and is perceived separately by the human listener. The noise separates at the cognitive level as a separate object. If dither is not used, then the errors are correlated and may be grouped to the signal "object" and modify its sound.

Distortions can be introduced at analog–digital–analog gateways, or in analog peripherals. However, once the signal is captured in a uniformly sampled, uniformly quantized digital channel, the bits maintain a precise 2:1 magnitude, and the potential for introducing distortion arises only in nontrivial signal processing, which increases the number of bits representing the data. Within an accumulator, or subsequently, this expanded representation of the signal will become too long and eventually the data require truncation. This truncation or requantization process can be made effectively linear by using an appropriate dither at the input to the quantization step, as described in [15].

When we consider high-resolution workflow (such as illustrated in Figs. 1 and 2), we would hope that extreme care be taken in the design and execution of each nontrivial step. In an ideal world, signal processing would be performed in an environment that guarantees adequate word length for all intermediate steps. This implies a higher internal precision both to permit correct use of dither at each stage of processing and to withstand amplified computational noise (for example, in recursive structures), with no audible impact on the noise floor inherent in the recording itself.

3.1 Lossless, Lossy, and Transparent Processing

In the preceding sections the rather difficult term "transparent" has been introduced in the context of processing (that is, modifying the signal) and loosely defined as a special case of high resolution. Transparency implies that whatever the processes, any errors introduced are such that the human listener cannot distinguish the input from the output. Transparency can be evaluated in listening tests. Such tests are complex, expensive, and notoriously difficult, but for the final analysis there may be no substitute. The approach taken here is to assume that errors threaten transparency, a methodology introduced in [12], 16]–[18]. We try to quantify errors to the degree they impact upon dynamic range (the addition of noise or limitation of headroom), linearity (the introduction of correlated distortions), changes in bandwidth (and therefore transient response), or temporal inaccuracies such as the introduction of wow, flutter, or jitter.

It is simpler to start by trying to define transparency in

a digital audio process such as the mixing or carrier blocks in Fig. 1. For the time being we sidestep the question of how we would determine transparency at analog-to-digital and digital-to-analog gateways, although, in fact, the principles are identical.

A lossless (bit-for-bit accurate) process, such as that shown in Fig. 4, will obviously be transparent so long as temporal inaccuracies are avoided. As soon as a nontrivial process is used, or if the audio coding is changed, as illustrated in Fig. 2, then the process is lossy and we need a way to understand its impact and ultimately to estimate whether it might be transparent.

Even trickier is to consider processes that clearly modify the signal in an intended way, such as illustrated in Fig. 5, but where we still want to maintain the concept of transparency when applied to the errors in the process itself. In other words, we would like to know, for example, that a filter we wanted to apply introduced no errors of its own—in effect sounding like a perfect analog filter.

One way to ensure that processing is inherently transparent is to carry it out in an environment that has a larger "coding space" than the original signal, that is in which the combination of bandwidth and dynamic range offered by the sample rate and/or word size exceeds that required to fully represent the signal.

On the other hand, common forms of lossy coding aim to (significantly) reduce the quantity of data to represent a signal. The less data on the output side, the more aggressive the techniques must be. Lossy bit-rate reduction schemes tend to introduce one or more of these characteristics to the output:

- The noise-floor is not constant (modulation noise or masked threshold),
- The noise-floor is psychoacoustically shaped (following either threshold or masking),
- Errors introduced are correlated with the signal.

4 PRECISION AND DYNAMIC RANGE

4.1 Dynamic Range in Uniform PCM Channels

Fig. 6 shows measurements of the level-dependent distortion produced in an undithered quantizer. The original signal (a 1-kHz sine wave) is attenuated in steps to show the effect of a fade when an undithered 16-bit quantizer is measured using an FFT. The graphs illustrate that at high levels the quantization error is noiselike, whereas at low levels it is highly structured. This distortion is more objectionable on lower level signals.[1]

At the gateway and for nontrivial processes we should aim to use appropriate dither at every step. Dithered quantization introduces uncorrelated noise, and although this

[1]To improve the understanding of this quantization effect, some audio examples are available for download at www.meridian-audio.com/w_paper/audio_coding.htm.

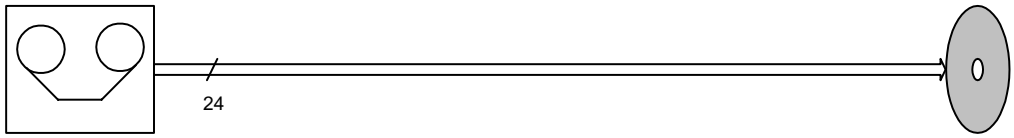

Fig. 4. Lossless process.

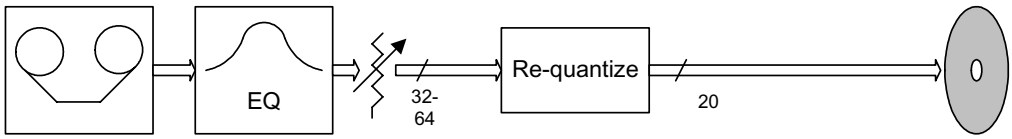

Fig. 5. More invasive "mastering" process.

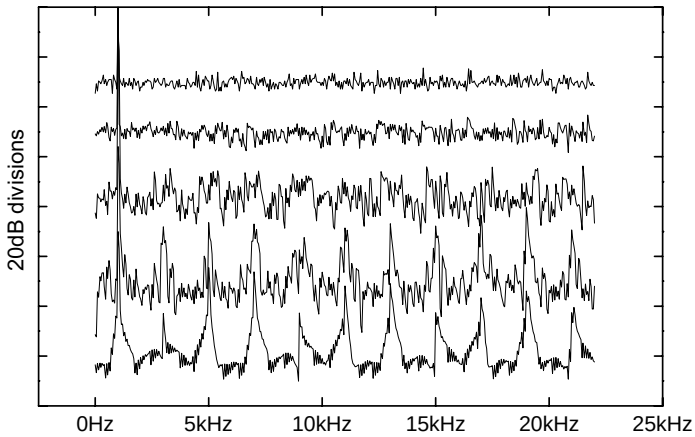

Fig. 6. FFT analyses of undithered 16-bit quantizations of 1-kHz tone at −20, −40, −60, −80, and −90 dBFS (top to bottom). Curves offset by 25 dB for clarity.

noise builds up with successive processing, it is essentially benign.

Fig. 7 shows the FFT measurements of a −90-dBFS 1-kHz signal subjected to 16-bit quantization with and without dither. In each case the 1 kHz signal appears at about the same level. With dithered quantization a smooth noise spectrum represents the benign sounding "error" in the operation. Without dither the resulting signal is rich in unwanted odd-harmonic components totaling 27%. Broadly speaking, truncated, rounded, or dithered quantizations introduce "errors" of similar power but of very different audible consequence.

Before analyzing quantization effects we illustrate some key auditory modeling concepts by considering the significance of the simple noise spectrum that results when a 24-bit channel is reduced to 16 bit using additive TPDF white-spectrum dither. The output noise is −93.32 dBFS in the Nyquist band (0 to one-half sampling rate). In our example the sampling frequency is 44.1 kHz, so the noise spectral density (NSD) will be uniform at −136.76 dBFS/Hz. The lower dashed curve in Fig. 8 is the NSD assuming the acoustic gain to be such that a full-scale digital signal would produce 114 dB SPL.

The intermediate curve is estimated from psychoacoustic modeling, and represents the intensity equivalent of the NSD in a way that allows it to be compared to the single-tone hearing threshold: wherever the noise curve is above the hearing threshold, it will be possible for the noise to be detected.[2] This type of analysis shows clearly the influence of the hearing threshold in determining the loudness and detectability of the noise spectrum.

Now if we were to change the system gain by ±5 dB, then a sound-level meter would indicate the measured noise level changing by the same amount. From auditory modeling in [7], Fig. 9 shows the specific loudness of the same spectrum at 114 ± 5 dB SPL and illustrates how the loudness of the stimulus varies with frequency and how "volume" changes with level. The loudness for the three

[2]Essentially this curve is derived by integrating the noise with a filter set that mimics the auditory filter bandwidth for the appropriate frequency and intensity.

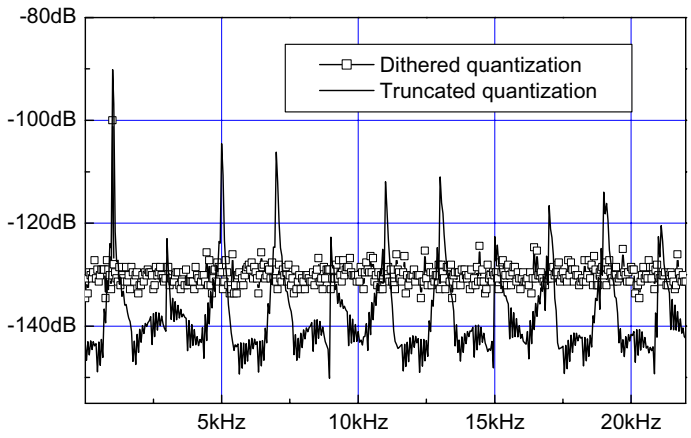

Fig. 7. FFT measurements of spectrum results when a −90-dBFS 1-kHz tone is quantized to 16-bit format with and without correct (triangular probability distribution) dither.

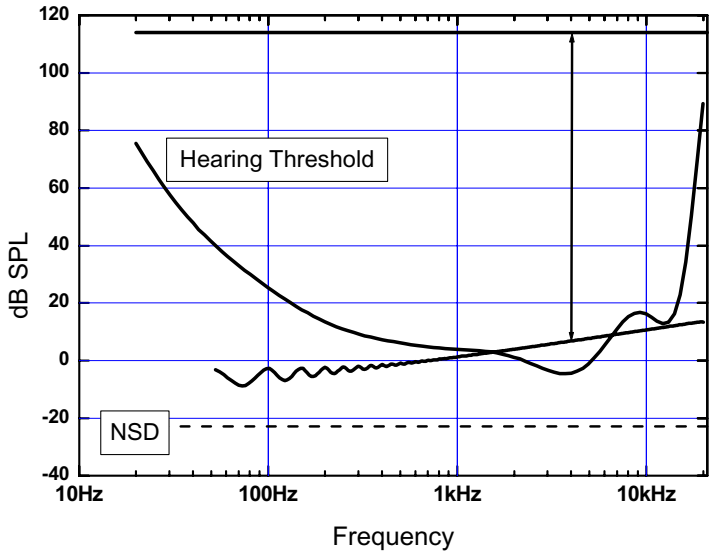

Fig. 8. Audible significance of reference spectrum (16-bit quantization using white TPDF dither), middle sloping curve. Acoustic gain is 114 dB SPL. (Note: Wiggles at low frequency in this and following graphs are precision artifacts of data supplied to model and are not significant.)

gain settings is estimated at 23, 32, and 41 phon—each 5-dB step yielding a 9-phon increase—illustrating nonlinear behavior.

Fig. 8 helps to explain why it is that we can hear sounds below the apparent noise floor. In the most sensitive area around 4 kHz, spectral components are detectable when they exceed the NSD by about 27 dB. Thus although the signal-to-noise ratio for a 16-bit channel may be 93 dB, in this example the spectral resolution is closer to 109 dB, some 16 dB greater.

4.2 CD Channel

The previous section showed how the acoustic gain of the overall system determines the audibility and character of errors such as channel noise. To determine transparency we need to establish a sensible maximum acoustic gain. Fielder [19] has suggested that to reproduce live music we should consider maximum playback levels as high as 126 dB SPL, and in studio situations the acoustic gain may be even higher. The remaining examples in this paper have used a somewhat lower gain that gives 120 dB SPL at the listening position for a sinusoidal signal encoded at full scale.

Fig. 10 presents the measurements shown in Fig. 7 in terms of audible significance. This plot is quite telling: it predicts that the harmonics generated by the undithered quantization will be significantly detectable right up to 15 kHz. The excitation curve shows that the distortion cannot be masked by the tone. It should also be noted that the harmonic at 5 kHz is nearly 30 dB above threshold, which implies that there may be circumstances in which the error can be detected at significantly lower acoustic gains.

Single undithered truncations at the 16-bit level have been regrettably all too common in practice. Not only can inadvertent truncations arise in the hardware filters of poorly designed converters, but the editing and mastering processes often include level shifts, mixing events, or dc filtering processes that in the past have not been dithered correctly. There have therefore been reasonable grounds to criticize the sound of some digital recordings—even though this particular defect can be avoided completely by combining good engineering with good practice.

Fig. 11 represents the audible significance of a channel in which a correctly dithered quantization (perhaps in a word-length reduction from 24 to 16 bit) is followed by a minor undithered process, in this case 0.5-dB attenuation. This figure shows how a single undithered process can degrade a correctly converted signal and illustrates the fallacy of the opinion (too often encountered) that once the dither has been added at the beginning, it will continue to work its magic downstream. Again the figure suggests that detection of the raised and granular noise floor is highly probable.

Fig. 12 represents the audible significance of the same −90-dBFS tone with all the errors introduced by an orig-

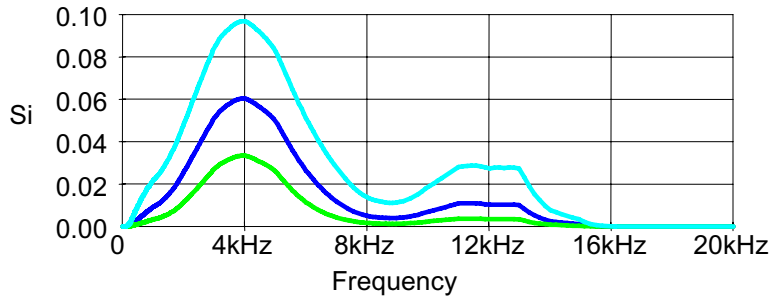

Fig. 9. Internal representation of reference spectrum (16-bit quantization using white TPDF dither) at replay gains of 109, 114, and 119 dB SPL (bottom to top). Specific loudness S_i represents excitation along cochlea after correction for responses, auditory filtering, and power-law adjustment. In this model an S_i of 0.02 would be detected.

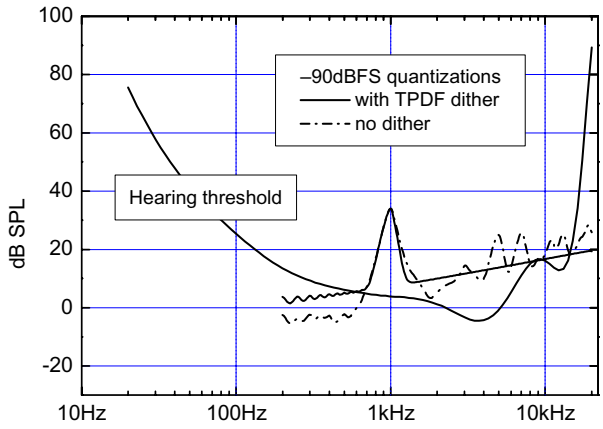

Fig. 10. Audible significance of dithered and undithered 16-bit 44.1-kHz sampling of 1-kHz −90-dBFS (i.e., 30 dB SPL) tone. (0 dBFS ≡ 120 dB SPL.)

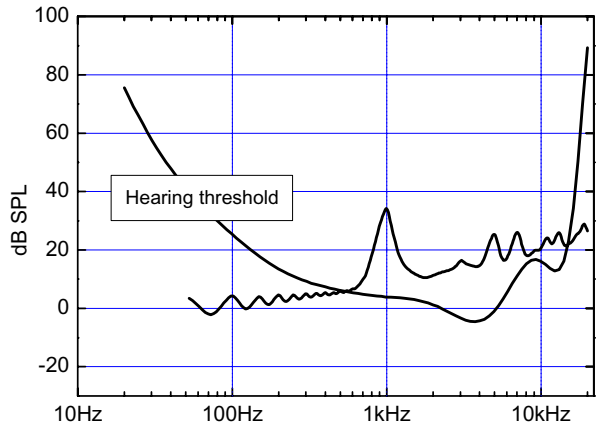

Fig. 11. Audible significance of undithered 16-bit 44.1-kHz sampling of 1-kHz −90-dBFS (i.e., 30 dB SPL) tone on signal already correctly quantized to 16 bit.

inal "correct" 16-bit quantization followed by four undithered signal-processing operations. Four operations may be taken as a baseline of bad practice in CD recording/replay where flawed mixing and poorly designed converters are used. This significance is also put in historical context. The figure includes the audible significance of the playback noise in a silent LP groove.

This analysis of the dynamic-range capability of the 16-bit 44.1-kHz channel makes it very clear that it cannot be considered transparent. Even in the absence of quantization distortion introduced by defective processing, the benign noise introduced by quantization and dither is audible from modest acoustic gains (around 100 dB SPL). Furthermore, undithered quantizations produce distortions that are extremely likely to be detectable and are likely to be unpleasant since they include high- and odd-order harmonics on low-level signals.

4.3 Beyond CD

Fig. 13 shows the human audible significance of the noise introduced by a single dithered quantization process in 44.1-kHz 16-, 18-, 20-, 22-, and 24-bit channels alongside the average hearing threshold. Wherever the noise curve is above this threshold it will be possible for the channel noise to be detected. The degree and frequency range of the suprathreshold spectrum indicate how it will sound. In the 16-bit example the component of noise between 700 Hz and 13 kHz should be audible, whereas audibility is predicted between 2 and 6 kHz for the 18-bit channel.

Fig. 13 suggests that the 20-bit channel noise would be inaudible, and indeed it may be sufficient so long as 20-bit representation is used only on a distribution format (as a bottleneck). Fig. 14 investigates the suitability of a 20-bit mastering chain. The channel's basic noise is shown together with the minimum steady increase in the noise floor that would take place with two or five dithered operations on the signal. Four stages of subsequent processing reduce the dynamic range by 1 bit; sixteen by 2 bit. The operations imagined are minor gain changes or simple filtering. However, within mastering or mixing, more invasive processes are sometimes used that will require some internal arithmetic shifts to prevent overload. Provided the processor has a wider word size than the incoming signal,

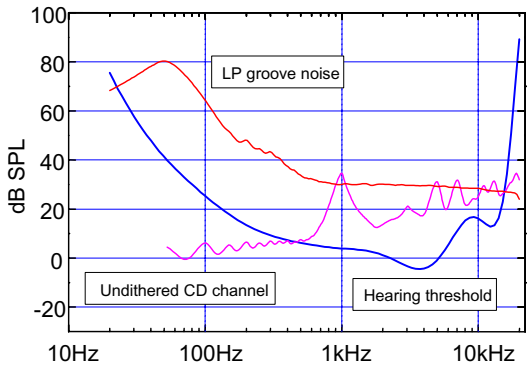

Fig. 12. Audible significance of four successive undithered 16-bit 44.1-kHz resamplings of 1-kHz, −90-dBFS (i.e., 30 dB SPL) tone on signal already correctly quantized to 16 bit, contrasted with audible significance of noise floor measured on silent LP groove.

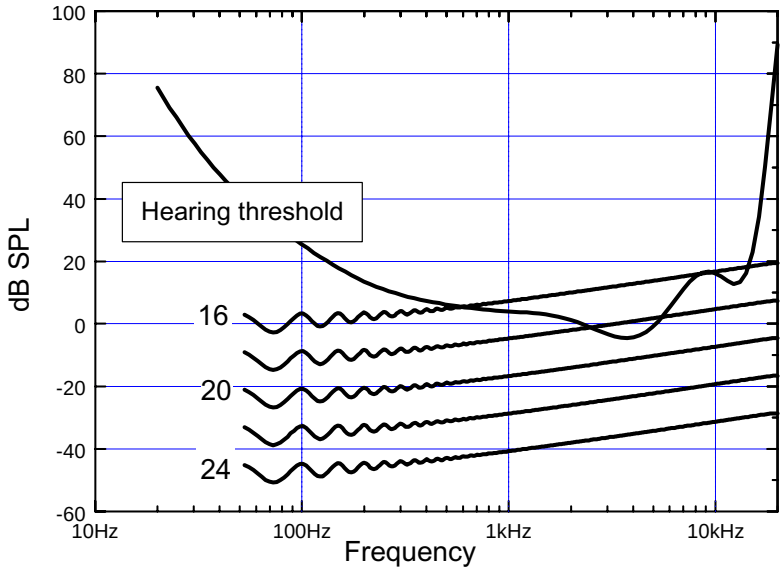

Fig. 13. Audible significance of noise created by single white-spectrum TPDF-dithered quantization in channels using 16, 18, 20, 22, and 24 bit at 44.1 kHz. Audibility has been plotted against average human hearing threshold assuming that a full-scale signal can attain 120 dB SPL at the listening position.

this need not impact on the noise floor, but it should be recognized that some signal processing may increase the noise floor more than indicated by these examples.

Fig. 15 shows the effect of one undithered quantization on a −90-dBFS tone in 16-, 20-, and 24-bit channels. As the channel precision is raised, the error becomes less structured and the power of the error decreases. We saw in Fig. 6 that at higher signal levels, that is, when more LSBs are available to represent the data, the quantization error is more noiselike.[3]

There is no excuse for undithered quantizations. The designer of a high-resolution system should have a zero-tolerance policy for this error. Despite the fact that by using more bits we can show the error to probably be inaudible, nevertheless the errors are correlated with the content, the noise floor will vary with the audio, and these effects propagate downstream.

[3]A −90-dBFS signal in a 24-bit channel behaves rather like a −42-dBFS signal quantized to 16 bit.

Figs. 14 and 15 do, however indicate the wisdom of using sample words as large as 24 bit for the capture, mixing, and mastering stages shown in Fig. 1. The 24-bit coding gives sensible working headroom for DSP processes and to "forgive" any inadvertent quantization that may happen through the workflow. This analysis also suggests that if there is a bit-budget decision to be made for the carrier, then 24 bit is probably excessive to ensure transparency—a topic explored in a later section.

4.4 Thresholds and Room Noise

The analysis so far has considered dynamic range in the context of the standard hearing threshold described in [20]–[22]. However, individuals can exhibit somewhat different thresholds. The minimum audible field has a standard deviation of approximately 10 dB, as shown in Fig. 16. Individuals can be found whose thresholds are as low as −20 dB SPL at 4 kHz, and although the high-frequency-response cutoff rate is always rapid, some can detect 24 kHz at high intensity.

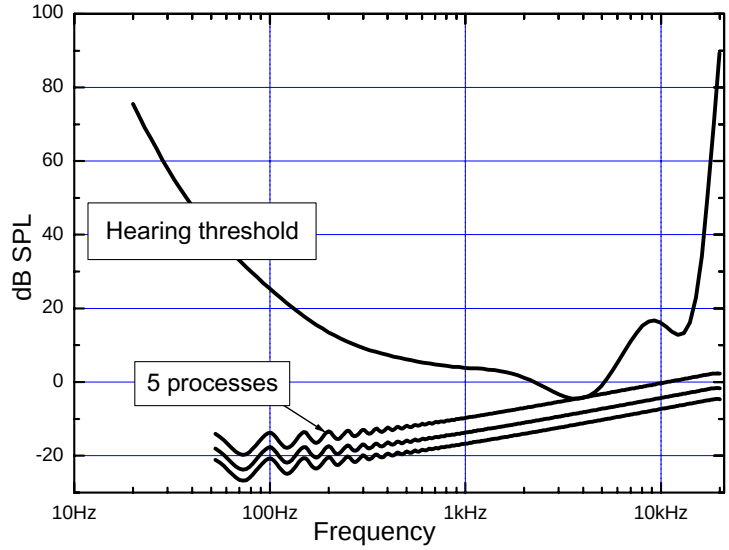

Fig. 14. Audible significance of noise created by 1, 2, and 5 (bottom to top) successive TPDF-dithered quantizations in 20-bit channel.

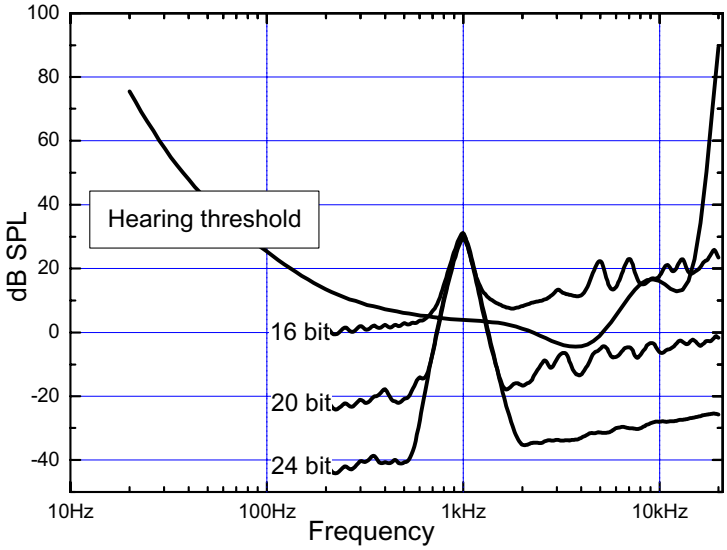

Fig. 15. Audible significance of one undithered quantization when resampling 1-kHz −90-dBFS (i.e., 30 dB SPL) tone in 24-, 20-, and 16-bit channels.

If low-level sounds turn out to be swamped by noise in the playback or recording environments, then we might risk overspecifying the coding required for transparency. While there may be many noisy recording or replay environments, there is no shortage of recording environments where the room noise is lower than the hearing threshold. [19], [23], [24].

Home listening rooms vary more in noise level than recording venues, but the quietest have noise curves below the hearing threshold. In an interesting survey, Fielder [19] found average room noise to be 10–15 dB above the threshold over the range of 50 Hz to 6 kHz (see Fig. 16). The directional properties of our hearing mean that we can in fact discriminate sounds up to 15 dB below the diffuse room noise. In listening tests Fielder determined that noise at a level corresponding to the absolute threshold can be detected in such rooms [19].

The inevitable conclusion is that we cannot reduce the dynamic-range requirement for transparency on account of room noise. On the contrary, certain individuals in quiet rooms may be able to detect the noise floor of a 20-bit channel.

4.5 Recording Noise

It is all too easy to consider that dynamic range may be increased arbitrarily. However, there are some fundamental physical limitations that show up in analog electronics (such as thermal and shot noise) and in the air itself. The human hearing system, in common with that of many mammals, is extremely sensitive. It is thought that one fundamental limit of sensitivity derives from Brownian motion of molecules within the cochlear fluid around the hair-cell receptors [25]. Such is the efficiency of the outer ear that the midrange limit for hearing is also close to that which would reveal the noise of Brownian motion in the air itself.

Fellgett derived the fundamental limit for microphones, based on the detection of thermal noise [26], and this result is plotted in Fig. 17 for an omnidirectional microphone at 300 K.

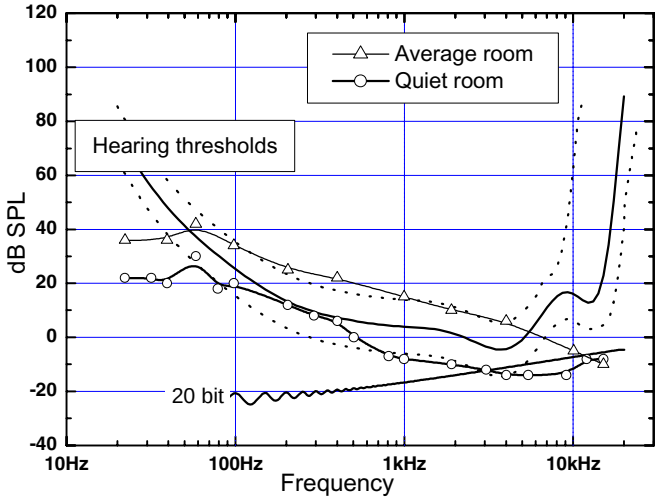

Fig. 16. Standard deviation of hearing thresholds (dotted curves), significance of noise from one 20-bit dithered quantization, and spectra of average and quiet rooms. (Data from [19].)

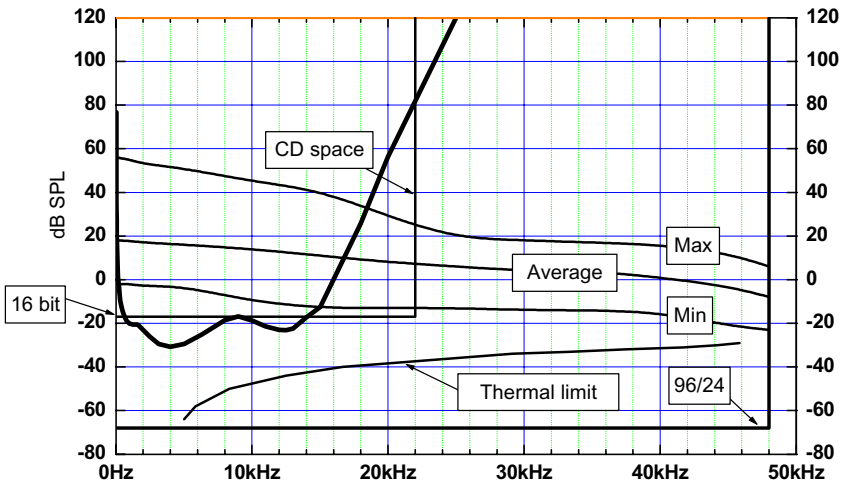

Fig. 17. Survey of inherent noise in 96-kHz 24-bit recordings. Once again assuming that a full-scale signal can attain 120 dB SPL at the listening position, noise spectra are plotted for worst (max), best (min) and average recordings. Thermal-noise limit for an ideal omindirectional microphone is included. Also shown are uniformly exciting threshold noise (described in Section 5.2) and coding spaces for both CD and a properly dithered 96-kHz 24-bit channel.

In [23] and [19] Cohen and Fielder included useful surveys of the self-noise for several microphones. Inherent noise is less important if the microphone is quite close to the instrument and multitrack mixing techniques are used, but for recordings made from a normal listening position it turns out that the microphone is the major limiting factor on the dynamic range—more so if several microphones are mixed. Their data showed one microphone with a noise floor below the human hearing threshold, but other commonly used microphones show midband noise 10 dB higher in level than just detectable noise.

Fig. 17 also shows results taken from an analysis of the background noise in a selection of twelve high-resolution 96-kHz 24-bit recordings. The recordings were chosen to range from reissues mixed from 40-year-old unprocessed analog tape through to modern digital recordings. The highest, lowest, and average noise spectra are shown on a decibels versus linear frequency plot. Also shown are the uniformly exciting noise at threshold (see Section 5.2) and the coding spaces for both CD and 96-kHz 24-bit channels. Obviously these analyses embody not only the microphone and room noise of the original venue, but in the average and higher cases, also analog tape-recorder noise. Even the best analog tape recorder has a noise floor above that of an ideal 16-bit channel. The curve labeled Min is of a recording made in a Dolby screening room using a B&K 4006 microphone. Data for both the room and the microphone are given in [23], and the analysis of the recording is in good agreement. The rise in noise at lower frequencies is due to the room.

In almost all cases the noise has a "pink" characteristic, that is, it declines with increasing frequency, although there are some examples in which the noise floor rises with frequency. The quietest recording still has an inherent noise floor that would be detectable at high listening gains.

It is worth noticing that the coding space provided by 96-kHz 24-bit PCM is not only more than adequate to contain these recordings, but has arguably excessive precision above 5 kHz when thermal noise is taken into account.

5 FREQUENCY RANGE

Nowadays high-resolution recordings are made with sample rates 1, 2, 4, or even 64 times the "base rates" of 44.1 or 48 kHz. Recording and mastering engineers and listeners tend to much prefer the sound when higher rates such as 96 kHz are used. Why? Is it because we really are sensitive to sounds beyond the single-tone threshold? Or is it that by running our systems at higher rates we end up with fewer problems in the lower frequency ranges?

5.1 Psychoacoustic Data to Support Higher Sampling Rates

The frequency response of the outer and middle ear has a fast cutoff rate due to combined rolloff in the acoustics of the meatus and in mechanical transmission. The cochlea provides frequency selectivity through a dispersal mechanism or auditory filtering. This filter action has been extensively studied through air-conducted tests [27].

The cochlea operates top down, and so the first auditory filter, formed by receptors at the basal end (closest to the eardrum), responds to the highest frequencies. Modeling with air-conducted stimuli suggests that this highest filter is centered on approximately 15 kHz, and extrapolation from known data suggests that it should have a noise bandwidth of approximately 2 kHz [28], [29]. It is possible that in some ears a stimulus of moderate intensity but of wide bandwidth may modify perception or detection in this band, so that the effective noise bandwidth could be wider than 2 kHz.[4] Middle-ear transmission loss seems to prevent the cochlea from being excited efficiently above 20 kHz, it is also thought that this region of the cochlea could respond to higher frequencies; in fact response in the range 28–100 kHz has been suggested [30].

There is some merit in the hypothesis that the hair-cell receptors at the basal end may respond to ultrasonic stimulus if it can be made to arrive. Bone-conduction tests using ultrasonics have shown that supersonic excitation ends up in this first "bin." All information above 15 kHz that manages to find its way to the cochlea ends up exciting this region and will accumulate toward detection. Bone-conducted ultrasound is often perceived with the same pitch as the highest audible air-conducted frequency (that is it sounds like a tone in the 15–24-kHz region) and the perceived pitch can be different for each ear.[5] There is some speculation that ultrasound may not (only) be transduced in the cochlea but by direct action on the brain itself.[6]

There is a large body of literature relating to the audibility of bone-conducted ultrasonic sound. In fact it can be used both to provide speech understanding for the profoundly deaf [31] and to mask tinnitus [32]. While bone-conducted ultrasonics can be detected, it should be emphasized that the intensities used or necessary for threshold detection are often quite high, and in fact, since the threshold for perception meets the threshold for pain at these extreme frequencies, there is risk of permanent damage to the cochlea if it is exposed to intense ultrasonic stimulus [33], [34]. There is, however, no evidence that the human can perceive these ultrasonic stimuli *as sound* when they arrive on air [35]. In the wider psychoacoustic literature there is little evidence to suggest that it might be important to reproduce sounds above 25 kHz.

One set of experiments by Oohashi and coworkers has, however, indicated some measurable brain response (but

[4]In this context the late Michael Gerzon surmised that any in-air content above 20–25 kHz may derive its significance from nonlinearity in the hearing transmission, and that combinations of otherwise inaudible components could be detected through any resulting in-band intermodulation products. However, music spectra that have content above 20 kHz tend to exhibit that content at quite low SPL. It is therefore less likely that the (presumed) lower SPL difference distortion products would be detectable and not masked by the main content.

[5]Ultrasonic dental equipment can sometimes be heard by the patient as a loud high-frequency whistle.

[6]Alternative transduction sites might be distinguished by comparing the effect of high-frequency filtering on sound which is either airborne or delivered with headphones.

not auditory response) to program material when the system frequency response is extended beyond 26 kHz [36], [37].[7] In contradiction to Oohashi, Yoshikawa et al. [38] suggest that the superposition of supersonic content (inaudible when played alone) modifies the percept of some music.

In Section 1.2 it was pointed out that the human hearing system exhibits a number of nonlinearities at both the perceptual and the cognitive levels. One implication of a nonlinear system is that linear relationships, such as the interchangeability of time and frequency, need not hold; the ear does not perform a perfect Fourier transform.

It has been suggested that perhaps higher sampling rates are preferred because, somehow, the human hearing system will resolve small time differences which might imply a wider bandwidth in a linear system. In considering this it is important to distinguish between perceiving separate events which are very close together in time (implying wide bandwidth and fine monaural temporal resolution) and those events which help build the auditory scene, for which the relative arrival times are either binaural or well separated [39]. In the first case, wider bandwidth is required to discriminate acoustic events that are closer together in time. This seems to be an alternative statement of the problem to determine the maximum bandwidth necessary for audible transparency.

For binaural time differences the errors to avoid are differential dispersion, delay, or time-quantization between channels. If this can be ensured, then the binaural cues will not be disturbed.

For well separated monaural events it is obviously important that the time scale be not itself quantized. However, the limit to such resolution is not so much frequency response as signal-to-noise ratio. For a perfect detector the ultimate ability to measure a time interval is noise, and therefore noise must be minimized in a system seeking to present fine temporal resolution. The lowest limit for temporal resolution in human hearing for both monaural and binaural events seems to be around 10 μs. [40–43] Obviously 10 μs is less than half the sampling interval at 44.1 kHz, and at first sight that may give rise to concern.

In a digital audio system either the sampling rate has to be sufficiently high to capture the content accurately, or it is necessary to limit the bandwidth of the signal to half the sampling frequency (or less). As pointed out in Section 3 and illustrated in [15], provided this bandwidth requirement is met, and provided TPDF dither is applied at the correct level, then the system resolution of both amplitude and time are limited only by the benign noise floor introduced by the dithered quantizer. Events in time can be discriminated to within very fine limits, and with a resolution very substantially smaller than the sampling period. This point is crucial because provided we treat all channels identically to ensure no skew of directional information, there is no direct relationship between the attainable temporal resolution and the sampling interval.

No matter which sample rate is selected, there must be a low-pass filter before the quantizer, and all questions come down to the same point: the bandwidth limitation is either audible or not.

5.2 What Should the Sampling Rate Be?

Up to now we have used an auditory-modeling technique that transforms noise spectra so they can be graphically compared with tonal errors and the hearing threshold for single tones. In [12] the author described an alternative transform which is more useful when comparing noise spectra directly and when considering coding spaces. In this technique the threshold is modeled as a masking threshold caused by internal noise.

Fig. 18 shows the auditory threshold transformed into this uniformly exciting noise at threshold, as described in [12]. The meaning of this curve is that a noise exhibiting this spectral density will become either undetectable or equally detectable at all frequencies as its level is lowered

[7]Unfortunately Oohashi's setup used 1-bit recording, which may have introduced high-frequency noise (see Section 7.8). It is unclear whether in his experiments it was necessary for the supra-26-kHz content to be correlated with the audio-band information to attain a response. Obviously there exists the possibility that some unrelated supersonic stimulus may modify our cognition.

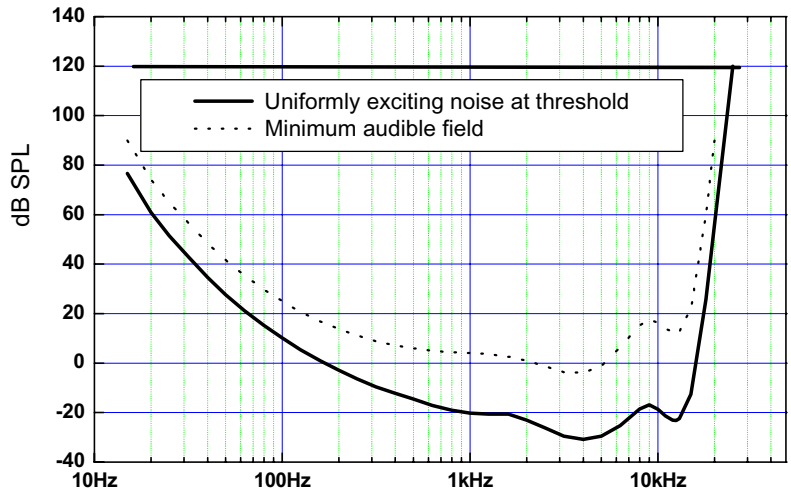

Fig. 18. Derivation of uniformly exciting noise at threshold (lower curve) and minimum audible field tonal threshold.

and raised. The threshold NSD curve is an analogy to the internal noise of the hearing system.

What should the sampling frequency be for transparent systems? To gain perspective on this question, refer to Fig. 19, which replots the auditory threshold on a decibel versus linear frequency Shannon plot. The area bounded by noise floor, maximum level (headroom), and maximum frequency in such a plot is a measure of the information or data capacity of the channel. When the noise floor and headroom are flat, we call it a "rectangular channel."

According to Shannon's theory and to the Gerzon–Craven criterion for noise shaping [4], it is possible to noise-shape a channel of 11 bit at a sampling rate of 52 kHz to obtain a noise spectrum equal to the uniformly exciting noise at threshold shown in Fig. 18. This straightforward analysis, of course, overlooks the fact that if only 11 bit is used, there will be no opportunity for any processing whatsoever and no guard band to allow for differences in system frequency response or between human listeners. In a sense the 52-kHz 11-bit combination describes the minimum PCM channel, using noise shaping, capable of replicating the information used by the ear. This simple analysis implies that 52 kHz is the minimum desirable sampling frequency. For comparison, Fig. 19 shows the coding space offered by both CD and 96-kHz 24-bit coding.

From the information-theory viewpoint the minimum rectangular channel necessary to ensure transparency uses 19-bit linear PCM and has a sample rate higher than 52 kHz. The dynamic range should be increased according to the total number of processes taking place before and after a carrier, and the number of channels feeding into the room. Since higher sampling rates are advocated and enjoyed, we need to look further to find guidance on the optimum sample rate.

5.3 High-Frequency Content of Music

Fig. 20 illustrates the high-frequency region of a CD channel. Superficially the average listener would find little to criticize in the in-band amplitude response. To acute listeners, a 44.1-kHz sample rate (even with the extremely narrow transition band shown) means a potential loss of extreme high frequency (between 20 and 22 kHz), and raising the sampling rate to 48 kHz does a lot to remedy

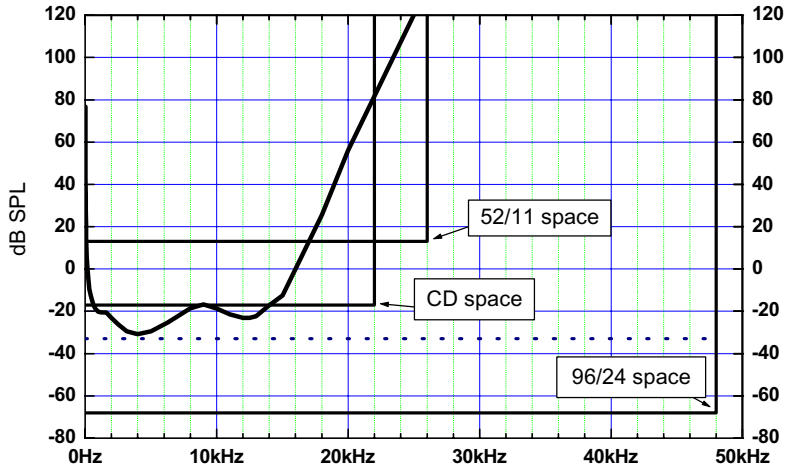

Fig. 19. Shannon space for human hearing and for three channels: CD (44.1 kHz 16 bit), 96 kHz 24 bit, and 52 kHz 11 bit. · · · · noise-spectral density of 18.2-bit channel sampled at 96 kHz.

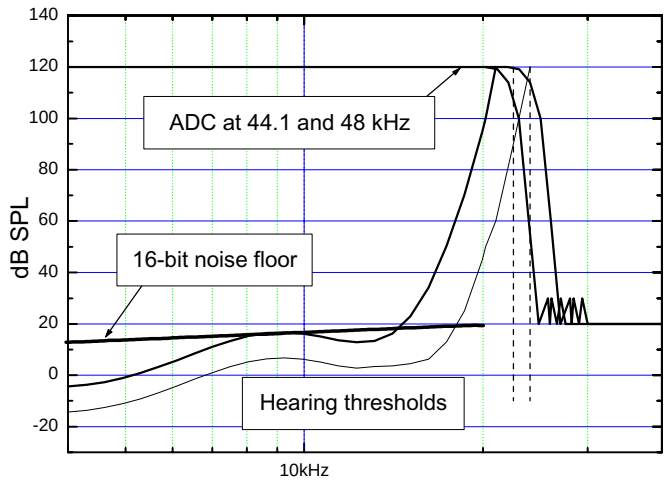

Fig. 20. Useful upper frequency region of low-rate 16-bit channels. Frequency response at 44.1 and 48 kHz is shown against audible significance of noise floor. Average and acute (lower curve) hearing thresholds are also plotted.

this. However, although there is an area of intersection between the channel frequency response and the hearing thresholds, this region is all above 90 dB SPL and the author is unaware of program material that has any significant content above 20 kHz *and* 90 dB SPL.

There is, however, significant content above 20 kHz in many types of music, as an analysis of high-rate recordings has revealed. Fig. 21 shows the spectral envelope for a cymbal crash recorded at a range of about 4 m using 96-kHz 24-bit PCM. The recording used is the same analyzed in [44, fig. 17], and at the highest peak it exhibits the most extended high-frequency content, which can only be captured with sample rates higher than 48 kHz.

A cymbal was chosen for illustration because, according to Boyk [45], it contains more content above 20 kHz than any other instrument, with up to 40% of its power in that range. In his experiment, in which the microphone was much closer to the instrument, at a range of approximately 0.5 m, there was no sign of the supersonic content declining at his measurement limit of 102 kHz. Boyk also gives details of the close-range high-frequency spectra of several instruments, including (in descending order of high-frequency power): cymbal, rim shot, claves, trumpet, speech, triangle, violin, piano, and oboe. He found components above 20 kHz in all of these, but the power above 20 kHz is less than 2% for both trumpet and speech and less than 0.05% for strings, piano, and woodwind. One notable and common characteristic of musical instrument spectra is that the power declines, often significantly, with rising frequency.

The recording used for Fig. 21 was sampled at 96 kHz. If it had been captured at 192 kHz, then we might expect to see the spectrum continuing above the recording noise floor, and continuing to decline in level up to 96 kHz (the Nyquist limit). Fig. 21 gives a useful measure of the SPLs involved.

Fig. 22 shows that the spectral level of 55 dB SPL at 20 kHz is right on the hearing threshold for noise. Obviously we are not 100% certain to what extent the broad-band energy of the cymbal crash might accumulate in the detection "bin" at the top of the cochlea, but it is reasonably clear that as we look up to 30 or 40 kHz, the spectral level is well below any recognized threshold for airborne sound.

Fig. 22 shows the most aggressive instrument played

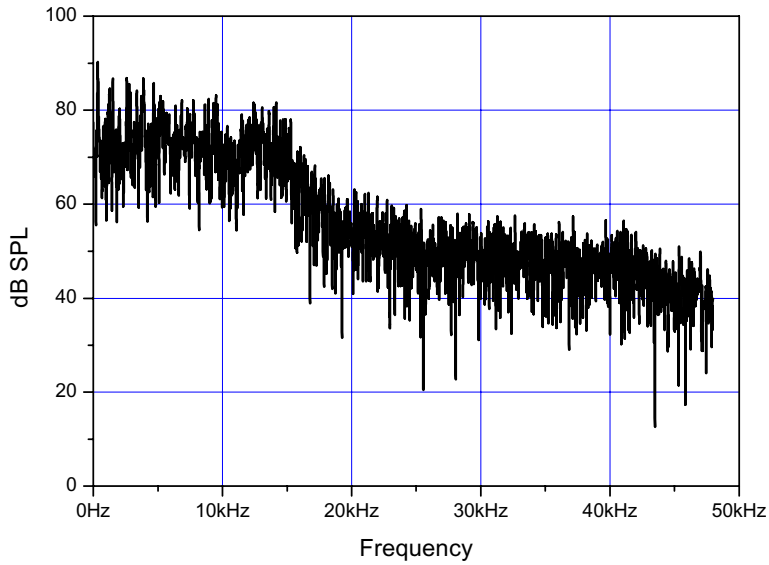

Fig. 21. Spectrum of cymbal crash. Signal is normalized to an acoustic gain of 120 dB SPL. Peak level of this section would be 113 dB.

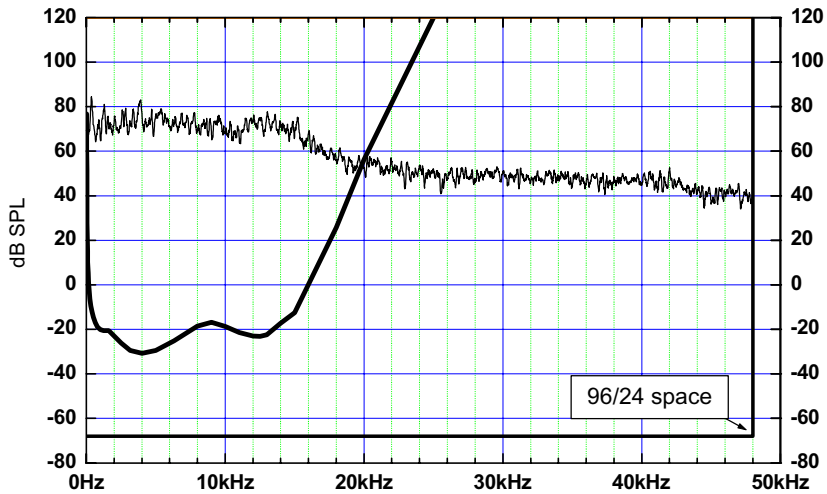

Fig. 22. Spectrum of cymbal crash from Fig. 21 alongside uniformly exciting threshold noise described in Section 5.2.

back at a high level. If the gain were reduced or if we were to analyze other instruments, we would see this supersonic content well below where we understand the threshold to be. We will return to this point in Section 7.

Even though some musical instruments produce sounds above 20 kHz, 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.

5.4 Time-Response Considerations

Even though we cannot strongly argue on the basis of standard psychoacoustics that a system needs to reproduce sounds above 20 kHz, experience shows, and anecdotal evidence suggests that higher sample rates "sound better." Typical observations are that with higher sampling rates the sound is clearer, smoother, has improved low-frequency definition, and is more "natural." In the author's experience higher sample rates can lead to better foreground/background discrimination. "Objects" are better separated from the acoustic and therefore sound clearer and more "complete." This is an indicator that complex percept changes permit clearer grouping at the cognitive level.

Significantly, many of the listening experiences in which a preference has been shown for higher sampling rates have involved somewhat band-limited material, loudspeakers without significant supersonic response, and even listeners with a self-declared lack of acuity at very high frequencies. It therefore seems probable that we should concentrate our attention on the methods used to limit the bandwidth, rather than spending too much time considering the rapidly diminishing potential for program content above 20 kHz.

A distinct feature of established PCM practice is the type of antialias and anti-image filters used for analog-to-digital, digital-to-analog and sample-rate conversion. Particularly at low sample rates, the guard band between 20 kHz and the Nyquist limit ($f_s/2 = 22.05$ kHz in the case of the CD) is narrow. To avoid audible aliasing products it is crucial that above $f_s/2$ the response be extremely low. Chasing "blameless" specifications has tended to encourage designs with tight in-band flatness limits (such as less than 0.01-dB ripple) and to be no more than 0.1 dB down at 20 kHz. The commonly used filter that meets this specification is the linear-phase brick-wall filter, which has a symmetrical impulse response such as those shown in Fig. 23. For efficient implementation in silicon a half-band filter is often used, although it can be less effective at avoiding aliasing or image components.

Although such filters have excellent measured response, they are nevertheless a relatively new item for audio and have no equivalent in the analog world. Analog filters do not have preresponse, and human hearing produces much less premasking (backward masking) than postmasking (forward masking). There is legitimate concern that the preringing of such filters may not be masked and that indeed this artifact may be unexpectedly easy to detect.

There are other concerns with these digital filters, including the subtle effects of ripple in the in-band response and, quite important, the fact that so many of these filters (in chip converters, for example) have not been correctly dithered.

We must discriminate between the *result* of the filtering (genuine listener response to audio content above, for example, 20 kHz in air), aliasing, or imaging effects caused by a combination of high-frequency content and less than adequate stopband attenuation, and any *side effects* that the filtering method itself may introduce, such as preresponses, ripple, or even nonlinearity.

Fig. 20 indicates a typical frequency response for an analog-to-digital converter. While the stopband attenuation of 80–100 dB seems impressive, if we invert this curve, we can see that a detectable in-band alias product may be generated by signals in the transition region

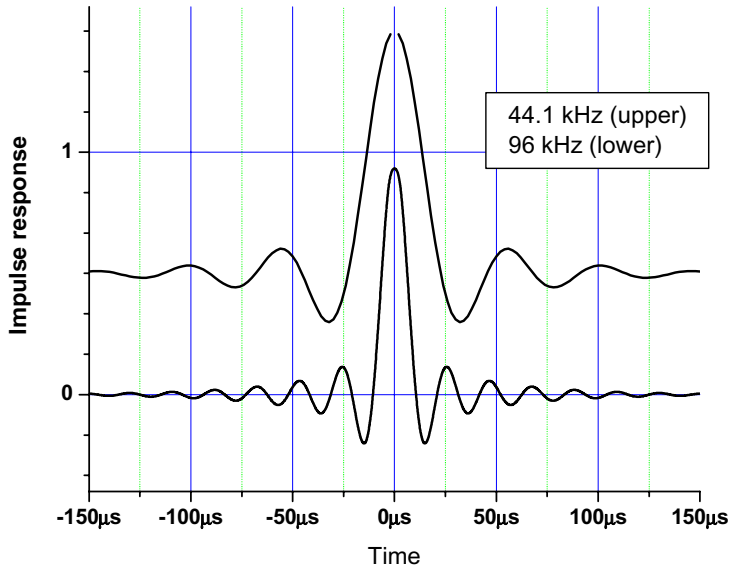

Fig. 23. Impulse response for typical brick-wall linear-phase filter. Responses are shown for 44.1 and 96 kHz. Note that scale is linear, and in fact response extends much longer backward and forward in time. Note also that at higher sample rate the energy is contained within a shorter period.

between 23 and 28 kHz, which are 20 dB below full scale. If, however, the sampling rate is 96 kHz, then components that alias need to be above 48 kHz and are much less likely to arise because sufficient stopband attenuation is simpler to achieve with an octave separation.

It is possible that higher sample rates are preferred because the impulse response of the filters is narrower and the margin between 20 kHz and the Nyquist frequency is so much larger that aliasing can be avoided.

The trend to higher sampling rates seems to have involved apparatus and integrated circuits that operate at higher rates but otherwise identically to established practice at 44.1 or 48 kHz. Higher sample rates do open up the opportunity for a complete review of the best way to design antialias and anti-image filters. At higher sample rates we could roll off the response more gently somewhere above 20 kHz and, possibly substantially, improve the sound of the channel. For this reason an investigation into more appropriate filters was undertaken and is described by Craven elsewhere in this issue [46]. A class of apodized filters has been developed that exhibits minimal preresponse and which have the fascinating property of removing the negative effects of more conventional linear-phase filters elsewhere in the chain.[8]

Fig. 24 compares the impulse response of an apodized filter designed for use at 192 kHz with a fifth-order 40-kHz analog Butterworth filter. The Butterworth filter shows a better transient performance than we expect from analog tape recorders, but is typical of the filters required to limit the ultrasonic noise arising in either oversampling digital-to-analog converters or oversampled channel-coding systems using between 1 and 8 bit (see Sections 7.4 and 7.7). The apodized filter also has an excellent transient response, and since the only remaining concern about PCM systems relates to transient performance, it seems very likely that a high-resolution chain including one of these filters will be free of this problem and can perform better than an analog system in every respect. It is very clear that if preresponses are significant, then the apodized filter is radically better than the almost universally used brick-wall filter illustrated in Fig. 23.

[8]Recently such filters have been deployed in mastering DVD-Audio titles, and initial reactions are very positive.

6 REPLAY CHANNEL

The section marked Replay in Fig. 1 is sometimes ignored, yet decisions made about coding in earlier stages can have a marked effect on both the performance and the implementation complexity of this phase. All too often it is imagined that data from a disc (carrier) are passed to a digital-to-analog converter and that is assumed to be the end of it. In fact, today's replay systems can be very sophisticated and need to cope with a number of different inputs (not just one particular playback channel). There is increasing interest in surround material, and if the user does not possess full-range surround loudspeakers then some form of bass management will be required. Bass management typically collects low frequencies from input channels, combines them with any specific low-frequency energy input from the source, and, after suitable protection or limiting, distributes the low-frequency energy among those loudspeakers capable of reproducing it. A generic schematic for bass management is shown in Fig. 25.

Bass management involves filtering, gain changes, additions, and even protection processing. Each of these steps results in an increase in the word size representing the data. Periodically within the processing or at the end, the data need to be quantized to a workable size and, just as in the mastering stages, correct dither must be applied. Such processing is inherently multibit, which is why players designed to play back 1-bit code (such as SACD) routinely convert to PCM before undertaking this process.

Bass management is not the only process used to optimize replay. Modern high-performance systems may also offer room-acoustic correction, loudspeaker equalization, loudspeaker mapping algorithms (to better match the incoming signals to the loudspeaker array), user tone controls, time-alignment, and so on.

In order for these processes to retain high resolution or transparency (as defined in Section 3.1) they need to be carried out in a coding space that exceeds the original signal. This requirement is most readily satisfied by a PCM representation of the waveform having a sufficient number

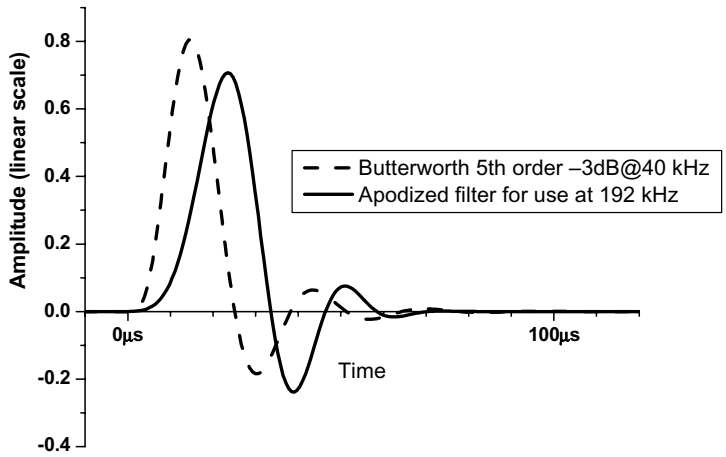

Fig. 24. Impulse response for fifth-order 40-kHz Butterworth filter (– – –) compared with apodized design intended for use at 192 kHz.

of bits, say 22–24, and these processes should take place in a rectangular channel (that is, not one that uses noise shaping). In the author's opinion the requirements for high-performance replay can be no less stringent than those for processing during mixing or mastering.

Since it is obviously a higher performance option to keep the signal in the digital domain until the last possible moment, the ideal replay channel should accept the digital signal from the carrier and postpone conversion to analog until the very last stage. For this reason it is a matter of great practical importance that the channel coding used on a carrier be compatible with all other sources that the user will access. Since CD, digital radio broadcast, DVD, satellite, download, lossy coding algorithms, and so on are based on multibit PCM, it is highly desirable that this coding be used in any new carriers aiming to offer high-resolution playback in the modern context.

7 CHANNEL CODING FOR CARRIERS

The previous sections have concentrated on signal processing used in the development and playback of recordings and on the overall properties of the chain shown in Fig. 1. For high-resolution recordings, ideally, the techniques selected lead to overall transparency even though identical signal representation is not used at every stage. One important section for consideration is the carrier itself. Here there are definite bit-budget tradeoffs between the amount of data and the playing time or room for additional contents. This section explores methods that can be used to reduce the data and data rate on a carrier without losing transparency in the system taken as a whole.

7.1 Lossless Compression

Lossless compression or packing of PCM is the preeminent method to reduce data size and rate on a carrier. Because the decoded data are identical to the input there is no impact on audio quality; it is perfect by definition. Fig. 26 shows the opportunity for carriers that support 24-bit channels to deliver "master quality."

Lossless compression can employ predictive algorithms to encode a PCM data stream efficiently while offering bit transparency across the encoder and decoder. Any stream representing coded audio information is in principle compressible because audio that conveys meaning to human listeners does not continuously occupy the full capacity of a coding channel and has structure that can in part be predicted.

The MLP lossless compression system, which is used

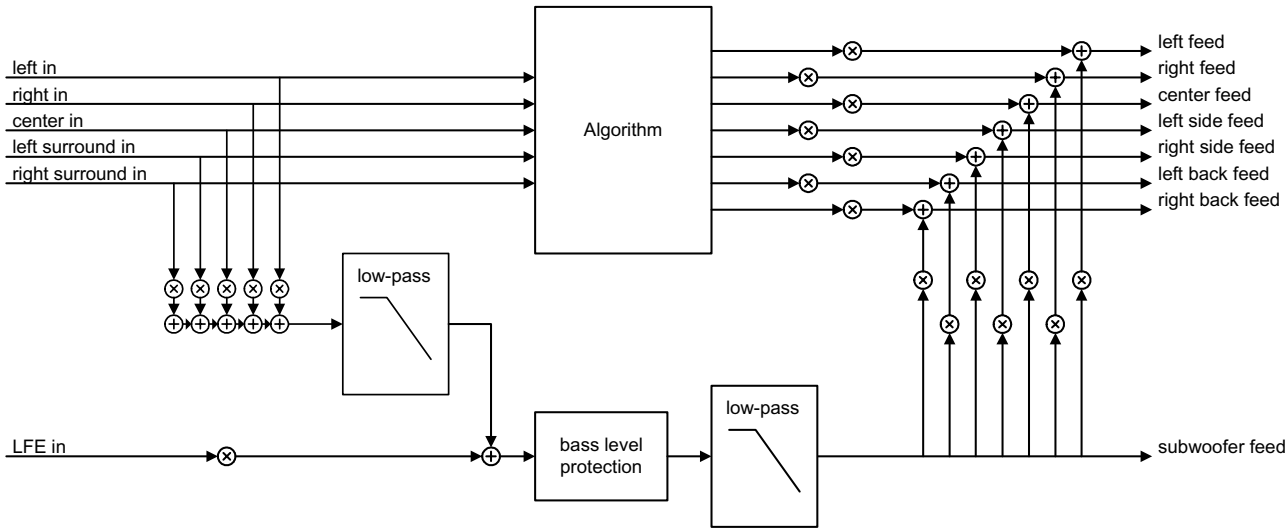

Fig. 25. Typical processing block diagram for bass management in surround system.

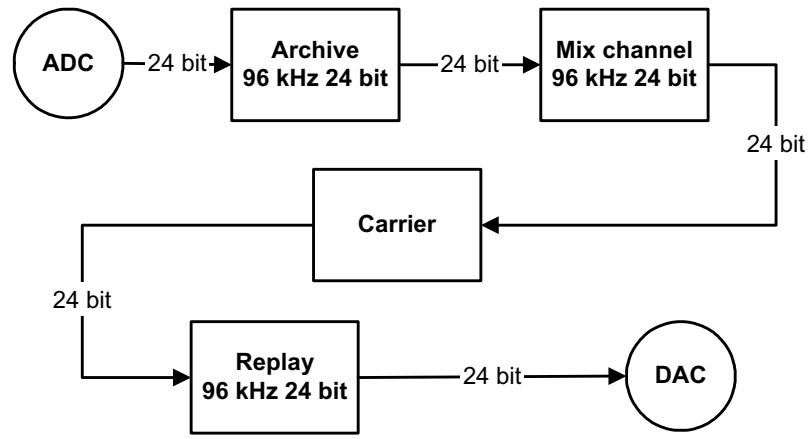

Fig. 26. Direct work flow for carriers support 24-bit coding.

for DVD-Audio and DVD-Audio Recordable, is optimized for carrier channels to provide reduction in both data rate and quantity. The background to MLP is described in detail elsewhere in this issue [44].

On multichannel material a lossless compressor such as MLP can save on average up to 12 bit per sample per channel. This means that for a certain content a 16-bit channel could be compressed up to 4:1 whereas a 24-bit channel could be halved in size. However, lossless compression cannot reduce the coded information below the rate needed to convey inherent noise in the source, so recordings with noiselike spectra or that contain high levels of background noise (such as those worse than the average case in Fig. 17) will compress less.

With high-resolution audio, for good practical reasons, the sample rates have been increased by factors of 2 to include 88.2, 96, 176.4, and 192 kHz. Doubling or quadrupling the data rate to convey less than twice the information is very inefficient. Lossless compression elegantly circumvents this problem since unused coding space is reclaimed and the process of packing PCM becomes more efficient as the sampling rate is increased. MLP automatically discovers the word size of the incoming audio, and the prediction filters adapt to wide-ranging inherent noise.

A mastering engineer may wish to further reduce the data used on a disc to increase playing time or to include different content. In such cases the audio can first be adjusted using the techniques for word-length reduction that are described in Sections 7.2 to 7.4 before feeding it to the lossless encoder. Provided the resulting PCM is multibit (preferably more than 8 bit) then it can be represented with any precision up to 24 bit. The mastering engineer can adjust in 1-bit increments and differently across channels (for example, giving higher precision to left, center, and right compared with a subwoofer or surrounds). Since MLP always delivers a 24-bit word from the decoder there is no requirement to flag any preprocess. Fig. 27 shows examples of how the playing time on a disc can vary as precision is adjusted.

7.2 Word-Size Reduction

The obvious way to reduce the data rate on a carrier is to reduce either the sample rate or the precision (word size). For transparent transmission of a high-resolution recording downsampling is not a serious option. However, properly dithered requantization to reduce the word size is a legitimate technique. In Section 4 we saw that while 24-bit representation is extremely sensible for capture, mixing, mastering (and playback), such channels can convey substantially more precision than is present in the final recording. More to the point, we saw in Fig. 17 that above 5 kHz, 24-bit channels have more dynamic range capability than the difference between the loudest tolerable sound and Brownian motion of the air, that is, absolute fundamental limits.

Consider the examples shown in Fig. 17. The worst example, shown as Max, has a noise floor that could be replicated by a properly dithered and appropriately shaped 11-bit carrier channel. It is therefore wasteful to represent this with 24 bit if the data space can be put to another use. Even the minimum spectrum in Fig. 17 remains above the inherent noise of a 16-bit channel—although we would not advocate reducing this recording to 16 bit if the result is to be transparent.

How far below the noise floor do we have to place a quantization for the result to be inaudible? Auditory modeling suggests that adding an uncorrelated noise that is more than 10 dB below the inherent recording noise will be inaudible. A useful rule of thumb would be that high resolution could be maintained so long as the channel provided 2 bit more resolution than that implied by the self-noise of the recording within the audible range. To take the examples shown in Fig. 17, space could be conserved by using channels between 14 and 18 bit.

7.3 In-Band Noise-Shaping

It is possible to exploit the frequency-dependent hearing threshold by shaping the quantization and dither so that the resulting noise floor is less audible. This subject has been covered extensively in [3]–[7] and [10].

Fig. 28 shows how an in-band (44.1-kHz sampling rate) noise shaper can allow a 16-bit transmission channel to have a subjective noise floor more equivalent to a 20-bit “simple” channel. If such a channel is to be useful, the resolution of the links in the chain before and after the noise-shaped channel must exceed the maximum resolution targeted. (In this example at least 20-bit resolution would be required.)

The error-shaping technique, when combined with TPDF dither, ensures linearity and gives the potential to not impact on the input noise floor. In fact, the noise spectrum from the dithered quantization process can be shaped using psychoacoustic criteria and can reduce the added noise in the sensitive midband sufficiently to give audible benefits of as much as 18 dB with 44.1-kHz sampling.

One problem with all shaped noises, and particularly those similar to that shown in Fig. 29, is that while the noise floor is definitely less audible, if the gain is increased sufficiently, then the noise that is exposed is quite colored and unnatural sounding. Signal processing at playback may expose the rapidly rising noise at high frequencies. In-band noise shaping at 44.1- or 48-kHz sampling rates has been used extensively to issue better sounding CDs, but has not been widely used in higher resolution work.

However, this technique is extremely powerful for

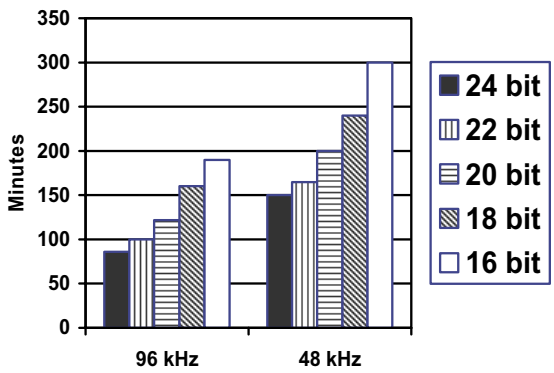

Fig. 27. Playing time obtainable on one layer of a DVD disc using MLP to compress a 6-channel orchestral recording. Incoming precision is varied between 16 and 24 bit at both 48- and 96-kHz sample rates.

improving the subjective dynamic range of channels sampled higher than 48 kHz. Since the higher sampling rate allows the bandwidth of the channel to exceed the high-frequency cutoff of human hearing comfortably, there are different options for noise shapers. In [47] Stuart and Wilson give examples that can provide perceptual gains of up to 6 bit in a 96-kHz channel, one of which is illustrated in Fig. 29. If we can provide a perceptual gain of even 4 bit then, in principle, a 24-bit recording can be conveyed transparently using a 20-bit carrier channel.

The unique advantage of using noise shaping as a coding method to minimize the data rate, or for maximizing the perceptual performance of a channel, is that it requires neither equipment changes for replay nor a decoder.

7.4 Oversampling with Noise Shaping

One strategy to gain the advantage of a higher sampling rate, while limiting the increase in the rate and quantity of data on a carrier, is to oversample and to reduce the word size, with noise shaping. Significant oversampling, for example at four times the CD rate, creates a large amount of coding space above 20 kHz into which quantization and dither noise can be shaped, thereby increasing the dynamic range available at audio frequencies.

Provided the word size used is large enough to support TPDF dither for any quantization steps, then oversampled noise-shaped schemes can provide very high efficiency with no modulation noise or correlated errors, that is, have the potential for transparency. One scheme that has attracted interest uses bytewide (8-bit) coding at quad rates. 8-bit coding has a number of attractive features when it comes to designing effective hardware and signal processing and was highly recommended by ADA [13].

Fig. 30 shows an example of an 8-bit shaper designed for use at 192 kHz. Again setting the acoustic gain to be 120 dB SPL, the noise spectral density for TPDF 8-bit PCM is shown at 25 dB SPL per hertz. The shaped noise is reduced below 35 kHz and has a 6-dB margin from the uniformly exciting threshold curve. This design was optimized to be inaudible at acoustic gains up to 126 dB and to minimize the total power of the shaped noise. Shaping has not changed the coding space of the 8-bit channel, however, the dynamic range above 35 kHz has been traded for a much more useful range below 20 kHz.

The equal-area property of the shaper (to meet the Gerzon–Craven criterion [4]) is more evident in the upper curve of Fig. 31, which also shows the resulting noise spectra for playback systems using low-pass filtering. The examples illustrated are a third-order Butterworth 50-kHz filter and the apodized filter whose impulse response is shown in Fig. 24.

This shaper design provides inaudible noise at the tar-

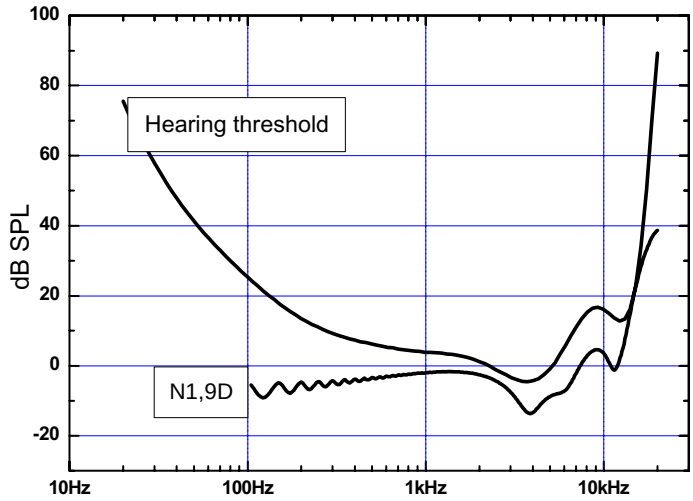

Fig. 28. Audible significance of simple 16-bit channel, with example from [6] of audible significance of noise shaping in 16-bit channel.

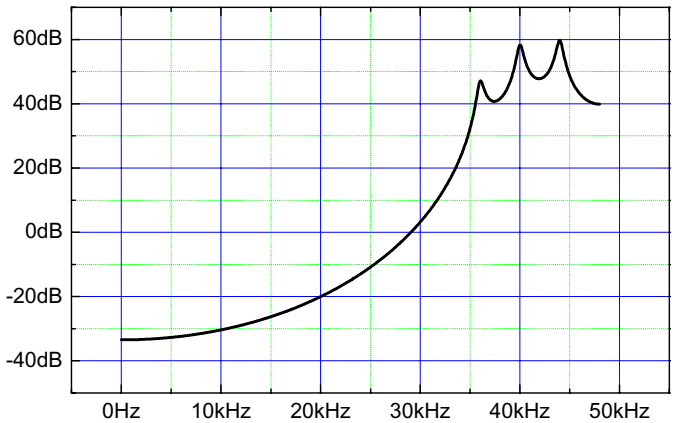

Fig. 29. Noise shaper from [47] for use at 96-kHz sampling.

get acoustic gain. Overall noise gain is 21.8 dB in the raw state, falling to −0.8 dB with the apodized filter. Since the inherent noise of an 8-bit channel is at −45 dBFS, this implies that even in the apodized case the wide-band signal-to-noise ratio of the system will not reach 50 dB even though it exceeds 120 dB below 20 kHz. This may or may not be significant.

What is certain is that ultrasonic noise of this magnitude presents a design and management issue for all downstream stages, and filtering is absolutely necessary. For example, if the replay system needed 200 W to attain 120 dB SPL, then without filtering the tweeter would have to dissipate almost 1 W continuously, whereas with the apodized filter this load drops to below 6 mW.

What is fascinating about this system is that it attains a very high dynamic range in the audio band, has an excellent transient response if an apodized filter is used, is wide-band, and yet the data rate is low at 1.536 Mbit/sample (the same as 96 kHz 16 bit and almost half that of the 1-bit example covered in Section 7.7).

It should be pointed out that this coding is considered for a carrier only. While it is possible to cascade processing in the 8-bit domain—and to do so linearly—a replay system that processes the signal from this carrier should operate with a minimum of 20- and preferably 24-bit precision.

Although less suitable for a carrier, oversampled noise-shaped multibit systems may have some merit for storage or mastering systems if the sample rate is increased. For example doubling the sampling rate to 384 kHz would allow a shaper, that had more dynamic range below 20 kHz with lower noise gain in the shaper and, of course, the replay filter can remove proportionally more of the noise. However, as we see in Section 7.8, this class of noise-shaped channels may be unsuitable for the highest quality work because supersonic components in the signal can be obscured in noise.

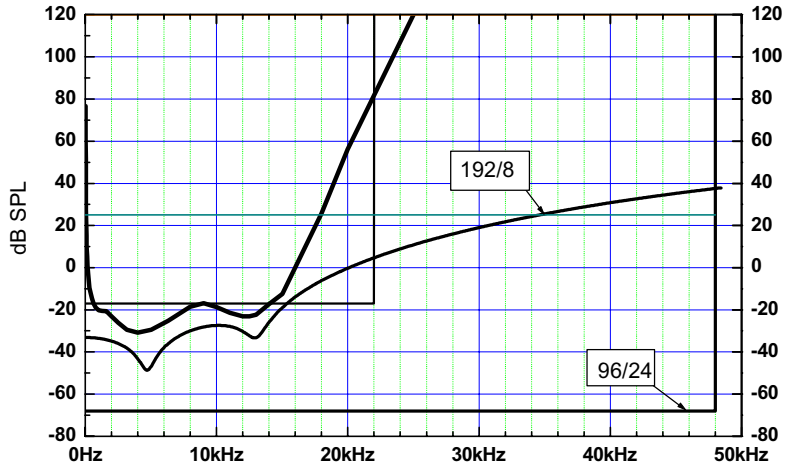

Fig. 30. Noise floor of coding system using a fifth-order noise shaper on a 192-kHz 8-bit channel. Acoustic gain is set to 120 dB SPL. Also shown are noise spectrum for unshaped channel (arrow identifies where they cross at 35 kHz) and coding spaces for CD and 96-kHz 24-bit PCM.

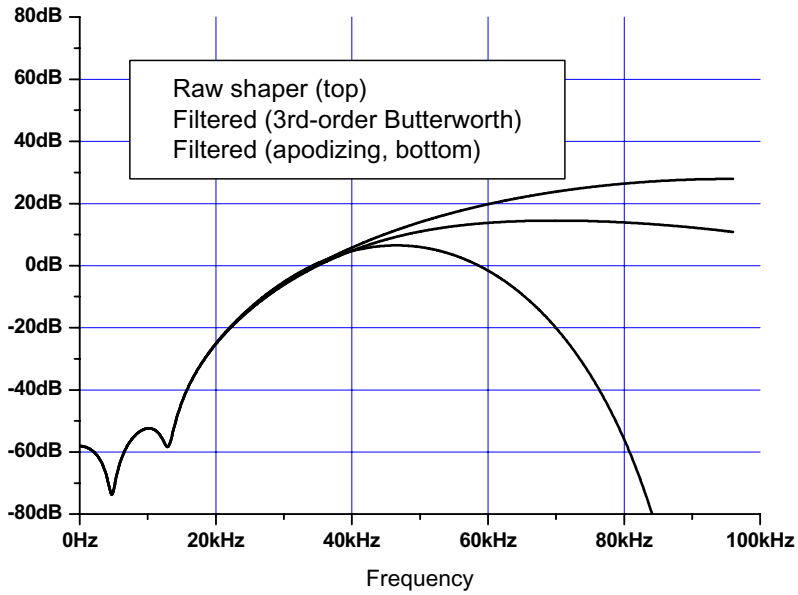

Fig. 31. Noise spectrum for example fifth-order 192-kHz 8-bit shaper raw (top) and when filtered at playback using third-order Butterworth (center) and apodized filter (bottom).

7.5 Preemphasis and Noise Shaping

The use of pre- and complementary deemphasis as a signal-processing method to optimize the subjective dynamic range of analog channels is quite familiar to audio engineers. The method has been used with particular success in cases where the analog noise level increases with frequency, as with magnetic tape, shellac, vinyl grooves, or FM broadcast. In each case a well-documented property of music and speech is exploited: when material of acoustic origin is microphoned at normal listening positions, the average and peak spectrum levels decline with frequency above a few kilohertz. It is therefore efficient to preemphasize high-frequency signals to make it more likely that they will occupy the channel's capacity. Deemphasis is applied on replay or reception and has the dual benefit of reducing both noise and distortion from the preceding chain.

So far all linear PCM standards for digital audio have permitted the use of 50/15-µs pre- and deemphasis (included in Fig. 32). However, this preemphasis has not been overwhelmingly popular with the recording industry, principally because it uses too much in-band high-frequency headroom and presents a mastering management issue because its use has to be flagged.

Designers of future carriers should bear in mind the very interesting possibilities that exist for preemphasis of material recorded at 88.2 kHz or higher. One scheme based on original work by the late Michael Gerzon (described in [48]) combines preemphasis specified in the digital domain with a matched noise shaper in the preemphasis filter, as shown in Fig. 32.

When preemphasis is applied to a channel with a flat noise floor, it is normal for the deemphasis used on replay to result in a final noise spectrum that falls at high frequencies. Gerzon proposed that the encoder should in addition incorporate noise shaping in order that the final noise spectrum should be flat. The lower curve in Fig. 32 shows the noise-shaping curve which in fact is parallel to the preemphasis curve. Fig. 33 shows the headroom and the final replay noise spectrum, which is flat. Note that in exchange for a small amount of headroom at the top of the audio band (3 dB at 20 kHz) this system delivers a perceptual gain of 2.1 bit.[9]

Compared with 50/15-µs preemphasis this scheme offers much improved high-audio-frequency headroom. (It is reduced by only 2 dB at 15 kHz, compared to 9 dB in the current standard.) The preemphasis method involves a noise shaper that gives a 2.1-bit increase in overall audio dynamic range when used as a word-length-reduction device and, because the noise shaper has the same shape as the preemphasis curve, the output (deemphasized) noise spectrum is "white."

This scheme can be combined usefully, at the user's discretion, with an appropriately chosen high-advantage noise shaper such as that shown in Fig. 34. Fig. 35 clarifies the way in which this noise shaper combines with the suggested preemphasis to provide increased dynamic range. The headroom curves at the top show the deemphasized response normalized for 16-, 20-, and 24-bit channels. The lower curve represents the noise spectrum of the shaper used (Fig. 34), after correction to allow for the gain achieved by the preemphasis scheme. The coding space (area between headroom and noise floor) remains equal to that of a 16-bit channel, but it has been redistributed to be more useful. This figure shows how a 16-bit channel at 96 kHz can have an effective dynamic range of 23 bit in the critical 4-kHz region while still offering 19-bit performance at 20 kHz. A key feature of pre- and deemphasis is that coding space can be redistributed, trading unneeded headroom in one region for lower noise.

[9]Analyzing this in terms of coding space, we see that an area has been removed from region A and used to lower the noise floor below 0 dB uniformly.

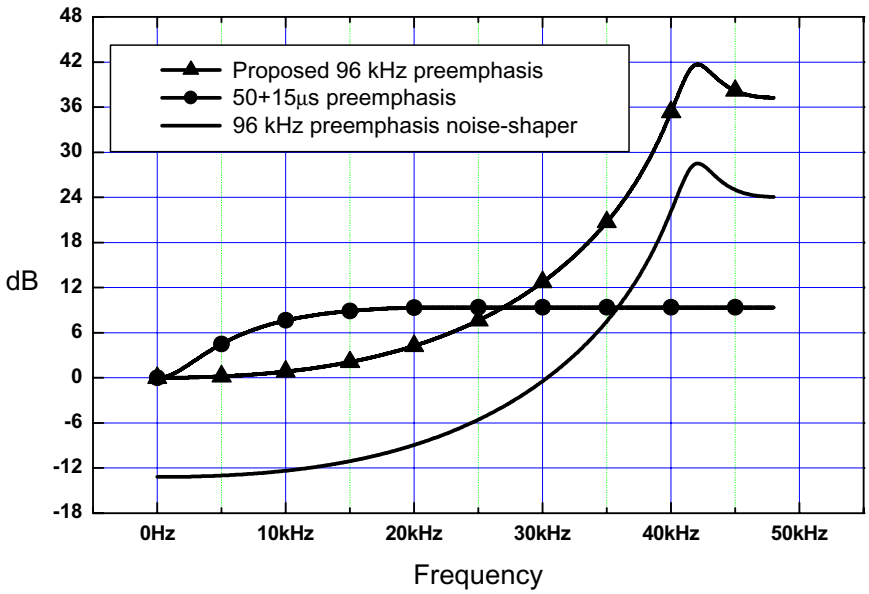

Fig. 32. Gerzon preemphasis scheme compared with 50/15-µs standard for CD, and noise spectrum resulting from preemphasis-matched noise shaper.

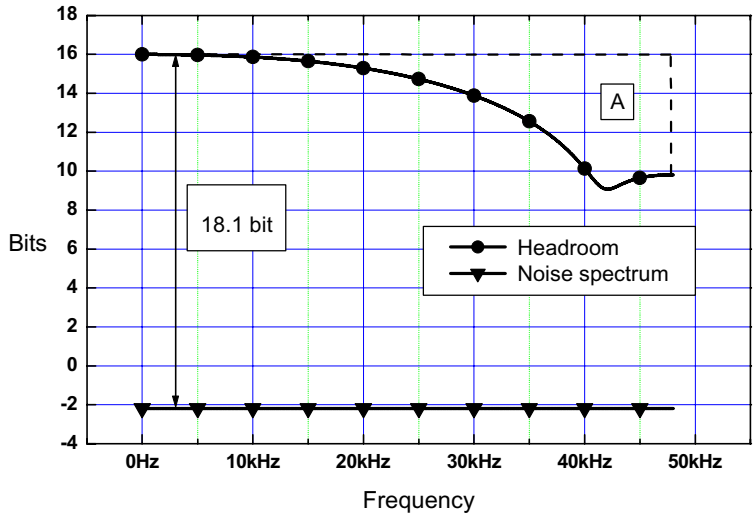

Fig. 33. Output noise spectrum and headroom for channel after application of proposed pre- and deemphasis. Example illustrates a capacity of 18.1 bit at 4 kHz for a 16-bit channel, i.e., a perceptual gain of 2.1 bit.

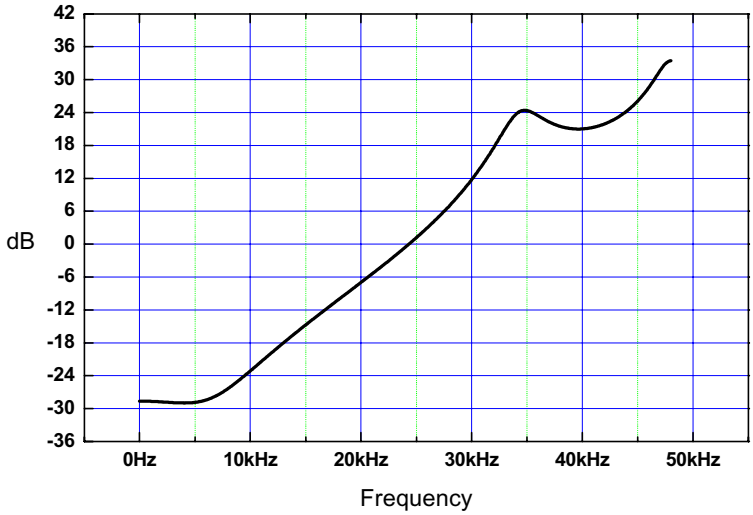

Fig. 34. Example of sixth-order noise shaper that can be combined with preemphasis scheme.

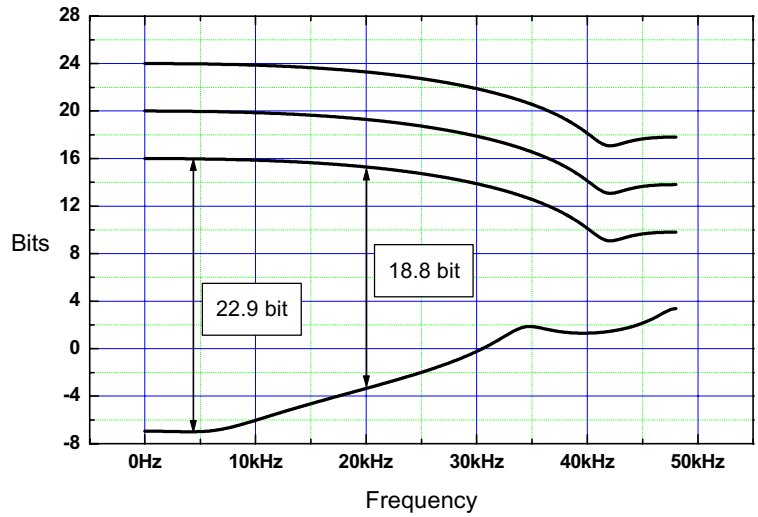

Fig. 35. Output noise spectrum and headroom for channel after the example sixth-order noise shaper has been combined with the proposed pre- and deemphasis. Example illustrates a capacity of almost 23 bit at 4 kHz for a 16-bit channel, i.e., a perceptual gain of 7 bit. Also shown are headroom curves for 20- and 24-bit channels.

7.6 Lossy Encoded PCM

In general, lossy compression schemes operate on the assumption that some input data can be ignored as irrelevant to human listeners, either because they fall below the hearing threshold, or because they will be safely masked by louder, closely cotemporal content.

Such coding does not set out to convey the exact data, or even the waveform, through the channel. Instead it attempts to convey its "sound." These schemes are to some extent successful, and the author freely admits that at some point in the future a lossy psychoacoustically based codec may prove to be audibly transparent. At the moment, however, the use of significant lossy compression in high-resolution systems cannot be advocated.

7.7 One-bit Coding

One-bit coding is a unique case of PCM and belongs to the family of oversampled noise-shaped coders described in Section 7.4. It is discussed in detail elsewhere in this issue by Lipshitz and Vanderkooy [15] and Reefman [49].

The concept of using the output of a highly noise-shaped single-bit quantizer for distribution on a carrier has its origins in the digital audio technology of the early 1990s. It was thought that the output of the then-standard single-bit modulator (operating at 64 times 44.1 kHz) in an analog-to-digital converter might be an appropriate way to avoid the ills of antialias and anti-image filters (discussed in Section 5.4).

As mentioned in Section 2, events rather overtook this idea since high-performance converters evolved to use multibit quantizers in order to avoid the severe problems of lack of linearity, modulation noise, spurious tones, and the high levels of supersonic noise that are unavoidable features of these undithered single-bit quantizers (see [15]).

One attraction of 1-bit 64 times coding is that it has a wide bandwidth and the potential for good transient performance, albeit with very low signal-to-noise ratio at higher frequencies. However, for the same reasons that were shown in the example of the oversampled 8-bit system (see Section 7.4), it is absolutely necessary to use postfiltering on replay to reduce the high-frequency noise from the shaper to an acceptable degree. In consequence the ultrawide bandwidth does not accrue. In fact playback systems tend to require a steep 50-kHz filter so as to not overload downstream equipment, and such a filter is specified for SACD. Real-world implementations of single-bit converters also may not exhibit the ideal transient performance that is expected [50].

There is no standard noise shaper for 1-bit delta–sigma coding. Fig. 36 includes two examples running at 64 times 44.1 kHz taken from [51]. The lower curve (LIP7ZP) is also the example used in Lipshitz and Vanderkooy [15]. The figure shows that the noise floor of either shaper should be inaudible for acoustic gains up to 120 dB SPL. The 1-bit code has a per-channel data rate of 2.822 Mbit/sample and in the region up to 48 kHz, LIP7ZP provides a coding space of approximately 64% of that provided by the 96-kHz 24-bit PCM channel—which uses the lower data rate of 2.304 Mbit/sample. Fig. 37 compares the noise spectrum of LIP7ZP with the 192-kHz 8-bit example from Section 7.4 which runs at 54% of the data rate. Of course the 1-bit system codes audio to above 1 MHz, but with further diminishing dynamic range. Overall this is not the most efficient way of providing the audio coding space we require.

In the dithered multibit case the signal and dithered quantization noise are uncorrelated. However, in the singe-bit case, the noise floor must change with the signal because the total power of the signal plus noise is always constant, that is, modulation noise is unavoidable.

Because the quantizer only offers two levels, TPDF dither cannot be applied. Without dither, the 1-bit system must introduce correlated errors, in the form of birdies and modulation noise. This is an important defect because it sets a limit on the perfectibility of the channel. [51]

Single-bit coding also presents a particular problem if any processing is to be performed on the signal. Typically cascaded processing will include low-pass filtering between stages to reduce the prospect of overloading the modulator, and each stage involves a new quantization step that builds up both correlated errors and supersonic

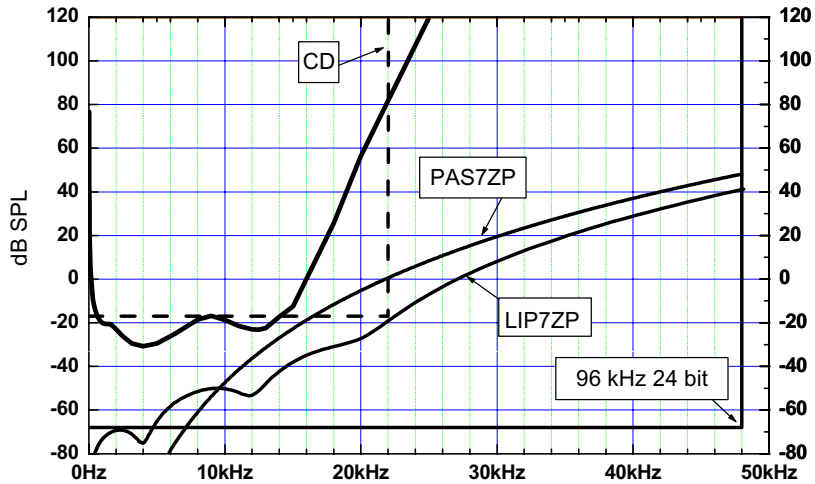

Fig. 36. Noise spectra of two single-bit noise shapers from [51] presented at acoustic gain of 120 dB SPL. Also shown are effective hearing threshold and coding spaces for CD and 96-kHz 24-bit PCM. The common reference that aligns the 96-kHz 24-bit noise floor and the 1-bit noise shaper undithered noise spectrum is the maximum sine-wave power, taken to be 120 dB SPL.

noise. Successive processing risks raising the high-frequency noise of the channel to the point of detectability. For this reason postprocessing (such as bass management) should be carried out in the multibit domain, with sufficient coding space to both contain the signal and prevent further deterioration, as illustrated in Fig. 38.

7.8 Some Comparisons of Channel Coding

Currently studies of high-resolution coding for music delivery tend to be focused on either high-rate multibit PCM or highly oversampled and noise-shaped 1-bit coding. In the multibit case we have illustrated that for sampling rates above 88.2 kHz, the traditional brick-wall antialias and/or anti-image filters are probably a poor choice. Instead we advocate designing the high-frequency rolloff in such a way as to provide a better overall transient performance for the channel.

So far as providing bandwidth is concerned, there is very little difference between the deliverable high-frequency responses of 96-kHz sampled multibit PCM and 2.8224-MHz 1-bit coding; both require to be filtered close to 50 kHz.

In Section 5.3 we see evidence that musical instruments can emit sounds up to 100 kHz and beyond. The only carrier in widespread use that can replicate these sounds is the 192-kHz PCM coding available on DVD-Audio. It is by no means clear that we need to convey audio, on air to the listener, up to 100 kHz to maintain transparency, an assertion also explored in Section 5.3. The author suspects that we are not quite ready to answer the question about whether, when it is well engineered, PCM sampled at 96 kHz offers less transparency than that sampled at 192 kHz. We hope that some useful experiments can be carried out now that apodized filters are becoming available.

Fig. 39 gives an opportunity to compare the noise floor of the recordings described in Section 4.5 with the channel capability of the 1-bit shaper described in the previous section and with 96-kHz 24-bit PCM.

The multibit channel has a dynamic range that exceeds that of the recording by more than 50 dB at all frequencies up to 48 kHz. In Section 4.5 we point out that this precision may be excessive for a carrier since the channel noise is well below the thermal limits for air itself. By contrast, the 1-bit system has a rapidly rising noise floor that, in the case of the highest resolution example, exceeds the background noise of the recording itself above 24 kHz.

Fig. 40 extends the frequency range to 100 kHz and includes data on a closely recorded high piano note from [45]—in this case an envelope of the piano spectrum and the background noise of the recording. Also included in

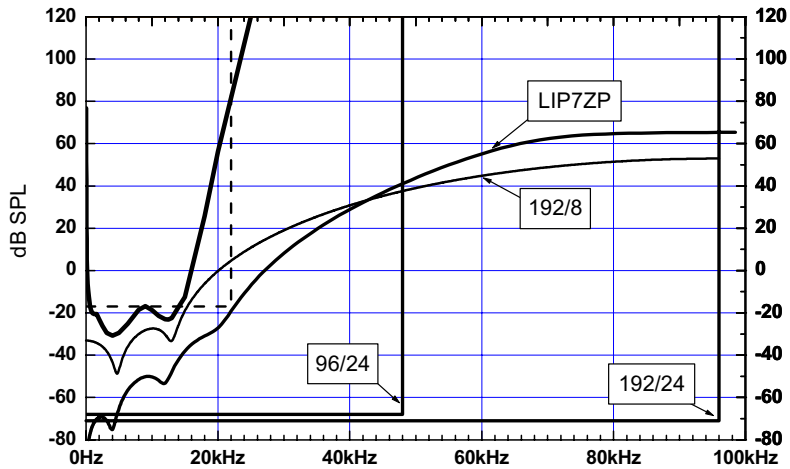

Fig. 37. Comparing two noise shapers: single-bit LIP7ZP and 192-kHz 8-bit example from Section 7.4. Also shown are coding spaces for 96- and 192-kHz 24-bit PCM and uniformly exciting threshold noise.

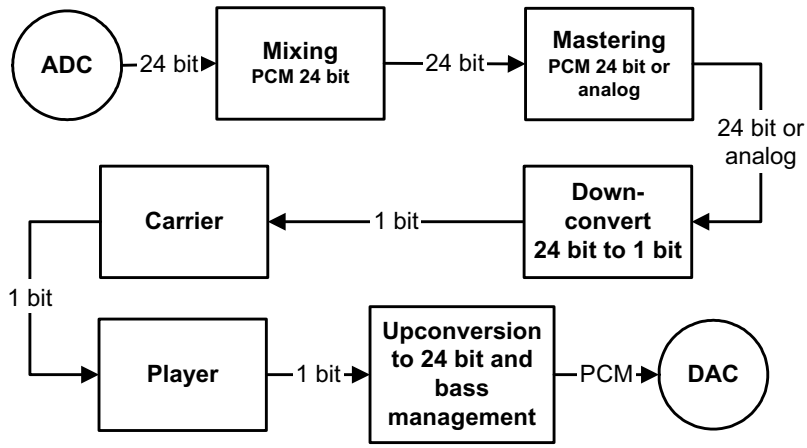

Fig. 38. Recommended work flow for content to be issued in 1-bit form.

the figure is the coding space for 192-kHz PCM.[10, 11] Once again, PCM is more than adequate to contain and replicate the information, regardless of which components may not be audible. However, the noise floor of the single-bit system swamps both the high-frequency components of the signal above 38 kHz and the background of the recording above 28 kHz.

This comparison raises a very important question. If supersonic content is significant to high-resolution audio, why would we cover it up with noise? The guidelines evolved in this paper for high-resolution audio, and for transparency, suggest that the carrier should use coding that provides a space that is larger at all frequencies than the source recording. By this criterion neither of the oversampled systems illustrated in Fig. 37 can be regarded as suitable for the highest resolution work.

Perhaps the clearest way to consider this question, evident from Fig. 40, is that 1-bit coding would be a totally unsuitable choice for a series of recordings that set out to identify the high-frequency content of musical instruments, despite claims for its apparent wide bandwidth. If it is unsuitable for recording analysis then we should also be wary of using it for the highest quality work.

[10]Note that at 192 kHz the quantization noise power is distributed over twice the bandwidth compared with 96 kHz; hence the NSD is 3 dB lower.

[11]In this figure the typical 50-kHz replay filter has not been shown in the 1-bit case. Obviously this playback filter would attenuate both the recording and the channel noise.

8 CONCLUSIONS

This issue gives a unique opportunity to highlight and compare many aspects of the current technology for high-resolution digital audio. This paper has not focused on techniques but instead takes an overview investigating broader questions such as what is high resolution?; can we predict transparency?; how should we choose sampling rates and bit depths?; as well as seeking guidelines for signal processing at different points in the recording and playback chain.

The paper makes an attempt to set the properties of various channel-coding techniques in the context of the properties of human hearing, of musical instruments, and of room and recording noise. Auditory modeling techniques have been employed to enable some useful insights and

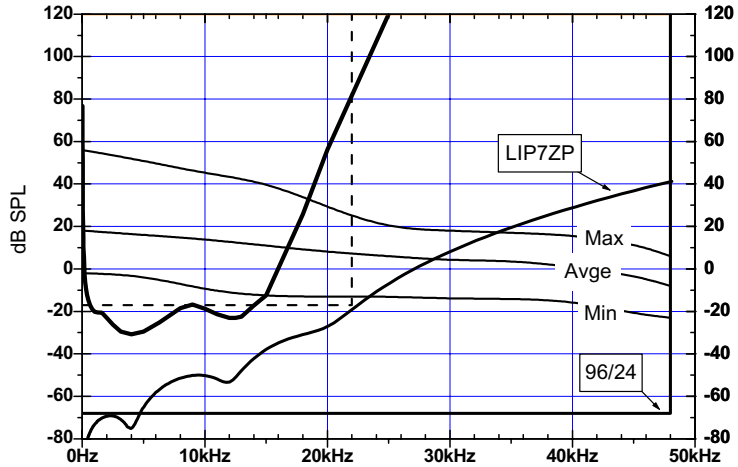

Fig. 39. Noise floor of single-bit coding channel compared to examples of recording noise spectra discussed in Section 4.5.

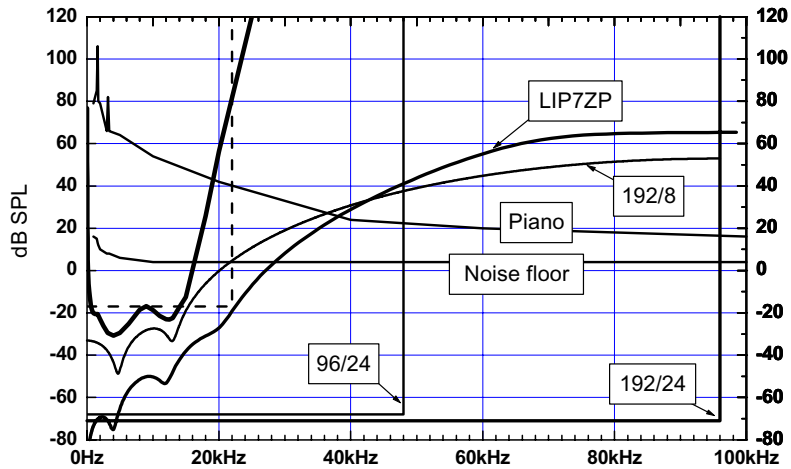

Fig. 40. Envelope of spectrum for a closely recorded high piano note (G-sharp 72) and recording background noise are compared to hearing threshold and noise floors of four different channels: 1 bit 2.4224 MHz, 8 bit 192 kHz (described in Section 7.4), 24 bit 96 kHz, and 24 bit 192 kHz, using data from [45]. Acoustic gain is set to 120 dB SPL and retains level of the recording.

guidelines; these techniques have been used to illuminate the concept of "coding space."

In considering the mastering, mixing, and playback phases it is concluded that these operations should be performed within a coding space that is larger at all audio frequencies (preferably up to at least 48 kHz) than the original signal. To remain below the noise floor of the recording and to ensure that no correlated errors are introduced, this essentially implies a rectangular channel (that is, without noise shaping) using coding such as 96- or 192-kHz sampling with 24 bit.

When recordings are issued on carriers for distribution it is not always necessary or efficient to use the massive coding space afforded by 24-bit PCM. Lossless compression is the best method for reducing space occupied. Other methods such as word-size reduction, various noise-shaping techniques, and also pre- and deemphasis are examined.

For the highest resolution work it is recommended that the noise floor of the coding method used on the carrier be up to 2 bit below the self-noise spectrum of the recording at all frequencies. In fact tools can be designed to automate the process of analyzing the recording and recommending appropriate word size case by case. The requirement that the channel noise be below the self-noise of the signal effectively rules out oversampled low-bit systems for the highest quality work.

For future standards the author commends the study of new possibilities for pre- and deemphasis since this technique can result in a lower, yet flat noise spectrum, trading unused supersonic headroom for in-band dynamic range.

The paper also examines the questions of frequency response and transient performance and suggests that for high-resolution work an apodized filter should be employed in the chain. Since the recording chain may involve sample-rate conversions or addition of material in mixing, it is preferable to use apodized anti-image filters in digital-to-analog converters. For the time being, until such filters become available, it is recommended that they be tried at the mastering stage and flagged in the bit stream. (MLP has this capability.) When such chains are widely available we may be able to reach a firmer conclusion on whether sampling at 192 kHz offers any sonic advantages over 96 kHz.

9 ACKNOWLEDGMENT

The author would like to thank Malcolm Law for providing Figs. 2, 4, and 5; Rhonda Wilson for optimizing the noise shaper used to illustrate Section 7.4, for generating the data for Fig. 34, and for providing Fig. 25; Peter Craven for providing Fig. 24; and Stanley Lipshitz for the data on shapers used in Section 7.7. He would also like to thank Peter Craven for useful comments on earlier drafts and Stanley Lipshitz, John Vanderkooy, and Brian Moore for friendly support over many years of inquiry on this topic.

10 REFERENCES

[1] M. Akune, R. Heddle, and K. Akagiri, "Super Bit Mapping: Psychoacoustically Optimized Digital Recording," presented at the 93rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 40, p. 1044 (1992 Dec.), preprint 3371.

[2] P. G. Craven and M. A. Gerzon, "Compatible Improvement of 16-Bit Systems Using Subtractive Dither," presented at the 93rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 40, p. 1039 (1992 Dec.), preprint 3356.

[3] J. Vanderkooy and S. P. Lipshitz, "Digital Dither: Signal Processing with Resolution Far below the Least Significant Bit," in *Proc. AES 7th Int. Conf. on Audio in Digital Times* (Toronto, Ont., Canada, 1989), pp. 87–96.

[4] M. A. Gerzon and P. G. Craven, "Optimal Noise Shaping and Dither of Digital Signals," presented at the 87th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 37, p. 1072 (1989 Dec.), preprint 2822.

[5] M. A. Gerzon, P. G. Craven, J. R. Stuart, and R. J. Wilson, "Psychoacoustic Noise-Shaped Improvements to CD and Other Linear Digital Media," presented at the 94th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 41, p. 394 (1993 May), preprint 3501.

[6] J. R. Stuart and R. J. Wilson, "Dynamic Range Enhancement Using Noise-Shaped Dither Applied to Signals with and without Preemphasis," presented at the 96th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 42, p. 400 (1994 May), preprint 3871.

[7] J. R. Stuart, "Auditory Modeling Related to the Bit Budget," *Proc. of AES UK Conf. on Managing the Bit Budget* (1994), pp. 167–178.

[8] A. W. J. Oomen, M. E. Groenwegen, R. G. van der Waal, and R. N. J. Veldhuis, "A Variable-Bit-Rate Buried-Data Channel for Compact Disc," *J. Audio Eng. Soc*., vol. 43, pp. 23–28 (1995 Jan./Feb.).

[9] M. A. Gerzon and P. G. Craven, "A High-Rate Buried Data Channel for Audio CD," presented at the 94th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 41, p. 402 (1993 May), preprint 3551.

[10] J. R. Stuart and R. J. Wilson, "A Search for Efficient Dither for DSP Applications," presented at the 94th Convention of the Audio Engineering Society, *J. Audio Eng. Soc*. *(Abstracts)*, vol. 40, p. 431 (1992 May), preprint 3334.

[11] Acoustic Renaissance for Audio, "A Proposal for High-Quality Application of High-Density CD Carriers," private publication (1995 April) www.meridian-audio.com/ara; reprinted in *Stereophile* (1995 Aug.); in Japanese in *J. Japan Audio Soc*., vol. 35 (1995 Oct.).

[12] J. R. Stuart, "Noise: Methods for Estimating Detectability and Threshold," *J. Audio Eng. Soc*., vol. 42, pp. 124–140 (1994 Mar.).

[13] Advanced Digital Audio, "Proposal of Desirable Requirements for the Next Generation's Digital Audio," presented at the Advanced Digital Audio Conf., Japan Audio Society (1996 Apr.).

[14] M. Story, "Audio Analog-to-Digital Converters," *J. Audio Eng. Soc.*, this issue, pp. 145–158.

[15] S. P. Lipshitz and J. Vanderkooy, "Pulse-Code Modulation—An Overview," *J. Audio Eng. Soc.*, this issue, pp. 200–215.

[16] J. R. Stuart, "Predicting the Audibility, Detectability and Loudness of Errors in Audio Systems," presented at the 91st Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 39, p. 1010 (1991 Dec.), preprint 3209.

[17] J. R. Stuart, "Estimating the Significance of Errors in Audio Systems," presented at the 91st Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 39, p. 1011 (1991 Dec.), preprint 3208.

[18] J. R. Stuart, "Psychoacoustic Models for Evaluating Errors in Audio Systems," *Proc. Inst. Acous.*, vol. 13, pt. 7, pp. 11–33 (1991).

[19] L. Fielder, "Dynamic Range Issues in the Modern Digital Audio Environment," in *Proc. of AES UK Conf. Managing the Bit Budget* (1994) pp. 3–19.

[20] D. W. Robinson and R. S. Dadson, "Acoustics—Expression of Physical and Subjective Magnitudes of Sound or Noise in Air," ISO131-1959.

[21] D. W. Robinson and R. S. Dadson, "A Redetermination of the Equal-Loudness Relations for Pure Tones," *Brit. J. Appl. Phys.*, vol. 7, pp. 166–181 (1956 May).

[22] R. S. Dadson, and J. H. King, "A Determination of the Normal Threshold of Hearing and Its Relation to the Standardization of Audiometers," *J. Laryngol. Otol.*, vol. 66, pp. 366–378 (1952).

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

[24] D. J. Meares and K. F. L. Lansdowne, "Revised Background Noise Criteria for Broadcast Studios," BBC Research Rep. RD1980/8 (1980).

[25] G. G. Harris, "Brownian Motion in the Cochlear Partition," *J. Acoust. Soc. Am.*, vol. 44, pp. 176–186 (1968).

[26] P. B. Fellgett, "Thermal Noise Limits of Microphones," *J. IERE*, vol. 57, pp. 161–166 (1987).

[27] B. J. C. Moore, Ed., *Frequency Selectivity in Hearing* (Academic Press, New York, 1986).

[28] S. Buus, et al. "Tuning Curves at High Frequencies and Their Relation to the Absolute Threshold Curve," in B. J. C. Moore and R. D. Patterson, Eds., *Auditory Frequency Selectivity* (Plenum Press, New York, 1986).

[29] M. J. Shailer, B. J. C. Moore, B. R. Glasberg, N. Watson, and S. Harris, "Auditory Filter Shapes at 8 and 10 kHz," *J. Acoust. Soc. Am.*, vol. 88, pp. 141–148 (1990).

[30] M. L. Lenhardt, "Human Ultrasonic Hearing," *Hearing Rev.*, vol. 5, no. 3, pp. 50–52 (1998).

[31] M. L. Lenhardt, R. Skellett, P. Wang, and A. M. Clarke, "Human Ultrasonic Speech Perception," *Science*, vol. 253, pp. 82–85 (1991).

[32] M. L. Lenhardt, "Ultrasonic Hearing in Humans: Applications for Tinnitus Treatment," *Int. Tinnitus J.*, vol. 9, no. 2 (2003).

[33] F. J. Corso, "Bone Conduction Thresholds for Sonic and Ultrasonic Frequencies," *J. Acoust. Soc. Am.*, vol. 35, pp. 1738–1743 (1963).

[34] B. H. Deatherage, L. A. Jeffress, and H. C. Blodgett, "A Note on the Audibility of Intense Ultrasound," *J. Acoust. Soc. Am.*, vol. 26, p. 282 (1954).

[35] R. J. Pumphrey, "Upper Limit of Frequency for Human Hearing," *Nature*, vol. 166, p. 571 (1950).

[36] T. Oohashi, E. Nishina, N. Kawai, Y. Fuwamoto, and H. Imai., "High-Frequency Sound above the Audible Range Affects Brain Electric Activity and Sound Perception," presented at the 91st Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 39, p. 1010 (1991 Oct.), preprint 3207.

[37] T. Oohashi, E. Nishina, Y. Fuwamoto, and N. Kawai, "On the Mechanism of Hypersonic Effect," in *Proc. Int. Computer Music Conf.* (Tokyo, Japan, 1993).

[38] S. Yoshikawa, S. Noge, M. Ohsu, S. Toyama, H. Yanagawa, T. Yamamoto, "Sound-Quality Evaluation of 96-kHz Sampling Digital Audio," presented at the 99th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 43, p. 1095 (1995 Dec.), preprint 4112.

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

[40] J. O. Nordmark, "Binaural Time Discrimination," *J. Acoust. Soc. Am.*, vol. 35, pp. 870–880 (1976).

[41] B. G. Henning, "Detectability of Interaural Delay in High-Frequency Complex Waveforms," *J. Acoust. Soc. Am.*, vol. 55, pp. 84–90 (1974).

[42] R. G. Klump and H. R. Eady, "Some Measurements of Interaural Time Difference Thresholds," *J. Acoust. Soc. Am.*, vol. 28, pp. 859–860 (1956).

[43] K. Krumbholz and R. D. Patterson, "Microsecond Temporal Resolution in Monaural Hearing without Spectal Cues?," *J. Acoust. Soc. Am.*, vol. 113, pp. 2790–2800 (2003).

[44] M. A. Gerzon, P. G. Craven, J. R. Stuart, M. J. Law, R. J. Wilson, "The MLP Lossless Compression System for PCM Audio," *J. Audio Eng. Soc.*, this issue, pp. 243–260.

[45] J. Boyk, "There's Life above 20 kilohertz! A Survey of Musical Instrument Spectra to 102.4 kHz," private publication www.cco.caltech.edu/~boyk/spectra/spectra.htm (2000).

[46] P. G. Craven, "Antialias Filters and System Transient Response at High Sample Rates," *J. Audio Eng. Soc.*, this issue, pp. 216–242.

[47] J. R. Stuart and R. J. Wilson, "Dynamic Range Enhancement Using Noise-Shaped Dither at 44.1, 48, and 96 kHz," presented at the 100th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 44, p. 646 (1996 July/Aug.), preprint 4236.

[48] Acoustic Renaissance for Audio, "DVD: Preemphasis for Use at 96 kHz or 88.2 kHz," private publication www.meridian-audio.com/ara (1996 Nov.).

[49] D. Reefman and E. Janssen, "One-Bit Audio: An Overview," *J. Audio Eng. Soc.*, this issue, pp. 166–189.

[50] C. Anderson, "Poking a Round Hole in a Square

Wave," /www.smr-home-theatre.org/surround2002/technology/page_07.shtml.

[51] S. P. Lipshitz, J. Vanderkooy, "Why 1-Bit Sigma-Delta Conversion is Unsuitable for High-Quality Applications," presented at the 110th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 49, p. 544 (2001 June), convention paper 5395.

THE AUTHOR

J. Robert Stuart was born in Belfast, Northern Ireland in 1948. He studied electronic engineering at the University of Birmingham, UK, where he was awarded First Class Honours. While at Birmingham he studied psychoacoustics under Professor Jack Allinson, which began a lifelong fascination with the subject.

After a year working for the Marconi Instrument Company, he received an M.Sc. degree in operations research from the Imperial College, London, in 1971. Following three years as a consultant in the audio industry, he co-founded Boothroyd Stuart Limited, which began manufacture of the Meridian brand of audio components. He is now chairman and technical director of the Meridian Group, which includes Meridian Audio Limited, Meridian America Incorporated, and MLP Limited.

Mr Stuart's professional interests are the furthering of analog and digital audio and developing understanding of the human auditory perception mechanisms that are relevant to live and recorded music. His specialities include the design of analog and digital electronics, loudspeakers, and optical disc players.

As an active member of the DVD Forum, Mr. Stuart has contributed to the DVD-Audio and DVD-Audio Recordable standards. He has also served on the technical committee of the National Sound Archive. He has a deep interest in music and spends a good deal of time listening to live and recorded material. He is a fellow of AES, a member of the ASA and IEEE, a visiting fellow at the Essex University, and the chairman of the Acoustic Renaissance for Audio.