A High-Rate Buried-Data Channel for Audio CD*

MICHAEL A. GERZON, *AES Fellow*

XtraBits, Oxford OX4 1XX, UK

AND

PETER G. CRAVEN, *AES Member*

XtraBits, East Challow, Wantage, Oxon OX12 9SG, UK

Subtractive buried data is a new proposal for conveying a high-data-rate data channel (with up to 350 kbit/s or more) compatibly within the data stream of an audio CD without significant impairment of existing CD performance. This proposal uses pseudorandomized data as noise-shaped subtractive dither for the conventional audio. The new data channel may be used for high-quality data-reduced related audio channels, or even for data-compressed video or computer data, while retaining compatibility with existing audio CD players.

0 INTRODUCTION

This paper describes a new proposal for burying a high-data-rate data channel compatibly within the data stream of an audio CD. The maximum rate that can be buried without significant impairment of existing CD performance is on the order of 220–350 kbit/s, or even more (over 500 kbit/s) if variable data rate techniques are used. The subtractive buried data proposal in this paper replaces a number of the least significant bits (LSBs) of the audio words (typically up to four per channel) by other data and uses the psychoacoustic noise-shaping techniques associated with noise-shaped subtractive dither to reduce the audibility of the resulting added noise down to a subjective perceived level equal to that of conventional CD.

Simply replacing the LSBs of existing audio data would, of course, cause a drastic audible modification of the existing audio signal for two reasons:

1) The word length of existing signals would be truncated to, say, only 12 bit, which would not only reduce the basic quantization resolution by 24 dB but it would also introduce the problems of added distortion and modulation noise caused by truncation (for example, see [1]–[4]).

2) In addition the replaced last, say, 4 LSBs would themselves constitute an added noise signal, which itself may not have a perceptually desirable random-noise-like quality and will also add to the perceived noise level in the main audio signal, typically increasing the noise by a further 3 dB above that due to truncation alone, giving in this case as much as 27-dB total degradation in noise performance.

This paper describes the following methods of overcoming all these problems in replacing the last few LSBs of an audio signal by other data.

1) Using a pseudorandom encode–decode process, operating only on the LSB data stream itself without extra synchronizing signals, to make the added LSB data effectively of random noise form so that the added signal becomes truly noise-like.

2) Using this pseudorandom data signal as a subtractive dither signal (for example, see [1]–[4]), so that simultaneously it does not add to the perceived noise and it removes all nonlinear distortion and modulation noise effects caused by truncation. This step is the essence of the subtractive buried data process. Remarkably, and unlike in the ordinary subtractive dither case [3], a special subtractive dither decoder is not needed, so that the process works on a standard off-the-shelf CD player.

3) Furthermore, at the encoding stage, incorporating psychoacoustically optimized noise shaping of the (subtractive) truncation error, thereby reducing the perceived

* Presented at the 94th Convention of the Audio Engineering Society, Berlin, Germany, 1993 March 16–19; revised 1994 November 2.

truncation noise error further by between 9 and 17 dB, depending on the psychoacoustic tradeoffs chosen.

Subtractive buried data using methods 1) and 2) can be applied with or without the noise shaping of method 3).

The overall effect of combining these three processes is that if one incorporates data into the last few LSBs, then the effects of distortion, modulation noise, and perceived audible patterns in the LSB data are completely removed, and the resulting perceived steady noise is reduced by around 15–23 dB below that of ordinary unshaped optimally dithered quantization to the same number of bits.

For example, using the most extreme noise-shaping strategy, when the last 4 LSBs of the 16-bit CD word length is used for buried channel data, the perceived signal-to-noise ratio is around 91 dB—approximately the same as ordinary 16-bit CD quality when unshaped dither is used. With a more moderate noise-shaping strategy, the last 2½ LSBs of the 16-bit CD word length can be used for buried channel data, without degrading the perceptual quality of the 16-bit CD medium.

The result of this process is that as much as $2 \times 4 = 8$ bit of data per stereo sample is available for buried data without significant loss of audio quality on CD, giving a data rate of $8 \times 44.1 = 352.8$ kbit/s. With more moderate noise-shaping strategies, $2 \times 2\frac{1}{2} = 5$ bit of data per stereo sample is available for buried data without significant loss of audio quality on CD, giving a data rate of $5 \times 44.1 = 220.5$ kbit/s.

The Appendix provides a detailed discussion of the tradeoffs in data rate versus perceptual quality for different data rates of buried data and for various options in noise shaping and choice of preemphasis for the CD medium. The main body of the paper is concerned with the technical method of implementing the buried data channel.

While the subtractive buried data process achieves potentially high data rates for the buried channel, it does of course reduce the room for improvements in CD audio quality, approaching 20-bit effective audio quality, as described in [3], [4]. However, there is no reason why the process should only be used with one fixed number of LSBs. By reducing the data rate of the buried channel to a smaller number of LSBs, one correspondingly improves the resolution of the audio—for example, achieving an effective perceived signal-to-noise ratio of around 103 dB for a system using 2 LSBs of data per signal channel sample, with a data rate still of 176.4 kbit/s.

One can even make the number of LSBs used fractional, say ¼, ½ or 1½ LSBs per sample. This may be used either to match the buried channel to a desired data rate precisely, or to minimize the loss of audio quality, especially at very low data rates.

In addition by including in the LSB data channel itself low-rate data indicating the number of LSBs "stolen" from the main audio channels, it is possible to vary the number of LSBs stolen in a time-variant way, so that, for example, more LSBs can be taken by the buried channel when the resulting error is masked by a high-level main audio signal. The noise shaping can also be varied adaptively at the encoding stage so that at high audio levels, the noise error is maximally masked by the audio signal. These ideas have been further explored by Oomen et al. [25],[1] who quote an average bit rate of 500 kbit/s, increasing to nearly 800 kbit/s in loud passages.

A variable-data-rate approach to transmitting data in an audio waveform, for use with the NICAM system, has also been described by Emmett [26]. Here the shape of the error spectrum is adaptively changed to be masked by the audio signal. This may or may not have some common features with the present proposal, as the details of Emmett's proposal are not clear from his published preprint.

It is also shown in this paper that with stereo signals it is possible to code data jointly in the least significant parts of the audio words of the two (or more) channels, using a multichannel version of the data-encoding process, involving the use of vector quantizers and subtractive vector dithering by a multichannel pseudorandom data signal for the dithering. The basic theory of vector dithering is described in Section 5, although readers may find it best to omit these technically difficult aspects on first reading. It is shown that the vector multichannel version of the data-coding process ensures left–right symmetry of any added noise in the audio reproduction and an advantageous noise performance.

The approach described in this paper is substantially different from an alternative method of burying data described in [7], which involved a process of splitting the audio signal into subbands, replacing the LSBs of the subbands with data based on auditory masking theory, and then reassembling the resulting data by recombining the subbands. Not only is that process very complicated, with a considerable time-delay penalty in the subband encoding–decoding process, but it has to be done with extraordinary precision to prevent data errors in the band splitting and recombining process. By contrast, the present process involves little time delay, involves relatively simple signal processing, and further is such as to guarantee the lack of audible side effects due to nonlinear distortion, modulation noise, or data-related audible patterns.

1 USES OF BURIED DATA

1.1 Advantages over CD ROM Media

The availability of a buried data channel with data rates on the order of 350 kbit/s without significant loss of audio quality on audio CDs, fully compatible with conventional playback on standard audio players, opens up prospects for many new products. Unlike standards such as CD-I based on CD-ROM, the additional data can be added without destroying compatibility with playback over tens of millions of existing audio players. This

[1] Oomen et al. [5] only consider subtractive buried data channels stealing an integer number of bits from each of the stereo channels. If stereo parity buried data methods are used, as described in Sections 2.2 and 6.3, the available data rate is typically further increased by around 22 kbit/s.

means that the new data channel can be added while still giving the CD the advantages of mass-market economies of scale of production, thanks to the existing audio-only market. Thus applications using the new data channel should result in much lower prices than for media where the number of players is limited.

1.2 Application to Multichannel Sound

One application of the new data channel is using the additional bits to add, by means of audio data compression, additional audio channels for three- or more loudspeaker frontal stereo or surround sound, as described, for example, in [8]–[10]. Because CDs have higher quality than available data-compression systems (despite claims of "transparency" or "CD quality" by some less cautious proponents of such systems), care must be taken that the additional channels are not too compromised in quality by the data-compression process, which means that a rather lesser degree of compression is desirable than for DAB or film surround sound. However, since two of the transmitted audio channels are the standard CD audio channels and the design of the buried channel avoids nonlinear or modulation noise effects on these main channels, all the data rate in the buried channel can be used solely for the additional channels, giving each a higher data rate than if the buried channel were used to transmit the whole audio signal. In using the buried channel to transmit additional directional audio channels, it is important to design the codec error signals so that they do not become audible through the mechanism of directional unmasking described in three of one of the authors' references [11]–[13].

The data rate available using the most extreme noise-shaping strategies is sufficient to transmit a Dolby AC-3 or MUSICAM surround five-channel surround-sound signal, but these systems involve a quality compromise with the data rate so that this is not a preferred procedure. Such systems are preferably used in a manner such that the main two channels conveying a stereo-compatible mix are conveyed as standard CD audio, with only the three or more supplementary channels in data-compressed form.

High-quality data-compressed additional audio channels can, unlike existing data-compression systems, minimize the risk of destruction of subtle auditory cues such as those for perceived distance (see [14]), thereby maintaining CD digital audio as the preferred medium for high-quality audio while adding additional channels. For high-quality (and especially musical) use it may be preferred to use additional buried audio channels either for frontal-stage three- or four-loudspeaker stereo or for three-channel horizontal or four-channel full-sphere with-height [15], [16] ambisonic surround sound (see [9], [10], [17], rather than for the rather cruder theatrical "surround-sound" effects considered appropriate for cinema or video-related surround-sound systems. However, systems have been proposed for intercompatible use of both kinds of systems [9], [10].

Since the main audio channels in this proposal convey high-quality audio, it is possible to use the spectral envelope of the main audio channels to convey most or all of the dynamic ranging information used for the subbands in data-reduction systems for related subsidiary channels conveyed in the buried data channel, especially if the main audio channels incorporate a mixture of all the transmitted channels so that no direction is canceled out. This saves the data overhead of conveying ranging data, which in high-quality systems may save on the order of 60 kbit/s as compared to a stand-alone data-compression system. This will allow a system conveying n related channels using 4 LSB per main CD audio channel to give a performance equivalent to that of a stand-alone data-compression system conveying $n - 2$ channels in about 410 kbit/s. For three-channel systems, such as horizontal B-format surround-sound or three-channel UHJ [17] or frontal-stage three-channel stereo, this quality is unlikely to be audibly distinguishable from an uncompressed data channel. For four-channel systems, the results will still subjectively approach that of critical studio-quality material, and even for five-channel material, the results will be considerably less compromised than that for DAB or cinema surround sound, using a data rate for the additional channels of well over twice that used in those applications.

1.3 Video and Computer Data

Alternatively, the buried data channel can be used for conveying related computer data, such as graphics, multilingual text, or track copyright information. Because of the high available data rate, this can be done with very much higher quality than is possible on the subcode channels of CD, conveying, for example, with JPEG image data compression on the order of one high-quality color photographic image per second. A data rate of 350 kbit/s is even enough to convey a reasonable video image. Using the existing MPEG standard, this would have very low resolution (although certainly good enough for moving inserts within a still image), but near-future image data-compression methods based on using the highly non-Gaussian nature of images are expected to make consumer-quality video available within this data rate.

1.4 Dynamic-Range Data

Another use would be to convey dynamic-range reduction or enhancement data, such as a channel conveying the setting of a gain moment by moment. This would allow the same CD to be played automatically with different degrees of dynamic compression according to the environment by choosing the gain adjustment channel appropriate for that environment. This would include the possibility of completely uncompressed quality for high-quality use, without making the CD incompatible for more normal use, such as in broadcasting. An advantage of providing the dynamic-range gain data in the data subchannel rather than using automated dynamic-range modification algorithms is that one can always do a much better subjective job using manual intervention based on a knowledge of the music and its needs, but at the expense only of considerable time and effort. This

effort can be recorded for consumer use in the buried data channel. If automated algorithms are used for the dynamic-range gain conveyed by the buried data channel, these can be of a much more sophisticated and subtle nature than those normally available to the consumer (for example, [18]).

1.5 Frequency Range Extension

A further use related to the original audio would be to add in the subchannel data-reduced information allowing information above 20 kHz to be reconstructed. (See, for example, Komamura [19], who uses buried data for this purpose, but not subtractive buried data.) One of the limitations of CDs is that the frequency range is limited to 20 kHz. Although the ears' sine-wave hearing is for all, except a small minority of (generally young and often female and/or asthmatic) listeners, limited to below 20 kHz, this does not mean that there is no loss of perceived quality caused by the sharp band-limiting to 20 kHz. It is widely noted that there is a significant loss of perceived quality when comparing high-quality digital signals sampled at, say, 44.1 kHz as compared to 88.2 kHz.

From a quality viewpoint it may be more important to use an extended bandwidth to provide a more gentle rolloff rate than to provide a response flat to 40 kHz since, unlike the brickwall filters used with ordinary CD, such gentler rolloffs are similar to those encountered in natural acoustical situations.

The extended bandwidth can be provided by using a high-order complementary mirror filter pair of the kind described in Regalia et al. [20] and in Crochiere and Rabiner [21] to split an 88.2-kHz-rate sampled digital signal into two bands sampled at 44.1 kHz. The filters involved will overlap, although using a high-order filter [20], the region of significant overlap can be reduced to about 1 kHz. Within the overlap region there will be aliasing from the other frequency range, although the reconstruction of the full bandwidth [20], [21] will cancel out this aliasing. The band below 22.05 kHz can then be transmitted as the conventional audio, and the band above 22.05 kHz can be transmitted in data-reduced form in the buried data channel at a reduced data rate of, say, between 1 and 4 bit per sample per channel, using known subband or predictive coding methods. Phase compensation inverse to the phase response of the low-pass filter in the complementary filter pair may be employed to linearize the phase response of the main sub- 22.05-kHz signal for improved results for standard listeners, with the use of an inverse phase-compensating filter in the decoding process of reconstructing the wider bandwidth signal.

1.6 Airplay Mixes

The buried data channels on a CD can be used to convey in data-reduced form alternative mixes of the musical material in the main track. For example, the buried-data audio channel might be an "airplay mix" designed for optimum effect when heard over AM or FM radios. At present such airplay mixes have to be distributed separately for promotional purposes, whereas buried data allow these versions to be distributed within the standard CD release.

1.7 Remixable CD

One potentially important use for buried data in audio CDs is for remixable CDs, where the end user has the option of changing the mix from that given by standard audio playback. This may be done by using the buried data to convey in data-reduced form additional audio signals, representing differences between the main-channel mix and alternative mixes.

For example, in library music applications, where musical material suitable for use as backing to radio, TV, audiovisual, and multimedia productions is provided on CD, the main stereo audio can be used for a "standard mix" of three stereo components,

$$m_1M + h_1H + r_1R$$

say a melody line M, a harmony line H, and a rhythm line R. The buried data channels can be used to convey alternative mixes $m_2M + h_2H + r_2R$ and $m_3M + h_3H + r_3R$, where some of the mixing coefficients may be zero or negative. Then a new mix can be derived by recovering M, H, and R separately by inverse matrixing and then mixing them together using a conventional user-adjustable mixing process.

Besides use for library music applications, the remixable CD can also be used in applications where hearing-impaired listeners, who form a significant proportion of the public, can raise the level of vocal lines for enhanced intelligibility. Further consumer applications include allowing consumers to prepare their own mixes, removal of "spot" microphones in classical music recordings, and multilingual recordings where the vocals can be provided in several languages. A buried audio channel can also be used to add or subtract vocals in music for Karaoke applications.

1.8 MIDI Applications

Another musical application would be to convey MIDI control data in the buried data channel. This can be used to control additional musical lines from MIDI modules as part of an overall musical mix that may be user adjustable. Each MIDI channel requires a data rate of 31.25 kbit/s, for example, allowing four MIDI channels to be conveyed by culling 1½ bit from each of the two stereo channels on the CD.

1.9 Combined Applications

Any or all of these uses can, of course, be combined, subject only to the restrictions of the data rate, so that the buried data channel could be used, for example, to convey one additional audio channel, a dynamic-range gain signal, extended bandwidth, and additional graphics, text (possibly in several languages), copyright, and even insert video data, as appropriate.

A specific audiophile example is the possibility of using the extra data to convey three-channel frontal-

stage stereo or three-channel ambisonic surround sound where all channels have extended bandwidth. This involves conveying one extra channel of audio in data-compressed form plus three channels of extended bandwidth data.

For historical material, where the dynamic range may be significantly less than 90 dB, it may even be possible to increase the data rate available further by allocating even more bits to the buried data channel since an increased noise level may not be significant. For this reason, it may be desirable to allow the possibility of allocating as many as 12 or even 16 bits of audio data (say, bits 10 to 15 or even 8 to 15 of each audio channel) to the buried data channel.

2 PSEUDORANDOM CODING OF DATA

2.1 Pseudorandomized Data

It is essential, if the LSBs of an audio signal are to be replaced by data, that the replacing data should truly resemble a random noise signal (albeit perhaps one that may be spectrally shaped for psychoacoustic reasons). Most data signals, when listened to as though they were digital audio signals, have some degree of systematic pattern which may well prevent them from sounding or behaving truly like random noise. Such departures from random noise-like behavior are generally much more perceptually disturbing or distracting than a simple steady noise.

Also, if we can ensure that added data behave like a noise signal with known statistical properties, one can use all that is known in the literature on dither and noise shaping (see [1]–[4], [22]–[25]) to optimize the perceptual properties of the added data to minimize their audible effects.

The data signal is rendered pseudorandom with predictable statistics in our proposal by a data encode–decode process, the encode process having the effect of pseudorandomizing the data signal, and the decode process having the effect of recovering the original data signal from the pseudorandomized data signal, as illustrated in Fig. 1. From a practical point of view it is highly desirable that the encode and decode process require no use of an external synchronizing signal, but that the decode process should work entirely from the pseudorandomized data sequence itself.

The simplest way of constructing such an encode–decode pseudorandomizing process for data is to use a cyclic pseudorandom logic sequence generator separately on each bit, as was realized in 1967 by Savage [26] and implemented in a commercial data-scrambling product by Hewlett-Packard in 1971 [27]. For example, if its input is zero, Fig. 2(a) shows a well-known binary pseudorandom logic sequence generator using feedback around three logic elements and a total shift register delay of 16 samples. (A one-sample delay is denoted by the usual notation z^{-1}.) Provided that the logic state in the 16 samples stored in the shift register is not all zero, this binary sequence generator has the 16 logic states cycle through all $2^{16} - 1 = 65\,535$ nonzero states in a pseudorandom manner.

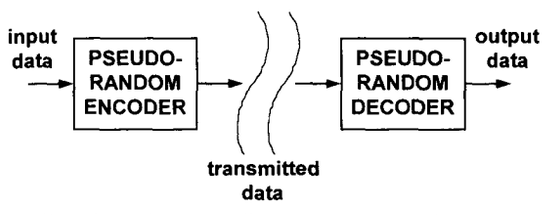


Fig. 1. Pseudorandom encoding and decoding of data transmitted via CD channel to ensure noise-like behavior.

If, instead of using a zero input, the pseudorandom sequence generator of Fig. 2 is fed with a binary data stream s_n, then it has the effect of a pseudorandomizer for the input data. This encoding scheme is based on the recursive logic

$$t_n = s_n \oplus t_{n-1} \oplus t_{n-3} \oplus t_{n-14} \oplus t_{n-16} \quad (1)$$

where t_n is the output binary logic value of the network at integer sample time n, s_n is the input binary logic value of the network at integer sample time n, and $\oplus$ represents the logic EXCLUSIVE-OR or Boolean addition operator (with truth table $0 \oplus 0 = 1 \oplus 1 = 0$, $0 \oplus 1 = 1 \oplus 0 = 1$).

Conversely, if exactly the same arrangement of logic gates is fed with the pseudorandomized data t_n, then the effect of the EXCLUSIVE-OR gates on the input signal is to restore the original data stream. This is achieved by the inverse decoding logic process

$$s_n = t_n \oplus t_{n-1} \oplus t_{n-3} \oplus t_{n-14} \oplus t_{n-16} \quad (2)$$

illustrated in Fig. 2(b).

Thus by using a logic network recursively with delay of a total of $L = 16$ samples and only four EXCLUSIVE-OR gates, a binary data stream can be pseudorandomized, and the same network can decode the data stream back to its original form. For constant signals there is a one in 65 536 chance that the undesirable nonrandom zero state will be encountered, but this low probability is probably acceptable, given that even a single binary digit change of the input is likely to "jog" the system back

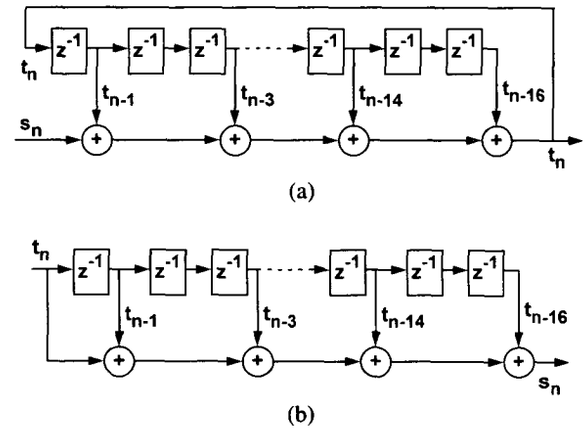


Fig. 2. Binary pseudorandom sequence generator using shift-register logic, with input EXCLUSIVE-OR gate for encoding and decoding of binary data stream.

into a pseudorandom output state.

Other well-known pseudorandom binary sequence logic generators with shift registers of length L longer than 16 samples can be used for encoding and decoding in the same way, with their fed back output given by subjecting the delayed sequence output and the input to as SUM logic gate. Such length L sequences will have, for a constant input, only one chance in $2^L - 1$ of giving an unrandomized output, and will have a sequence length of $2^L - 1$ samples.

Although the pseudorandom binary sequence generator described in Eq. (1) and Fig. 2 is a maximum-length sequence for a zero input, it has a shorter length for an all-1 constant input, and in general, the precise behavior with, say, periodic inputs is hard to predict. Partly for this reason it is not absolutely essential to use a maximum-length sequence generator, provided that the length of the sequence is not too short for constant inputs.

It will be noted that the network of Fig. 2 only has $L = 16$ samples of memory, so that when used as a decoder, any data errors in the input will only propagate for L samples, and then the output will recover. This lack of long-term memory in the decoding process means that there are no special requirements on the error rate of the transmission channel. Because of the small number of logic elements in Fig. 2, a single sample error in the received data stream will only cause five sample errors in the decoded output.

As shown in Fig. 3, typically, for use with CDs, the data will first be arranged to form a number of bits of data per sample of each audio channel, for example, 8 bit of data constituting bits 12 to 15 of the left and right audio channels [where bit 0 is the most significant bit (MSB) of a 16-bit audio word and bit 15 the LSB].

Then each of these (say 8) bits will, separately, be encoded by a pseudorandom logic such as that of Fig. 2 to form a pseudorandom sequence, and the resulting pseudorandomized bits will be used to replace the original bits in, say, bits 12 to 15 of the left and right audio channels. The resulting noise signals in the left and right audio channels will be termed the (left and right) data noise signals.

Alternatively, instead of pseudorandomizing individual bits of the audio words representing data separately, they can be pseudorandomized jointly by regarding the successive data bits of a word as being ordered sequentially in time, and applying a pseudorandom encoder such as that in Fig. 2 to this sequence of bits. For example, 8 bit of data per audio sample can be ordered sequentially before the next 8 bit of data corresponding to the next audio sample, and the pseudorandom logic encoding can be applied to this time series of bits at eight times the audio sampling rate.

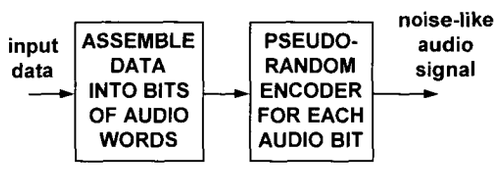


Fig. 3. Schematic of processing of data to form audio noise-like signal.

An advantage of this strategy is that errors in received audio samples propagate for (in this example) only one-eighth of the time, as in the case where each word bit is separately pseudorandomized.

M-level data signals, taking one of M possible values, conveying $\log_2 M$ bits per sample, can also be pseudorandomized by a direct process involving congruence techniques, whereby the coded version w'_n of the current sample M-level word w_n is given by

$$w'_n = w_n + \sum_{j=1}^{L} a_j w'_{n-j} \,(\mathrm{mod}\, M) \tag{3}$$

where the a_js are (modulo M) integer coefficients chosen (if necessary by empirical trial and error) to ensure that all M possible constant inputs result in a pseudorandomized output with reasonably long sequence lengths. The inverse decoding of the pseudorandomized M-level words is

$$w_n = w'_n - \sum_{j=1}^{L} a_j w'_{n-j} \,(\mathrm{mod}\, M)\,. \tag{4}$$

The logic techniques described with reference to Fig. 2 are just the special case when $M = 2$ of this more general congruence technique. The congruence technique can result in sequence lengths for constant inputs of length up to a maximum of $M_L - 1$ samples, so that in general the larger the value of M, the smaller L need be with a consequent shortening of the time duration of propagation errors.

A slightly more complex pseudorandomization of data will provide an initial pseudorandomization of M-level data by a method such as one of those described here, and follow it by an additional one-to-one map between the M possible data values. The decoding will first subject the M levels to an inverse map before applying the inverse of the above pseudorandom encodings.

There are many similar but more complicated methods of pseudorandomization of data streams. As we have seen these need have no coding delay or increase in data rate after coding, and they can limit the duration of any errors in received data in the inversely decoded output to not more than a few samples after the occurrence of an erroneous audio sample.

As audio signals, the resulting pseudorandomized data noise signals have a steady white-noise spectrum and a (discrete) uniform or rectangular probability distribution function (PDF) in the example case described, having 16 levels in each of the left and right channels. Such discrete noise does not have the ideal properties of rectangular dither noise, although Wannamaker et al. [22] have shown that it approximates many of these desirable properties in a precise mathematical sense. However, adding to it an extra random or pseudorandom white rectangular PDF noise signal with peak level $\pm 1/2$ LSB converts it into noise with a true rectangular PDF, with

peak levels (in this example) of ±8 LSBs. In this case the added noise to convert from a discrete to a continuous PDF is at a very low level, being 24 dB below the level of the data noise signal.

2.2 Stereo Parity Coding

Although in the preceding example we described data being conveyed separately on each audio word bit of the data signal, it will be realized that data can alternatively be conveyed by more complicated combinations of the LSBs of audio words (in any numerical base M, not just the binary base 2)—for example, on the Boolean sum of the corresponding bit in the left and right audio signals.

For example, consider the case where a data rate of only 1 bit per stereo audio sample is required. Such a signal can be conveyed as the Boolean sum of the LSB in the left and right audio channels, leaving the values of the LSB in individual audio channels separately unconstrained. Conveying a data channel using the Boolean sum of the corresponding bits of the left and right audio signals is herein termed stereo parity coding.

It is of course desirable that the effect on the conventional audio of reallocating bits to a buried data channel should be left–right symmetrical. In particular, if a buried data channel is used with a data rate of just 1 bit per stereo sample (BPSS), then one does not wish to code the data in the LSBs of only one of the two stereo channels. If the values of the respective Nth bits of the respective left and right channel signals are denoted by L_n^N and Rr_n^N at time n, then one codes a pseudorandomized 1-bit per sample data channel t_n^N as

$$t_n^N = L_n^N \oplus R_n^N . \tag{5}$$

This encoding can be accomplished by flipping, if necessary, the parity of the Nth bit in either of the two stereo channels to ensure the desired value of t_n^N. The added error noise is minimized by flipping that channel whose quantization was closer to a decision threshold, as described more fully in Section 6.3.

If desired, an additional second pseudorandomized 1-bit per sample data channel u_n^N can be encoded in the Nth bit of the stereo audio signal, say, as

$$u_n^N = L_n^N \tag{6}$$

in which case the data can be encoded via $L_n^N = u_n^N$, $R_n^N = L_n^N \oplus t_n^N$ and decoded via $u_n^N = L_n^N$, $t_n^N = L_n^N \oplus R_n^N$. Alternatively u_n^N can be encoded as R_n^N. The use of stereo parity encoding allows the separate 1-BPSS data channels to be decoded separately while maintaining left–right symmetry in the audio when an odd number of 1-BPSS channels is used.

One could standardize a basic 1-BPSS data channel as being conveyed via the parity (Boolean sum) of the LSBs (that is, bit 15) of the left and right audio channels. Information about the way other data channels conveying more BPSS are coded will, in such a standardization, be conveyed by this basic data channel. By this means, a data decoder can read from the basic 1-BPSS stereo parity data channel how to decode other data channels present, if any. In particular, this allows, if desired, moment-by-moment variation of the data rate, either adaptively to the amount of data needing transmission or adaptively to the audio signal according to its varying ability to mask the error signal caused by the hidden data channels.

For example, in loud passages in pop or rock music, the data rate allocated to, say, a video signal could be increased, allowing quite high-quality video images in, say, heavy metal music.

2.3 Fractional Bit Rates

There is no reason why the buried data channels should be restricted to data rates of an integer number of BPSS, although this may be a convenient implementation. Several methods can be used to allocate less significant parts of audio words to data at fractional bit rates.

One method conveys $\log_2 M$ bits for integer M in the less significant parts of audio words by conveying data in the M possible values of the remainder of the integer audio word after division by M, whereas the rounding quantization process used for the audio involves rounding to the nearest multiple of M. For M a power of 2, this reduces to conventional quantization to $\log_2 M$ fewer bits.

In Eqs. (3) and (4) we described how such M-level data channels can be pseudorandomized by pseudorandom congruence encoding and decoding. Alternatively, if M can be expressed as a nontrivial product of $K = 2$ or more integer factors, $M = \prod_{j=1}^{K} M_j$, then one can uniquely expand the M level data word w in the form

$$w = \sum_{k=0}^{K-1} w_{(k)} \prod_{j=1}^{k} M_j \tag{7}$$

with $w_{(k)}$ an integer between 0 and $M_{k+1} - 1$. Eq. (7) is the generalization of the expansion of a number to base M_0 in the case $M_j = M_0$ for all $j = 1, \ldots, K$. Each of the expansion coefficients $w_{(k)}$ can, if desired, be pseudorandomized separately before the final length M word is formed. Again, this generalizes the binary case described where the M_j equaled 2.

A second method for fractional bit rates especially suitable for very low data rates of $1/q$ BPSS for integer q is to code data only in one out of every q audio samples. The encoding schemes are as before, but with a data sampling rate divided by q, and decoding involves the decoder trying out and attempting to decode each of the q possible subsequences until it finds out (for example, by confirming a parity check encoded into the data) which one carries data.

For integers $p < q$ a data of p/q BPSS can similarly be obtained by encoding data in the LSBs of p out of every q samples (for example, samples 1 and 3 out of every successive 5 samples for $p = 2$ and $q = 5$).

A third method for fractional bit rates also codes data in the LSBs of q successive samples, but codes the data into different logical combinations of all q bits. For

example, a data rate of $1/q$ BPSS can be obtained by encoding data as the parity (Boolean sum) of the q LSBs. It turns out that this option is often capable of significantly less audio noise degradation than the simpler scheme of the second method. A part of the advantage is that if one needs to modify the parity, then one can choose to modify that sample out of the q successive samples that will cause the least error in an original high-resolution audio signal, rather than being forced to alter a fixed sample.

We shall see in the following that, for all three kinds of fractional bit rate data encoding, it is possible to use a subtractive dithering technique by a data noise signal to eliminate unwanted modulation noise and distortion side effects on the modified waveform data. The advantages of the subtractive buried data process are not confined to integer bit rates per sample.

3 SUBTRACTIVELY DITHERED NOISE SHAPING

3.1 Subtractive Dither

Here we briefly review the ideas of subtractively dithered noise shaping, detailed by the authors in [1], [3], and [4]. In this paper, by a "quantizer" we mean a signal-rounding operation that takes higher resolution audio words and rounds them off to the nearest available level at a lower resolution. We assume that the quantizer is uniform, that is, the available quantization levels are evenly spaced, with a spacing or step size denoted by STEP.

The quantizer rounding process introduces nonlinear distortion, but this distortion may be replaced by a benign white-noise error at the same typical noise level by using the process of subtractive dither shown in Fig. 4. The process comprises adding a dither noise before the quantizer and subtracting the same dither noise afterward. Provided that the statistics of the dither noise are suitable, it can be shown (see [1], [2]) that this results in the elimination of all correlations between the error signal across the subtractively dithered quantizer and the input signal. One such suitable dither statistic is what we term RPDF dither, that is, dither where each sample is statistically independent of the other samples and with a rectangular probability distribution function having peak levels of $\pm\frac{1}{2}$ STEP.

An audio word of B bits, each of which is a pseudorandom binary sequence, is a 2^B-level approximation to a signal with RPDF statistics, so that the data noise signals considered earlier may be used as dither signals for dithering audio to eliminate nonlinear quantization distortions and modulation noise. Similarly, the M-level data noise signals described in Section 2.3 using the remainder modulo M for data, if made to be of a pseudorandom form by a pseudorandom data encoding–decoding process, can be used as an M-level approximation to RPDF noise.

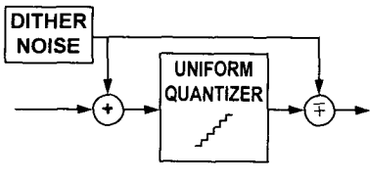


Fig. 4. Subtractive dither around uniform quantizer.

Although data noise signals are discrete approximations to RPDF noise, they can be converted to continuous RPDF noise statistics by the simple process of adding to them an additional smaller RPDF noise with peak levels $\pm\frac{1}{2}$ LSB, where LSB is the step size of the LSBs of the transmitted audio words (as distinct from the step size STEP = M LSBs of any rounding process used in encoding hidden data channels). This is shown schematically in Fig. 5.

Conventionally, as described in [1] and [3], use of subtractive dither requires the use of a decoding process in which, during playback, the original dither noise added before the quantizer is reconstructed before being subtracted. This requires either the use of synchronized pseudorandom dither generation algorithms, or an encode–decode process in which the dither noise is generated from the LSBs of previous samples of the audio signal [3]. However, in the application of this paper, as will be seen, no special dither reconstruction process is required for the discrete dither since this is already present in the transmitted LSBs.

3.2 Noise Shaping

A white error spectrum is not subjectively optimum for audio signals, where it is preferred to weight the error spectrum to match the ears' sensitivity to different frequencies so as to minimize the audibility or perceptual nuisance of the error. The spectrum of the error signal may be modified to match any desired psychoacoustic criteria by the process of noise shaping, discussed, for example, in [1], [4], [23]–[25].

Noise shaping may be static (that is, adjusting the spectrum in a time-invariant way) and made to minimize audibility or optimize perceptual quality at low noise levels, or alternatively it can be made adaptive to the audio signal spectrum so as to be optimally masked by the instantaneous masking thresholds of audio signals at a higher level. The latter option is particularly valuable in the present application, where loud audio signals may well allow an increased error energy to be masked, thereby allowing a higher data rate to be transmitted in the hidden data channels during loud audio passages.

The form of noise shaping with subtractive dither used

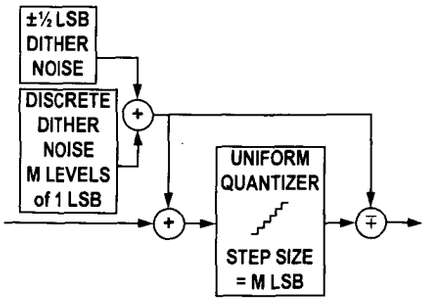


Fig. 5. Subtractive dither using a combination of discrete and continuous RPDF dither.

in this paper is indicated in the schematic of Fig. 6. It will be noted that, while it is equivalent to some of the forms described in [1], it is not the arrangement described previously by the authors in [3] in that here we put the noise-shaping loop around the whole subtractive process. With the arrangement of Fig. 6 the output of the quantizer itself differs from the noise-shaped output of the whole system by a spectrally white dither noise, so that in this arrangement, unlike those suggested in [3], the spectral shapes of the quantizer output error and the system output error are not identical.

With the noise-shaped subtractively dithered quantizer of Fig. 6 the error feedback filter $H(z^{-1})$ must include a one-sample delay factor z^{-1} in order to be implementable recursively, and the originally white spectrum of the subtractively dithered quantizer is filtered by the frequency response of the noise-shaping filter,

$$1 - H(z^{-1}) \tag{8}$$

which is preferably chosen to be minimum phase to minimize noise energy or a given spectral shape [1], and may be chosen to be of any desired spectral shape.

Other implementations of noise shaping around a dithered quantizer system are possible. Alternative implementations are reviewed in [4]. By way of example, Fig. 7 shows an alternative "outer" form of noise-shaping architecture described in [4], which is equivalent to Fig. 6 if one puts

$$H'(z^{-1}) = \frac{H(z^{-1})}{1 - H(z^{-1})}. \tag{9}$$

The application of noise shaping around a subtractively dithered quantizer will not result in any unwanted non-

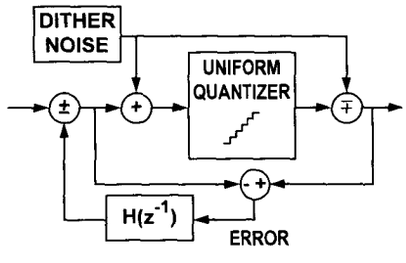


Fig. 6. Noise-shaped subtractively dithered uniform quantizer.

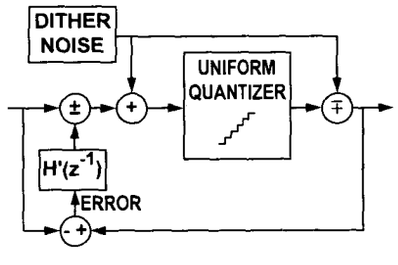


Fig. 7. "Outer" form equivalent to that of Fig. 6 for noise-shaped subtractively dithered uniform quantizer, where $H'(z^{-1}) = H(z^{-1})/(1 - H(z^{-1}))$.

linear distortion or modulation noise, provided that the dither noise added in Fig. 6 or 7 is RPDF dither matched to the step size STEP of the quantizer.

4 APPLICATION TO BURIED DATA CHANNELS

4.1 Noise-Shaped Subtractively Dithered Buried Channel Encoding

Either the arrangement of Fig. 6 or that of Fig. 7 can be applied to obtain subtractively dithered noise-shaped audio results when the last digits of an audio signal word (whether the last N binary digits or the remainder after division by M) are replaced by buried data bits.

The procedure is now simple to describe. The data are first pseudorandomized and then used to form a data noise signal as described. This data noise signal has (discrete M-level) RPDF statistics, and may be used as the dither noise source in Fig. 6 or 7. This is shown in Figs. 8 and 9, where the quantizer is simply the process of rounding the signal word to the nearest integer multiple of M LSBs, or the nearest level if the levels are placed uniformly at other than the integer multiple of M LSBs. The process shown in Fig. 8 or 9 subtracts the data noise signal from the audio at the input of the uniform quantizer (which has step size STEP = M LSBs) and adds it back again at the output of the quantizer so as to make the least significant digits of the output audio word equal to the data noise signal. Noise shaping is performed around this whole process.

For best results using the algorithms of Figs. 8 or 9 (or equivalent algorithms such as that in Fig. 10), it is best if the input audio word signal is available at a higher

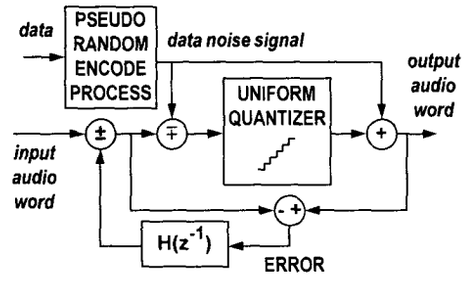


Fig. 8. Application of noise shaping around the pseudorandom data noise signal encoding of data into the audio word. Standard noise-shaper form.

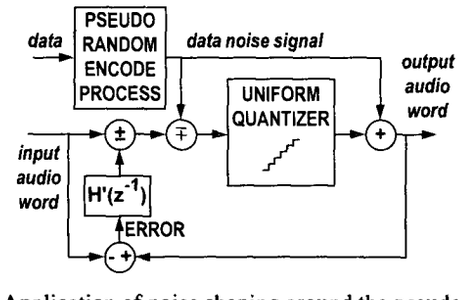


Fig. 9. Application of noise shaping around the pseudorandom data noise signal encoding of data into the audio word. "Outer" noise-shaper form equivalent to Fig. 8 if $H'(z^{-1}) = H(z^{-1})/(1 - H(z^{-1}))$.

resolution or word length than that used in the output, since this will avoid cascading the rounding process used in Fig. 8 or 9 with another earlier rounding process. By making the input signal available at the highest possible resolution, any overall degradation of the signal-to-noise ratio is minimized.

Since the output equals the output of the quantizer plus the data noise signal, the noise shaping has no effect on the information representing the data in the output audio word, but merely modifies the process by which the quantization of the audio is performed so as to minimize the perceptual effect of the added data noise on the audio. It is remarkable that this output signal, being the output of a noise-shaped subtractively dithered quantizer, automatically incorporates all the benefits of noise-shaped subtractive dither without the audio-only listener needing any special subtractive decoding apparatus.

Moreover, because the information received by the data channel user is not dependent on the noise-shaping process, the noise shaping can be varied in any way desired without affecting reception of the data, provided only that no overflow occurs in the noise-shaping loop near peak audio levels. (Fitting a clipper in the signal path before the quantizer to prevent this may be desirable.) Thus the noise-shaping process does not affect the way the signal is used by either audio or data end users of the signal, and so does not need any standardization, but may be used in any way desired by the encoding operative to achieve any desired kind of static or dynamic noise-shaping characteristic.

Other equivalent noise-shaped dithering architectures may be used in place of those shown in Figs. 8 and 9 for encoding the data signals into the output audio word, using the kind of equivalent architectures discussed in [4]. Purely by way of example, Fig. 10 shows yet another implementation having performance identical to that shown in Fig. 8 or 9. It is also evident that in a similar way, the data noise signal can be added and subtracted outside the "outer" noise shaper of Fig. 9 rather than inside the noise shaper as shown.

4.2 Buried Channel Decoding

Optimum recovery of the audio channels involves no need for any kind of decoder in this proposal. Playback is conventional with the effect of subtractive dither by the data noise signal being automatic, as described.

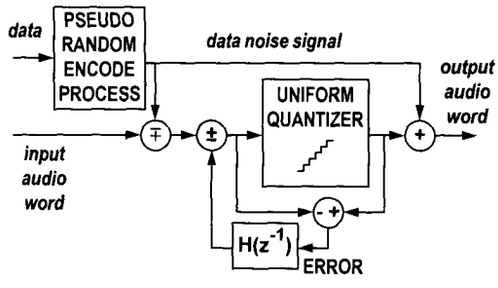


Fig. 10. Alternative application of noise shaping around the pseudorandom data noise signal encoding of data into the audio word.

Recovery of the buried data is also straightforward, simply being recovery of the data noise signal by rejecting the highest bits of the received audio word. In the case of M-level data, the inverse process to the encoding may be used, namely, reading the remainder of the audio word after division by M, that is, resolving the least significant digits of the audio word via modulo M arithmetic. This is followed by the inverse pseudorandom decoding process to recover the data before pseudorandomization, and then the data are handled as data in the usual way. This decoding process is shown schematically in Fig. 11.

In the case where the data are encoded as integer coefficients $w_{(k)}$ with more than one base M_j, as in Eq. (7), the data are recovered by K successive divisions by M_1 to M_K, at each stage discarding the fractional part, the K coefficients $w_{(k)}$ being the integer remainders of the division by M_{k+1}. This is the same process as that shown in Fig. 11, but with K stages of the modulo division.

5 VECTOR QUANTIZATION AND DITHER

5.1 Reasons for Digression

It may not be completely clear to the reader without further explanation that the preceding descriptions of the use of noise-shaped subtractive dithering apply to the case of stereo parity coding as well. To see this, we first need to look at vector quantization and vector dithering and to show that the exact same ideas for subtractive dithering, noise shaping, and data encoding can be applied to the vector quantizer case as to the scalar case described earlier. Because the description in this section may be found rather technical, we suggest that it be omitted on first reading.

The description here is given in greater generality than needed just for the stereo parity coding case, since it has applications to coding information in the parity of the corresponding bits in three or more channels in transmission media carrying more than two audio or image channels as, for example, in the three channels containing the three components of a color image.

5.2 Uniform Vector Quantizers

As briefly indicated in earlier papers [1], [3], [11], the concepts of additive and subtractive dither can be applied to vector as well as scalar quantizers. Vector quantizers quantize a vector signal y comprising n scalar signals $(y_1, \ldots, y_n)$ in geometrical regions covering the n-dimensional space of n real variables. As in the scalar case, we shall say that a vector quantizer Q is a uniform quantizer if the signal y is quantized to a point

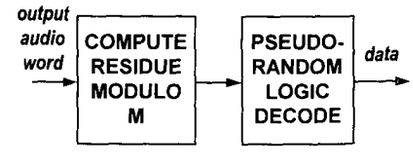


Fig. 11. Recovery of data signal from received coded audio word.

in a discrete grid G of quantization vectors $\{y_g: \ g \in G\}$ where there exists a region C around $(0, \ldots, 0)$ of n-dimensional space such that the regions $y_g + C = \{y_g + c: \ c \in C\}$ cover without overlap (except at their boundary surfaces) the range of the signal variables y being quantized. Thus a uniform vector quantizer divides the n-variable space into a grid of identical vector quantization cells that are translates of the cell C to the points of the grid G, and quantizes or rounds any point in the cell $y_g + C$ to the point y_g.

There are many examples of uniform vector quantizers. The simplest has a hypercubic cell C = the region $\{(c_1, \ldots, c_n): \ |c_1| \leq \frac{1}{2}\text{ STEP } \forall\, i = 1, \ldots, n\}$, that is, separate scalar quantization of the n variables. The grid G in this case consists simply of the points of the form $(m_1\text{STEP}, m_2\text{STEP}, \ldots, m_n\text{STEP})$ for integer m_j, and the associated vector quantizer is simply the one that takes $(y_1, \ldots, y_n)$ to m_j = round (y_j/STEP) for $j = 1, \ldots, n$, where "round" takes a number to the nearest integer. This case is trivial in the sense that it is equivalent to using separate uniform scalar quantizers on each of the n channels.

A more complicated but easily visualized example is the two-channel case where C is a regular hexagon in the plane as, for example, the region consisting of points (c_1, c_2) in the plane, such that

$$\begin{aligned} |c_1| &\leq \tfrac{1}{2}\text{STEP}, \\ \left|-\tfrac{1}{2}c_1 + \frac{\sqrt{3}}{2}c_2\right| &\leq \tfrac{1}{2}\text{STEP}, \\ \left|-\tfrac{1}{2}c_1 - \frac{\sqrt{3}}{2}c_2\right| &\leq \tfrac{1}{2}\text{STEP}. \end{aligned} \tag{10}$$

Here the grid G is the centers of the hexagons in the honeycomb grid covering the plane, that is, G is the set of the points

$$\left((m_1 + \tfrac{1}{2}m_2)\text{STEP}, \frac{\sqrt{3}}{2}m_2\text{STEP}\right) \tag{11}$$

for integers m_1 and m_2.

A uniform vector quantizer of particular interest and practical use in n dimensions is what we shall term the rhombic quantizer. This starts off with a conventional hypercubic grid G_c of points at positions $(m_1\text{STEP}, m_2\text{STEP}, \ldots, m_n\text{STEP})$, where STEP is a step size and $m_1, \ldots, m_n$ are integers, which, of course, includes the hypercube quantizer cell just described and corresponds to the use of n separate scalar uniform quantizers. However, we then produce a new grid $G \subset G_C$, which consists of just those grid points in G_C with $m_1 + \cdots + m_n$ having even integer values. This new grid only has half as many points as the original, and it can be equipped with a new vector quantization cell C as follows, which we shall term the n-dimensional rhombic quantizer cell.

The rhombic quantizer cell can be described geometrically by thinking of the original hypercubic cells as being colored white if $m_1 + \cdots + m_n$ is even and black if $m_1 + \cdots + m_n$ is odd, forming a kind of n-dimensional checkerboard pattern of alternately black and white hypercubes. Then attach to each white hypercube that "pyramid" portion of each adjacent black hypercube lying between the center of the black hypercube and the common "face" with the white hypercube. The resulting solid is the rhombic cell C.

It is evident, since together all the pyramid portions taken from adjacent black hypercubes are enough to form one black hypercube if pieced together, that the volume occupied by the rhombic quantizer cell is twice that occupied by the original hypercube quantizer cell, and that the versions of the rhombic quantizer cell translated by grid G indeed cover the n-dimensional n-parameter vector signal space.

For $n = 2$ the rhombic quantizer cell C is a diamond shape, being a square whose sides are rotated 45° relative to the channel axes, as shown in Fig. 12. For $n = 3$ the rhombic quantizer cell C is a rhombidodecahedron, a 12-faced solid whose faces are rhombuses. For $n = 4$ the rhombic quantizer cell C is a regular polytope unique to four dimensions, termed the regular 24-hedroid [28].

Calculations involving quite complicated multidimensional integrals, which we shall not detail here, show, for a given large number of quantizer cells covering a large region of n-dimensional space, that for $n = 2$, rhombic quantization has the same signal-to-noise ratio as conventional independent quantization of the channels, but that for $n \geq 3$, rhombic quantizers give a better signal-to-noise ratio than conventional independent quantization of the channels. The improvement reaches a maximum of about 0.43 dB when $n = 6$. This improvement in signal-to-noise ratio is maintained when additive or subtractive dither is used as described hereafter. (The hexagonal two-channel quantization described earlier gives a 0.16-dB better signal-to-noise ratio than independent quantization of two channels.)

Mathematically, the rhombic quantizer has grid G consisting of the points

$$(m_1\text{STEP}, M_2\text{STEP}, \ldots, m_n\text{STEP}) \tag{12a}$$

where the m_i have integer values with

$$m_1 + \cdots + m_n \text{ having even integer values.} \tag{12b}$$

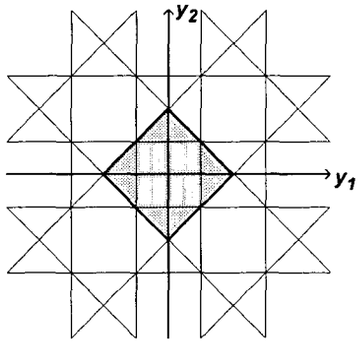


Fig. 12. Two-dimensional rhombic quantizer region (shaded square with sides tilted 45°) shown against a background (squares with horizontal and vertical sides) of conventional independent quantizers (whose square quantizer region is darkly shaded) on channels y_1 and y_2.

The rhombic cell C is that region of points $(c_1, \ldots, c_n)$ that satisfies the $n(n-1)$ inequalities

$$|c_i + c_j| \leq \text{STEP}, \quad |c_i - c_j| \leq \text{STEP} \tag{13}$$

for $i \neq j$ selected from $1, \ldots, n$. The associated uniform vector quantizer rounds a vector signal $(y_1, \ldots, y_n)$ by an algorithm whose outline form might be

$m'_i :=$ round(y_i/STEP)

If $m'_1 + \cdots + m'_n$ is even

then $m_i := m'_i$ for all $i = 1, \ldots, n$,

else $c_i := y_i/\text{STEP} - m'_i$,

(*) $d_j :=$ sgn(c_j) if $|c_j| > |c_i|$ for all $i < j$ and $|c_j| \geq |c_i|$ for all $i > j$

$d_i := 0$ for all other i,

$m_i := m'_i + d_i$ for all $i = 1, \ldots, n$.

End if. (14)

There are, of course, various equivalent forms for this kind of rhombic quantizer algorithm, a computationally demanding aspect on typical signal processors being the determination in line (*) of that j for which $|c_j|$ is biggest.

In the $n = 2$ case there is a simpler rhombic quantization algorithm as follows:

$x_1 := y_1 + y_2, x_2 := y_1 - y_2,$

$m'_1 :=$ round$(x_1/(\sqrt{2}\text{STEP}))$

$m'_2 :=$ round$(x_2/(\sqrt{2}\text{STEP}))$

$m_1 := m'_1 + m'_2, m_2 := m'_1 - m'_2$ (15)

which is based on the observation that the rhombic quantizer cell for $n = 2$ has the same shape as the square cell used for ordinary independent quantization of the two channels, but rotated by 45° and with an increase of the step size by a factor of $\sqrt{2}$ (Fig. 12).

5.3 Subtractive Vector Dither

The concepts of dithering developed in [1]–[4] for scalar uniform quantizers may also be applied to the vector case by using appropriate vector dithers. An n-signal dither noise vector $(v_1, \ldots, v_n)$ is said to have a uniform probability distribution function (PDF) in a region C of n-dimensional space if its joint probability distribution function is constant within the region C and zero outside it. This is the n-dimensional generalization of rectangular PDF dither for vector signals, and we denote the associated n-vector dither signal by r_C.

It can be shown (we omit any proofs here) that if the subtractive dither arrangement of Fig. 4 is used for modifying an input vector signal, where the uniform quantizer becomes a vector uniform quantizer with quantization cell C, and the dither noise becomes a uniform PDF vector dither r_C on the region C, then the output vector signal of the system is free of all nonlinear distortion and modulation noise effects (that is, the first moment of the output signal error is zero, and the second moment is independent of the input signal [4]). Moreover, this is still the case if any statistically independent additional noise is added to the uniform PDF dither noise r_C on the region C.

Moreover, noise shaping can be applied around such subtractive dither in exactly the same way as shown in Figs. 6 and 7, or in equivalent noise-shaping architectures, the only difference being that any filtering is now applied to n parallel signal channels. It is also possible, if desired, to use an $n \times n$ matrix error feedback filter $H(z^{-1})$ or $H'(z^{-1})$ in order to make the noise shaping dependent on the vector direction, for example, to optimize directional masking of noise by signals [11], [12].

It is possible to generate uniform PDF vector dither r_C over the rhombic cell C by an algorithm such as the following. First generate, for example, by the well-known congruence method n statistically independent rectangular PDF dither signals r_i $(i = 1, \ldots, n)$ with peak values $\pm$½STEP, and also generate an additional two-valued random or pseudorandom signal u with value either 0 or 1. Then the values of the noise signal $r_C = (v_1, \ldots, v_n)$ are given by

If $u = 0$

then $v := r_i$ for all $i = 1, \ldots, n$,

else $d_j :=$ sgn(r_j) if $|r_j| > |r_i|$ for all $i < j$ and $|r_j| \geq |r_i|$ for all $i > j$

$d_i := 0$ for all other i,

$v_i := r_i - d_i$ STEP for all $i = 1, \ldots, n$.

End if. (16)

However, in applications of subtractive dither, this algorithm may involve unnecessary complication, since it can be shown that with the subtractive dither arrangement of Fig. 4 with a uniform vector quantizer with quantization cell C, a uniform PDF vector dither signal r_D may be used for *any other* uniform quantization cell D sharing the same grid G, and it will still eliminate nonlinear distortion and modulation noise in the output. Whatever the shape of the other quantization cell D used for the dither signal, the resulting error signal from the subtractive dither arrangement of Fig. 4 is a noise signal with uniform PDF statistics on the quantizer cell C of the uniform vector quantizer used.

This can allow a much simpler algorithm to be used for generating the vector dither in which u STEP is added to (or subtracted from) just one of the n rectangular PDF noise components. For example, a uniform PDF vector

dither noise signal $r_D = (v_1, \ldots, v_n)$ given by

$$v_1 := r_1 - u\,\text{STEP}$$

$$v_i := r_i \text{ for } i = 2, \ldots, n. \quad (17)$$

may be used to subtractively dither this rhombic quantizer.

5.4 Nonsubtractive Case

Although we shall not need to use the nonsubtractive vector dither case in the hidden data channel application of this paper, it is easy to note the extension of the preceding to the nonsubtractive case. As in the scalar case reported in [2], it can be shown that a uniform vector quantizer with quantizer cell C can be made to give an output suffering from no nonlinear distortion or modulation noise if dither noise is added before the quantizer that has the form of the sum of two statistically independent uniform PDF vector dithers, each of the form r_C over the region C.

Such a dither is a vector analog of the triangular PDF dither [2] used in the scalar case, and may similarly be subjected to noise shaping of the dithered vector quantizer without introducing nonlinear distortion or modulation noise effects. As in the scalar case, such nonsubtractive dithering with no modulation noise gives a noise energy three times as large as does substractive dithering.

6 REFINEMENTS OF BASIC PROPOSAL

6.1 Further Developments

The encoding process described will work well as it stands, but it does not incorporate various desirable refinements, which we shall now describe. These include methods to take account of the fact that the data noise signal has a discrete and not a continuous PDF dither, and applications involving stereo parity coding.

6.2 Nondiscrete Dither

The fact that the dither given by the data noise signal has an M-level discrete probability distribution function rather than a continuous RPDF means that there is still unwanted quantization distortion at the level of the LSB of the audio word which is not properly dithered. Preferred methods of adding "nondiscrete" dither (or, strictly speaking, dither at a significantly high arithmetic accuracy such as implemented using 24- or 32-bit arithmetic) are now described. The method of adding such dither shown in Fig. 5 is not preferred for three reasons:

1) Optimum playback requires subtractive decoding of the $\pm\frac{1}{2}$ LSB RPDF dither signal, with all the usual problems of implementing subtractive dither [1], since unlike the discrete data noise signal, this is not explicitly transmitted in the audio word.

2) The $\pm\frac{1}{2}$ LSB RPDF dither signal added before the quantizer does not eliminate modulation noise in nonsubtractive playback, having the wrong statistics for this purpose [2].

3) If the whole system is noise shaped as in Figs. 6 or 7, the nonsubtractive listener will hear the $\pm\frac{1}{2}$ LSB RPDF dither signal as having a white spectrum not affected by the noise shaping, and thus will perceive an increase in noise level.

A correct way of adding extra dither to avoid nonlinear quantization distortion and modulation noise at the $\pm\frac{1}{2}$ LSB level is shown in Fig. 13. The dither used has a triangular PDF with peak levels ± 1 LSB (so-called TPDF dither) with independent statistics at each discrete time instant, so as to eliminate modulation noise in nonsubtractive playback [2]. It is added before the quantizer in the noise-shaping loop, but not subtracted in the noise-shaping loop. This ensures that the added noise in nonsubtractive playback is noise shaped.

Subtractive playback of the extra dither is done, also as shown in Fig. 13, by reconstituting the triangular ± 1 LSB PDF dither at the playback stage, passing it through a noise-shaping filter $1 - H(z^{-1})$ and subtracting the filtered noise from the output audio word. Subtractive playback of course reduces the extra noise energy caused by the nondiscrete dither by a factor of 3, although this will only be highly advantageous in the case where the data noise signal has fairly low energy, such as at a data rate of 1 BPSS.

The triangular dither signal may be generated, in encoding, as proposed in the "autodither" proposal of [3] by means of a pseudorandom logic look-up table (or a logic network having the effect of a pseudorandom look-up table) from the less or least significant parts of the *output* audio word in the K previous samples, where typically K may be 24, and can be reconstructed from the same audio word at the input of the system by the same look-up table or logic in the decoding stage. This is shown in Fig. 14 for the system of Fig. 13.

Although Figs. 13 and 14 are shown for the particular noise-shaping architecture of Fig. 6, similar ways of adding the extra triangular dither can be used with any other equivalent noise-shaping architecture such as the outer form of Figs. 7 and 10—again by adding the triangular dither just before the quantizer and subtracting it again, via a noise-shaping filter $1 - H(z^{-1})$, only at the output of the decoder. It is clear that the points at which dither signals are added can be shifted around in various ways without affecting the functionality.

6.3 The Stereo Parity Case

Suppose we have two-channel stereo signals in which data are encoded pseudorandomly in bit N for all $N = 15$ to, say, $15 - h + 1$ (where the integer h may typically be any integer from 0 to perhaps 6 or 8, 0 being the case of no bits being encoded) of the left and right audio words. Data are also being encoded in the stereo parity (Boolean sum) of bit $15 - h$ of the left and right audio words, as described in Section 2.2.

Based on the results on uniform vector quantization and subtractive vector dither of Section 5, the noise-shaped subtractive encoding of the data described in the scalar case for individual audio channels may be applied to this case too with just two reinterpretations:

1) The uniform quantizer used in Figs. 6–10 now

becomes a uniform two-dimensional rhombic quantizer [such as described in algorithm (15) and illustrated in Fig. 12] with STEP = 2^h LSB.

2) The "data noise signal" used for dithering is given, for example, by Eq. (17), where r_i is the data noise signal of the last h bits of the ith-channel audio word (with the first channel being, say, left and the second channel right), and u being the parity of bits $15 - h$ of the left and right audio words. In units of LSB, the data noise signal for the left channel is then $L_0 - 2^h u$ and for the right channel it is R_0, where L_0 and R_0 are the respective integer words represented by the last h bits of the audio word formed by the data in the two channels.

Any alternative data noise signal may be used that represents an appropriate uniform PDF vector dither as described in Section 5.3, such as that given by algorithm (14).

The residual nonlinear distortion and modulation noise effects at the LSB level caused by the fact that the vector data noise is discrete rather than continuous can be removed by using exactly the same technique as that described in Section 6.2 and Figs. 13 and 14 by adding and, where appropriate, subtracting ±1 LSB triangular PDF dither in each channel separately, the only difference being that the uniform quantizer has become a rhombic vector quantizer and the data noise signal has a modified vector form, as just described.

The particular case $h = 0$, where data are transmitted only in the parity of the LSB of the audio word in two channels, simply uses the parity signal itself at the LSB level as a "data noise signal" in one of the two channels in the encoding process—it does not matter which of the two channels is chosen. With subtractively dithered playback it turns out that the use of properly designed stereo parity coding of data, using a rhombic vector quantizer in the encoding process, gives a total noise level 1 dB lower than would be obtained by the process of coding the data into the LSBs of the words of just one of the two audio channels. Thus stereo parity coding at low bit rates not only ensures audio left–right symmetry for added noise, but gives a significant noise level advantage.

6.4 Generalized Stereo Parity Coding

There are various generalizations of the particular stereo parity coding case just described. We outline these briefly to show the applicability of these ideas to other cases.

A first obvious generalization is that the same process may be applied to other audio word lengths besides the 16-bit wordlength of CDs—for example, the 10-bit word lengths of NICAM encoded digital signals or the 20- or 24-bit word lengths used in some professional audio applications when it is desired to hide data in the audio words. For example, in [3] the authors described a proposal to add data at the 24th bit in studio operations on signals to detect whether or not they had been modified, and the data encoding techniques of this paper can be used in that application to minimize the audibility of the modification of the signal proposed there.

The second generalization is that one can also apply stereo parity coding to the case where one replaces the 2^h-level data in the last h bits by an M-level case for any integer $M > 1$. In this case, data are coded into the residue of the audio words of the two channels after division by M, and the stereo parity data channel is coded into the Boolean sum of the binary LSB in the two channels of the integer parts of the audio words divided by M. This case is handled identically to that in Section 6.3, except that 2^h is replaced throughout by M, and the phrase "last h bits" is replaced by "residue modulo M."

A third generalization instead considers n channels rather than two. As before, this uses a rhombic quantizer in the encoding process for STEP = M LSBs. However, now one uses the n-dimensional rhombic quantizer described in Eqs. (12)–(14) and a vector data noise signal comprising the n M-level data noise signals generated for the residue modulo M data conveyed in each of the n audio channels, to just one of which at each instant is added or subtracted u STEP, where u is the parity (Boolean sum) of the binary LSB in the n channels of the integer parts of the audio words divided by M. Other than replacing the ordinary uniform quantizers with step size STEP by a rhombic quantizer and using the modified

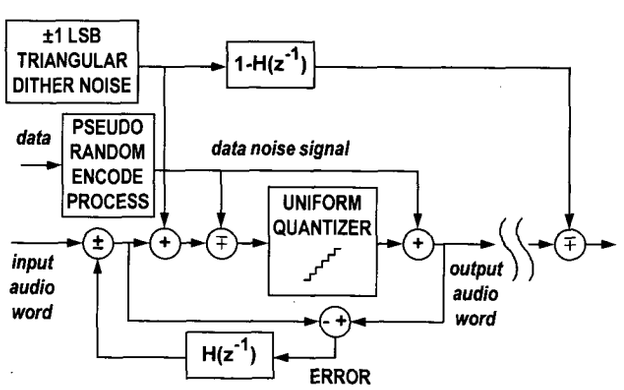


Fig. 13. Use of extra subtractive dither to eliminate nonlinear distortion and modulation noise at LSB level, using noise-shaped triangular PDF dither having ±1 LSB peaks to achieve good results in both nonsubtractive reproduction of output audio word and (shown) subtractive reproduction.

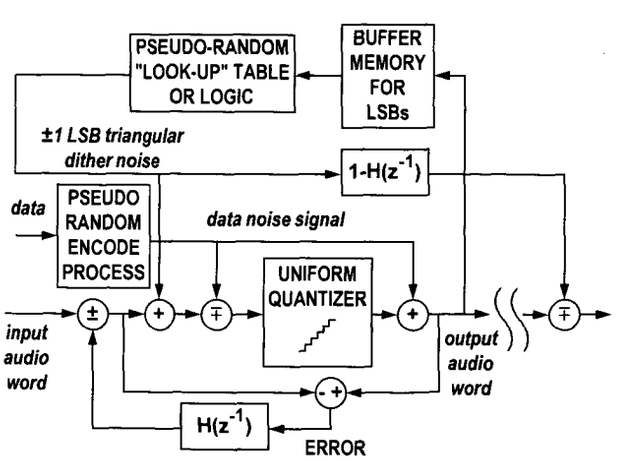


Fig. 14. Use of autodither with Fig. 13 to generate triangular dither in encoder and decoder.

data noise signal, the descriptions given earlier for coding data still apply to this case.

Note that the choice of which channel of the vector data noise signal to add or subtract u STEP, and the choice of whether to add or to subtract, can be made freely, and that this choice can be made adaptively instant by instant to minimize data noise energy if desired, such as by making the choice that minimizes the maximum of the data noise signals in the n channels at each instant. This choice is a discrete approximation to that described in Eq. (16) for uniform PDF vector dither over a rhombic quantizer cell.

6.5 Low-Bit-Rate Case

If one has n transmitted channels of audio, then the parity of their LSBs can be used to transmit a 1-bit per n-channel-sample data channel with remarkable little loss of signal-to-noise ratio, especially in the case where full subtractive dithering is used at the LSB level. One might expect a loss in signal-to-noise ratio of $6.02/n$ dB because the loss is shared among n channels, but for $n > 2$, one gets a smaller loss, typically between 0.3 and 0.4 dB better, because of the fact noted in Section 5, namely, that rhombic vector quantization has a better signal-to-noise ratio than independent channel quantization for a given density of quantization points in the quantization grid. For $n = 6$, a 1-bit per n-channel-sample subtractively dithered buried data channel causes a degradation in signal-to-noise ratio of less than 0.6 dB compared with a properly dithered case with no buried data channel.

Exactly the same techniques can be used to convey data via q successive samples of a monophonic signal—for example, by coding into the parity of the LSBs of each successive block of q samples, as described in Section 2.3. What we have now shown is that by using the parity signal as a subtractive dither for any one sample with a q-dimensional rhombic quantizer, plus normal triangular additive or subtractive dither, this fractional-rate channel can be coded with a very small loss in signal-to-noise ratio (for example, 0.6 dB for a block length $q = 6$), and yet with no nonlinear distortion or modulation noise in either nonsubtractive or subtractive reproduction.

This kind of efficient low-bit-rate culling of data capacity could be used, for example, with successive samples within individual subband channels of a subband data compression system. Its application is not confined to audio. Culling, say, 1-bit per six 10-bit video samples in a digital video recorder with a video data rate of 200 Mbit/s would give a data rate typically enough for four 16-bit audio channels or a consumer-grade additional data-reduced video signal while losing only 0.6 dB in video signal-to-noise ratio in the original video channel.

7 CONCLUSIONS

7.1 Audio Quality Considerations

Anyone concerned with the future potential of the audio art will have some concern about using information originally allocated to a high-quality audio signal to transmit other data instead, as in the proposal in this paper. In order to encourage progress in the audio art, there is a need for at least one widely available consumer medium without built-in serious quality compromises, such as CDs (unlike data-reduced digital systems) offer, so that the market is there in which recordings with improved quality can be made, heard, and sold. Without such a medium, we shall find ourselves permanently locked into limitations many of which will only become apparent as the art of recording, psychoacoustics, and studio production develop further.

Even the best theoretical models of the ears are still extremely crude, for example, not describing the effect of hearing multiple events with individually low but jointly high detection probabilities, especially for nonstationary or transient signals. Many of the musical subtleties of the best "purist" recordings probably reside in these areas of our technical ignorance.

We have therefore been concerned with devising buried channels that satisfy far more stringent requirements than simply crude masking models, which we feel still have limited applicability to state-of-the-art recording quality. This conservative attitude means that—although the option is there with adaptive data rates and noise shaping for our proposal to code data if desired to satisfy existing masking models—such masking models are in no way assumed in the standard. It is a matter of judgment on a case-by-case basis of individual recordings whether such signal manipulations of the error are subjectively acceptable.

In cases where such compromises are not acceptable or are considered too risky (especially for material with high or serious artistic intent), our proposal allows the hidden data channels to produce the most benign kind of error—a steady-noise error free of all nonlinear distortion or modulation noise, and having any desired spectral shape. Unlike previous proposals, this allows avoidance of all psychoacoustically disturbing patterns in the error signal, whether related to the audio signal or to patterns in the transmitted data.

The beauty of this proposal is that by incorporating noise shaping and subtractive dither, it avoids adding any more error noise to the audio signal than is strictly necessary to handle the desired data rate, typically allowing up to 20 dB better perceived signal-to-noise ratio than would be achieved simply by replacing the relevant audio word bits by data, and typically allowing up to 25 dB better perceived signal-to-noise ratio than would be achieved were one also to attempt adding dither in a simple replacement scheme to avoid nonlinear distortion and modulation noise.

Particularly at low data rates, the audio performance of our proposed scheme will typically be comparable to or better than some of the better noise-shaped dithering systems currently on the market, that is, a CD carrying the hidden data channels is likely to sound better than current CDs without the data channel, since the encoding standard incorporates properly designed dithering (and optional properly designed noise shaping). Even at the

higher data rates, the use of proper dithering may well mean a better sound than is currently the norm.

All other things being equal, an audiophile listener would not choose any degradation of audio quality, even if this takes the form of a smooth steady noise free of unwanted modulation and nonlinear distortion effects. But things are not equal since the data channels can be used to convey additional audio channels in a fully compatible way. Provided the coding of these additional audio channels is done with sufficient care to avoid audible data-reduction artifacts, we believe that the overall improvement obtained by adding at least one extra audio channel, either for horizontal B-format ambisonic surround sound or for three-channel frontal-stage stereo, may subjectively more than make up for the relatively benign loss in signal-to-noise ratio (compared to the best noise-shaped dithered performance of which CDs are capable) of the added data channels.

Alternative audio uses of the additional data channel include compatible frequency-range extension without the audible degradations of quality heard in existing commercial schemes for this, and the transmission of level-alteration information to allow dynamic-range adjustment of the recording for users equipped with data decoders.

7.2 Summary

In this paper we described a method of forcing the least significant information in the audio words to conform to the data values of data channels, while ensuring that the effect on the audio is that of adding a noise-shaped steady pattern-free random noise at a level no greater than would be expected from Shannon information theory from the number of bits "stolen" from the audio for an optimally noise-shaped subtractively dithered system.

These techniques involve a process for pseudorandomizing the data so that the audio sees it as a random noise signal which is optimized for subtractively dithering the audio, to eliminate both nonlinear distortion and modulation noise. Not only is the subtractive dithering automatically operative in ordinary playback, but in addition full noise shaping can be applied to the data dither as well.

This paper has further extended this technique not just to the encoding of data in individual audio signals, but to a technique—stereo parity coding—that allows efficient coding of data jointly into two or more audio channels, by using a vector quantization and subtractive vector dithering process. The joint coding process not only ensures symmetry of the way noise is distributed among the audio channels, but in addition gives a substantial improvement in noise performance, especially at low data rates in the data channels. The attainable noise performance approaches the theoretical Shannon limits for the combined Shannon data rate of the audio and buried data channels.

In describing these techniques, a brief account has been given of the generalization of the ordinary theory of subtractive dither to the vector quantizer and vector dither case.

Possible uses of the resulting benign hidden data channels have been described, including additional audio channels for multiloudspeaker stereo or surround sound, audio bandwidth extension, dynamic range control, as well as obvious data applications such as graphics, text or lyrics, copyright, track information, and even data-reduced video.

Unlike previous approaches, no assumptions have been made regarding the masking abilities of the ears. Rather the design aim has been to ensure that the only effect on the existing audio of adding data is to cause a minimal increase in steady background noise, ensuring no compromise with other audio virtues of CDs. If a noise performance comparable to good current CDs is acceptable, this allows data rates of up to 350 kbit/s to be transmitted in the buried data channel, although much more stringent noise requirements can be met at the expense of a reduced data rate.

8 ACKNOWLEDGMENT

The authors acknowledge useful and relevant discussions with many persons over the years on topics related to this paper, including Dr. Geoffrey Barton, Dr. Raymond Veldhuis, and J. R. Emmett.

9 REFERENCES

[1] 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.

[2] S. P. Lipshitz, R. A. Wannamaker, and J. Vanderkooy, "Quantization and Dither: A Theoretical Survey," *J. Audio Eng. Soc.*, vol. 40, pp. 355–375 (1992 May).

[3] 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.

[4] M. A. Gerzon, P. G. Craven, J. R. Stuart, and R. J. Wilson, "Psychoacoustic Noise-Shaped Improvements in 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.

[5] A. W. J. Oomen, M. E. Groenewegen, R. G. van der Waal, and N. J. Veldhuis, "A Variable-Bit-Rate Buried-Data Channel for Compact Disc," *J. Audio Eng. Soc.*, this issue, pp. 23–28.

[6] J. R. Emmett, "Buried Data in NICAM Transmissions," presented at the 92nd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 40, p. 438 (1992 May), preprint 3260.

[7] W. R. Th. ten Kate, L. M. Van de Kerkhof, and F. F. M. Zijderveld, "A New Surround–Stereo–Surround Coding Technique," *J. Audio Eng. Soc.*, vol. 40,

pp. 376–383 (1992 May).

[8] M. A. Gerzon, "Hierarchical Transmission System for Multispeaker Stereo," *J. Audio Eng. Soc.*, vol. 40, pp. 692–705 (1992 Sept.).

[9] M. A. Gerzon, "Hierarchical System of Surround Sound Transmission for HDTV," presented at the 92nd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 40, p. 445 (1992 May), preprint 3339.

[10] M. A. Gerzon, "Compatibility of and Conversion Between Multispeaker Systems," presented at the 93rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 40, p. 1050 (1992 Dec.), preprint 3405.

[11] M. A. Gerzon, "Problems of Error-Masking in Audio Data Compression Systems," presented at the 90th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 39, p. 380 (1991 May), preprint 3013.

[12] M. A. Gerzon, "Directional Masking Coders for Multichannel Subband Audio Data Compression Systems," presented at the 92nd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 40, p. 439 (1992 May), preprint 3261.

[13] M. A. Gerzon, "Problems of Upward and Downward Compatibility in Multichannel Stereo Systems," presented at the 93rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 40, p. 1050 (1992 Dec.), preprint 3404.

[14] M. A. Gerzon, "The Design of Distance Panpots," presented at the 92nd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 40, p. 447 (1992 May), preprint 3308.

[15] M. A. Gerzon, "Periphony: With-Height Sound Reproduction," *J. Audio Eng. Soc.*, vol. 21, pp. 2–10 (1973 Jan./Feb.).

[16] M. A. Gerzon, "Practical Periphony: The Reproduction of Full-Sphere Sound," presented at the 65th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 28, p. 364 (1980 May), preprint 1571.

[17] M. A. Gerzon, "Ambisonics in Multichannel Broadcasting and Video," *J. Audio Eng. Soc.*, vol. 33, pp. 859–871 (1985 Nov.).

[18] N. H. C. Gilchrist, "DRACULA: A Dynamic Range Audio Controller with Unobtrusive Level Adjustment," in *Managing the Bit Budget, Proc. AES UK Conf.* (London, 1994 May 16–17), pp. 99–106.

[19] M. Komamura, "Wide-Band and Wide-Dynamic-Range Recording and Reproduction of Digital Audio," *J. Audio Eng. Soc.*, this issue, pp. 29–39.

[20] P. A. Regalia, S. K. Mitra, P. P. Vaidyanathan, M. K. Renfors, and Y. Nuevo, "Tree-Structured Complementary Filter Banks Using All-Pass Sections," *IEEE Trans. Circuits & Sys.*, vol. CAS-34, pp. 1470–1484 (1987 Dec.).

[21] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1983), ch. 7.

[22] R. A. Wannamaker, S. P. Lipshitz, J. Vanderkooy, and J. N. Wright, "A Theory of Non-Subtractive Dither," submitted to *IEEE Trans. Signal Process.* (1991).

[23] S. P. Lipshitz, J. Vanderkooy, and R. A. Wannamaker, "Minimally Audible Noise Shaping," *J. Audio Eng. Soc.*, vol. 39, pp. 836–852 (1991 Nov.). Corrections to be published in *J. Audio Eng. Soc.*

[24] S. P. Lipshitz, R. A. Wannamaker, and J. Vanderkooy, "Dithered Noise Shapers and Recursive Digital Filters," presented at the 94th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 41, p. 396 (1993 May), preprint 3515.

[25] J. R. Stuart and R. J. Wilson, "A Search for Efficient Dither for DSP Applications," presented at the 92nd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 40, p. 431 (1992 May), preprint 3334.

[26] J. E. Savage, "Some Simple Self-Synchronizing Digital Data Scramblers," *Bell Sys. Tech. J.* (1967 Feb.).

[27] E. S. Donn, "Manipulating Digital Patterns with a New Binary Sequence Generator," *Hewlett-Packard J.*, vol. 22, pp. 2–8 (1971 Apr.).

[28] H. M. Coxeter, *Regular Polytopes* (Methuen, London, 1948; 2d ed., Macmillan).

[29] 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.

[30] S. P. Lipshitz and J. Vanderkooy, "High Pass Dither," presented at the 4th Regional Convention of the Audio Engineering Society (Tokyo, 1989 June); in *Collected Preprints* (AES Japan Section, Tokyo, 1989), pp. 72–75.

[31] J. R. Stuart, "Predicting the Audibility, Detectability, and the Loudness of Errors in Audio Systems," presented at the 91st Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 39, pp. 1010, 1011 (1991 Dec.), preprint 3209.

[32] L. Fielder, "Dynamic Range Issues in the Modern Digital Audio Environment," in *Managing the Bit Budget, Proc. AES UK* Conf. (London, 1994 May 16–17), pp. 3–18.

10 PATENT NOTE

The authors have applied for patents on various techniques described in this paper, which are now assigned to XtraBits.

APPENDIX
CALCULATION OF HOW MANY BITS CAN BE BURIED

Here we discuss whether it is possible to quote a figure for how many data can be buried while maintaining a signal-to-noise ratio psychoacoustically equivalent to "conventional" CDs. We take the conventional CD as

being dithered with white nonsubtractive TPDF dither, as advocated by Lipshitz et al. [2] some years ago. This has a signal-to-noise ratio of approximately 93.3 dB.[2] This is chosen because it can be shown to be [2], [4] the least "white" dithering that avoids completely, with nonsubtractive reproduction, the effects of low-level nonlinear distortion and modulation noise due to quantization effects.

A plausible starting point is to see how much psychoacoustic improvement can be obtained by optimal noise shaping to add the 4.77-dB advantage from subtractive dither, and to express the result as the maximum number of bits per channel that can be "culled" while retaining the original performance.

In a recent paper Stuart and Wilson [29] examine the psychoacoustic advantage in some detail, and in Table 1 of that reference we see a threshold advantage of, for example, 15.3 dB for the "N2,9" noise shaper under "MAFM" listening conditions. Stuart and Wilson also draw attention to the synergy between noise shaping and pre- or deemphasis. Conventional deemphasis is inefficient at noise reduction because the greatest reduction occurs in spectral regions (at 10 kHz and above) to which the ear is relatively insensitive. The provision of additional noise shaping allows this advantage to be transferred to the critical band around 4 kHz.

Referring to the same table [29, table 1], we see that the standard 50/15-μs deemphasis curve reduces the perceived level of white noise by 3.4 dB, but it reduces the threshold for the "N2,9" noise-shaped noise by 6.4 dB, giving a total improvement of 21.7 dB relative to white TPDF dither. An even greater advantage is obtainable if the noise shaping is reoptimized with the preemphasis in mind. In [29], table 3 we see an advantage of 22.4 dB quoted for the noise shaper "M2449P."

It is unfair to quote 22.4 dB as the advantage over the conventional CD, since the 3.4-dB preemphasis advantage has been available right from the launch of CDs. Furthermore, the use of preemphasis entails providing extra headroom due to increased high-frequency levels, and this provision may be comparable to or in excess of 3.4 dB (which is one of the reasons why preemphasis has not found universal favor so far). We therefore consider it more correct to quote 22.4 − 3.4 dB = 19 dB as the advantage of this scheme over conventional technology.

To this we add 4.18 dB subtractive dither advantage—not quite the 4.77 dB quoted for rather subtle reasons [4] concerning the fact that the "N2,9" and "M2449P" curves assume use of "Lipshitz high-pass dither" [30] rather than white TPDF dither. Thus the total advantage of the new technology could be claimed as 19 + 4.18 dB = 23.18 dB.

From this it could naively be concluded that we could cull 23.18/6.02 = 3.85 bit from each channel and still claim performance equivalent to conventional CDs. If fractional-bit data rates are feasible, this translates to a rate of 3.85 × 44.2 × 2 = 339.57 kbit/s.

However, there are some points that this discussion has overlooked.

1) As pointed out in [29] and elsewhere, the threshold reduction advantage for a noise-shaping curve (as quoted) is not the same as the reduction in loudness if the replay level is such that the hiss is actually audible. In principle, this will be the case for some CDs played at realistic levels [31].

2) If the noise-shaped noise is loud enough to be audible, its subjective quality is also important, and heavy HF boosts involved in some of the noise shapers have been questioned in this regard.

3) There is also a "risk" element. It is possible that certain listeners with exceptional HF hearing may find the standard noise-shaping curves (optimized for average or typical listeners) extremely objectionable. To the authors' knowledge, this point has not been investigated experimentally. There could also be problems with tweeters with a resonant peak within the audible spectrum, or with listeners who use EQ for increased treble.

4) As is well known, although noise shaping reduces the audibility of the noise, the total noise power is increased, typically by around 20 dB for the noise shapers discussed. This noise power does become audible, however, when data errors on cheap CD players cause noise sidebands in the spectral regions where the ear is more sensitive. If we apply the same noise shaping to a channel where 4 bit have been culled for buried data, the total noise power is now about +44 dB relative to an ordinary disk (about −49 dB relative to peak level), and it seems quite plausible that this could cause operational problems in domestic playback.

In order to address these problems, one of us (MAG) (unpublished) has devised a family of "moderate" noise-shaping curves, which provide a more modest noise-shaping advantage, but also much lower levels of HF boost. Furthermore, the subjective quality of the noise has also been taken into account. Our preferred option is a curve that provides 9.2 dB of psychoacoustic advantage in the nonpreemphasized case, with an increase in total noise power of only about 6.6 dB. The corresponding advantage in the preemphasized case is 12.5 dB.

Taking into account the 4.77-dB potential advantage from subtractive dither, we have a total advantage of 13.97 dB in the nonpreemphasized case, or 17.27 dB in the preemphasized case. Hence it seems safe to claim a culling of 2.5 bit per channel, giving a total buried data rate of 220.5 kbit/s, with a signal-to-noise ratio 2 dB better than the conventional preemphasized CD and no significant disadvantages.

For the chosen "moderate" psychoacoustic noise shaper, the break-even point for data giving conventional CD subjective performance is thus about 2⅝ bit culled from each of the two stereo channels. Clearly for other choices of noise shaper with different tradeoffs between perceived noise reduction and data-error-noise risk, the amount that could be culled may be larger.

Insight into the various tradeoffs can be gained by

[2] If the point of reference for conventional CDs were to be taken as that using a quantizer with white Gaussian dither, as was the case a few years ago, then the reference would have a signal-to-noise ratio of 92.1 dB.

perusal of Table 1. In the first row we see displayed the conventional 93.3-dB signal-to-noise ratio, the 3.4-dB perceived improvement from deemphasis, and the 4.77-dB improvement from subtractive dither, implemented using autodither [3] or by other means.

The second row explores the possibility of adding just enough noise shaping to restore the final spectrum to flat, in the deemphasized case. A simple noise-shaping filter with just one pole and one zero will do this, and we see a further improvement of 3.6 dB. The reduction in noise level relative to the nonpreemphasized case is 6.7 dB. However, Table 1 does not take account of extra headroom requirements of preemphasized signals, so it is probably safer to quote 3.6 dB as the practical advantage.

The authors believe that this possibility of a preemphasized disk, but with a flat noise spectrum on replay, deserves a wider consideration. It will have none of the operational problems listed earlier (save possibly for 4) in a very mild form), and thus one obtains a 3.6-dB advantage "for free" on all signals except those with exceptionally high peak treble energy content. Such exceptional signals have low probability according to the data of Fielder [32].

The third row of Table 1 introduces the "moderate" psychoacoustic noise shaping discussed earlier. It can be implemented on its own or, in the preemphasized case, "on top of" the noise shaping that flattens the deemphasized noise spectrum. In either case the (further) advantage is 9.2 dB, with still a very low risk of operational problems.

Subsequent rows of Table 1 show the effect of culling up to 3 bit of data from each channel in ½-bit increments (using, for example, stereo parity coding, Sections 2.2 and 6.3). In the autodither decode case the signal-to-noise ratio is worsened by 3 dB for each ½ bit culled, as might be expected. For the listener with an ordinary player, the slope is less steep. Here we see the advantage of subtractive buried data—the culled bits behave as subtractive dither, and by the time 3 bit have been culled, the signal-to-noise ratio is only 0.2 dB worse than for the listener with a fully subtractive player.

Referring to the entry in the second column and penultimate row of Table 1, we see that a preemphasized signal-to-noise ratio of 98.7 with 2.5 bit culled per channel gives a signal-to-noise ratio of 98.7 dB for the ordinary listener, compared to 96.7 dB for the ordinary preemphasized disk. This is the basis for our claim of a signal-to-noise ratio 2 dB better than conventional practice with a data rate of 220.5 kbit/s total.

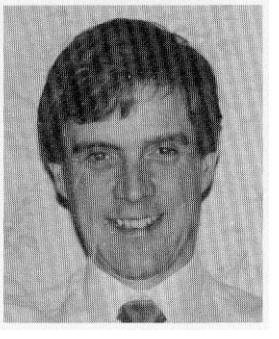
Table 1. Psychoacoustic signal-to-noise ratio (in dB) of 16-bit-per-channel stereo carrier with and without buried data.*

	Compatible Listener		Autodither Listener	
Buried Data Culling Rate	No Emphasis	With Emphasis	No Emphasis	With Emphasis
0 bit, standard TPDF dither	93.3	96.7	98.1	101.5
0 bit, noise shaped to flat	93.3	100.0	98.1	104.8
0 bit, "moderate" psychoacoustic noise shaping	102.5	109.2	107.3	114.0
½ bit = 44.1 kbit/s	101.3	108.0	104.3	111.0
1 bit = 88.2 kbit/s	99.5	106.2	101.3	108.0
1½ bit = 132.3 kbit/s	97.3	104.9	98.3	105.0
2 bit = 176.4 kbit/s	94.7	101.4	95.3	102.0
2½ bit = 220.5 kbit/s	92.0	98.7	92.3	99.0
3 bit = 264.6 kbit/s	89.1	95.8	89.3	96.0

* The buried-data case assumes use of moderate psychoacoustic noise shaping. The figure quoted is based on the ratio of the power in the psychoacoustically equivalent white noise, compared to the maximum sine-wave power at low frequencies. No allowance has been made for loss of headroom in the preemphasized cases.

THE AUTHORS

M. A. Gerzon



P. G. Craven

Michael A. Gerzon was born in Birmingham, England, in 1945 and received an M.A. degree in mathematics at Oxford University in 1967, where he did postgraduate work in axiomatic quantum theory.

His interests in audio stemmed from interests in music, sensory perception, and information theory. Since 1967 he has been active in sound recording live music, recording artists as diverse as Emma Kirkby, Michael

Tippett, Pere Ubu, and Anthony Braxton, and has recorded music for around 15 LP and CD releases.

Arising from this interest, since 1971, he has earned his living from consultancy work in audio and signal processing. He was one of the main inventors of the Ambisonic surround sound technology, working in the 1970s and early 1980s with the British National Research Development Corporation.

With Dr. Peter Craven, he co-invented the Soundfield microphone. He also developed mathematical models for human directional psychoacoustics for use in the design of directional sound reproduction systems, and was made a fellow of the Audio Engineering Society in 1978 for this work. He was awarded the AES gold medal in 1991 for this work on Ambisonics. With Dr. Craven, he also developed the basis of the noise-shaped dither technologies now widely used for resolution enhancement of CDs, and continues to work actively in this area.

He has published over 90 articles and papers in the field of audio and signal processing, including numerous papers on stereo and surround sound systems. He has so far been granted 9 British patents and corresponding applications internationally. He has published papers on practical and theoretical aspects of linear and nonlinear signal processing and systems theory, digital reverberation, room equalization, data compression, spectral analysis, and noise-shaping and dither technologies.

He is currently working on digital signal processing algorithms for professional digital audio with ks Waves, on multiloudspeaker directional reproduction technologies with Trifield Productions Ltd., and on dither and buried data technologies for XtraBits.

He has many current research interests, including the development of general methods for designing complex signal processing systems based on the methods of *-algebras and category theory, image data compression, and advanced mathematical modeling of auditory perceptual effects relevant to audiophile quality.

Other interests include writing poetry, the history of contemporary musics, and the mathematical and conceptual foundations of theoretical physics.

•

Peter Craven attended Oxford University during the period 1966–74, from which he received an M.A. in mathematics and a Ph.D. in astrophysics.

Much of this time was spent not on study but on designing audio equipment and, with friends, making "purist" recordings of the choirs of Oxford. He co-invented (with Michael Gerzon) the Ambisonic Soundfield microphone. This was followed by university teaching posts at Liverpool (computing) and Essex (electronics). In 1979 he started to write a compiler for Algol68, a purist computer language. This work has continued and more recently has formed the basis for a Fortran 90 compiler.

Dr. Craven became an independent consultant in 1982, specializing in digital audio and in high-level methods for DSP code generation. Work on noise-shaping and related issues, jointly with Michael Gerzon, was started at this time and continues to spawn inventions.

He has consulted extensively for B&W Loudspeakers, first in connection with their room equalizer project, and more recently in connection with analog-to-digital and digital-to-analog conversion.

Despite a professional involvement with digital sound, Dr. Craven retains a fondness for early electrical 78-rpm recordings.