Keith O. Johnson and Michael W. Pflaumer
Pacific Microsonics, Incorporated
Berkeley, CA 94710, USA

**Presented at
the 101st Convention
1996 November 8–11
Los Angeles, California**

AES

AN AUDIO ENGINEERING SOCIETY PREPRINT

"–

Compatible Resolution Enhancement in Digital Audio Systems

Keith O. Johnson, AES Fellow and Michael W. Pflaumer, AES Member

Pacific Microsonics, Incorporated
Berkeley, California, USA

Abstract
A conjugate record and playback system is described which enhances the resolution and sonic performance of digital audio recording based on linear pulse code modulation, while remaining compatible with standard linear playback. Using additional information encoded entirely in the program material, it addresses both the amplitude resolution and sampling rate limitations of the current commercial recording standards.

Introduction

As the compact disc (CD) has become a widely used medium for distribution of high fidelity audio, it has become apparent that its sound quality is not as high as original expectations, which were based primarily on conventional distortion and frequency response measurements. This view has been expressed both by those involved in the production of CD's [1] and by fidelity conscious consumers. Dissatisfaction with the fidelity of CD digital recordings when compared to analog master tapes of the same recording sessions prompted the authors to investigate the factors responsible for the loss of fidelity and to devise approaches for improving the situation.

A conjugate digital encode/decode system is described which allows recording an audio signal of greater than standard resolution and sonic accuracy on standard linear pulse code modulation (PCM) recording formats, such as the compact disc (CD). The conjugate decoding process is used to recover the full improvement in fidelity, but compatibility and some sonic benefits are maintained for standard linear playback on non-decoding equipment. A brief examination of the general requirements for a high fidelity recording system will form a basis for understanding the rationale behind this system.

Requirements of a transparent system

In the majority of cases, the ultimate use of an audio recording system is to reproduce sounds at some later time. In any case where the reproduced sound will be listened to, the measure of fidelity of the system should be predicted from system performance but can only be judged by listening. The goal of a high fidelity recording system should be to be as transparent as possible, to change the sound as little as possible. In other words, if one were listening to an analog audio signal on a monitor system, the ideal recording system would be one which did not produce any change in the sound, other than a delay in time, when inserted into the signal path for any given source on any monitor system for any listener. Of course, such an ideal system does not exist, as most recording engineers will attest. A high quality live microphone feed has a quality and fidelity which can only be approximated by the best recording systems, regardless of technology.

Over the years, a great deal of work has been done to try to specify in objective terms the parameters of a recording system which closely approaches the subjective ideal. Without

such objective measurements, it is not possible to design a real system. Target values have been proposed and updated for dynamic range, signal to noise ratio, allowable levels of distortion, and frequency range. As the understanding of human hearing advances and as actual systems have been constructed, the target values for the parameters and measurement systems used to evaluate them keep changing.

The peak dynamic range requirement for professional recording has been shown to approach 130 dB [2]. Even conservative estimates produce numbers greater than 120 dB [3]. While the capabilities of an average home playback system cannot cover this range because the average home speaker system cannot reach the necessary peak sound pressure level (SPL), there will always be some systems which can. In addition, edit situations and listeners who change gain during a program pose added dynamic requirements. Therefore, these numbers remain a valid target.

Simple signal to noise ratio has been shown to be a poor description of system performance, given the understanding of critical band theories of human hearing. This is even more true for allowable levels of distortion. The simple total harmonic distortion and intermodulation distortion measurements devised in the days of vacuum tube electronics do not do a very good job of predicting the sonic performance of transistor amplifiers, and they produce essentially meaningless numbers when applied to current digital recording systems. Even so, they are still offered as a standard part of every data sheet.

Research [2] has shown that distortion products other than low order harmonics must be kept at least 120 dB below peak levels in the presence of complex signals in order to achieve a satisfactory level of accuracy in a critical listening environment. Since interactions with different listening systems may bring out different problems, this level of performance is required to be safe for all cases. This is especially true for intermodulation difference products and fold-over distortions in digital systems, in which high frequency complex signals can produce non-musical lower frequency distortion products which appear unmasked in the frequency range where hearing is most sensitive. The requirements can be relaxed somewhat for the more naturally musical second and third harmonic distortions at high signal levels.

Most current methods of measuring distortion involve the use of simple steady state tones, which do not stimulate the system in the manner in which music does, combined with averaging measurements, which do not respond well to small distortion products having a high peak to average ratio. Our current research indicates that infrequent distortion products with peak levels in the -120 dB full scale (dBFS) range are potentially audible [4]. Measuring these distortions at the edge of audibility can be a truly challenging problem.

The frequency range requirement for a recording system is a very controversial issue. Problems at the low frequency end seem to arise primarily from phase distortions, and for digital systems there is normally no difficulty extending the frequency response as low as one wishes.

The requirement for high frequency extension involves a whole different set of concerns. Since the data storage requirements for non-compressed digital audio are directly proportional to the sampling rate, bandwidth is expensive. The conventional argument is that since human hearing cannot detect steady state tones with a frequency greater than about 20 kHz, a recording system need not store frequencies above this limit. There is some evidence to suggest that this convenient assumption may not be strictly true [5]. Many real world acoustic instruments generate plentiful energy above 20 kHz, and it might be argued that to reproduce their sound transparently involves reproducing all of it or at least preventing the consequences of bandwidth restrictions. Whether or not there is a real

need for frequencies above audibility is still an open question. However, it is now clear that the very sharp cutoff "brick wall" filter required to represent an audio signal in digital form with a sampling rate only slightly higher than twice the audible frequency range does cause problems. This issue will be discussed in more depth below.

Having examined some basic requirements for a high fidelity recording system, it is instructive to see how the current compact disc measures up. It is interesting to start with some subjective evaluations of the sound, and to relate those to objective measurements.

Limitations of current standard

CD's have been described as having a "digital" sound, which includes varying degrees of such attributes as a dull lifeless sound, an edgy harsh sound, especially on cymbals, a sound with artificially silent pauses and a loss or collapse of hall ambiance, a sound stage which is two dimensional, like a curtain stretched between the speakers, and a sound characterized by a loss of individuality for the massed voices of a chorus or a violin section. In addition, the timbre of individual instruments is altered by digital recording using the current standards. Bass violins sometimes sound as if the body of the instrument were made of cardboard. The attack of the notes of a piano frequently takes on a pingy quality and the body of the instrument disappears leaving a thin sound. Flutes may have a rough edge with exaggerated breath artifacts. Reed instruments sometimes take on an almost synthesized buzzing sound. Of course, these artifacts are most blatant on CD's made on early equipment, but the effects can still be found to some extent on recent recordings.

A question which arises is how much of the loss of fidelity is caused by implementation errors, such as distortion in the analog to digital (A-D) and digital to analog (D-A) conversion processes, and how much is due to the limitations of the CD format. One of the tools which the authors have used to explore these questions is an A/B switch box (Figure 1) which allows a listening comparison between a high quality analog source connected directly to a monitoring system versus the source going through an entire digital encode/decode chain and then to the monitor. This setup allows exploring the effects on the sound of different digital systems inserted into the signal path and of different digital algorithms within a given system.

Great care was taken to minimize extraneous factors which could influence the result. The switch box itself is composed of sealed rhodium relays in both the active signal leads and the grounds, with careful isolation of the control signals. The relay operation sequencing is timed to minimize switching transients. The line amplifiers on the original analog source and the output of the digital system were of similar design, and the load impedance of the input of the digital system was matched by that of the passive attenuator in the monitor system. The power amplifiers used for most of the tests were very fast with very low levels of transient intermodulation (TIM) distortion for frequencies well outside the audio range. Of course, levels were matched carefully. This basic setup was repeated with different power amplifiers and loudspeakers, and with different analog sources. The musical selections used were usually out-takes of first generation analog master tapes and were repeated many times for each group of listeners so that they were "over learned" by all listeners. It is important when using music for listening tests that the listeners be very familiar with it so that they are not still learning it as the A/B comparisons are going on.

The results of the listening tests were used in combination with objective measurements in an iterative fashion to determine which factors in the measurable performance of the system

were responsible for each of the sonic aberrations of the system as a whole. In many cases, there were several different objective situations which could produce similar sounding problems, and it was only by addressing all of them that the problem could be substantially reduced. (Of course, as each layer of distortion was dealt with, the next level became more obvious.) The starting point of the research was to build a high resolution, oversampled multi-bit A-D converter and a matching D-A converter. These could be chained together without attempting to store the intervening digital data to explore the sonic effects of converter problems, including various non-linearities, clock jitter, digital to analog crosstalk, etc. The converters formed the framework into which various digital signal processing (DSP) algorithms could be introduced to reduce the voluminous digital data down to the sampling rate and word length of the CD standard and interpolate it back again.

Sonic Effects on Program Material

Many sonic shortcomings of digital audio are attributable to inadequate conversion accuracy. Often these limitations are difficult to predict using only standardized tests. Systems need to be tested using hearing related evaluation procedures specific to the behavior of a design and its components. Our tests employed program material that was easily and immediately perceived as distorted by specific design changes affecting the conversion process. However, some design changes, particularly those which affected the settling time of sampled analog signals, sounded transparent with short term A/B listening tests yet displayed shortcomings after prolonged listening. Frequently, changing a design parameter produced a clear perceptual correlate that eluded quantification by measurement. We found that additive distortions were much easier to examine and quantify than those resulting from subtractive information loss, while distortions producing time related perceptual changes presented a greater challenge. During development, program material chosen because it was most audibly altered by identified sonic artifacts was used. Conversion parameters thought to be relevant were then altered and the results were evaluated in repeated listening sessions. In time, our perception of the problems became keener and we were better able to predict appropriate corrections and quantify system behavior related to perception. The program material which we ended up using for A/B listening tests was in many ways similar to that employed by others, and contained examples of problems widely criticized in digital recordings.

The following discussions will first consider conversion performance and its sonic effects, as well as the specifics of listening tests. Development of the compatible resolution enhancement system will then be described.

High peak value, low probability, pulse or "glitch" type conversion distortions were found to create sonic artifacts such as "hash" on cymbals and "grainy" exaggerated breath sounds on flute. Often the cause of these distortions was digital to analog crosstalk or the ranging and ladder mismatches that can occur with parallel converter topologies. Even when the average distortion level was very low, the damage to the sound of the flute was still substantial, particularly when large amplitude bass notes in the program material swept higher frequency instrument harmonics through troublesome conversion levels. Various amounts of dither added prior to and subtracted after conversions were found to reduce such "grainy" artifacts by randomizing them to noise. Subtractive dither with levels high enough to randomize parallel converter ladder errors around the A-D conversion process became a standard part of the converter setup.

We also experimented with adding large amounts of high frequency in band dither at the output of the A-D converter to help linearize the D-A converters. Unfortunately, the reproduction of intense high frequency sounds such as cymbals and jangling keys through a slower, less sophisticated power amplifier was degraded as a result. Both crossover and TIM components were higher in our test example. Since uncompromised playback was required using any equipment, we chose a conservative dither level at the output of the record processor. It is appropriate to apply high amplitude dither designed to linearize D-A conversion at the converters, not in the program material. For satisfactorily transparent converter operation, both harmonically rich and high peak value, low duty cycle distortions had to be at least ll0 dB below peak level for dynamically changing signals traversing most of the conversion range.

The familiar digital distortion of roughness or "broken" edges added to the sound of massed groups of similar instruments was also examined. We chose violins, mixture ranks in pipe organs, and vocal choruses from both live mike feeds and master analog tapes as source material. Listening tests verified the necessity of keeping the amounts of foldover,

transient intermodulation and other envelope or activity related distortion products very small. Attempting to precisely quantify the cutoff level for audibility of these distortions was not worthwhile, since listening off axis to loudspeakers would increase frequency response irregularities and alter the high frequency content of signals, thereby unpredictably changing the distortion perception threshold. Since high frequency program material content can occur without midrange maskers, we felt that to be safe, these sonically discordant distortions should be well below the noise level. Tests designed to substantiate this requirement used clusters of high frequencies with characteristics chosen so that possible envelope, difference, and sample rate related foldover distortions generated would all appear in the perceptible middle audio range. When distortion components were l20 dB below maximum peak level, further reductions became inconsequential as the levels then approached hearing thresholds and fell well below other limitations in the reproduction chain. However, reduction of these otherwise transparent oversampled signals to a 44.l kHz data rate using standard decimation filters still produced some subtle "broken" artifacts and a slightly dull or smeared sound.

The very good steady state performance of a particular single bit converter prompted us to examine using that approach. Unfortunately, -90 dB foldover distortions in the upper midband frequencies were found, probably due to slew error integration and dynamic phase limitations produced by analog circuits within the conversion loop. Listening tests confirmed the measurements. In one demonstration, the fine, experienced artistic control of a highly distorted electric guitar became shockingly broken, unpleasant and confused as a result of the converter artifacts.

Subtractive distortions manifesting as a loss of ambient information and sound staging were observed when data was reduced to l6 bits. Without dither, quantization noise drew the listeners attention to the loudspeakers, and small amplitude incremental information became lost. Both effects seriously damaged spatial illusions even when signal levels were substantially above quantization step resolution. Triangular probability density function (PDF) dither with an amplitude of two least significant bits (LSB's) peak-to-peak removed many distracting signal related artifacts, but not all sense of ambient information was restored. Many different signal sources were used to explore these effects and to optimize parameters for low level bit extension processes. However, a few types of program material had much greater sensitivity to subtractive loss. These were large scale productions with very large dynamic range and spatial contrasts between close and distant instruments. Several had useful first reflection timings in the 40 ms range, typical of performance stages. Such reflection timings could define distant instruments by reasonably certain perception. Yet frequency response or level changes of these small reflected energies could alter whole instrument sounds [6]. Other recordings had delay panned spotlighting using auditory inhibition to achieve instrument placement [7]. These examples were sensitive to phase and timing changes. Both localization methods rely on delicate spatial cues easily inhibited by frequency response and standing wave problems in the audition environment. As these factors became better resolved through room treatment and component placement, the limitations of dither became more apparent. An effect became evident in which the sound shifted more to the center with less sense of sound stage as the conversion level dropped. Raising the program level through the converter chain while maintaining a constant listening level restored the proper perspective. Examination of small amplitude signals containing ambient information and instrument harmonic decays posed questions that might explain these observations.

Small amplitude signals that produce instrument timbre and ambient decay perception can have midband levels below the thermal noise of many microphones and electronic devices. A dynamic gain structure which is adjusted to suit signal dynamics can achieve low distortion at high levels while still preserving the sense of ambience. Spectral alterations to

ambient signals and signal related peak noise modulations such as those produced by amplifier crossover distortion also appeared to be perceptible below random noise levels. As the signal frequencies increase, and the hearing thresholds increase, the same characteristic relationship appears to hold at higher levels. However, we did not study these perceptual effects at the highest audible frequencies. We also found that small signal perception does not change when noise having large peaks and a low average level is added. Our concern was how small "broken" pieces of signal observed in the higher peak quantization noise might sound. A spectrum analyzer will average these over time and present a clear unchanging display. For this type of signal to remain timbrally neutral, the auditory nerve is required to average burst differentials made up from quantization levels. Averaging takes time to present the full stimulus. [8], [9], [l0] Consequently, small changing events may be gone before auditory thresholds have been reached. We know that the potassium gate responses of hair cells are very fast molecular events that occur even with small stimuli in tens of microseconds. [9] These quick initialization responses can alter thresholds which change nerve firing patterns in response to later stimuli. Whether these events, which change time and amplitude thresholds for input to hair cells, will alter perception is an important question. One could argue that low level transient information is unlikely to occur from instrument and hall decay sounds. However, background sounds providing ambient cues are complex and the auditory brain responds to their changes. A second consideration involves higher peak energy noise in the presence of the "broken" signal. The auditory filter provides high Q response sharpening possibly by electromechanical nerve feedback from outer to inner hair cells [11], [12] and by inhibitory transmitters which stiffen hair cells through actin molecules at their tips. [10] Statistical nerve firing of cells near each other shows phase coherence to the stimulus at the beginning of bursts. These responses which try to sensitize low level perception may be disrupted by the combination of a broken stimulus and noise. If this does happen, some of the ambient or decay signal would be perceived with a spectral balance altered by which parts of the signal start losing low level phase coupling. We have observed a subtle dulling of the sound of a plucked bass when its identifiable harmonics approached quantization levels.

Converter and process limitations and their resulting distortions are responsible for many timbral errors. "Buzzing" reed sound caused by conversion discontinuities and "broken" piano attack resulting from signal correlated jitter are examples. Even with better conversion performance, lesser degrees of these problems still exist. For example, recording engineers mixing down from multi-track digital recordings often make spot equalization adjustments to each channel. Thus digital colorations are usually somewhat compensated for along with changes made for production effect, creating a fairly neutral composite. However, what concerned us was the unavoidable trade-off between timbral accuracy and resolution inherent in conventional systems.

Early development work started with comparisons of analog tapes to 18 bit A-D conversion with 24 bit processing. The system employed had performance levels approaching those previously described. Live microphone feeds, higher precision conversion, and a recently completed 88.2 kHz recording system have been added to the roster of comparison programs. Analog signals could be converted and decimated to 88.2 kHz for listening to directly as compared to decimating down to 44.1 kHz and interpolating back up. Both systems used 8 times oversampling D-A converters with 20 bit precision. For listening tests at this level of accuracy to have been meaningful, the elaborate ground and switch control isolation, as well as the source versus output line amplifier matching described earlier was necessary. Comparison of decimation and interpolation filter cascades operating with 44.l kHz 24 bit data was made to unprocessed 88.2 kHz data or analog signals. From these experiments, it became apparent that the introduction of an anti-alias decimation filter followed by an interpolation filter in the signal path had a considerable

effect on the sound. It also became apparent that it was important to consider these cascaded filters as a system, even though they are normally separated in both time and space.

In the process of optimizing the system, like the mixdown engineer we found conflicting sonic effects with different instruments. Some were altered nuances which seemed to change the sense of an instrument's material construction and operation. Others were resolution losses characterized by the dulling or smearing of subtle detail. We observed a "paper" quality with snare drums, and the beater sound was smeared. Cymbals sounded heavier, as if made of zinc, and had a less airy ring. A perceived "hardening" of piano hammers was combined with reduced fast note delineation. We discarded the oversimplifications of traditional approaches to filter design and improved the arithmetic precision and high frequency rejection of the interpolation filter. This improvement reduced hardness, particularly the "paper" drum quality. However, the dulling tendency of the cascaded filter system then became more exposed. Experiments with sharp attack and steady ensemble sounds verified the conflicting situation improperly handled by specific equalization to the auditioned track. Resolution of inner detail transients was compromised while at the same time other sounds sounded broken and brighter.

Changing the transition band response of the anti-alias filters to achieve faster settling around transient edges substantially improved the resolution problem but created timbral alteration. Reducing the transition rate and slightly peaking the frequency response of the cascaded filters improved timbral errors but degraded resolution. Filters whose frequency response was allowed to ripple in the region of the transition band produced excellent timbre, but did not do well with transient edges. Further detailed listening tests verified these perceptions but also revealed that specific losses create a persistence of hearing or mental impression like an improper anchor effect [13] which lingers and colors perception of reference signals which do not have the sonic problems. When the filters were switched silently, it took some time for the change in sonic character to be appreciated.

The persistence of a sonic impression created when certain cues in the program material are effectively reproduced led to the development of an automatic switching system. This system would select different anti-alias filters depending on the content of the program material in order to attempt to provide the optimum one for each condition. Fortunately, the program conditions dictating the filter requirements were usually non-conflicting. The result was like having the best of both worlds. It appeared that as long as the cognitive listening process was given enough cues to relate to the full experience, the sense of it was there without actually having the full experience. Further study into the physiology of the auditory nerve and psychophysiology of hearing revealed some collaborative support for these filter choices and sonic observations.

Electro-mechanical motility of outer hair cells or nerve feedback through efferent paths has been thought responsible for heightened selectivity of the auditory nerve or cochlea. [11] Other actions external to the auditory nerve, in time, create alterations in hair cell stiffness and physical tuning. [10] Motions of the basilar membrane suggest sounds emanating from the ear are related to these activities. Other observations show threshold inhibition of hair cells near tuned sites of the basilar membrane and enhancement at other sites from below perception stimuli. All of these appear instrumental in creating a high "Q" critical band filter. These actions are predictable when stimulus levels are low, frequencies don't conflict or beat, [14] and the stimulus frequencies and levels influencing the tuning remain long enough for each process to stabilize. Phase locking or coupling and positive feedback sometimes describe these perception enhancing responses. For small to moderate signals, none of these tuning actions are instantaneous molecular level events. Yet threshold changing gate responses at hair cell tips are very fast. Consequently, one can reasonably

predict that a transient envelope or energy burst can produce different amounts of short term peak energy at basilar membrane sites dependent on envelope shape. If waveforms are non-repetitive or infrequent, thresholds affecting more delicate tuning will have relaxed and become subject to rapid change thereby altering future nerve response. As a result, transient sounds having the same spectral energy but different envelope shapes can produce different threshold patterns. This will change timings of subsequent tunings which are likely to alter perception. [15] Other factors like beats between close frequencies create higher peak energies at tuning sites on the basilar membrane. These create greater masking loss than average levels can predict. Indeed, signals below perceptual limits can create beats with other signals that can be perceived. One can predict that fast cut off low pass filters could be problematic. The sharp removal of higher frequency sidebands creates response ripples from rapidly swept sine waves. These will alter peak energy at basilar membrane sites. In addition, ringing responses near transient events can mix with low level harmonics in the signal to create short beats. [17] One can perform simple experiments to demonstrate that both of these are audible.

The conclusions which have come from our research indicate that because hearing is non-linear, it is sensitive to the shape of the envelops of sounds even at high frequencies. The necessity of having a sharp cut-off brick wall anti-alias filter in a digital system in order accommodate a sampling frequency near the minimum necessary for the audio band is bound to create some distortions to sound envelopes. This can be easily demonstrated: If one measures the frequency response of such a low pass filter with a very slow frequency sweep, one gets the classical rectangular envelope. If, on the other hand, one increases the sweep speed, the envelope becomes full of ripples because the faster sweep produces sidebands which fall above the cutoff frequency of the filter. Removing the filter restores the rectangular envelope even for the fast sweep. Listening tests have shown that sweeps with different envelopes but the same spectral content do sound different, although it is very difficult to eliminate all system related sources of confusion in these tests.

Music can contain very complex high frequencies which produce complicated envelopes. It is a logical extension of the above results that if removing high frequencies, which in themselves may not be audible, produces changes in the envelope of audible sounds, then a change in those sounds can be heard. Changes in the characteristics of the filters in the neighborhood of their transition regions can have an effect on the behavior of signal envelopes.

Further evidence of hearing sensitivity to high frequency envelopes resulted from experiments with interpolation filters cascaded with anti-alias filters when treated as a system. If one takes a conventional good brick wall filter whose passband extends out over 20 kHz, cutting off by Nyquist, and cascades it with a conventional interpolation filter which is exactly 6 dB down at the Nyquist frequency, one gets a peculiar result. For frequencies near the transition region of the input, the interpolation produces alias components of fairly high amplitude which beat with the input signal. One can see the potential for this in the frequency domain by taking the response of the anti-alias filter and creating a mirror image around Nyquist, which is what happens in interpolation. The resulting response is cascaded with the interpolation filter response to produce Figure 2. Notice the alias peak above Nyquist. In the time domain, even steady state sine waves near the cutoff will produce modulated envelope beats. Listening tests on music have shown that these beats are associated with a papery sound on snare drums and a hard sound on cymbals and reeds. When the frequency response of the interpolation filter is altered slightly to eliminate the alias peak, the papery sound goes away.

The results of our research, which relied heavily on listening tests combined with objective system measurements, led us to a conjugate system of encoding and decoding which is now commercially available, patented# , and is known as the HDCD* process.

Summary of HDCD

The encoding process (Fig. 3) starts with a high resolution analog-to-digital (A-D) conversion of the analog input signal, typically yielding a digital signal at twice the final sampling frequency and a 24 bit word length. This over-sampled high resolution signal is analyzed in real time to determine which aspects of its sonic character will be most compromised by reduction to the standard PCM signal at any given time, and pairs of conjugate processes are chosen dynamically to minimize the sonic damage on playback reproduction. Areas of sonic compromise which the system addresses include dynamic range, amplitude resolution, timbre, and anomalies normally caused by a brick wall anti-alias filter with a cutoff frequency near the top of the audio band.

The over-sampled signal, delayed long enough to allow the analysis and process choice, is decimated to the final sampling rate using one of several filters chosen based on program content. The resulting signal has its dynamic range reduced using a combination of reversible soft peak limiting for infrequent peaks, and average level based compression for low level signals. The signal then has high frequency dither added and is quantized to the final wordlength.

All parameters and processing choices made during the encoding phase are inserted into the data stream as pseudo-random noise encrypted control signals inserted into the least significant bit of the audio data as part of the dither, on an as-needed basis. The encrypted control signal allows the decoder to apply accurately timed conjugate processes during playback without using media format dependent sub-codes. The decoder (Fig. 4) restores the limited peaks, accurately expands the compressed low levels, and applies a conjugate interpolation filter to match the anti-alias filter of the encoder.

The system remains compatible because the corrections are only applied for signal extremes, increasing the average modulation with minimal sonic alteration of undecoded playback. Since the encoding parameters and exact timing are conveyed in the control channel, the recording engineer can select a combination of process parameters suited to the program material, and the decoder accurately follows, providing the correct conjugate reconstruction.

A-D performance

The A-D converter used in the encoder is a multi-bit type running at a multiple of the final sampling frequency of 44.1 kHz. Subtractive dither with a fairly high level is used around the conversion process in the classic textbook fashion to help linearize the conversion. Special attention is paid to minimizing conversion jitter and crosstalk between digital and analog signals. The converter output is decimated down to 88.2 kHz in the classic fashion. The result has approximately 19 bit precision.

The most important aspect of the converter's performance is the extremely low level of distortion products in the presence of complex signals. These products are typically on the

order of -120 dBFS, in other words, at about 1 part per million. An example of the response to an 8 tone cluster test described above is shown in Figure 5.

This high precision, wide band digital signal forms the starting point for the HDCD process.

Decimation filter switching

The first step in the process is decimation to the final CD sampling frequency of 44.1 kHz. One of the important results of the research discussed above is the fact that "brick wall" decimation filters have a considerable effect on the sonics of the resulting signal. Since virtually all modern playback equipment uses a digital interpolation filter to produce an oversampled signal to drive the D-A converters, it is important to consider the cascade of these filters as a system. It became apparent that no one anti-alias filter cascade was sonically neutral under all program conditions. The solution was to use a set of filters which could be switched dynamically in response to the content of the program material on a moment by moment basis. The persistence of cognitive hearing allows the result to sound very much as if there were no filter cascade there and the system had a higher sampling frequency.

Implementation of this system involves analyzing the high resolution signal in real time in order to decide which filter to use at any given time. The analysis looks at average mid-band energy, both peak and average high frequency energy and overall signal level. Ratios of peak to average high frequencies and high frequency to mid-band levels, as well as total level, are used to identify situations which call for one filter or another. The choice also involves a hysteresis mechanism so that a given choice remains in effect long enough for hearing mechanisms to respond to it. The chosen set of filter coefficients is then used to decimate the 88.2 kHz signal to 44.1 kHz.

All of the filters used are symmetrical finite impulse response (FIR) types of the same length, so that they have the same constant group delay. They all have very flat frequency response below 16 kHz and excellent stopband attenuation above Nyquist. The differences between them are small changes in frequency response between 16 and 22 kHz, which are enough to influence their subtle sonic behavior. Also, in order to ensure compatibility when not decoded, all anti-alias filter choices have been evaluated when cascaded with conventional interpolation filters used in standard equipment, in addition to the complementary decode filters. Because the filters are so much alike, it is possible to change from one to the other with a simple switch: no fade or merge is required.

Amplitude resolution

The result of decimation to the output sampling frequency is a signal with considerably higher amplitude resolution than the CD standard 16 bits. In the current implementation of the encoder, it is a 44.1 kHz, 24 bit signal. It is generally agreed that the best overall results occur when a signal is kept at high precision throughout the editing process, with a final quantization to 16 bits done at the end. To accommodate this, the encoder makes this signal available either as a 24 bit signal, or dithers and quantizes it to 20 bits for editing. The edited signal is then run through the encoder again for the final reduction to 16 bits.

The approach to preserving as much as possible of the sonic benefits of the high amplitude resolution is to handle the cases of the two signal extremes in special ways. Since different program material varies tremendously in dynamic range requirements, these cases involve options which can be chosen at the time that the signal is reduced to 16 bits. The first of these options involves the handling of peak levels in the program material. Many kinds of acoustic music have infrequent peaks in level which, if they are to be preserved, end up setting the overall record level for the rest of the program low enough to accommodate them. Analog tape has a built in soft limiter in the form of tape saturation, which limits the peak extremes in a fairly benign way, or at least one which people are used to. Digital systems, on the other hand, saturate suddenly or run out of bits, which causes a hard clip that is more audible and usually undesirable. For this reason, some form of limiting is frequently used ahead of the final digital quantization.

The peak option, called ìPeak Extensionî, is an instantaneous soft limit which has a one-to-one mapping so that it can be restored in the decoder. The operation is done digitally, so that it has a precise and stable curve. (Figure 6) For levels below the onset of the limit, there is no effect on the signal other than a constant gain factor, which for the current system is a factor of two, allowing the average signal level to be increased as much as 6 dB or 1 bit for material with very high but infrequent peaks. The curve, with its maximum limit of 6 dB, starts its effect very gradually at 9 dB below maximum level. In effect, it squeezes the top 9 dB of the signalís dynamic range into 3 dB on the final 16 bit medium. The shape of the curve was chosen to mimic the effects of tape saturation, and to have minimal audible distortion for music signals when not restored. Of course, as with any limiter, the degree of penetration into the curve before the resulting distortion becomes unacceptable varies with the program material, and must be determined for each case. For material which is limited upstream, or which naturally has little dynamic range, this feature can be turned off.

In the decoder, the peak limit curve is known, and can be expanded using a complimentary operation. There is a larger quantization error for the segments of the waveform which are limited, but in practice, this is not a problem because the system is normally only used for brief peaks. In the interest of compatibility when not decoded, the amount of limiting is restricted to that which does not cause severe audible distortion in single ended mode (encoding only). Research has shown that for many types of music which are complex, this limiting is essentially inaudible [16]. The peak extend feature provides the advantages of a limiter with good characteristics for undecoded playback combined with full restoration of dynamic range when the decoder is used.

Because recordings made with the peak extend feature can have peaks which are as much as 6 dB higher when they are decoded, there is a difference in average level at the output of the D-A converters of 4 to 6 dB between these signals and those encoded without the feature. The converters must be able to reproduce those peaks. Consequently, the decoder requires a 6 dB level adjustment which can be switched in when the peak extend feature is in use, which provides an average playback level when decoded which matches that when not decoded, as well as matching the level to recordings made without the process. This level adjustment can be done in the digital domain or in the analog domain after the D-A conversion.

Low level

The other extreme of signal amplitude which is handled as a special case is that of low average levels. The system uses a form of average signal based low level compression which very gradually raises the gain when the average level drops below a threshold. The gain is determined using a control signal derived from an average of the broad middle

frequencies in the signal, with low and high frequencies attenuated. The main signal itself has only its gain modified, with no alteration of its frequency response, to minimize artifacts for undecoded playback. A control signal is derived for each channel, but the lower gain value of the two is applied to both channels so that there is no left-right image shift for undecoded playback. There is also a hysteresis in the control so that the average level has to change by more than a given factor in order to trigger a change in the compressor gain. The main signal is also delayed long enough so that the gain control can ìlook aheadî to see any coming change in level, so that the system can restore unity gain before the onset of a large signal transient. In this way, normal musical transients do not have grossly distorted leading edges caused by compressor overload, which is typical of many analog compression systems.

The nominal gain is changed in half dB increments and the actual gain ramps logarithmically between designated values to prevent clicking. The parameters for threshold, total available gain increase, and range over which the gain is applied, can be chosen to suit the program material. The normal low level threshold is for an average signal level of 45 dB below full scale and increases the gain only 4 dB over a drop in input level of approximately 20 dB. The maximum gain increase which the current system supports is 7.5 dB, although using this level may produce audible shifts in noise floor for some material when not decoded. The encoder provides several choices of parameters for this option, which may be chosen to suit the program material. Of course, one choice of parameters is to turn the option off, which is appropriate for material such as old analog recordings which have high levels of tape noise.

The nominal gain is sent to the decoder using the hidden code described below. For each gain change, the decoder can perform an exactly complementary adjustment. The decoder has the same ramping algorithm and is instructed when to apply it for each step. This allows the encoder to have any set of threshold parameters for gain changes which are appropriate for the given program material and degree of compatibility desired, and the decoder simply follows orders to produce an exactly timed complementary operation. The hidden code side channel eliminates any issues of compressor mis-tracking.

High frequency dither

The final step in the reduction to 16 bits is to add high frequency weighted dither and round the signal to 16 bit precision. There has been considerable research on dither types, some of which have already been discussed. The benchmark 2 LSB peak-to-peak triangular probability density function (PDF) white dither has approximately a 5 dB noise penalty. The use of very narrow band dither close to the Nyquist frequency has been suggested, but suffers from the problem that the narrower the bandwidth, the less random it becomes, and the higher the level required to eliminate quantization errors. Adding high amplitude high frequency dither may stress the behavior of some analog circuits, causing audible TIM distortion, which is also a problem associated with in-band noise shaping. It may also cause beats with the program material.

The approach which was arrived at after numerous listening tests was to use dither filtered to confine it to the last critical band of hearing, above 16 kHz, at a level sufficient to eliminate quantization errors. Since its bandwidth is fairly large, it has a good random character and does not require very high levels. Stuart [17] and others have suggested that this type of dither should have low audibility, which was confirmed by our research. Rectangular PDF dither is generated using a random number generator and then filtered with a high pass filter to limit it to the frequency range from 16 kHz to Nyquist at 22.05 kHz. The result has approximately a Gaussian PDF and leaves the noise floor flat below 16

kHz, where the critical bands of hearing are associated with tonality. The noise floor below 16 kHz has an almost 5 dB lower level than with the benchmark white triangular PDF dither. The only down side to this form of dither is that it is computationally expensive to produce it. With the cost of processing power decreasing, this is becoming less of a problem.

Hidden Code

As part of the final quantization, a hidden code side channel is inserted into the LSB when it is necessary for the encoder to inform the decoder of any change in the encoding algorithm. It takes the form of a pseudo-random noise encoded bit stream which occupies the least significant bit temporarily, leaving the full 16 bits for the program material most of the time. Normally, the LSB is used for the command function less than five percent of the time, typically only one to two percent for most music. Because the hidden code is present for a small fraction of the time and because it is used as dither for the remaining 15 bits when it is inserted, it is inaudible. This was confirmed experimentally with insertion at several times the normal fraction of time.

The design of the whole system was done with a view to minimizing the bandwidth required in the command stream. The algorithms described above for enhancing the amplitude resolution and dealing with filter artifacts were done in such a way that most of algorithm is known by pre-arrangement between the encoder and decoder, so that little information need be passed between the two for the system to work. The information carried in the hidden code consists of the filter choice used for decimation to the final sampling frequency for each channel, whether the peak extend algorithm is on or not, and the nominal value of the gain for the low level compression algorithm. The command protocol is expandable, so that more information could be added at a later time.

Using the LSB of the program to carry the side channel commands has several advantages. First, it is available without regard to which medium the audio is stored on or transmitted over, as long as the data is preserved. Also, the command data automatically follows the audio data from one storage medium to another over a digital audio transmission medium, without requiring special equipment which "knows about" the commands. Another very important reason for inserting the commands in the audio is that command timing relative to the audio is guaranteed to remain accurate to the sample. This is essential for the complementary decoding operations, such as gain changes, to work properly.

The disadvantage of using the LSB in this way is that the audio data must remain intact. Any operation, such as gain scaling, which changes the digital data values will destroy the code. Of course, changing the gain will also cause the peak extension restoration to mistrack, so destroying the code will protect the audio from being mis-decoded. The normal procedure for manipulating the audio data requires that it be decoded and the full dynamic range restored before gain scaling or filtering are performed. While this may seem to be a disadvantage, it may not be in many cases. It allows one to detect undesired or unintentional manipulation of the data by a CD manufacturing plant or other tampering with the data.

The mechanism which allows insertion of commands only when needed consists of encapsulating the command word and parameter data in a "packet". A synchronizing pattern is prepended to the data and a checksum is appended. The resulting packet is then scrambled using a feedback shift register with a maximal length sequence and inserted serially, one bit per sample, into the LSB of the audio data. The decoder sends the LSB's

of the audio data to a complementary shift register to unscramble the command data. A pattern matching circuit looks for the synchronizing pattern in the output of the de-scrambler, and when it finds it, it attempts to recover a command. If the command has a legal format and the checksum matches, it is registered as a valid packet for that channel.

The arrival of a valid packet for a channel resets a code detect timer for that channel. If both channels have active timers, then code is deemed to be present and the filter select data is considered valid immediately. However, any command data which would effect the level of the signal must match between the two channels in order to take effect. The primary reason for this is to handle the case where an error on one channel destroys the code. In such a case, the decoder will mistrack for a short time until the next command comes along, which is much less audible than a change in gain on only one channel, causing a shift in balance and lateral image movement. If either of the code detect timers times out, then code is deemed not to be present, and all commands are canceled, returning the decode system to its default state. If the conditions on the encoder side are not changing, then command packets are inserted on a regular basis to keep the code detect timers in the decoder active and to update the decoder if one starts playing a selection in the middle of a continuous recording.

Since the decoder is constantly scanning the output of the de-scrambler shift register for valid command packets even when none are present, the possibility exists that there may be a false trigger. For audio generated by the encoder, this possibility is eliminated in the absence of storage and transmission errors by having the encoder scan the LSB of the audio data looking for a match. If a match to the synchronizing pattern is found, the encoder inverts one LSB to destroy it.

Modern digital storage and transmission media incorporate fairly sophisticated error detection and correction systems. Therefore, we felt that only moderate precautions were necessary in this system. The most likely result of an error in the signal is a missed command, which can result in a temporary mis-tracking of the decoding, as mentioned above. Given the low density of command data, and the small changes to the signal which the process uses, these errors are seldom more audible than the error would be in the absence of the process. The chances of a storage error being falsely interpreted as a command are extremely small.

For material not recorded using the encoder, a small probability for a false trigger does exist. Given a moderate length for the scrambling shift register so that its mapping behaves in a noise-like fashion and a choice of synchronizing pattern which avoids patterns likely to appear in audio data with a higher than average probability, susceptibility to false triggers can be made arbitrarily small by increasing the length of the part of the packet requiring a match. In the case of the current system, the combination of the synchronizing pattern with the bit equivalence for all valid commands plus check sum results in a required match equivalent to 39 sequential bits. For a stereo signal, in which a match must occur in both channels within a one second interval and the commands in both channels must specify the same gains, this amounts to an expectation of one event in approximately 150 million years of audio.

The scrambling operation uses a feedback shift register designed for a maximal length sequence in which data taken from taps in the register are added using modulo two arithmetic, equivalent to an 'exclusive or' operation, and fed back to the input of the register. For a given register length there are certain configurations of taps which will produce a sequence of one and zero values at the output that does not repeat until 2N-1 values have emerged, where N is the length of the shift register. This corresponds to the number of possible states of the shift register minus one illegal state, and is called a

maximal length sequence. Such an output sequence has very noise-like properties and, in fact, is the basis of some noise generators. We use the noise-like behavior of the generator to scramble the command signals by adding them modulo two to the input of the shift register, as for example in Figure 7a. This has the advantage that a second similar shift register with taps in the same places but with only feed forward addition modulo two (Figure 7b) will reproduce the original input sequence when fed with the output of the first one. The fact that the decode side has no feedback means that the initialization requirements are limited to having N input samples prior to the beginning of decoding, which means that the decoder will "lock up" very quickly. In this scheme, the presence of a bit error anywhere in the length of a packet plus initialization sequence will completely scramble the data, preventing recovery. However, in practice, this has not been a problem for reasons described above.

Decode operations

The decode operations for the current version of this system have been realized in a monolithic integrated circuit (IC) designed to replace the digital filter used in most playback equipment. The individual aspects of the decode operation have been discussed in conjunction with the descriptions of the encode operations above. What remains is to detail their sequence, which is the reverse of the encode sequence. (See Figure 4.)

The first operation is the extraction of the hidden code from the LSB's of the audio data of both channels, followed by decoding the commands. This operation provides the parameter state information for the rest of the processes. One of the results of code extraction is a code detect signal which indicates the presence of the hidden code in the recording, thus identifying it as having been made using the process.

Following code extraction are the processes complementary to the amplitude modifications of the encoder, which restore the dynamic range reductions made there. These involve the expansion of the instantaneous soft limiting of the signal peaks done in the encoder if that option was on, and the expansion of the low level gain compression based on average signal levels. The hidden code provides commands for exactly complementary gain changes in the later operation, timed to the sample, so that there is no problem with tracking and the operation is transparent.

Finally, the signal is interpolated to twice the sampling frequency using a filter which is complementary to the anti-alias filter used in the encoder. This signal is available as the output of the process, or, in the IC, it can be further interpolated to a four or eight times oversampled signal to drive common D-A converters. The IC also incorporates features designed to improve the performance of multi-bit converters, such as selectable levels of supersonic dither and output timing designed to reduce conversion timing jitter.

Compatibility issues

In any resolution enhancement system involving complementary encode and decode operations, the question of compatibility when not decoded arises. This includes the fundamental question of what compatibility means. It certainly does not mean identity. Our goal in designing this system was to produce one in which any artifacts of the encoding process which might be objectionable when not decoded would be outweighed by

improvements in the overall fidelity for most listening situations. Without success in this area, the system would never be adopted by a significant number of users.

The fundamental strategy for remaining compatible is to alter the signal only at its extremes. For the majority of time for any given program material, the process is not doing anything in the amplitude domain which differs from normal PCM encoding with good high frequency dither. In terms of signal amplitude, high levels can be peak limited in a benign way, which allows the average level to be increased for the whole program, resulting in better resolution. For very low levels, the gain is normally increased slightly, which actually makes up for the lack of low level accuracy in many inexpensive playback systems, in addition to improving the wideband resolution of the format at low levels. It is also an improvement for playback in noisy environments. Because both of these features are controlled by a hidden control channel, the parameters of their use are under the control of the recording/mastering engineer. Unlike older analog systems in which all parameters are fixed, the tradeoff between performance when decoded and artifacts when not decoded is in the hands of the engineer making the recording.

In the time/frequency domain, all changes made only effect the frequency response above approximately 16 kHz while improving envelope distortion effects and settling time for very short events. All encoder filter choices were analyzed with regard to their performance when combined with conventional Nyquist type interpolation filters used by most playback equipment and carefully auditioned with those filters. They were thought to result in a clear improvement in timbre and sense of bandwidth over standard anti-alias filters when played back on conventional equipment.

Finally, the hidden code, which allows the flexibility of system configuration and makes possible exact conjugate decoding operations, is not audible. It has a very noise-like random character, it is used as dither for the remaining 15 bits when it is inserted, and it occupies the LSB for a small percentage of the time. Code packets are only slightly more than one millisecond long and are typically inserted at intervals of several tens of milliseconds. The persistence effect of hearing results in no audible loss of resolution, and the packets themselves are not detectable in the dither even in an otherwise silent signal condition.

Conclusion

A flexible conjugate encode and decode system has been described which addresses limitations of the current digital audio recording standards, such as CD, while remaining compatible for undecoded playback. The system deals with limitations both in the area of amplitude resolution and effective frequency response.

The system is available commercially as the HDCD process, with both encoders and decoders in use. There are hundreds of recordings available which have been made using this process, many of which have received favorable reviews.

Our goal with this paper has been to set forth some of the research which led to the development of the process, as well as details of the process itself, in order to satisfy the technical concerns of users and potential users.

References

1. Moe, Pete "Analog Resurfaces as Mastering Standard", Pro Sound News, June 1996

2. Fielder, Louis D. "Human Auditory Capabilities and Their Consequences in Digital Audio Converter Design" Conference Paper, Audio and Digital Times, Audio Engineering Society, l989, May l4 - l7

3. R. A Greiner and Jeff Eggers, "The Spectral Amplitude Distribution of Selected Compact Discs," J. Audio Eng. Soc., Vol 37, No 4, pp. 246 - 275, (l989, April)

4. Fielder, Louis D. "Dynamic Range in Digital Audio", J. Audio Eng. Soc., Vol. 43, No 5, pp. 322 - 339, (l995, May)

5. T. Oohashi, E. Nishina, N. Kawai, Y. Fuwamoto and H. Imai, "High-Frequency Sound above the Audible Range Affects Brain Electric Activity and Sound Perception," AES 91st Convention, Preprint 3207, October 1991

6. Benade, A. H. "From Instrument to Ear in a Room: Direct or via Recording", J. Audio Eng. Soc., Vol. 33, No 4, pp. 218 - 233, (l985, April)

7. Bekesy, Georg von "Auditory Backward Inhibitions in Concert Halls", Science, Vol. 171, No. 3971, 12 February 1971

8. Sporer, T. et al "Evaluating a Measurement System", J. Audio Eng. Soc., Vol. 43, No 5, pp. 353 - 363, (l995, May)

9. Hudspeth, A. J. "The Hair Cells of the Inner Ear", Scientific American, 1975?

10. Franklin, Deborah, "Crafting Sound from Silence", Science News, Vol. 126, 20 October 1984

11. Evans, Edward F. "Basic Physiology of the Hearing Mechanism", AES 12th International Conference, June 1993

12. Stuart, Robert J. "Noise: Methods for Estimating Detectability and Threshold", J. Audio Eng. Soc., Vol. 42, No 3, pp. 124 - 140, (l994, March)

13. Letowski, Tomasz R. "Anchor Effect in an Optimum Timbre Adjustment", J. Audio Eng. Soc., Vol. 40, No 9, pp. 706 - 710, (l992, September)

14. Trahiotis, Constantine & Robinson, Donald E. "Auditory Psychophysics", Annual Reviews, Psychology, Vol. 30 pp. 31-61

15. Stodolsky, David S. "The Standardization of Monaural Phase", IEEE Transactions - Audio and Electroacoustics, September 1970, pp. 288-298

16. Krause, Manfred & Petersen, H. "How Can the Headroom of Digital Recordings Be Used Optimally?", J. Audio Eng. Soc., Vol. 38, No 11, pp. 857 - 863, (l990, November)

17. Stuart, Robert J. "Estimating the Significance of Errors in Audio Systems", AES 91st Convention, Preprint 3208, October 1991

United States Patent No. 5,479,168 and patent pending in the rest of the world.
* HDCD and High Definition Compatible Digital are registered trademarks of Pacific Microsonics, Inc.

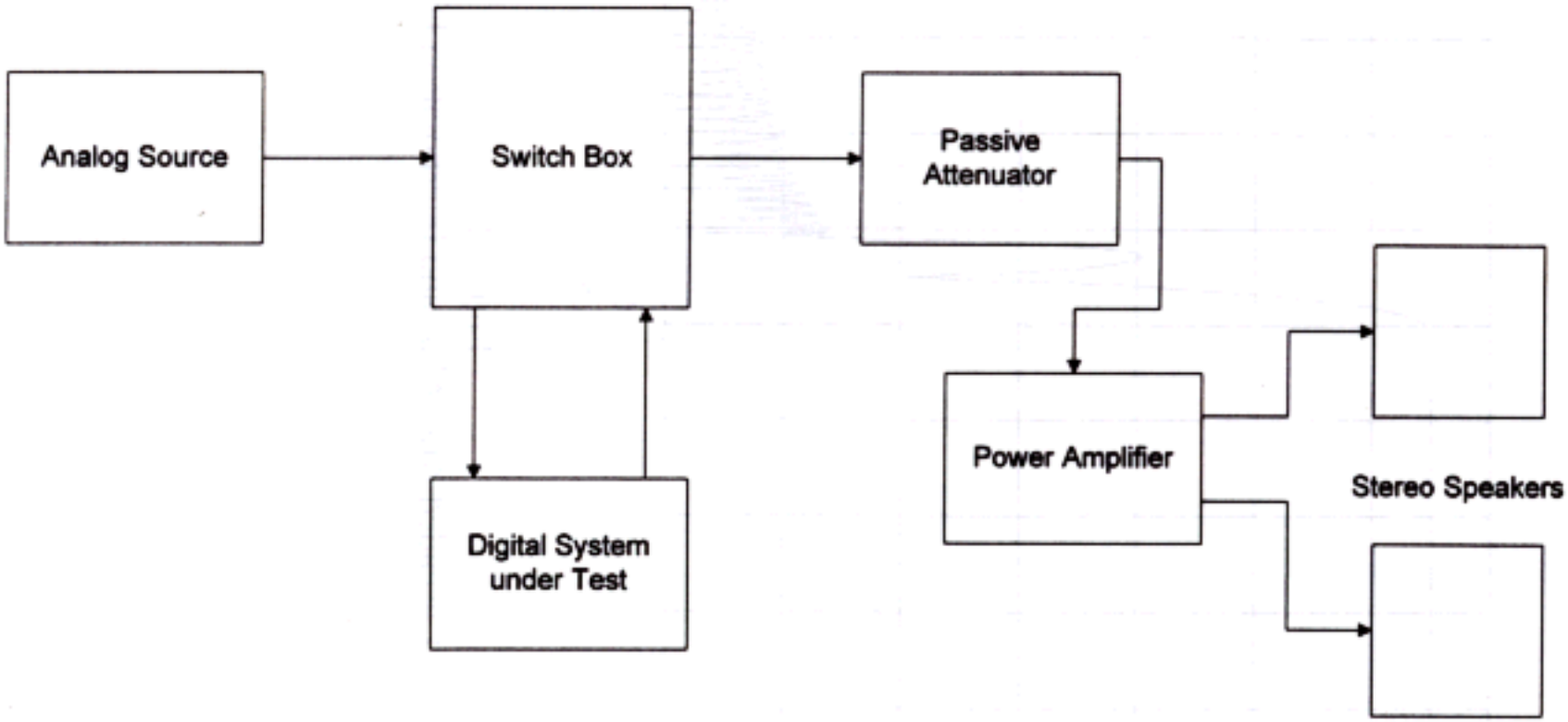

Figure 1
Basic setup for listening tests.

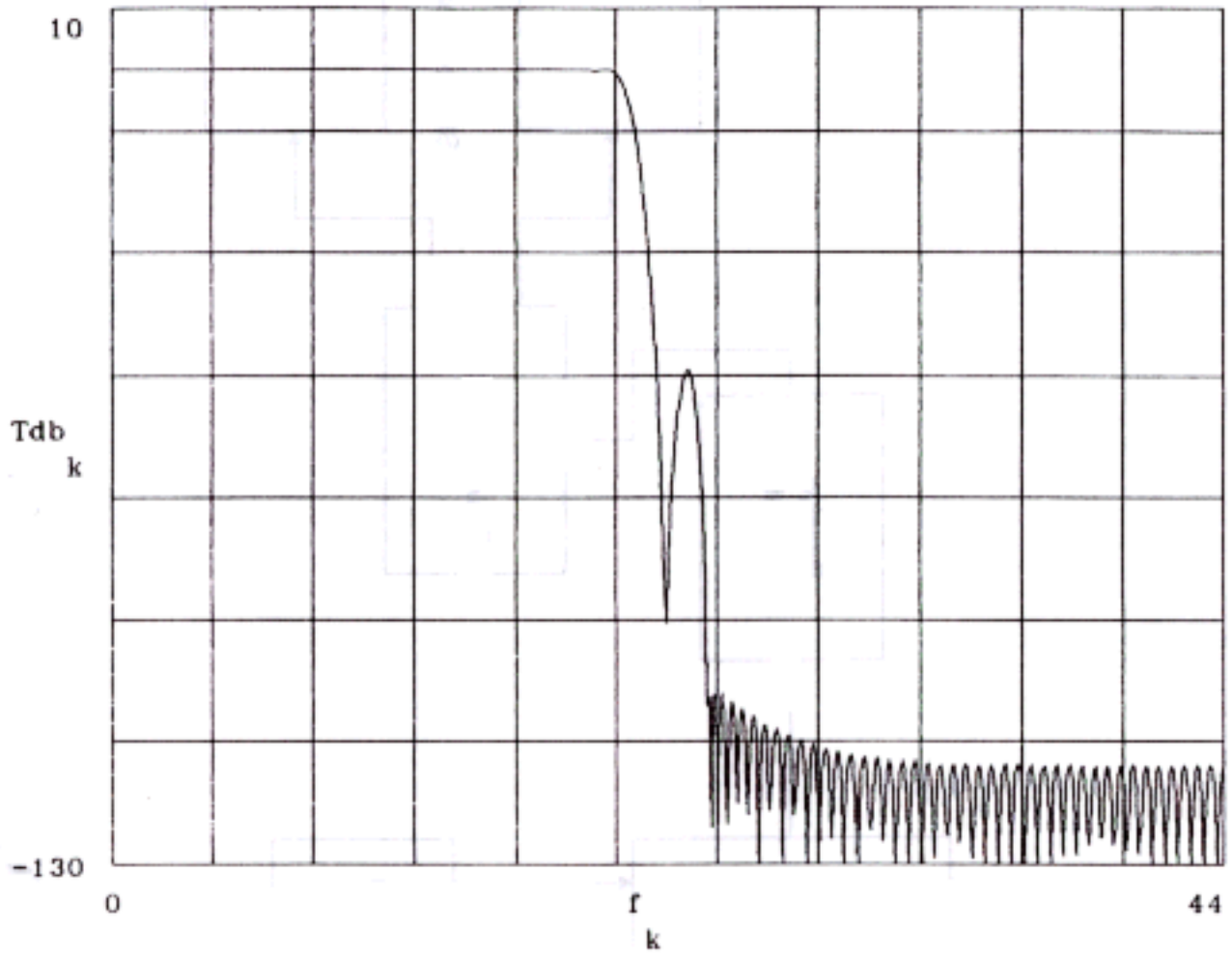

Figure 2
Cascaded Anti-Alias and Interpolation Filters

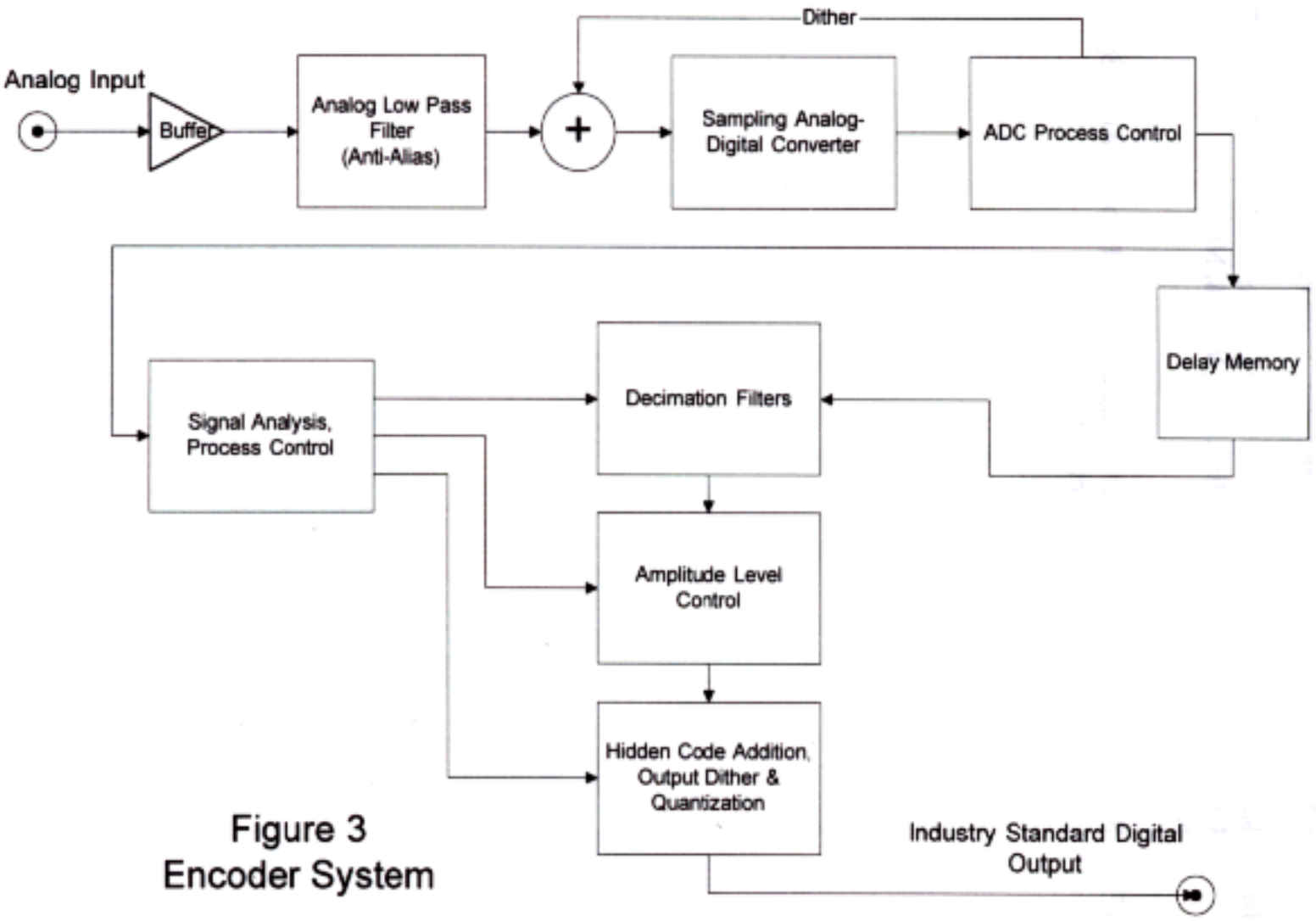

Figure 3
Encoder System

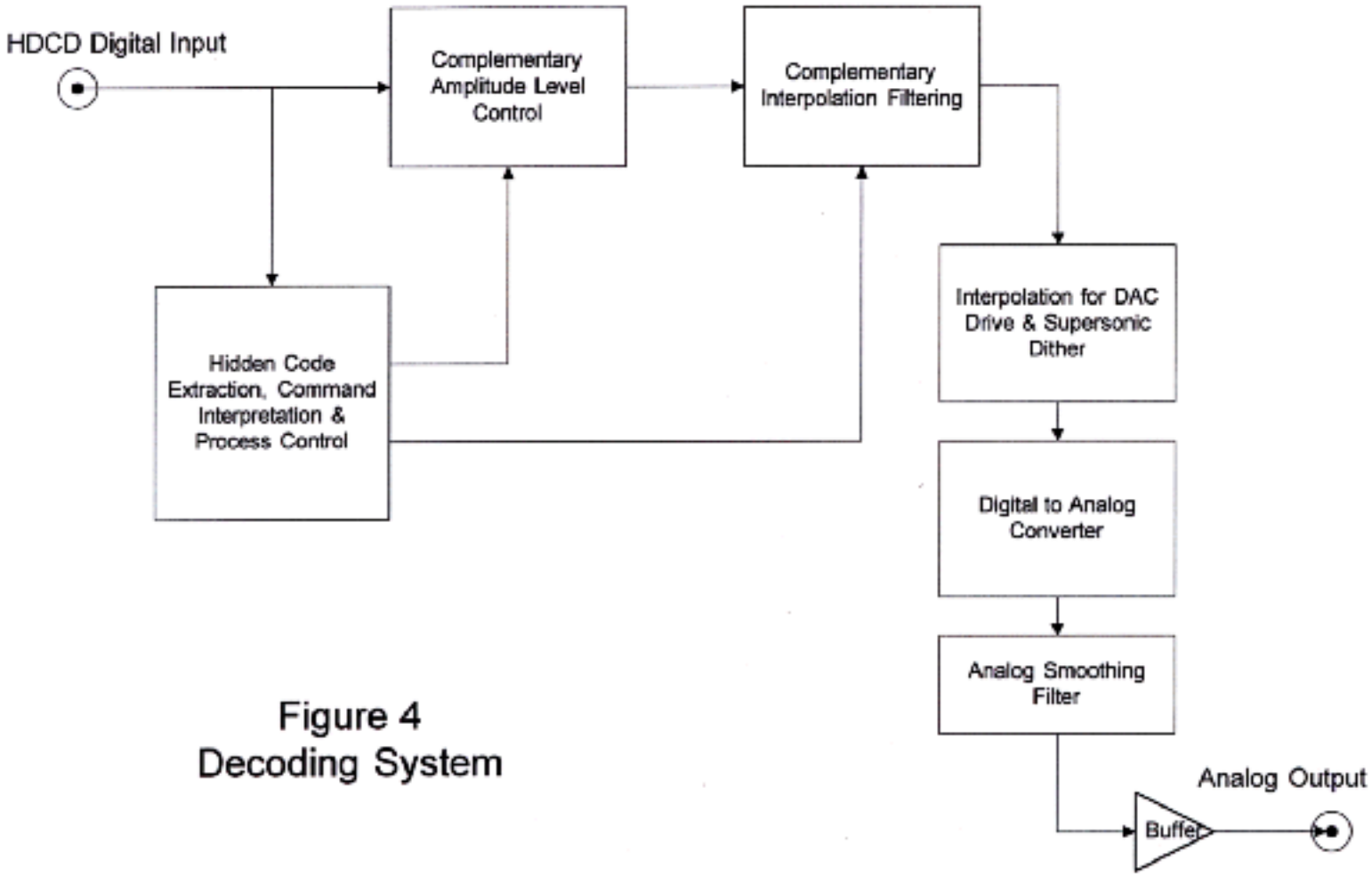

Figure 4
Decoding System

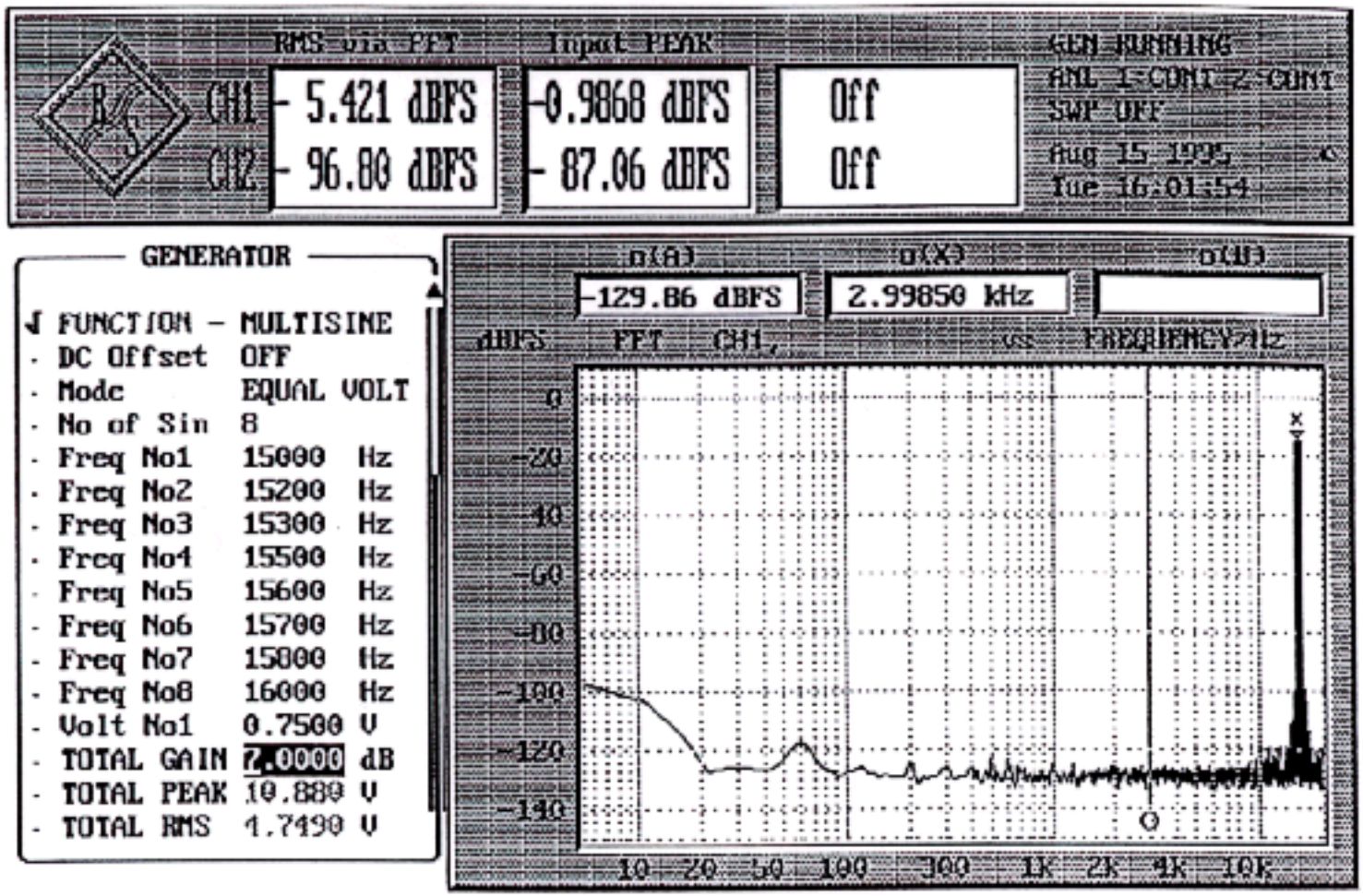

Figure 5

A-D Converter 8 Tone Cluster Test

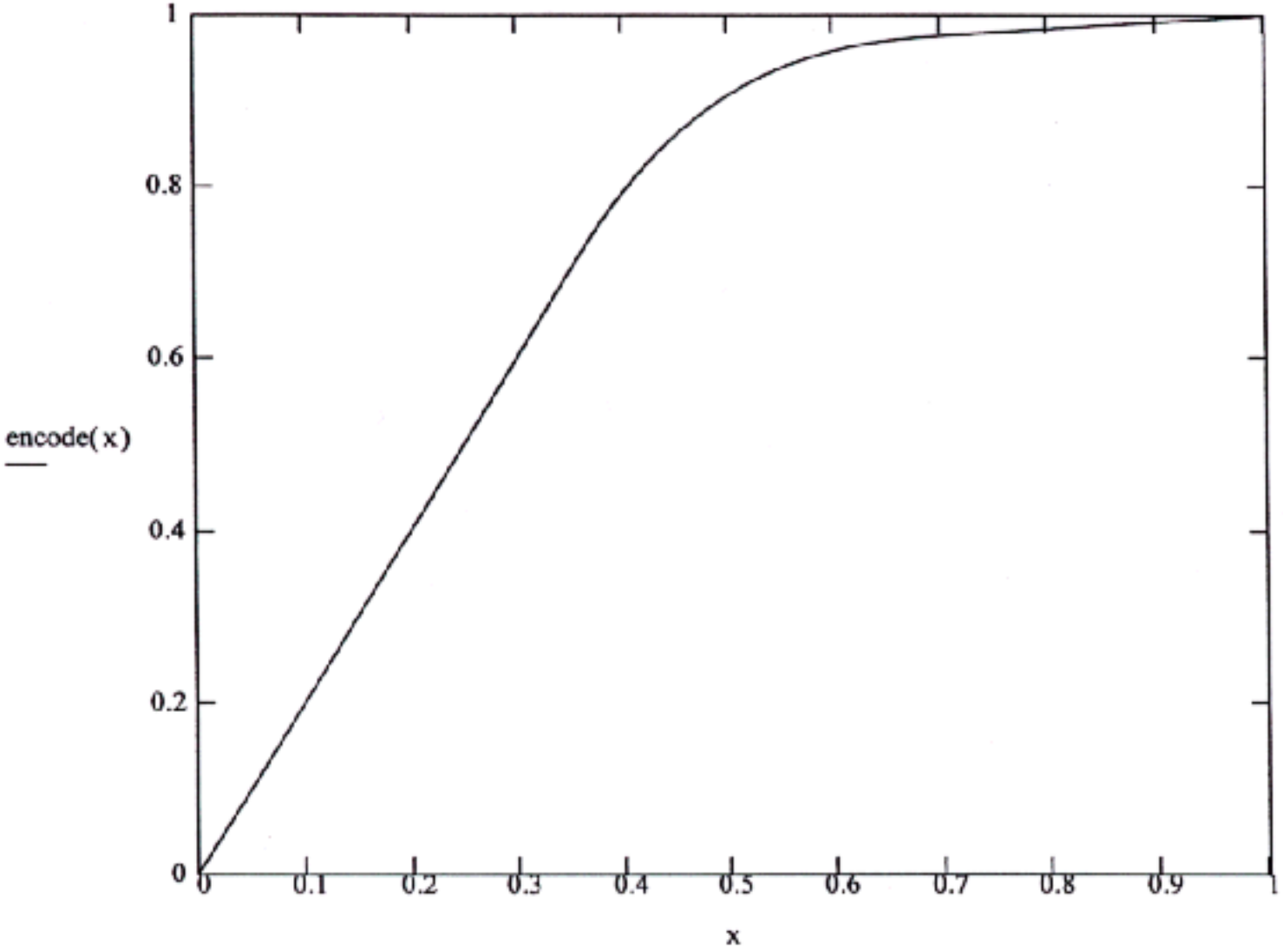

Figure 6 - Peak Extension Limit Curve

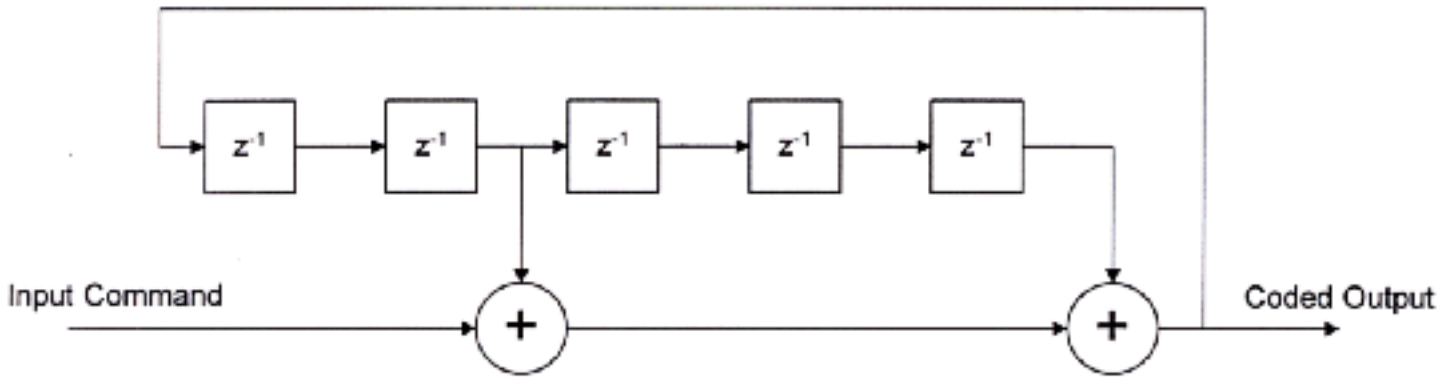

Figure 7a
Example Shift Register Encoder

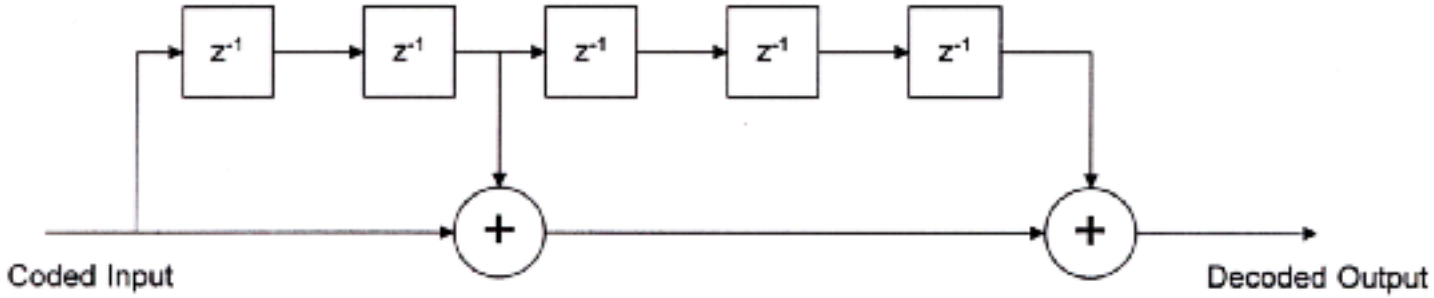

Figure 7b
Example Shift Register Decoder