Digitization of Audio: A Comprehensive Examination of Theory, Implementation, and Current Practice*

BARRY A. BLESSER

Blesser Associates, Raymond, NH 03077

One of the central issues in digitized audio is the conversion between the analog and digital domains. In many ways the conversion process determines the final system quality because a digital system can be made arbitrarily high quality by using additional resolution bits in the digital word. The theory of conversion in terms of perceptual and mathematical degradation is reviewed for different kinds of conversion systems. State-of-the-art implementation questions are also examined in terms of the types of degradations which are produced.

INTRODUCTION: Only since the introduction of a digital delay system [1] in 1971 has the audio profession been living with commercial digital equipment. Since that time, however, there has been a very rapid growth in this field. Currently digital technology has been applied to such traditional analog areas as reverberation [2], tape recording [3]–[5], long line transmission [6], [7], restoration [8], mixing consoles [9], and level indicators [10]. The reader may refer elsewhere [11] for a review of the state of the art in 1976.

The audio industry has now had time to become familiar with some of the special considerations which are particular to digitization. This learning phase has produced more than a few surprises because the classical digital signal-processing field has usually viewed performance and degradation in terms of mathematical quantities. The audio profession, in contrast, is more concerned with the perceptual manifestations. Numeric values for degradation measurements are useful only when they relate to well understood phenomena.

It is probably safe to say that the audio profession, and the author, were unaware of many of the central issues when the first commercial equipment was designed. Some years later, Stockham presented a table of degradations [12]. Because of the extensive work which has been done by numerous scientists and engineers, the understanding of digitization has increased dramatically. It is the goal of this paper to both summarize and analyze the current state of knowledge. The paper is organized in three distinct sections. Section 1 presents a philosophic overview and an introduction that attempts to unify the discussions of all conversion systems in terms of information transformation between the analog and digital domains. Section 2 offers a review of alternative conversion architectures including the classical pulse-code modulation, and it provides a method for comparing the different systems. Section 3 discusses the major design considerations in terms of the technical defects and their perceptual consequences.

Because this paper is written for both the neophyte and the sophisticate, the author must ask for a certain degree of

* Manuscript submitted July 14, 1978; revised August 24, 1978.

tolerance on the reader's part. Some sections will clearly be too simple for the seasoned designer who has had extensive experience in conversion systems, whereas some of the issues will be too complex for the reader who is using this paper as an introduction.

1. GENERALIZED ANALOG-TO-DIGITAL CONVERSION

A digital audio system can be viewed as containing up to five distinct sections: input analog, analog-to-digital conversion, digital processing or digital storage, digital-to-analog conversion, and output analog. Although the two conversion subsections can be designed using any number of conversion techniques, they can all be analyzed as information transformations between the analog and digital domains. This provides a unifying structure for examining conversion without regard to implementation.

1.1 Quantization Error

The analog domain can be viewed as a voltage (or current) which can take on any value between some maximum and some minimum. Because of inherently and arbitrarily large resolution in the analog domain, the voltage 0.5568983, for example, is truly a different voltage than 0.5568984. However, if we add the concept of analog noise, then the resolution cannot be any better than the noise value. In this example, if the noise magnitude was on the order of 0.0001 volt, we might elect to say that the two voltages in our illustration could be considered as having originated from the same "true" value and that the difference can be attributed to noise. The additive noise limits the analog resolution in that the ability to distinguish two voltages is impaired by noise.

In the digital domain all the information is represented by the bit values (one and zero) in a word of n bits. There are 2^n uniquely different words. This is analogous to an ordered series of coins, each of which can have the value head or tail. Three coins, for example, allow one of eight possible words to be specified: HHH, HHT, HTH, HTT, THH, THT, TTH, and TTT. With a 10-bit word there are 1024 possible word values. The word 1001000111 is a uniquely different word from 1001000110 because of the difference in value of the least significant bit (LSB).

In order for each digital word to represent a signal that originated in the analog domain, each word is assigned to a region of the analog signal range. This requires that the analog domain be divided (quantized) into the same number of regions as there are digital words. Consider an analog signal ranging between +1 and −1 volt which is to be mapped on to a 10-bit word. This requires that the 2-volt range be divided into exactly 1024 regions, as illustrated in Fig. 1. Since we made each of the quantization regions the same size, in this example the levels are spaced at intervals of 0.001953 volt. Any voltage between 0.9961 and 0.9981 volt would be assigned a unique word such as 0000000001. This process is repeated until the 1024 regions are each assigned one of the 1024 digital words. Generally, the digital word is viewed as representing the voltage at the center of the quantization interval. This would mean that the digital word 0000000001 is defined as being exactly equivalent to the analog voltage 0.9971. Since all other voltages in this interval are represented by the same word, we can say that the quantization process creates an error, called quantization error. Clearly, adding another bit to the digital word would allow twice as many levels to be specified, and the quantization error would be cut in half.

Increasing the number of bits can reduce this error significantly, but there must always be an error since there are a discrete number of exact analog voltages represented by the digital words but an infinite number of analog voltages. The act of quantization destroys information in the same way that adding analog noise destroys precision. Because there is a good analogy between additive analog noise and quantization error, the error is often called quantization noise. However, the properties of this quantization error are such that it may sound identical to white noise or it may sound much worse. The auditory qualities of the noise are discussed later.

The algorithm for distributing the quantization levels and the technique for mapping these to the digital words determine the type of converter architecture. In the example of Fig. 1, all the levels are equally spaced (linear); but they could have been nonuniformly distributed (nonlinear). The mapping between the intervals and the digital words is monotonic, but it could also have been assigned in terms of magnitude and sign. Some systems use a many-to-one mapping such that different intervals are coded with the same digital word, while others use a one-to-many mapping such that the same interval can be represented as different words.

Because the linear pulse-code modulation (PCM) is the most classical, and because it offers the highest possible quality, we will view this system as the reference when considering other types.

1.2 Discrete Time Sampling

The previous discussion considered the mapping of a single analog voltage into a single digital word; however, the audio signal is time varying. This requires that we partition the continuous time variable into a discrete series of time points. At each of the time points, referred to as sampling times, the analog voltage is converted into a digital word. Thus a sequence of digital words is generated at the same rate as the sampling.

The concepts of *discrete* time and *quantized* amplitude are not the same. Quantization describes the process of collapsing a group of voltages into a single value, whereas discrete means that only specific values of the time variable are being considered. All changes in the analog signal between discrete sampling times are ignored. Fortunately if the analog signal is band-limited relative to the sampling rate (Nyquist rate), the information in the sampled analog values is identical to that contained in the complete unsampled analog signal. Even though the sampling process ignores all signal changes between samples, no information is lost. Time sampling can be a lossless process, whereas amplitude quantization always destroys

information.

Mathematically we can show the lossless nature of the sampling process by considering an analog signal, $a(t)$ and its spectrum $A(f)$ as shown in Fig. 2a. This signal is defined to have a spectrum which is band-limited to f_{max}, that is to say, there is absolutely no energy above this frequency (and by definition none below $-f_{max}$). We can think of the sampling signal $s(t)$ as being composed of a series of pulses appearing at the sampling rate. This is shown in Fig. 2b along with its spectrum $S(f)$. Sampling is equivalent to a multiplication of $a(t)$ and $s(t)$ since the sampling signal only preserves the information in $a(t)$ at times which are multiples of T.

The spectral manipulation corresponding to multiplication of two time-domain signals is convolution as shown in Fig. 2c. Observe that the spectrum of the sampled analog signal contains the same spectrum as the original unsampled signal, except that it is repeated at multiples of the sampling rate. If this signal were low-pass filtered, the resulting spectrum would be identical to that of the original. It is the low-pass which is used to convert the sampled information back into a continuous analog signal.

We must also note that the ability to recover the original is predicated on the fact that the multiple versions of the spectrum which are repeated at f_s (sampling rate) do not overlap. The spectrum at the origin spans the frequency region from $-f_{max}$ to $+f_{max}$; the second version of this centered at f_s spans the region from $f_s - f_{max}$ to $f_s + f_{max}$. In order that there is no overlap, $f_s - f_{max}$ must be higher than f_{max}. We can restate this condition as f_{max} must be less than $0.5 \times f_s$. The limiting signal frequency is sometimes referred to as the Nyquist frequency.[1]

An example where this condition has been violated is shown in Fig. 3. Observe that the low-pass operation will not recover the original signal. In fact, no technique is available to recover the original signal if the sampling rate is less than twice the highest frequency of the analog signal. We can describe this process in other terms by considering an input signal sine wave at frequency f_1. The sampling process creates new frequencies at f_i, $f_s + f_i$, $f_s - f_i$, $2f_s + f_i$, $2f_s - f_i$, $kf_s + f_i$ where k is any integer. With a 50-kHz sampling frequency, a signal frequency of 20 kHz will create components at 20 kHz and 30 kHz, etc. However, a 30-kHz input will also create components at these same frequencies. Thus the presence of components at 20 and 30 kHz means that the input contained energy at either 20 or 30 kHz or both. If we reinstate the band-

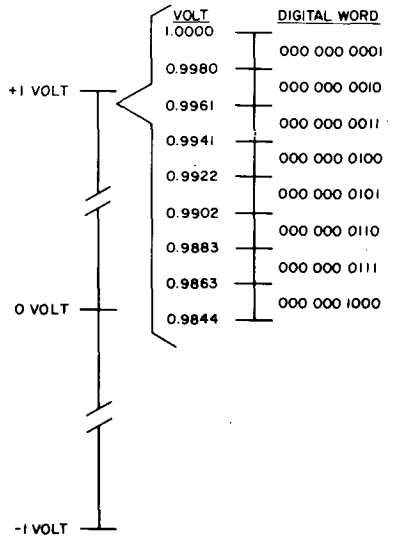


Fig. 1. The analog range from +1 to −1 volt divided into 1024 quantization intervals. Each quantization level is separated by 0.001953 volt, and the interval between two levels is assigned a unique digital word.

[1] The Nyquist rate is the required sampling frequency for a given audio signal, whereas, the Nyquist frequency is the maximum frequency of an audio signal for a given sampling rate. The Nyquist rate is twice the Nyquist frequency.

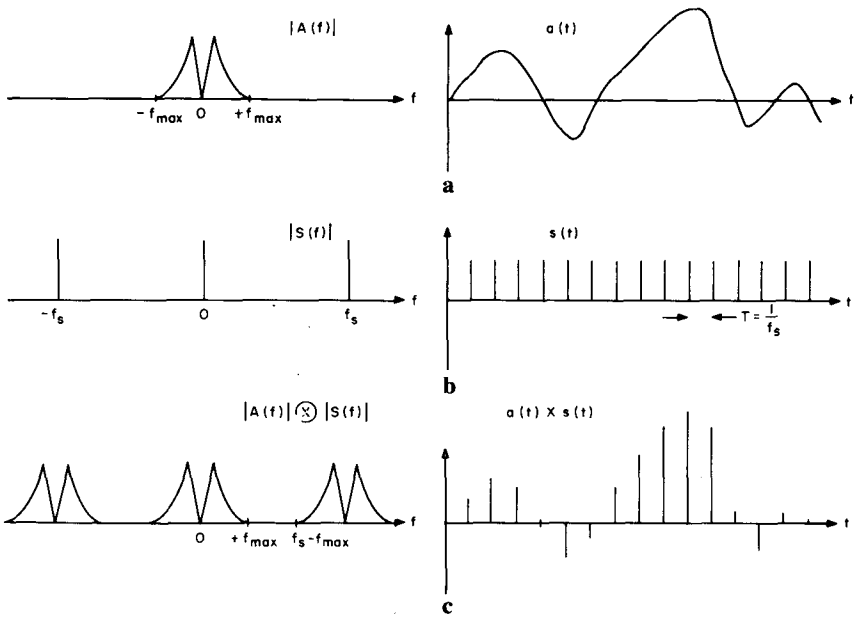


Fig. 2. Time and frequency domain signals in the sampling process. **a.** Input signal $a(t)$ and its spectrum $A(f)$. **b.** Sampling signal $s(t)$ and its spectrum $S(f)$. **c.** Sampled input signal and its spectrum.

limitation conditions, this requires that there be no energy in the original above 25 kHz; then the components can only have appeared from a 20-kHz input.

The only way that the analog signal can be band-limited is by the use of a very sharp low-pass filter before the sampling process. It is therefore the low-pass filtering which destroys information (bandwidth reduction) rather than the sampling process. This is preferable since the low-pass filter merely removes higher frequency components above the Nyquist frequency, whereas the sampling process would generate new frequencies. Specifically, if we allowed a 49-kHz component to enter a 50-kHz sampling process, then a 1-kHz tone would be created.

1.3 Complete Conversion System

A complete digitized audio system is shown in Fig. 4. The incoming analog signal is low-passed with a very sharp filter to restrict the bandwidth to a frequency below the Nyquist frequency. This signal is then sampled, and each sample is held to allow the analog-to-digital (A/D) converter time to convert the information into a digital word. Once in the digital domain, the digital processor can perform any number of functions such as delay, transmission, storage, filtering, compression, reverberation, etc. At the output, the reverse process takes place. A sequence of digital words is converted to a discrete series of analog voltages by the digital-to-analog (D/A) converter. An output low-pass filter smoothes the discrete analog samples to produce a continuous-time waveform.

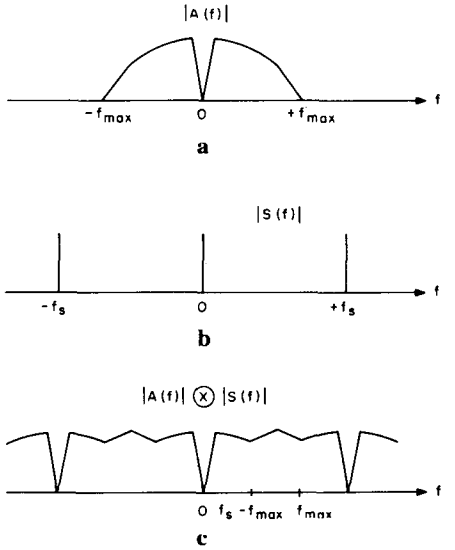


Fig. 3. Spectra of an input signal (**a**), sampling signal (**b**), and sampled input signal (**c**) which is similar to that in Fig. 2. The spectrum of the input signal is not band-limited to the Nyquist frequency, and the low-passed sampled output is not identical to the input signal.

If we ignore technological imperfections, the only sources of degradation are the low-pass filtering and the quantization process at the input. It is not the digitization process which creates the degradation: a band-limited, time-sampled, quantized *analog* signal has the identical information as the sequence of digital words.

1.4 Signal-to-Quantization Error Ratio

One of the important measures of quality for a digital conversion system is the ratio of the maximum signal to the quantization error. This ratio is a function of the number of bits in the conversion. The quantization error originates from the fact that any analog voltage in a quantization interval is represented by the center voltage. Thus the maximum quantization error occurs with an analog signal which is at the edge of a quantization interval; this produces a peak error of one half the quantization interval size. The quantization error becomes statistically random from sample to sample when the analog signal is high level and spectrally wide band. The error has an equal probability of being any particular value between $+ Q/2$ and $- Q/2$, where Q is the size of the quantization interval. This is represented mathematically as the probability density function which is shown in Fig. 5a. Because there is no predictive relationship between the error of one sample and the error of the next, the spectrum of the error is flat with equal energy at all frequencies. This energy density function is shown in Fig. 5b. We must emphasize that these conclusions are only valid for high-level complex input signals which result in independent and uncorrelated quantization errors.

When these assumptions are valid, the quantization noise has the same auditory perceptual qualities as analog additive white noise. For this reason the error is often referred to as quantization noise. There are, however, many types of audio signals that violate this assumption, and there are some technical imperfections which also result in more complex noise patterns. Nevertheless, we can calculate a figure of merit for the conversion system based on the maximum sine-wave signal (below clipping) to the rms value of the quantization error.

For a signal quantized with an n-bit ideal converter, the peak value of the maximum signal will be $2^{n-1} Q$, that is, half the 2^n intervals can be used for each polarity. The maximum rms sine wave can thus be calculated as

$$V_{\text{signal}}\,[\text{rms}] = \frac{Q\,2^{n-1}}{\sqrt{2}} \qquad (1)$$

Similarly, we can calculate the energy in the quantization error signal. This is derived by taking the energy for a

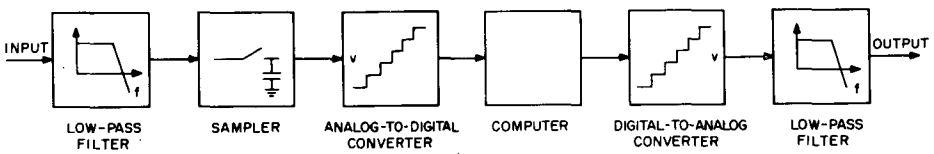


Fig. 4. Block diagram of complete digitization audio system containing low-pass filter, time sampler, analog-to-digital converter, digital processing, digital-to-analog converter (with implied hold circuit), and output low-pass filter.

given error X and multiplying this energy by the probability of that value of X occurring, $\Pr(X)\,dX$. Summing (integrating) for all possible values of error gives the expression

$$\text{energy}_{\text{noise}} = \int_{-Q/2}^{+Q/2} X^2 \Pr(X)\,dX \tag{2}$$

which becomes

$$V_{\text{noise}}\,[\text{rms}] = \frac{Q}{\sqrt{12}}. \tag{3}$$

This gives the well-known result that the signal-to-noise ratio (SNR) can be expressed as

$$\text{SNR} = \sqrt{1.5}\quad 2^n \tag{4}$$

which becomes the following, in decibels,

$$\text{SNR}\,[\text{dB}] = 6.02n + 1.76. \tag{5}$$

Each bit contributes 6 dB to the system performance.

1.5 Quantization Noise

With a typical audio program such as music and speech, the assumptions which led us to the conclusion that the quantization error could be considered to be white noise are generally valid. There are, however, some very important exceptions which can make the digital systems sound much worse than its degradation measurements would indicate. We can clearly demonstrate this by considering a low-frequency sine wave whose amplitude (peak) is slightly less than the size of a quantization interval. Let us also assume that the dc reference is centered exactly at a quantization level. Since the analog signal only crosses one level, the quantization results in one of two possible words depending on the sign of the analog signal. The signal represented by the digital word is a square wave; this suggests that the quantization process is equivalent to a hard limiter.

Such a nonlinear operation is better described in terms of the distortion that it produces rather than in terms of adding quantization noise. The difference between the input signal (a sine wave) and the output signal (a square wave) is the odd harmonics. There is no noise in the analog sense. Not only is the energy density spectrum of the error coherent, but the probability density function is no longer flat. For the low-level, low-frequency sine wave, the quantization process produces a degradation that is analogous to that produced by crossover distortion in a power amplifier. If we consider the case of a high-level bass note fading out, the quantization error will begin as noise and change into distortion.

To further appreciate the quantization action, we can examine other properties of limiters. A limiter will capture the larger of two signals and tend to suppress the weaker [13] which is called capture; and the noise will decrease when a narrow-band signal appears [14] which is called quieting. To be truly equivalent to a limiter, the incoming signal must be small enough such that only one quantization level is crossed. However, even when the signal is somewhat larger, say three levels, these phenomena still manifest themselves, but to a lesser degree.

Observe that any harmonics generated by the quantization process can be at frequencies above the Nyquist frequency since these components are introduced after the initial low-pass filtering. Even though these components are generated by the quantization process, which follows the sampler, they are effectively sampled. The order of quantization and sampling can be reversed since the sampling is a linear time-dependent process and the quantization is a no-memory nonlinear operation. Sampling the output of a limiter or limiting the output of a sampler are indistinguishable since the resulting signal is a sampled square wave. A model of the process is shown in Fig. 6, and the reader is referred elsewhere for an expanded discussion [15].

In this example we consider the case of a 9.333-kHz sine wave at low level in a sampling system with a 30-kHz sampling frequency. The third harmonic of the 9.333 kHz is 29 kHz, which creates a 1-kHz component in the sampled signal. The same kind of aliasing occurs with the fifth harmonic which becomes 13.333 kHz after sampling. Fig. 6c (bottom) shows the spectrum of the error, where the numbers in parentheses refer to the harmonic number of the original sine-wave frequency.

This type of quantization error is neither noise nor distortion in the analog sense, since the new discrete frequencies are created but they are not harmonically related to the input frequency. The auditory system often tolerates relatively high levels of harmonic distortion in reproduction equipment because these components are at similar frequencies as the naturally appearing overtones. Moreover, the fundamental of the musical tone is usually large and can mask the higher harmonics. With the digital system, the new components are not masked since they appear at sensitive frequencies. Specifically, high-frequency tones can create low-frequency error signals.

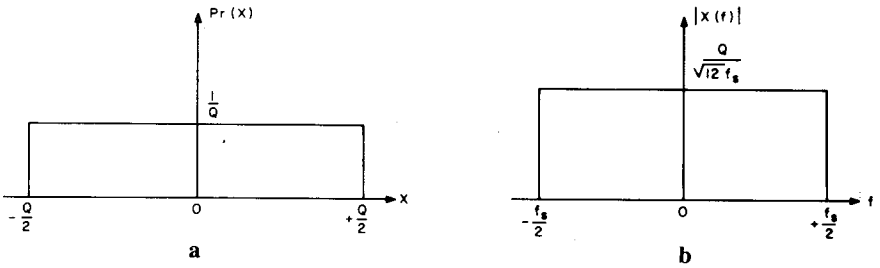


Fig. 5. **a.** The probability that the quantization error will have a particular value between X and $X + dX$ is given by $\Pr(X)\,dX$. **b.** The energy in a particular spectral region between f and $f + df$ is given by $|X(f)|^2 df$.

This kind of error process is often called "granulation" noise since it has the quality, to some listeners, of granular particles being rubbed together. If the harmonics are very close to a multiple of the sampling frequencies, the beat tones drift through the frequency origin to produce a sound which is sometimes called "bird singing."

Although the above discussion has assumed that the input signal was low level, the same phenomena appear for higher level signals when they are very narrow band. The extent to which the error produces noise or complex distortions is a function of the degree of statistical coupling between the successive error samples. This coupling is described as a correlation factor which decreases from 0.5 for low-level sine waves to about 0.01 for high-level signals [16]. However, for high-level signals which are narrow band, the absolute energy in the spectral components remains approximately the same [17]. Only when the signal is broad band is the correlation factor low and is the noise white. In all of its many manifestations, the quantization error has approximately the same energy, only the statistics and perceptual qualities change.

1.6 Benign Quantization Noise

Because of the extreme unpleasantness of the quantization noise under the conditions just discussed, the conversion system must suppress the digital quality of the noise. The most obvious way to achieve this is to reduce the quantization error to a low enough level such that it is completely inaudible regardless of its quality. Increasing the number of bits reduces the noise by 6 dB per bit. However, an excessive number of bits in the digitization results in an uneconomical system. We shall show later that the requirements of the audio profession are sufficiently high that the current technology is just barely able to supply the necessary electronic components. Based on a subjective study using very sensitive piano duets as test signals, Croll implies that an extra 2 bits are required to reduce the granulation noise relative to an equivalent amount of white noise [18].

Because this is a very expensive solution, an alternative approach is that of adding a small amount of analog white noise, referred to as dither, to the input analog signal. This technique was first presented by Roberts [19] in TV picture encoding, but it also has important applications in audio digitization. The effect of dither in the time domain is demonstrated in Fig. 7. Fig. 7**a** shows a low-level sine wave with a small amount of additive dither noise and Fig. 7**b** the resulting quantization. In this example the dither noise has a rectangular probability density function, and its peak-to-peak value is exactly equal to a quantization interval. Regardless of the sine-wave value, the composite signal is always traversing one quantization level. Thus the digitized signal is not a square wave, but a sine wave with noise. When the sine wave is at 0°, the composite signal is between levels B and C, and the conversion produces the word 001. As the sine wave increases, the composite is spending a larger percentage of time above level B; hence the quantized signal is spending more time at word 010. At 90° it is mostly at 010 and only seldom at 001. Observe that the *average* value of the quantized signal can move continuously between two levels. We may think of the quantized signal as containing the original unquantized signal plus noise. This is very much analogous to the signal produced by a pulse-width modulation amplifier. Amplitude changes are achieved by changing the percentage of time spent at fixed levels.

There are many different kinds of dither signals which could be used in order to produce a quantization noise which is independent of the signal. Schuchman [20] examined this subject mathematically to determine the properties of the general class of dither signals. For audio applications, however, the most practical is one which has the following properties: 1) the probability density function is rectangular, and 2) there is no statistical dependence between successive noise values (at sampling time). Such a dither will create a quantization noise which is constant and spectrally white. This makes it equivalent to analog noise: perceptually benign. We should note that this kind of dither signal has the same properties as the quantization noise for high-level broad-band signals.

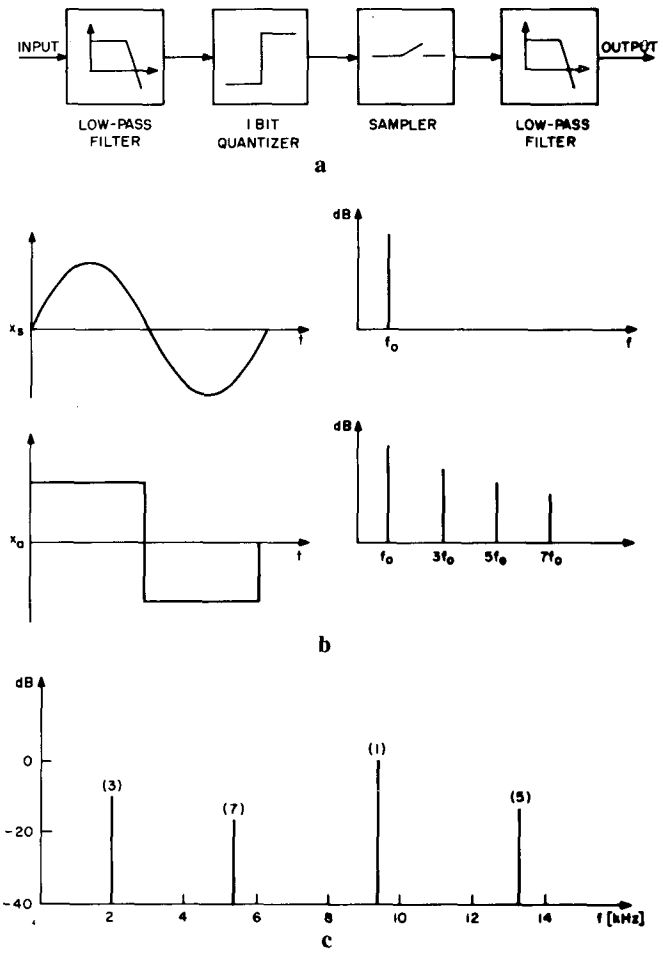


Fig. 6. Illustration of the process by which a low-level sine wave produces granulation noise. A 9.33-kHz sine wave whose amplitude is equal to a quantization interval is applied to a conversion system with a 30-kHz sampling frequency. **a.** Model of the process with the quantization and sampling reversed. **b.** Time and spectral representations for the input signal before and after quantization. **c.** Spectrum of the quantization error. Numbers in parentheses indicate harmonic number of input frequency.

Because the dither signal removes the limiter qualities of the quantizer, no digital artifacts can be observed. The only difference between the resulting noise and that found in typical analog equipment is that the probability density function is triangular rather than Gaussian. For the special case of a dither with a digital subtraction of the rounded dither noise, the probability density is rectangular. Peak or average reading meters calibrated with an rms scale will, however, give slightly different readings.

Generating a stable rectangular density-function noise signal, whose peak-to-peak value is exactly equal to one quantization level, is rather difficult. Most designers ignore this requirement and substitute a small amount of Gaussian noise with a standard deviation approximately equal to a quantization interval. Not only will this produce a small additional degradation in the noise, but it will also create a modest amount of noise modulation. This penalty is rather irrelevant since technical imperfections will result in some noise modulation in any case, even if a rectangular, probability-density-function dither is used.

Empirically, Croll [18] found that he needed a dither noise which was 2 dB greater than the quantization error in order to make the granulation effects inaudible to 50% of his listeners. This produces about a 4-dB total degradation. He also observed that the dither noise could be reduced by 6 dB if a small amount of square wave, at the Nyquist frequency, was also used. Empirically, the square wave was about 0.5 of a quantization interval. In general, the designer can expect something on the order of 1.5-dB degradation in quantization error when an additive dither process is used [21]. It is not uncommon to use the inherent analog noise of the input amplifiers and filters as a source of dither noise. Some specialized feedback conversion systems contain a limit cycle oscillation process which is the equivalent of dither [22].

2. ARCHITECTURE OF CONVERSION SYSTEMS

Although the mapping between the quantization intervals in the analog domain and the binary words in the digital domain is somewhat arbitrary, there are only a few methods that are preferred. The choice of mappings is usually defined by the particular architecture of the conversion system or by a need to perform certain types of arithmetic manipulations on the digital data.

The linear PCM technique is usually considered the designer's choice since it will provide the highest possible audio quality. The analog scale is divided into 2^n equal intervals, and the intervals are assigned digital words in monotonic order. This corresponds to current circuit technology for building A/D and D/A modules[2] and it provides a digital word in standard 2's complement format.[3] For these reasons we will use the PCM system as a reference when discussing the advantages and disadvantages of other architectures.

2.1 PCM Conversion Systems

A standard D/A circuit module, which is the essential element in both the A/D and the D/A converter, is built such that each bit corresponds to an analog switch. Each switch turns a specific current on or off, depending on the bit value. For example, a converter might have the most significant bit (MSB) correspond to a 1-mA current, the next bit to a 0.5-mA current, the next to a 0.25-mA current, etc. The currents are summed and turned into an output voltage. We can think of the analog voltage as being formulated by the following operation:

$$V_{\text{analog}} = (B_1 \times 1) + (B_2 \times 0.5) + (B_3 \times 0.25) + \ldots + (B_i \times 2^{-i}) \quad (6)$$

where B_i is the value (1 or 0) of the ith bit. Because each of the switched currents are factors of 2 apart, any of $2n$ possible analog voltages can be created. The reader is referred elsewhere for additional discussions on circuit implementations of D/A modules [23]–[25]. However, a representative D/A module circuit is shown in Fig. 8. This is a resistive ladder approach with the current in each leg being 0.5 of the preceding leg.

The A/D module is actually built using a D/A module in a feedback configuration. A special digital circuit, called a successive approximation register, attempts to find a digital word which results in the minimum difference

[2] The term "*module*" is used to indicate the actual electronic circuit component as opposed to a section of a block diagram. A D/A conversion section contains a D/A module. The former is a system block, the latter a circuit.

[3] The number system 2's complement refers to a representation of digital numbers. Negative numbers are formed by taking the complement of the positive equivalent and adding a 1 (LSB). The negative of 00010 (2) is thus (11110 − 2). This is illustrated in the following discussion.

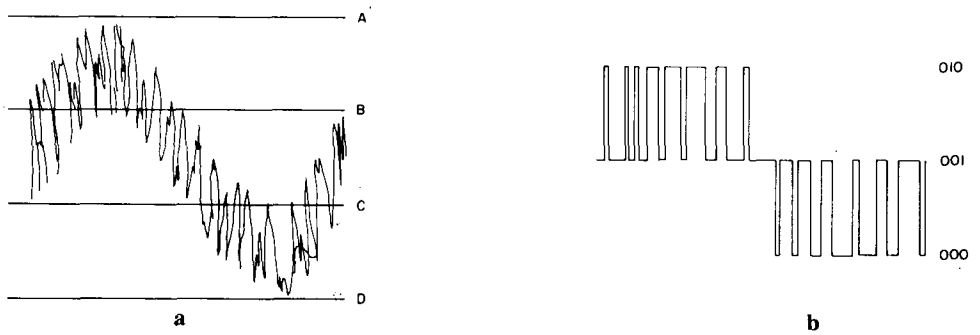


Fig. 7. **a.** Additive rectangular probability-density-function dither noise added to a low-level sine wave which is quantized at levels *A*, *B*, *C*, and *D*. **b.** The resulting digitized signal shows the effect of dither.

between the input analog signal and the D/A module's analog output. The successive approximation is an implementation technique for producing the conversion; it does not indicate the conversion format. Because the D/A module can create any voltage corresponding to a quantization level, the best choice will always be within ± 0.5 quantization intervals of the input signal. An extensive discussion, along with a block diagram in Fig. 23, is presented in Section 3.4.

Since the A/D and D/A systems both contain the same circuit element, the number systems for the digital words must be the same. Moreover, the currents increase monotonically with the digital words. Thus 000000 (all switches open) must correspond to the lowest value, and 111111 must correspond to the highest value. Since the currents are all unipolar, the range of the conversion is centered by offsetting the currents with a fixed half-scale shift. As a result, 111111 is the most positive value and 000000 is the most negative. This would correspond exactly to the standard 2's complement format except for the inversion of the sign bit (MSB). With the 2's complement, the most positive number is 0111111 and the most negative is 100000 with the first bit being the sign bit.

With the 2's complement, positive numbers increase sequentially from 000000 to 011111, whereas negative numbers are formed by taking the complement of the positive number and adding 000001. A negative number and a positive equivalent will always add to 000000. For example, 111100 (−4) is the negative of 000100 (+4). Because there are an equal number of positive and negative words, and because one of those positive words is treated as 0, the negative range is one quantization interval larger than the positive range. Although the ranges can be balanced by offsetting the scales by 0.5 quantization interval size, audio design ignores this issue. The conversion from the D/A format to the 2's complement format is only a matter of an inverter on the sign bit. Frequently, an extra inverted output is provided just for this purpose.

Theoretically, a well designed PCM system using 16 bits and the correct amount of dither noise will produce upward of 96 dB of dynamic range. Even if one allows a small degradation due to circuit imperfections, this kind of performance is usually adequate by even the most severe "golden ears" test. The noise is constant, and there is no inherent degradation process.

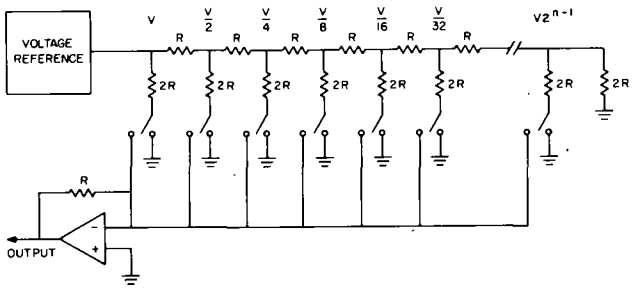


Fig. 8. Circuit diagram of binary resistive ladder configuration for implementing D/A module. Input digital data control the state of switches with each switch (from left to right) producing half as much current as the previous one.

2.2 Specialized Conversion Systems

Even though a linear 16-bit PCM system appears to be ideal, there are many other applications in audio where the quality may, or must, be degraded in order to achieve a significant reduction in conversion cost or processing power. A reduced quality requirement is especially applicable to secondary channel applications, popular or synthetic music, consumer equipment, broadcasting, and the audio portion of television or cinema. Quality compromises may be expected when the design engineer knows that the input or output of the digital system will be degraded by other technologies.

The cost ratio between a 12-bit converter module and a 16-bit version can be as much as a factor of 100. Even with the expected rapid advances in technology, the ratio will probably remain reasonably constant. A conversion architecture which was able to use a 12-D/A module to achieve the equivalent perceptual performance of a 16-bit system would be an extremely important advancement in the audio profession. Moreover, the field of digitized audio is now entering a phase where the cost, rather than the technology, is becoming a controlling variable.

The truly optimized conversion technique has not yet been discovered, but the examples to be presented offer interesting insights into the tradeoffs between system architecture and expected perceptual degradation. In many cases the degradation is extremely minor and easily acceptable for certain applications. The linear PCM system is used as a reference.

At this point we need to make a semantic distinction to avoid possible confusion since many of the specialized techniques themselves contain a linear PCM converter as part of the total system. In this context, we will use the phase "PCM element" to refer to the subsection whereas the phrase "PCM system" is used to refer to the complete system.

2.3 Floating-Point Converters

A floating-point converter produces an output digital word which is made up of two different parts: exponent and mantissa. The name "floating point" derives from computer technology where the decimal point (binary point) is specified by the exponent value, hence floating rather than fixed. Europeans refer to it as "flying comma" since "comma" translates as "point" in representing noninteger numbers. The exponent part of the digital word may be thought of as representing a kind of prescale gain which is applied to the input signal before conversion with a standard PCM element. The block diagram in Fig. 9 shows an implementation to create a floating-point conversion.

The input signal, after being low-pass filtered and sampled, is amplified or attenuated by one of the available gains. The switch-select algorithm chooses the largest gain (minimum attenuation) which does not result in the signal exceeding the PCM element range, at point A. Although the output word from the converter element (mantissa) is standard PCM for the signal at A, the true value of the input is this word divided by the selected gain (exponent).

To illustrate how this kind of floating-point number system can be used to represent analog signals, let us assume that the exponent contains three bits and that the mantissa is a 12-bit code. The three exponent bits can thus specify any one of eight possible gain values. Although not required, the gains are usually integer multiples of decibels, such as 0, 6, 12, 18, 24, etc. The exponent word 011 (decimal 3) thus indicates that the selected gain was 18 dB and that the actual analog signal was ⅛ of the number converted. In this example the three bits allow an additional 42 dB of dynamic range to be added to the basic PCM element range.

The most common exponent factors are 6 dB and 10 dB, where the former is much preferred if the digital word is subject to arithmetic operations. Depending on whether the exponent represents gain or attenuation, the 6-dB increments correspond directly to either a right or a left shift of the mantissa. If the gain selected was +18 dB, then the digital word must be right shifted by three binary places (2^{-3}). The floating-point number 011, 0111001 becomes 0000111001 when converted to the fixed-point representation. This word has ten binary places, or a resolving capability of a 10-bit converter, even though the mantissa was created by a 7-bit element.

With a standard PCM system, low-level signals span only a few quantization levels, producing a relatively high quantization error in relation to the signal. The floating-point system amplifies these signals in order that the signal still spans a large number of quantization levels. The quantization error thus increases and decreases with changes in the analog signal level because it is this signal which changes the gain. The absolute quantization error is constant in the mantissa, but the exponent shifts the mantissa right or left.

This is illustrated in Fig. 10 which shows the location of the quantization levels, as seen by the input signal, for different settings of the gain select. In this example a 5-bit mantissa (positive polarity only) is combined with a 2-bit exponent. Regardless of the exponent value, there are the same number of quantization levels, namely $2^5 = 32$. As the exponent increases in value, the quantization levels become more closely spaced; however, the dynamic range is also reduced by the same factor. An exponent of 001 would mean that the incoming signal was less than $0.5 \times V_{max}$ but greater than $0.25 \times V_{max}$.

In contrast to a PCM system, the floating-point format produces a mapping in which many different digital words can represent the same analog signal. In Fig. 10, points A, B, and C are all the same analog voltage; the words 00,00010, 01,00100, and 10,01000 are identical. The many-to-one mapping property means that the word format is not optimum in terms of bit efficiency. Although not optimum in an information sense, this redundancy is minimal and is usually ignored.

A more important issue is the changes in quantization noise occurring with changes in the signal. We must observe modulation noise as the signal forces the gain into a different state. Because of this property, we must now distinguish between the dynamic range and the signal-to-noise ratio. The dynamic range is defined as the ratio of the maximum signal (sine wave) to the quantization noise in the absence of signal. We adopt the notation SNR_{ns} where ns means noise measured with no signal. This case actually assumes that there is some small signal or an inherent analog noise which excites the low-order bits of the quantization system. For signal-to-noise ratio we will use the notation SNR_{ws} (S) where ws indicates noise measured with signal and the S is the signal level. With a 3-bit exponent and a 10-bit mantissa, the SNR_{ns} is an exceptionally large 102 dB, but the maximum SNR_{ws} is still only 60 dB. The full $SNR_{ws}(S)$ is shown in Fig. 11.

The modulation noise can be extremely disturbing, especially when the analog signal is low frequency. These signals, which can occur with high amplitudes and low loudness, force the switch-select algorithm to use a small gain, that is, the leftmost scale of Fig. 10. The quantiza-

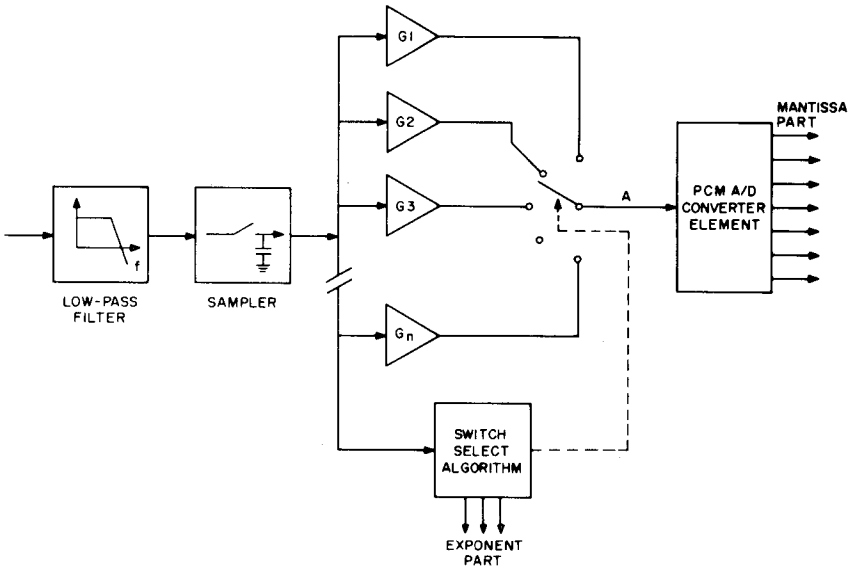


Fig. 9. Block diagram of generalized floating-point conversion system. Conversion is performed with standard PCM converter element (mantissa), and the switch-select algorithm determines the optimum preconversion gain (exponent).

tion noise, being wide band, is not masked by the signal. Moreover, there can be 42 dB of gain change and hence 42 dB of quantization noise modulation. A strong preemphasis which suppresses the low frequencies can be of great help. Nevertheless, a minimum of 12 bits for the mantissa is usually required [26], even though 10 bits can be acceptable under medium-quality conditions [27].

Although the gain-select algorithm must always be able to decrease the gain instantaneously (within a single analog sample), the return to higher gain when the signal drops can be delayed. There are in fact three basic algorithms in common usage for controlling the gain: syllabic, instantaneous, and block. An instantaneous floating-point system has the optimum gain selected anew for each analog sample which is to be converted [28]. Although this would appear to be optimum, there are several disadvantages not the least of which is implementation. It is not possible to align the gains and offsets exactly and we must expect small errors when the gain changes state. Thus when a sine-wave signal is converted, the gain is maximum until the signal reaches an amplitude which corresponds to the maximum for that gain setting. The gain is then decreased to the next setting until its maximum is reached. With a 3-bit exponent and a full-level sine wave, there will be eight different gain changes per quarter cycle. In order to avoid small discontinuities, the gain amplifiers and switches must be extremely carefully aligned and very stable. They should be accurate to within a quantization interval.

Because this kind of accuracy matching is technically equivalent to the problem of adding additional bits to the converter itself, an alternative approach must be used.

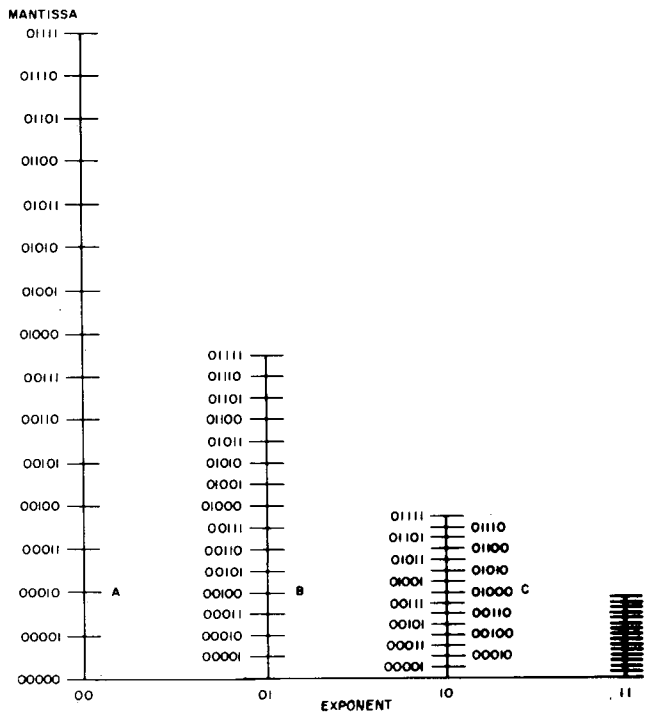


Fig. 10. Quantization levels for a floating-point converter as seen by the input analog signal for different states of the gain select (exponent).

Given the choice, one would always prefer to put additional accuracy into the converter so that the modulation noise problems of floating point are avoided. The solution to this problem is found in the "syllabic" floating-point algorithm. Like the instantaneous select algorithm, the gain can be reduced during a given sample. However, once the gain is reduced it is held at the low value for a predefined time period. If, during this time out, a new peak appears, the waiting period is again extended. Typical time out periods are on the order of 100 to 300 milliseconds. This is very much analogous to an automatic level control with a fast attack time and a slow release time.

Because the gain is effectively following the envelop of the peaks, the gain changes are infrequent and generally not more than a few per second. Moreover, the gain decreases are mostly occurring when the signal increases in level; the very small discontinuities produced by mismatch are masked by the transient change in level. With extremely low-frequency signals, the syllabic system is actually better than the instantaneous. The gain is held constant with a syllabic system, and hence the quantization noise is constant. In contrast, the instantaneous system will produce a quantization noise that follows an almost inaudible signal.

For both converter types the problem of granulation noise is the same as that discussed in Section 1.5 and some kind of dither is required. Because the quantization interval changes size as a function of the signal, the dither noise must be designed such that it tracks the gain changes. Alternatively, the dither noise may be added at point A in Fig. 9. Relative to the input signal, the dither is changing amplitude even though it physically remains constant.

The syllabic floating-point converter is much less expensive to build than a straight PCM system of the same dynamic range. However, the SNR_{ws} is always much worse. The implementation of a high-quality syllabic system is not trivial because the gain amplifiers must be extremely accurate, stable, low noise, and very wide bandwidth. The gain select must be able to determine if an overload will occur extremely rapidly. And the noise in the system must be less than the quantization error on the most sensitive gain state.

Although the motivation for using floating point is usually the economics of manufacture, the floating point also offers an economy of bits per word for a given dynamic range. A format of 3,10 (exponent, mantissa) produces the same dynamic range as 17 bits of PCM. This saving in bits can also be very important in applications which have a high cost associated with storage or transmission. For example, long-line transmission cost is proportional to the number of bits per second being transmitted. Data-rate compression may be extremely important and may dominate any construction costs of the conversion system. Under these conditions, a technique called block floating point is used, as shown in Fig. 12.

The analog data are converted to a digital format using a very-high-quality PCM system or a high-quality syllabic converter, for example, 16 bits or 2,13 bits. As the

converted words are read into a shift register memory, an amplitude monitor determines the largest word. Once the memory is full, the largest possible scale factor is chosen such that the largest word remains below full scale. All the data in the memory are digitally amplified (attenuated) by this scale factor as the words are read out of the memory. During the readout process, the next group of data is read in and the maximum level is again monitored.

Typically, the number of bits at the output is considerably smaller than the bits in the converted signal. The lower-order bits are thus truncated after the scaling process. To illustrate this technique, consider a 16-bit system which is to be block coded into 10 bits; moreover, assume that in a particular block the maximum signal was 0.4 V_{max}. The scale algorithm would thus conclude that all words could be scaled upward by a factor of 2 and the lower five bits would be discarded. If, however, the maximum signal was 0.1 V_{max}, the words could be scaled by a factor of 2^3 and only three bits would be discarded. Very small signal peaks would result in maximum amplification and no lower bits would be truncated.

The system is analogous to a syllabic floating-point converter, except that the peak for the block is used to determine the exponent and the amplification is accomplished digitally. This introduces no artifacts of implementation. Drift, noise, and bandwidth are no longer problems. Moreover, the scale information is valid for the complete block and need be transmitted only once per block. With the syllabic system, the exponent can change (downward) at any sample, and it therefore must be sent with each word. Since the exponent in a block system is sent infrequently, it contributes very little to the required data rates: only the mantissa determines the information capacity required. The SNR_{ws} determines the data rate, but increases in the SNR_{ns} do not change the rate. The BBC [29] has made very effective use of this technique for centralized broadcast distribution.

The cost of a block converter can be very high since it requires a very-high-quality conversion system before the scaling, and it also requires a digital memory with control logic. Of all the systems, this is the most expensive, but it has the highest data rate efficiency.

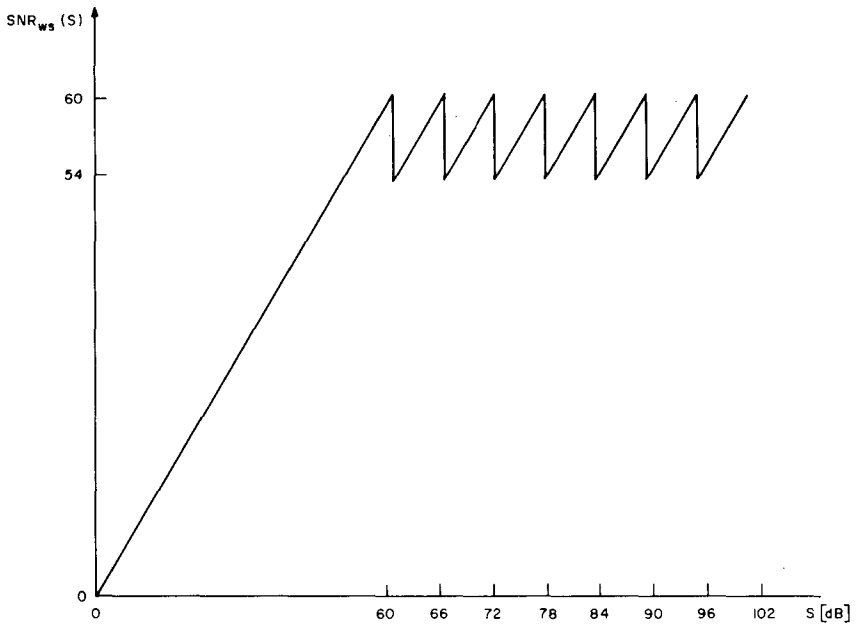


Fig. 11. $SNR_{ws}(S)$ for a floating-point converter with 10-bit mantissa and 3-bit exponent using 6-dB increments.

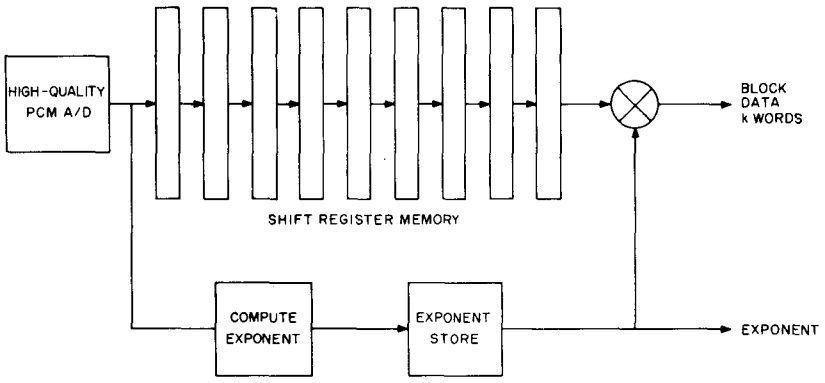


Fig. 12. Block floating-point encoder using linear PCM to feed shift register memory of k words. Exponent for scaling determined as data is read into memory and is then stored for output scaling when data exist from memory. One exponent value is valid for all k words of block.

2.4 Nonlinear Conversion Systems

Any conversion technique which has quantization levels that are not uniformly spaced belongs to the general class of nonlinear quantizers. Using this definition, the instantaneous floating-point system discussed previously is really a member of the nonlinear class since the levels are closer together for small signals. That particular nonlinear scale happens to be convenient for floating-point numbers since it can be easily reconverted to a straight binary 2's complement. The choice of quantization level spacings is arbitrary when one gives up this requirement.

Generally the quantization levels are spaced far apart for large signals and very close together for small signals. This gives a higher dynamic range (SNR_{ns}) and a more constant SNR_{ws}. These systems have found their widest application in voice quality systems since they have extremely low implementation cost [30]. However, experimental audio systems for quality transmission have also been investigated [31].

Regardless of the nonlinear quantization intervals, this kind of system can be modeled as a linear PCM system with a pre- and a post-distorter as shown in Fig. 13. The function $F(x)$ typically has a compression shape such that larger signal increments are required to go from one quantization level to the next. At the decoding part of the system, the reciprocal expansion function $F^{-1}(x)$ is used. There are two common functions in use for designing this kind of companding system, the μ law and the A law.

The μ law is given by

$$F(x) = \text{sgn}(x)\,\frac{\ln(1 + \mu|x|)}{\ln(1 + \mu)} \tag{7}$$

where x is the input, $F(x)$ the output, and μ is the fixed factor which determines the degree of compression. A similar function, the A law, is given by

$$F(x) = \text{sgn}(x)\,\frac{1 + \ln(A|x|)}{1 + \ln(A)} \tag{8}$$

for $1 \geq |x| \geq 1/A$ and

$$F(x) = \text{sgn}(x)\,\frac{A|x|}{1 + \ln(A)}$$

for $1/A \geq |x| \geq 0$, where A is another compression factor.

Increases in either the factor μ or the factor A results in a higher degree of companding: the SNR_{ns} can thus be made arbitrarily high. However, the SNR_{ws} becomes smaller. These systems are analogous to a companding noise reduction system for an analog system with limited dynamic range. Because the analog channel noise is added before the expander, it will also have the same kind of signal-dependent properties as the quantization noise of nonlinear quantizers.

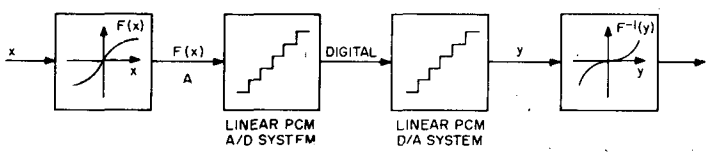


Fig. 13. Model of nonlinear conversion system with nonlinearities $F(x)$ and $F^{-1}(y)$ as companding elements.

It should be noted that the input low-pass filter, which is required to restrict the bandwidth, must precede the nonlinearity even though the sampling takes place afterwards. The nonlinearity *does* generate spectral components above the Nyquist frequency, but these components do *not* produce aliasing. The compensating nonlinearity $F^{-1}(x)$ cancels these components. We must recall that the sampling process creates the aliasing components, not the quantization process. The sampling process, which we have represented as belonging to the linear A/D block in Fig. 13, could be interchanged with the nonlinearity. Because the interchange is mathematically equivalent, the sampler would be operating on the undistorted signal.

If the system were implemented as shown, dither noise could be injected at point A in the block diagram. The quantization levels are constant when viewed from this point in the system. We would, however, expect extensive noise modulation as the signal moved the operating point up and down the reciprocal nonlinearity $F^{-1}(x)$. This is the same noise modulation as we observed in the instantaneous floating-point system.

Unfortunately the block diagram cannot be implemented directly since the nonlinearities are very difficult to match when built with analog circuits. The more conventional approach is to incorporate the nonlinearity directly into the conversion logic. New integrated circuits contain built-in piecewise linear approximations to either the μ or the A law. However, there is no natural point in the system for the introduction of dither noise. In very critical applications, the identical D/A nonlinear element can be used for implementing both the encoding and the decoding. This ensures a perfect match.

The German Post Office has conducted several experiments using the A-law compander to create a 14-bit dynamic range using a 10-bit conversion system [31]. They set $A = 87.7$ to provide 24 dB of gain reduction as the signal decreases from high level to low level. Even though the SNR_{ns} was 84 dB, the SNR_{ws} was a mere 50 dB. This is only marginally adequate [27]. To minimize noise modulation by low-frequency signals, a preemphasis of 13 dB was provided above 2 kHz. In comparison to the BBC block floating-point system with about 10.2 bits, this system may be judged inferior because of larger granulation noise with certain kinds of high-level signals.

2.5 Differential PCM Conversion

Data compression, that is, the reduction in the number of bits needed for a given quality, can also be achieved by digitizing the difference between two neighbor samples of the analog data. Proponents of this technique [32] have observed that low-frequency energy, which can have a large amplitude, has a very small derivative. Thus the dynamic range of the difference signal is small and a lesser number of bits can be used.

Although not obvious, differential encoding is actually a special case of the more general predictive encoding. This class of encoder has a section which creates a prediction for the current input signal based on the *transmitted* past data. The difference between this prediction and the actual

input is transmitted as a correcting signal. Since the prediction was derived from data common to both the encoder and the decoder, the decoder can also create the same prediction which was used by the encoder. The system output is thus the prediction plus the correcting signal. For a more general discussion of predictive encoding, the reader is referred elsewhere [33], [34].

For our example of differential PCM conversion, the prediction is simply the previous input signal and the correction is the difference between past and present signals. This prediction is also available at the decoder since the past value plus the current correction is the present value. This is stored and becomes the past value for the next cycle. Under certain circumstances the approach does provide a very significant reduction in the required converter dynamic range with no apparent degradation in noise performance. However, this is very misleading as we will show. A careful analysis of this approach is presented here because there appears to be a "naturally seductive" quality to this technique in professional audio applications. In the following discussions we will divide the analysis into two distinct topics: maximum signal amplitude capacity as a function of frequency, and noise spectrum.

The block diagram in Fig. 14 shows a differential PCM system which digitizes the difference between two samples of the analog signal. After being low-pass filtered and sampled, the incoming analog signal is delayed and subtracted from itself at the subtractor indicated as point A. This signal $y(t)$ is then quantized. At the decoding part of the system the quantized signal is converted back to analog and added to the previous value of the output at the summer indicated as point B. The signal $y(t)$ is a series of samples which is related to the input $x(t)$ by

$$y(t) = x(t) - x(t - T) \tag{9}$$

where $t = nT$ and T is the sampling period.

An alternative block diagram which achieves the same relationships is shown in Fig. 14**b**. The difference operation is replaced by an analog filter composed of a delay line of value T. This differencing filter has a Laplace transform given by

$$H(s) = 1 - e^{-sT} \tag{10}$$

which has a magnitude frequency response approximately given by

$$|H(f)| \cong 2\pi fT \sqrt{1 - \frac{(\pi fT)^2}{3}} . \tag{11}$$

This filtering function is operating in exactly the same way that a preemphasis filter would change the frequency response. Moreover, the preemphasis is rather extreme since Eq. (11) is almost a pure differentiation for low frequencies. At $f = 0.33 f_{Nyquist}$, the gain through the filter is unity; and for high frequencies the gain is a maximum of 2. The compensation deemphasis filter has the reciprocal characteristic such that the composite gain through the system is unity.

If we assume that the linear PCM conversion element produces spectrally flat quantization noise, then the spectrum of the noise at the output will be determined only by the deemphasis filter. Since the deemphasis is almost a pure integrator, the output noise at $x'(t)$ will have a $1/f$ spectrum. We observe that both the signal range *and* the noise are increased by the same factor at a given frequency. Thus the dynamic range is unaffected.

A predictive encoding implementation of the differential PCM system is shown in Fig. 15**a**. As we shall show, this version has the same preemphasis effect on the signal but the noise spectrum at the output remains spectrally flat. The signal to be quantized $y(t)$ is the difference between the current value $x(t)$ and the past value at the output of the system $x'(t - T)$. Observe that the output $x'(t)$ is recreated at the decoder and at the encoder from the same information: the digital word. Error components in $x'(t)$ cannot become large, especially at low frequency, because such

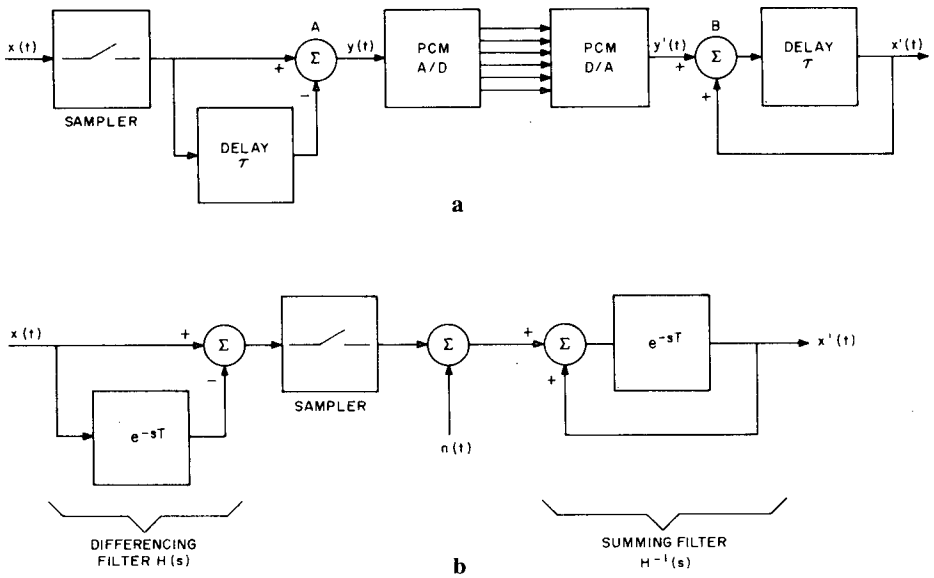


Fig. 14.. **a.** Block diagram of differential PCM system where difference operation is done in an analog format. **b.** Equivalent model of *a* with difference filter represented as preemphasis.

error would be reintroduced into the encoding feedback.

For noise analysis purposes, the modified block diagram in Fig. 15**b** is very useful. We need to compute the observed noise in $x'(t)$ based on the assumption that the quantization process introduced a noise $n(t)$. The Laplace transform for transmission from the noise input port to the output variable $x'(t)$ is given by

$$G(s) = e^{-sT} \tag{12}$$

Except for a phase effect, the gain is unity. Thus the noise appearing at the output of the total system is the same as that which would appear from a normal PCM system, yet the signal range capability has been greatly expanded at low frequencies. The only penalty for this improvement is a reduction in dynamic range of no more than 6 dB for frequencies above $0.33f_{\text{Nyquist}}$.

This is not a contradiction nor are we gaining something for nothing; the information capacity of the digital word is being redistributed. We should note that an extra 60 dB of dynamic range over the band from 0 to 20 Hz contains the same information as 0.1 dB dynamic range from 1 to 13 kHz. If we apply an information metric to Eq. (10), we find that it contains exactly the same information as a linear PCM of the same number of bits.

We can view the signal $x'(t)$ as a computed approximation to the current value $x(t)$ such that the difference is the error signal. Observe that the computed approximation is effectively a zero-order extrapolation: the present value is assumed to be the same as the past value. More complex predictor functions can be used which take into account higher order changes in the past values. We can represent this in the following terms:

$$x'(t) = k_1x'(t - T) + k_2x'(t - 2T) + k_3x'(t - 3T). \tag{13}$$

Such predictors have been optimized for statistically based input signals. A third-order system can reduce the quantization error in the output signal by more than 13 dB over the frequency region from 0 to $0.8f_{\text{Nyquist}}$ [35]. However, the error in the remaining frequency region from 0.8 to $1.0f_{\text{Nyquist}}$ increased by a factor of 60 dB. To be practical, the sampling rate would have to be increased by 25% in order for this excessive noise to be filtered out. The predictor technique is effectively similar to a Taylor extrapolation. For lower frequencies the approximation becomes extremely good using higher order systems. With higher frequencies, however, the approximation can become much worse than no approximation. Empirical results with speech-like signals also gave the same results with about 11 dB of improvement in signal-to-noise ratio [36]. If we are willing to assume that audio program material has dominant low frequencies, then such systems have an interesting appeal. This assumption is generally valid, but the exceptions make the predictive encoding totally unacceptable. The efficiency of this system is based on *a priori* knowledge about the signal. The absence of any assumptions of the signal type, or a violation of the assumptions, makes the system worse than a linear PCM system. Even a modest violation can result in a degradation of 3 dB [36].

In the preceding discussions we have used a linear time-invariant model for the analysis of the signal and noise properties. The quantizer and sampler were modeled as a unity gain transfer with additive quantization noise; this model is not entirely valid. The quantizer is actually a very complex nonlinearity and the sampler produces a frequency folding. The model in Fig. 15**b** could be used for frequencies above f_{Nyquist}, yet such frequencies do not exist in the actual block diagram (Fig. 15**a**) since there is a sampler.

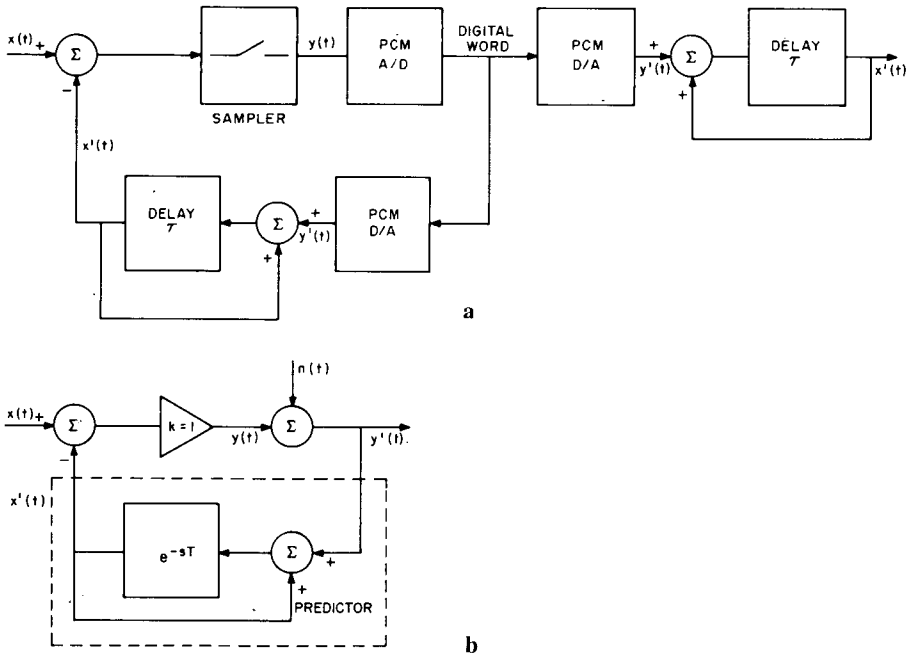


Fig. 15. **a.** Block diagram of differential PCM system where difference operation is performed between current value $x(t)$ and decoder approximation $x'(t)$. Observe that the decoder structure is contained explicitly in the encoder. **b.** Linear time-invariant model of encoder. The $x'(t)$ signal is represented as a prediction of $x(t)$, and the error $y(t)$ is used as a correction feedback.

The nonlinear quantizer can be modeled as an amplitude-dependent linear gain. Consider a sine-wave input to the quantizer with an amplitude of $0.1Q$; the output will be a square wave of amplitude $1.0Q$. The sine-wave component in this output signal is $4Q/\pi$. Thus, for this sine wave the gain of the quantizer is $40/\pi$. With an input signal amplitude of $1.1Q$, the output is a square-like signal which crosses two quantization levels. As the signal becomes larger, the effective gain of the quantizer comes closer to unity.

The Laplace transfer function from the noise input port to the output $x'(t)$ becomes the following rather than that given by Eq. (12):

$$G(s) = \frac{e^{-sT}}{1 + (1 - k)\, e^{-sT}} \tag{14}$$

where k is the effective gain of the quantizer. This system function contains poles in the right half-plane for values of k greater than 2. Thus if we assume that there is no input signal to the system, the value of k is effectively large and oscillations build up until the effective value of k is reduced to 2. The dominant pole occurs at frequencies near the Nyquist frequency. This kind of bounded oscillations is called limit cycle. We would expect the amplitude to be at least $1.0Q$ at the quantizer. However, the noise at $x'(t)$, resulting from limit-cycle oscillations, can be very large since Eq. (14) has a very large gain at this frequency.

Another major problem in our analysis is the omission of the sampling process. Our model has assumed that the feedback loop is closed continuously. However, the sampling process means that only discrete parts of the signal are fed back to the system. A detailed analysis is beyond the scope of this paper, and the reader is referred elsewhere for a more comprehensive examination of sampled-data feedback systems [33]. We must point out that the sampler can result in very large noise amplitude at $x'(t)$ even though the signal at sampling times is much smaller. This is especially true for noise which is narrow band at the Nyquist frequency. Because the zero crossing of this kind of noise occurs at approximately the same time as the sampling, the sampled data can be of much lower amplitude than the actual analog signal. This effect is much more pronounced for higher order systems.

An actual implementation of this kind of system usually uses a leaky integrator for the predictor $x'(t)$ rather than the accumulator (delay plus summing) as shown in the block diagram. Some experimenters have had somewhat better experience with a second-order integrator and zero compensation. An actual optimization is further complicated by the fact that the quantization levels of real converter elements are not perfectly accurate. This tends to introduce a further noise instability as the signal sweeps the operating point to a region where the quantizer intervals are too close or too far apart. The designer of such a system should not forget that the sampler results in a delay of at least T seconds around the feedback loop. This is 180° of phase shift at the Nyquist frequency. Because of these complexities, only the most sophisticated designers should attempt this method unless a significant degradation from the theoretical performance is acceptable. In practice, the author has observed 1.5 to 3 bits of information loss as a result of the limit-cycle oscillations.

Candy [22] has proposed an interesting variation of the classical differential PCM system. In his approach only a few bits are used in a coarse quantizer, but the system is run at a much higher clock frequency. The encoder output is then digitally low-passed to remove the noise at higher frequencies. This averaging process effectively creates the intermediate quantization levels which were not present in the coarse quantizer. Inherent in such a technique is the property that the quantization levels (derived) are stable and reasonably uniformly spaced. This results in a very simple structure with high performance except that the higher sampling rate and digital low-passing produce a significant inefficiency. This tradeoff is discussed in the next section.

2.6 Delta Modulation

The number of bits required to digitize a differential PCM signal was shown to be less than that required by a linear PCM system. The property derives from the fact that the digitization is applied to the error signal (difference) between the input and the approximation. Decreasing the sampling interval makes the prediction more accurate since the input signal cannot change value in a short time interval. The band-limited requirement means that signal changes are well behaved. In the limit with very high sampling frequencies, the error signal becomes vanishingly small. The delta modulation technique uses only a 1-bit digitization of the error signal. It is thus the limiting case of differential PCM.

The attraction of this kind of system, which is illustrated in Fig. 16, is the simplicity of technological implementation. The quantizer is a simple limiter which serves to determine the sign of the error; there is no input low-pass or sampler required since these are an implicit part of the system operation. The entire process can be integrated on a single integrated circuit. It is for this reason that delta modulation has been studied extensively for its application in voice grade communications [38]–[41], and the reader is referred elsewhere for an in-depth treatment of the subject subtleties. Recently, however, there has been a serious attempt to use this technique for higher quality audio [42], [43], but special adaptation algorithms must be added.

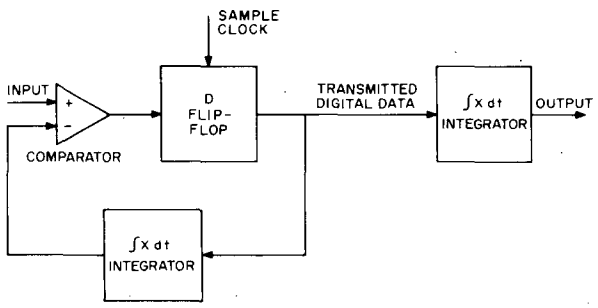


Fig. 16. Delta modulation encoder and decoder. The integrator output is compared to the input signal in order to generate a 1-bit correction pulse at the sample time.

At each sample time the sign of the error is determined. If the error is positive, an incrementing pulse is applied to the integrator in order to increase the signal. Negative errors result in the corresponding negative pulse. The integrator is the past approximation to the input, and the correcting pulse is the differential correcting error required to bring the integrator closer to the current value. This system needs to be analyzed in terms of the maximum signal-handling ability as a function of frequency and in terms of the inherent noise.

We observe that the integrator is rate bound since it cannot increase faster than one correction pulse per period. Thus the maximum derivative of the integrator is given by

$$\frac{dV}{dt} = Qf_s \tag{15}$$

where Q is the step size and f_s the sampling frequency. In order to appreciate the meaning of this equation, let us consider the required size of f_s to produce a 1-kHz sine wave with a peak value equal to $2^{15}Q$. The sampling frequency would have to be about 200 MHz in order to quantize the signal to the same accuracy as a 16-bit linear PCM system. Not only would current technology prevent such a high sampling rate, but the bit-rate efficiency is extremely poor. We should also note that a 10-kHz signal would need to be reduced by 20 dB in order to be encoded since the limit is the derivative. As we noted with differential PCM, the decrease in signal range for high frequencies is equivalent to preemphasis.

The noise behavior is somewhat better than one would expect since the quantization noise energy is spread over a much wider bandwidth. The above illustration would have a noise bandwidth of 100 MHz. The dynamic range of the system can be shown to be given by [44]

$$\mathrm{SNR_{ns}} = \frac{0.2f_s^{1.5}}{f_{in}W^{0.5}} \tag{16}$$

where f_s is the sampling frequency, f_{in} the signal frequency, and W the bandwidth of the system (for noise energy calculation). It should be clear from this equation that the performance of delta modulation is greatly inferior to PCM since a doubling in the number of data bits, by doubling the sampling rate, produces only a 9-dB $\mathrm{SNR_{ns}}$ improvement: 6 dB originates from a doubling of the slope limit, and 3 dB from a decrease in the noise energy in the signal band. For a PCM system, doubling the number of bits produces a doubling in the $\mathrm{SNR_{ns}}$ as expressed in decibels.

The basis for this inefficiency comes from our definition of signal bandwidth. The higher sampling frequency, which was motivated by a need for improved SNR, also provides the ability to digitize out-of-band signals. A sampling frequency of 200 MHz can digitize signals in the region from 20 kHz to 200 MHz, albeit at decreasingly smaller amplitudes. This information capacity is not being used.

As we noted, the implementation is extremely simple and the only major design issue stems from the limit-cycle granulation noise. This noise has a complex behavior and can be perceptually disturbing [45]. Rather than use a single integrator for the predictor signal, a double integrator offers somewhat improved performance. Its dynamic range is given by [44]

$$\mathrm{SNR_{ns}} = \frac{0.026f_s^{2.5}}{f_{in}W^{1.5}}. \tag{17}$$

We observe that the second-order system can yield a 65-dB $\mathrm{SNR_{ns}}$ in comparison to a 50-dB $\mathrm{SNR_{ns}}$ for a 1-kHz sine wave and a 14-kHz noise bandwidth [46]. This improvement is not obtained, however, without incurring additional stability problems [47]. Unlike the first-order system, this has a clipping limit on the second derivative of the input signal.

A nonadaptive delta modulation technique can never be expected to achieve professional level quality, but some consumer applications may be possible. Both the first- and second-order delta-modulation systems show a strong tendency to produce excessive slew-rate limitations for high-frequency signals and hence a high amount of transient intermodulation (TIM) distortion. A very considerable improvement can be achieved if the step size is made adaptive. This is discussed next.

2.7 Adaptive Strategies

Both the delta modulation and differential PCM have constant noise such that the $\mathrm{SNR_{ws}}$ is a maximum only for large signals. We can, however, incorporate a form of floating-point companding into these systems in order to increase the dynamic range. Fig. 17 shows an adaptive delta modulation where the step size can be selected depending on the transmitted data. In the simple form of the step-size algorithm, a series of continuous 1's or continuous 0's indicates that the integrator is in slope limit and the step size is increased. The larger step size allows the integrator to follow rapid signals. Correspondingly, an alternating series of 1's and 0's, which indicates a hunting behavior, is used to reduce the step size. In the simplest embodiment, two similar bits are used to increase the step size by a factor P or two dissimilar bits to decrease it by a factor Q. Extensive experiments indicate that the optimum values are $P = 1.5$ and $PQ = 1$ [48], [49]. Factors of 2, however, do not produce significant degradation.

More complex algorithms using 6 bits of historic information rather than 1 can produce an additional 8 dB of $\mathrm{SNR_{ws}}$ improvement [50]. Because the algorithm is using the bit stream for making its decisions, the decoder has the same information available as the encoder. This allows very complex algorithms to be used. In one variation of the delta-modulation technique, a binary 1 is used to indicate that the signal is continuing in the same direction, and a binary 0 is used to indicate a temporary no change when the basic direction is upward. The reverse is used for a downward direction [42].

The author built one such system at 500 kHz and was able to achieve a 70-dB $\mathrm{SNR_{ws}}$ for a 1-kHz sine wave with a 15-kHz noise bandwidth. The $\mathrm{SNR_{ns}}$ was over 95 dB since this is only dependent on the range of the step size. The $\mathrm{SNR_{ws}}$ did, however, decrease at the rate of 6 dB per octave for higher frequencies. The noise associated with

high-frequency signal transients was effectively masked; only high-frequency pure tones produced excessive noise.

One of the major problems with this technique is the requirement that the decoder synchronize with the encoder after a transmission-bit error has occurred. A single bit error could require as much as 10 ms of 500-kHz data in order to resynchronize. During this time the decoder produces very complex error signals. Unfortunately, almost all the literature on this subject concerns speech signals.

An analogous approach can be used for a differential PCM system which changes its quantization interval based on the previous digitized result. For example, with a 4-bit system, the highest and lowest quantization levels can be used to indicate that the signal is beyond the conversion range and that the quantization interval should be increased. Correspondingly, when the conversion results in a digitized value which is in the center of the converter range, the quantization interval can be reduced. Because there is more than 1 bit in the conversion, the adaptation is much more straightforward. The adaptive differential system has been used extensively for speech [49], [51]. However, we may only infer its performance for music. A 4-bit system gave a 20-dB SNR_{ws} for an 800-Hz sine wave with a 2.8-kHz bandwidth. Using these data extrapolated for music, we can expect that an 8-bit system at 50 kHz sampling frequency should yield a 64-dB SNR_{ws} at 800-Hz sine wave over a 15-kHz bandwidth. This compares favorably with a 10-bit floating-point PCM system sampled at 35 kHz. More sophisticated algorithms should probably give still better performance.

Another subjective test showed that an adaptive PCM system produces 1.5 bit less noise in comparison to an instantaneous nonlinear companding PCM system. Perceptually, the improvement was approximately 2.5 bits since a 4-bit adaptive system was judged better than a 6-bit nonlinear companded system [52]. It is uncertain if these results would hold for music signals.

2.8 Critical Summary

All conversion architectures have an information limit which is determined only by the number of bits per second being transmitted. This information can be distributed to optimize a specified class of input signals with some *a priori* class definition. The PCM reference system for example is optimum for signals that have an equal probability of spanning the complete amplitude frequency space. That is, any frequency is as likely to appear at any amplitude as any others.

The floating-point systems assume that high-level signals and low-level signals come in groups. The adaptive systems assume that the spectral weighting is dominated by lower frequency energies since the adaptation algorithms are not fast responding.

The designer must, therefore, make the judgment about the class of signals that he wishes to digitize. If he makes the assumption that there is no *a priori* information, as would be the case with synthetic music, then he must choose the linear PCM approach since its limits correspond to the amplitude-measuring indicators of a typical audio frequency.

One might speculate that the optimum conversion system would be one that was perceptually matched to the human auditory system since quantization error would be designed for minimum perceptual degradation [11]. Such a system would make extensive use of spectral masking [53], as well as forward and backward temporal masking [54], [55].

Until such time as this subject has been investigated, we recommend a 16-bit linear PCM system when the highest quality is required. When other constraints such as cost limit the choice, however, other types of architectures should be considered.

3. PERFORMANCE OF REAL CONVERSION SYSTEMS

All the previous discussions considered the performance of idealized conversion techniques where the degradation was an inherent part of the system structure. In this section we will consider degradations introduced by component imperfections and, when possible, we will relate them to their perceptual manifestations.

It is not possible to discuss the implementation of each of the conversion techniques because of space limitations; however, we will consider the 16-bit PCM linear conversion technique at a 50-kHz Nyquist rate as an example. This is a preferred choice since this technique is able to yield the optimum performance; moreover, this converter system is at the state of the art and is close to some of the theoretical limitation of signal processing. The reader can extrapolate the issues to his particular conversion technique since most of the phenomena are applicable.

The audio engineer should not expect to design many of the conversion system's circuits since this is a very specialized technology. Nevertheless, he must understand the details in order to specify and evaluate the system. Only by understanding the degradation mechanisms can the designer create the required testing procedure.

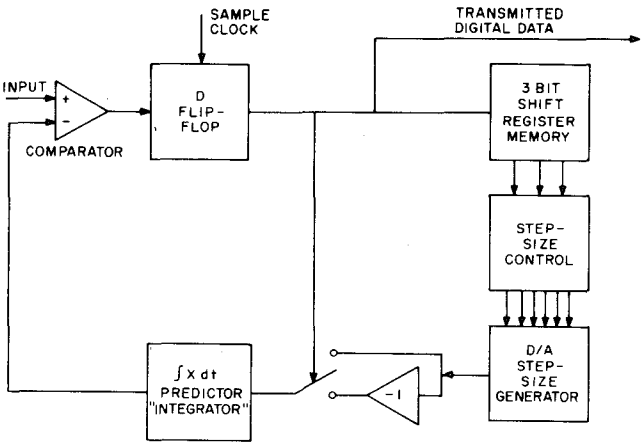


Fig. 17. Encoder part of an adaptive delta-modulation system. Three bits of transmitted data are stored in the shift register memory for use by the step-size control algorithm. The D/A converter element is used to create the actual analog step size. Either a positive or a negative correcting step is applied to the integrator depending on the sign of the comparator.

3.1 Quantization Accuracy

The D/A converter element is clearly one of the most important subsections of the conversion system since it does the actual mapping from the digital word and the analog voltage. Moreover, this element is also used explicitly in the A/D converter.

The technology involved in the design and manufacture of the D/A elements is so specialized that the audio engineer *must* buy the element from firms that have the required technology. Construction involves the use of matched monolithic integrated circuits, laser-trimmed resistors, or ultrastable switches and resistors. Considering that the requirements for high-quality audio digitization are almost at the limits of current technology, these elements must be purchased. The audio engineer may design the A/D element with a purchased D/A element, but this is only recommended for those reasonably sophisticated in high-speed analog signal processing. The main burden of performance thus reduces to the task of carefully reading and interpreting the manufacturer's specifications. To understand the meaning of the specifications in terms of total system performance, we need to review the degradation mechanisms.

Although the quantization levels should be uniformly spaced, there will be some deviations between the ideal and the actual. These deviations are characterized by absolute accuracy, relative accuracy, and monotonicity. For the audio engineer the relevant measure is the size of the quantization intervals rather than absolute accuracy. Typical errors affect the noise performance but not the distortion mechanisms. Fig. 18 shows different kinds of errors which can occur (exaggerated to show their effect). Scale **a** is ideal and is shown for reference. Scale **b** shows errors where the maximum error is no more than ± 0.5 LSB with 1 LSB being the ideal quantization interval size. Note that the interval between A and B can reduce to almost nothing, and that the interval between B and C can become almost 2 LSB. By definition, this error is the largest possible with the quantization levels still being monotonic, that is, a lower voltage level does not cross a higher voltage level.

Scale **c** shows an example where the errors are much larger than 0.5 LSB, approaching 1.5 LSB for the level C, yet the quantization intervals are well behaved. The largest is 1.5 LSB and the smallest is 0.75 LSB. The relative error is thus small even though the absolute error is large. Scale **d** also shows a large absolute error, but this is due to a gain change; all levels are equally spaced.

Because the quantization interval determines the noise, any changes in this interval will result in changes in noise, that is, modulation noise. In particular, low-frequency signals act to sweep the converter over its range, and any changes in the quantization interval will produce a proportional change in the noise. Moreover, the matching of dither noise to the intervals is severely impaired if the levels are not uniform. Since it is impossible to change the dither noise with changes in the interval, the designer must use a dither that is matched to the average interval spacing. When the levels are too far apart, as in B–C of scale **b**, the dither stops functioning and the system noise suddenly drops to zero. However, in the region from A to B the dither spans more than two levels.

Generally a converter element will not be specified as having better than ± 0.5 LSB accuracy. One could, however, order an element which had two more bits than were needed in order to have more accurate spacing on those bits that were used. The basis for this problem is clear: if the converter levels are too accurate, the manufacturer can add additional bits at the bottom of the range to increase the module's performance. Once the error exceeds the 0.5 LSB criterion, the converter element can become nonmonotonic.

All the accuracy burden is carried by the most significant bits (MSB) since these must be accurate to a tolerance determined by the lowest bit. The percentage requirement is extremely high. To contribute to the same error to a given quantization level, the MSB needs to be 2000 times more accurate than the LSB for a 12-bit converter element. Specifically, consider the converter region near the center of the range. When the converter switches between the levels represented by 10000000 and 01111111, all the switches are changing state. In other words, the current produced by the MSB must be 1 LSB

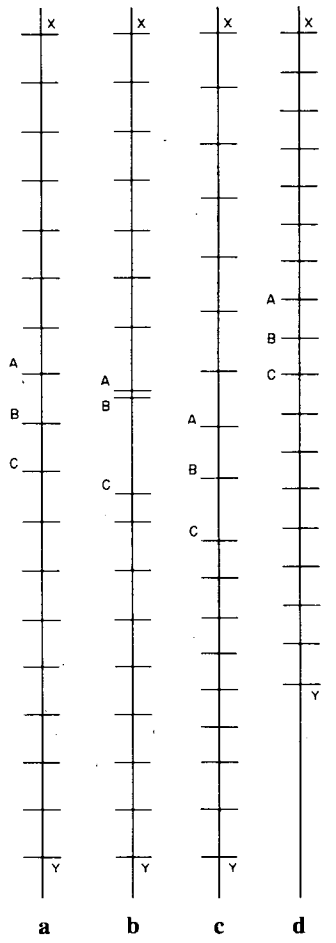


Fig. 18. Examples of quantization levels for a D/A converter element. **a.** The reference scale idealizes. **b.** An example of ±0.5 LSB errors. **c.** Large absolute errors with small relative errors. **d.** Effect of gain error on level locations. All scales are aligned at the peak value mark *X*. Levels A on each of the scales have the same digital word. Similarly for B, C, and Y.

more than the sum of all the currents produced by all the other bits.

We must pay particular attention to this phenomenon since this region is likely to have the largest quantization error; and it is this region which corresponds to the smallest audio signal. The center of the range corresponds to analog 0 volts since the range must be symmetrical about 0. The greatest modulation noise is likely to occur with the smallest signals. This region is also the least stable and the most prone to drift with temperature and time.

Some designers have added a small dc offset to the analog signal at the A/D converter in order to bias the no-signal operating point away from this critical region. If this bias were 10% of the full scale, then the critical region would first be encountered for signals which were −20 dB re full scale rather than for low-level signals. This shift will generally be benign, and the only degradation produced is a 1-dB restriction in the maximum signal due to the induced asymmetry of the range.

From this discussion we can see that the accuracy requirement must be increased by about 1 to 2 bits if modulation noise is to be avoided. A 16-bit converter may produce 96 dB of dynamic range, but if it is not perfect, it will also produce modulation noise. Disabling the lower two bits would remove the modulation by raising the absolute noise floor. That would be analogous to covering modulation noise with a higher amplitude fixed noise source.

In other types of conversion architectures the nonuniform level spacing will have different manifestations. With the differential PCM system the level spacing was an inherent part of the loop stability design in order to optimize the limit-cycle oscillations. These oscillations can change amplitude dramatically when the signal moves the converter into a high-error region. To some extent this can be reduced by sharing the same D/A element for both the encoding and the decoding. When the digital signal is subject to arithmetic processing, no form of shared D/A elements will provide any help since there is no direct relationship between the input signal and the output signal in terms of quantization errors.

Last, we should point out that the modulation noise will be the sum of that produced by the A/D converter element and the D/A element. In practice, an engineer may use a dither signal which is as much as 6 dB larger than that which would be expected. This does degrade the *average* dynamic range, but it also reduces the modulation noise. For the highest quality system, the noise may be sufficiently low that modulation remains inaudible. Thus no efforts need be devoted to a reduction of the modulation. Moreover, the granulation may also be inaudible for quality 16-bit systems.

3.2 Peak Overload

Excessive program level, which exceeds the amplitude range of an audio system, is a familiar problem. In the digital context, however, this problem takes on a special meaning. If the program exceeds the A/D converter's range, the conversion process will convert the signal to the maximum allowable digital word value (or minimum for negative overloads). This is equivalent to clipping *after* the low-pass filter. Those harmonics introduced will be aliased to a new frequency. This mechanism is exactly the same as that discussed with a low-level sine wave in an undithered conversion quantizer (see Section 1.5).

For example, symmetric clipping of a 5.5-kHz sine wave in a 50-kHz sampling system produces a 500-Hz beat tone since the ninth harmonic is 49.5 kHz. There is no mechanism to avoid this except to limit the signal before low-pass filtering. Clipping, which is done prior to the low-pass filter, must have a threshold value which is approximately 3 dB below the nominal 100% if this effect is to be avoided under extreme overload conditions. A square wave of peak 100% becomes a sine wave with a peak of 127% for frequencies above one third the low-pass cut-off frequency. This would be ultrasafe and it is probably unnecessary. Nevertheless, some provision for signal limiting should be provided. It should also be noted that an input signal which is below the maximum level can become a signal with a peak above this level as a result of nonlinear phase shift in the low-pass filter.

The problem of level setting can be avoided if the operator of a digital system has the opportunity to preview the loudest passages in the program material. Under these conditions, the level at the converter is set to about −2 dB for this worst-case example. Uncontrolled inputs present a much more difficult situation.

3.3 Input Low-Pass Filter

The input low-pass filter serves to remove all spectral components above the Nyquist frequency since these would become in-band signals if they were not removed. Realizable filters, in contrast to the ideal, do not produce infinite attenuation in the stop band. It is therefore necessary to create a specification for the filter performance. A finite stop-band attenuation means that if any spectral components are present in the signal, they will create aliasing components in the audio band. At this point in the discussion we need to specify both the required threshold for these components, and we also need to determine the initial level. The audio industry has not yet advanced to the point where these data are available.

We may argue that there are two distinctly different sources of high-frequency energy: 1) those created by technical artifacts such as a tape recorder bias signal or sampling energy from another digital audio system; and 2) those occurring naturally in the program signal such as transients from sharp percussive instruments or very high overtones from certain wind instruments. The specification for the first category needs to be an absolute threshold since the worst case will occur when the program is off. In contrast, the components with natural music are always relative to the main in-band signal; and the main signal provides extensive masking. There is actually a third category for sources of high-frequency energy, namely, synthetic music. We cannot make any *a priori* assumption about the spectral content nor the extent to which masking

will function.

The author has performed some informal experiments in this area in order to gain some insight into the filter specification. With natural music signals it is almost impossible to hear any aliasing components even without any low-pass filter with high sampling frequencies. Only some carefully selected program material can provide a just noticeable difference. However, the author has also observed that the interconnection of two audio digital systems will create extensive disturbance unless the out-of-band signals are extremely low. In fact, there is some indication that these kinds of tones need to be suppressed to the inherent noise floor of the system. Thresholds of −100 dB might be required in the highest quality system.

In order to continue with the discussion, we will assume that the out-of-band energy is typically more than 40 dB below the program maximum, and we will also assume that an attenuation of 40 dB is therefore the minimum required. The current generation of digital audio equipment tends to have stop-band attenuations of 40–60 dB, although some equipment is significantly better and some worse. We expect that the audio industry will discover that only a few applications require excessive attenuation; and these applications will be satisfied with supplemental filters. We should also note that systems which use a much lower sampling frequency may need a more severe specification.

A filter also has a frequency response specification for pass-band ripple. Since this specification is germane to all audio equipment, it will not be discussed here. Numbers such as± 0.1 dB are achievable if internal calibration controls are provided.

Between the end of the pass band and the beginning of the stop band there is the transition region or guard band. In a very conservative design the stop band begins at the Nyquist frequency and the transition region is considered as a kind of "no man's land." A large guard band will mean that a large percentage of the spectrum is wasted or that the sampling frequency must be that much higher. It is therefore very desirable to keep this region as small as possible. This is especially true where data rate efficiencies are important. However, narrow guard bands can result in extremely complex filters: a 30% guard band with a 50-dB stop attenuation requires a seventh-order elliptic filter [56].

In order to determine the true guard band, however, it is necessary to consider the output low-pass filter in combination with the input filter. The amount of 29-kHz energy which becomes 21 kHz in a 50-kHz system is determined by the attenuation of the input filter at 29 kHz and by the output filter at 21 kHz. This is graphically represented in Fig. 19 for an output filter that is identical to the input filter. The input frequencies and filter are shown using the standard scale **a**, whereas the output filter and frequencies are shown on scale **b**. The cascade combination of these two filter characteristics gives the frequency response for components which are subject to aliasing. In other words, it is the transmission attenuation for components which are moved to a different frequency at the output. Consider an input frequency f_{in} which is attenuated by an amount indicated by X from the input filter and by an amount indicated by Y from the output filter. The true guard-band transition rate is higher than that of the input filter alone. The attenuation at the Nyquist frequency can be 25 dB in order to produce a 50-dB total attenuation. Generally the stop band begins above the Nyquist frequency by the same amount as the pass band below the Nyquist frequency.

If the filter is implemented using active filter technology, there is an inherent limit to the minimization of the guard band size. This arises from the dynamic-range limitation of the highest Q stage. A high-order filter can be thought of as a cascade of second-order stages. Although there is no formal method for ordering the stages,[4] we can use the same approach as that which is used to optimize the level diagram of a console. We consider each stage in terms of its gain or attenuation at the edge of the pass band. As an illustration, consider a seventh-order elliptic filter with 53-dB stop band attenuation. It has three complex pole-pair stages with gains (at the band edge) of +14.739 dB (1), +5.971 dB (2), and −3.795 dB (3); three complex zero-pair stages with gains of −1.318 dB (4), −6.632 dB (5), and −4.264 dB (6); and a single-pole stage with a gain of −4.717 dB. Fig. 20**a** shows the optimum ordering for maximum dynamic range where the numbers in parentheses correspond to the stage numbers above. Fig. 20**b** shows the worst possible ordering.

The total dynamic range lost for the optimum case is 14.739 dB since the input level must be reduced to prevent overload at the output of stage (1) and the noise from stage (6) contributes the dominant noise. The nonoptimum case has a dynamic range loss of 19.392 dB as a result of the overload condition at the output of stage (2). With filter orders greater than seventh, the effect is even more pronounced. Using typical operational integrated circuits, the dynamic range would be on the order of 110 dB. Although this is significantly greater than the converter performance, even with 16 bits, the margin is not excessive. The above discussion serves mostly to illustrate the way in which the guard-band specification interacts with the SNR of the total system.

There is a growing concern among the "golden ears" of the audio profession about the perceptual degradation produced by using very sharp cut-off filters. Because most listeners cannot hear spectral energy above 20 kHz, the perception of these filters does not arise from the removal of high-frequency energy. There is some suggestion that these filters may produce perceived ringing. Mathematically, the filter cannot introduce spectral components near the band edge if one views these components in the Fourier sense. The usual mathematical definitions consider a time frame which is infinitely long.

The large amount of phase distortion in typical elliptic filters has also been considered; however, the phase is smoothly changing and the ear cannot hear phase distortion at high frequencies except in very special extreme cases. Generally, the rate of change of phase with increases in frequency determines the probability of it being

[4] The pairing of pole and zero pairs has been examined under various assumptions and the reader is referred elsewhere [62].

perceived. If phase distortion were the true issue, all-pass phase correctors could be introduced [57] such that the phase shift would be identical to pure delay.

A more interesting speculation about the perception of sharp filters probably lies in the direction of "time smear." An ideal low-pass filter with an infinitely sharp cutoff has an infinitely long impulse response as shown in Fig. 21. This response does *not* contain any more energy at frequencies near the cutoff than it does in any other band of frequencies. We must note, however, that the signal is spread in time such that, in a given time region, there is discrete energy at the cut-off frequency. We may consider three time regions: from $-\infty$ to $-1/2f_s$; from $-1/2f_s$ to $+1/2f_s$ and from $+1/2f_s$ to $+\infty$. The signal in the first and third regions is very much like a sine wave at a frequency equal to f_s. Mathematically it just so happens that the energy in the first and second regions cancels the excess energy in the third. To the ear, the spread may be over a large enough time interval such that the cancellation does not entirely function perceptually. With a 20-kHz filter the time spread is 7.5 ms if we consider the low-level parts of this impulse respose to −60 dB. It is conceptually possible that some "golden ears" can hear this.

If this speculation proves to be true, and if the profession decides to use gentle-slope cut-off filters of the Gaussian or Gaussian-hybrid form, then the guard-band region must be increased very significantly. Any filter which is gentle in the time domain must also be very slow in the frequency domain. A Gaussian filter of infinite order only achieves an attenuation of 30 dB at three times the cut-off frequency. Such a situation would be intolerable for digital systems since the sampling frequency would be almost three times higher than that which is required with very sharp elliptic filters.

3.4 Input Sampler

The input sampler, conventionally referred to as the sample-and-hold module, serves two independent functions: it samples discrete values of the input signal at a fixed period rate, and it holds those values constant during the interval for A/D digitization. These two functions are represented in the track-and-hold phase. During track the output of the sample-and-hold module follows the input. This track mode terminates at the sample-clock time, and that last value is held for the duration of the period. The module is built with a capacitor and FET switch which open for the hold phase; the capacitor is left with the analog value which occurred at the clock transition from track to hold. Each of these phases can produce its own type of degradation: timing error in the effective sampling time and changes in the hold value during digitization.

One form of time error originates from the fact that the time for the FET to respond to its clock is a function of the analog voltage. The aperture time is defined as the time interval between the sampling command from the clock and the actual "freezing" of the analog signal on the capacitor. If this aperture time were constant, the effect would be a fixed small delay in the clock signal. However, the time to turn off the FET is a function of residual charge in the circuit and the source-gate shut-off voltage. This voltage can be a function of the analog signal. For medium-quality samplers, the nonlinear component can be on the order of 5–100 ns. If the gate voltage control has a

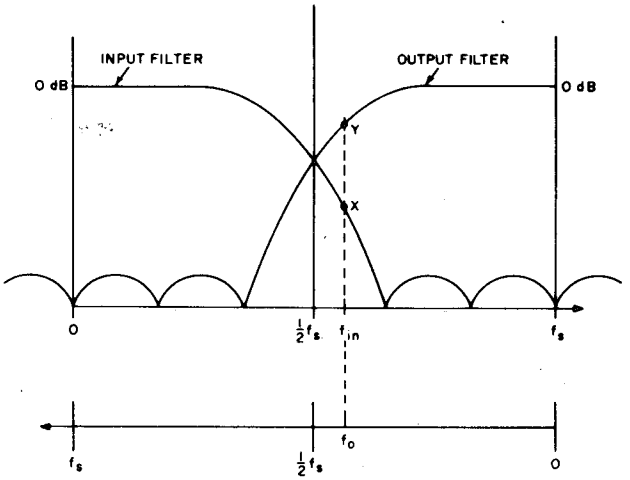


Fig. 19. Effective frequency response for aliasing components is determined from the cascade combination of the input and output low-pass filter . Because the frequency is changed, the output filter is reversed with its frequency scale running backwards. Point X is the attenuation produced by the input filter on a signal at frequency f_{in}, and point Y is the attenuation produced by the output filter.

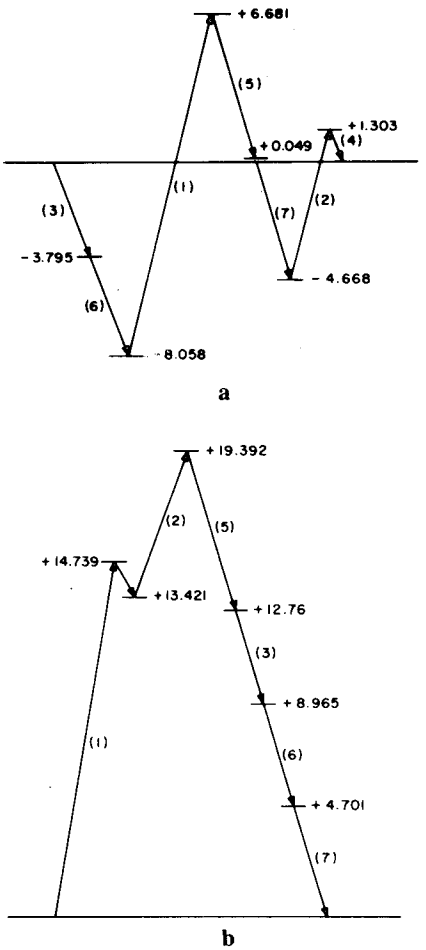


Fig. 20. **a.** Level diagram for ordering of filter stages for a seventh-order elliptic filter in an optimized order. **b.** Worst case order. Base line is reference 0 dB and gains (attenuations) indicated cumulatively.

long rise time, this effect is much worse; therefore, rise times of less than 100 ns are generally required on gate drives.

An error in the effective sample time means that the analog voltage which is held is different from the one which occurred at the true sample time. We can view this error as an additive voltage which is equal to the signal derivative multiplied by the error time. The first-order approximation is more than adequate for the small time errors which can occur. Because the error time is proportional to the derivative, the magnitude of the effect is largest near the Nyquist frequency. The significance can be appreciated by considering a numeric example: a 20-kHz sine wave, a 50-kHz sampling frequency, and a 50-ns change in aperture time as a function of signal voltage. The peak error is about 45 dB below the fundamental.

Because the nonlinearity is primarily an asymmetric effect (sign dependent), the error components will tend to be mostly second harmonic. Since this signal nonlinearity effect is after the low-pass filter, the harmonic components will beat with the sampling frequency. In some sample-and-hold modules, the relative magnitude of the error increases with increases in signal level because the nonlinearity is greatest for full-level signals. This distortion mechanism is essentially nonexistent for low frequencies since the magnitude increases at a rate of 6 dB per octave. To reduce the worst case peak error to that of an LSB for a 16-bit converter requires that the aperture time error be on the order of 200 ps for a full-level maximum-frequency signal.

Sampling errors can also originate from phase jitter in the main digital clock signals. Typical analog-based clock generators, excluding crystals, are very prone to jitter because of the inherent mechanism which converts additive voltage noise into phase jitter. Consider a hypothetical clock generator made up of a sine-wave oscillator followed by a comparator, as shown in Fig. 22. Furthermore, let us model all the analog noise sources in the oscillator and comparator as a single noise voltage in series with one of the comparator inputs. If we microscopically examine the zero crossings, we observe that they are shifted randomly depending on the value of the noise at the zero crossing time. The exact representation of the phase jitter statistics is rather complex [58] and not relevant. However, a 1-volt sine wave at 50 kHz can produce upwards of 300-ps time jitter when the noise voltage is on the order of 100 μV. Noise magnitudes of this amount are not uncommon in comparators and other very broad-band analog circuits. A 20-MHz bandwidth will produce 30 dB more noise than the typical 20-kHz analog bandwidth for the same noise energy density. The reader should also be aware of the fact that poor-quality digital integrated circuits can also contain noise sources which contribute to the clock jitter. Noise amplitudes of 50 mV are irrelevant to the function of digital circuits, but they may affect phase purity.

Clock jitter, unlike the nonlinear aperture time, results in an additive noise rather than distortion products. The error is random even though it is proportional to the signal amplitude and frequency. However, this noise only manifests itself as modulation noise which degrades the SNR_{ws} but not the SNR_{ns}. It is a broad-band noise controlled by program derivatives.

A general expression for the noise produced by Gaussian time jitter is given by

$$SNR_{ws} = -20 \log_{10} (2\pi f \Delta t) \tag{18}$$

where f is the signal frequency and δt is the standard deviation of the time jitter [63]. A significant amount of

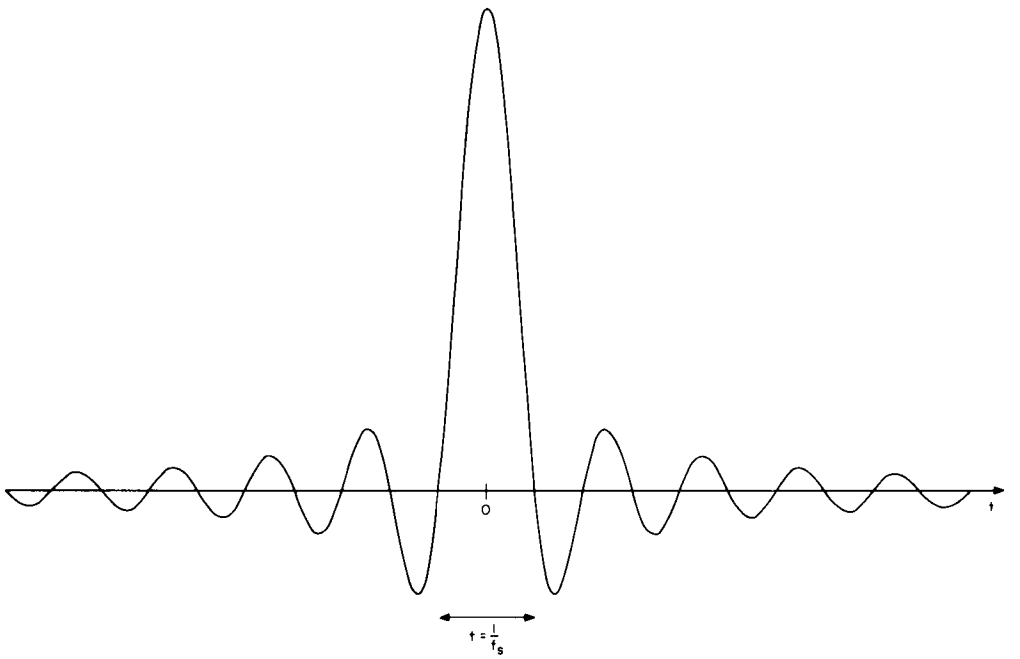


Fig. 21. Classical impulse response of ideal low-pass filter with cutoff at frequency f_s. Even though the energy density is constant over frequency, certain time regions have a large amount of energy at the cutoff frequency.

hum in either the clock generator or the sampling logic will create a systematic frequency modulation of the input signal. Fortunately the auditory system is less sensitive to small amounts of 60-Hz modulation [59].

For these reasons the highest quality conversion system must be driven with a crystal clock system, and the phase purity must be examined in the logic of the input sampler and output hold amplifers. The output clocking is just as critical as the input clock. Although it is not that difficult to achieve reasonably small sample-time errors, they are difficult to measure directly.

Empirical studies of perceptual jitter threshold indicate much less severe requirements. Manson [60] suggests that less than 5% of the population could hear 35 ns jitter if the jitter spectrum was in the maximally sensitive region above 2 kHz. However, since the tests were run with tape-recorded test material, the limiting factor may have been the inherent dynamic range of the listening facility. Stockham [12] recommends a more conservative 5–10-ns specification. In both cases, however, other degradations probably limited the detectability of small jitter values.

The hold phase of the sample-hold amplifier produces an entirely different error mechanism as a result of changes in the hold voltage. To understand this effect, we first need to examine the temporal dynamics of the A/D conversion cycle. A special logic device, called a successive approximation register (SAR), contains a hardware algorithm to find a digital word which, when converted to an analog signal by the D/A element, is the best approximation to the input analog signal. The SAR drives the D/A element to create an approximation to the input, while the comparator tells the SAR if this approximation is too high or too low. A block diagram is shown in Fig. 23.

At the same time that the input analog signal is held constant in the hold phase, the SAR is initialized such that all bits are set to 0 except the MSB which is set to 1. This creates the first approximation to the input, the word 10000 which is 0 volts. After a fixed time has elapsed such that the analog voltages have settled to their final values, the SAR tests the comparator output to determine if the current approximation is more positive than the input (clear MSB) or more negative (leave MSB set). For discussion purposes only, assume that the MSB was left set because the input signal was greater than 0. Once having made the decision about the MSB, the next bit is set. The new approximation signal is $+0.5V_{max}$ since the digital word is 11000. At test time this bit is set or cleared depending on whether the input was between 0 and $+0.5V_{max}$ or between V_{max} and $+0.5_{max}$. With each successive bit test, the maximum error between the input and the approximation is reduced by a factor of 2. We are now in a position to consider the case of droop in the hold signal which is caused by a small fixed leakage current flowing from the holding capacitor.

The temporal sequence of the SAR's approximation voltages for 5 bits is illustrated as a voltage versus time graph in Fig. 24. The initial approximation voltage, indicated as point A changes to either voltage B or C at test time t_0, depending on the sign of the difference voltage. The set of binary decisions at each of the five test times determines the path that is taken from the initial voltage to one of the possible 32 end voltages. The time behavior of two hypothetical input analog voltages, with initial values X and Y, is also superimposed on the same graph (dashed lines). Because the X signal is above the approximation voltage at t_0, the MSB is set to 1; however, with all subsequent tests, the approximation voltage is greater than the drooping input. Thus this signal is converted to the digital word 10000.

The Y signal, which started slightly lower than the X signal, follows a completely different path and ends at an approximation voltage corresponding to the digital word 01100. Not only did a small difference in the starting value between X and Y result in a large difference in digital words, but there is no possible input voltage which can be converted to the words 01111, 01110, or 01101. These are missing codes which correspond to a nonmonotonic displacement of those levels. This effect is most pronounced for signals at 0 volts because the MSB decision is critical and irrevocable. With the Y signal the approximation voltage could track the droop, whereas with the X signal the decision at t_0 prevents the approximation from ever

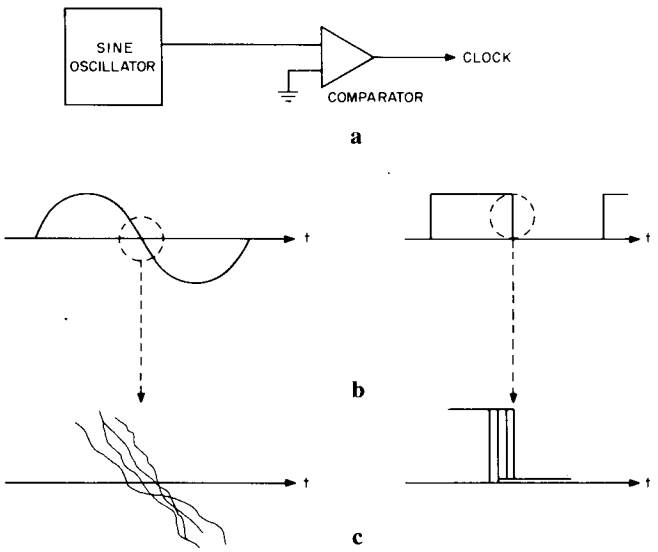


Fig. 22. Illustration of the process by which additive noise can get turned into time jitter. **a.** A sine oscillator's output is limited by a comparator. **b.** The sine wave and the resulting clock. **c.** Exploded view of the zero crossing region to show the way in which a small amount of noise can produce minor changes in the clock transition.

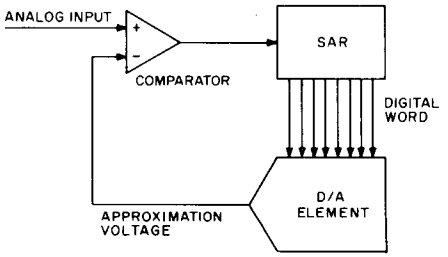


Fig. 23. Block diagram of classical A/D converter using a D/A element, successive approximation register (SAR), and a comparator. The SAR finds a digital word that minimizes the difference between the input and the approximation voltage.

going below 0 volts. The nonlinear displacement of certain quantization levels is equivalent to simple D/A element inaccuracies; both error mechanisms are static, signal independent, and severe at the major higher order bit transition regions. Moreover, the effects are additive in terms of total system errors.

The droop error mechanism can be kept to a minimum by either increasing the size of the capacitor or decreasing the leakage current. Let us consider the following specifications as an illustration: 16-bit conversion, 1-nF holding capacitor, and a requirement that the droop be less the 0.5 quantization interval over a 20-μs cycle. This gives a calculated leakage current of 5 nA. Only a few specialized integrated-circuit operational amplifiers will meet this specification since large bandwidths (fast settling times) imply high internal operating current. Moreover, the leakage for non-MOS can double for every 10° C increase in junction temperature. Alternatively, the holding capacitor can be increased; but this means that the driving amplifier must supply high charging currents during the tracking phase. A 10-volt step in 1 μs requires a current of 0.1 A for a 10-nF capacitor.

The acquisition of the input signal after completing the hold phase is determined by the bandwidth of the driving amplifiers. If we require the error to be within an LSB and assume a full-scale change in level, we must allow 11 time constants (using an RC model) for the sample-hold amplifier to settle to the required tolerance. This implies a full power bandwidth of over 350 kHz at 0.1 A in order to settle within an allotted 5 μs. Observe that this allocation only leaves 15 μs for the 16-bit tests in the SAR operation.

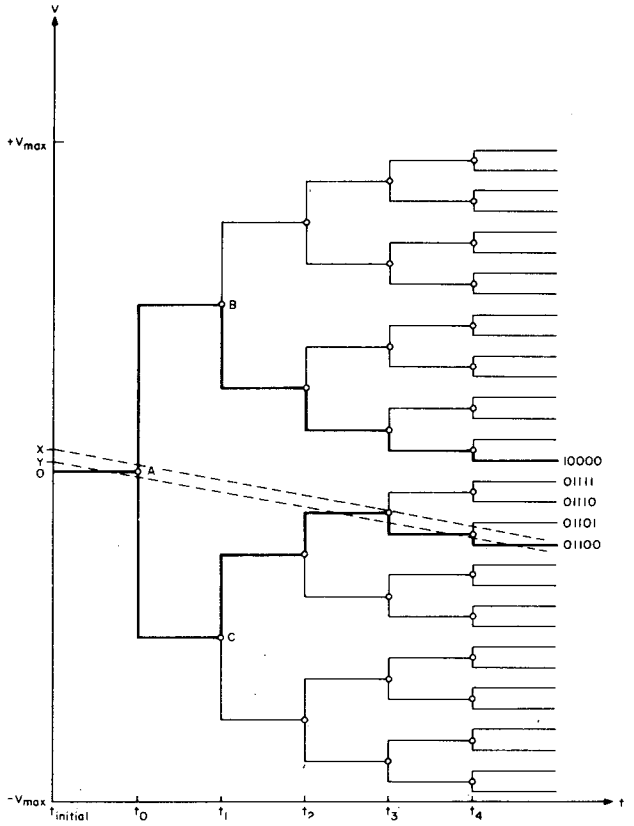


Fig. 24. Graph of the approximation voltage, as observed at the D/A output of Fig. 23, for all possible conversion sequences. At the discrete clock times (t_0, t_1, etc.) a decision is made about the sign of the difference between the current approximation and the input. Two possible input signals with droop are shown (dashed lines) and the conversion sequence for these two inputs is illustrated (heavy lines).

A final difficulty with the sampling process arises from the fact that the capacitor voltage is being changed in an extremely short time. Even the best capacitors exhibit a phenomenon known as dielectric absorption and hysteresis. Because the energy is stored in the dielectric material, the terminal voltage undergoes a relaxation as the dielectric reaches equilibrium. The effect of this change in voltage is similar to that produced by droop, namely, the hold voltage changes between the test times. Capacitors with low hysteresis must be used, such as polystyrene or polypropylene; mylar is not good and ceramic is absolutely unacceptable.

3.5 Bandwidth and Accuracy

Although the 16-bit PCM system at 50 kHz sampling is at the limits of today's technology, it is also not that far from the theoretical limits. This can be appreciated by considering the implicit thermal noise and noise bandwidths of the circuits used to implement the A/D conversion. Specifically, we must note that the process of digitization must take place in an analog medium because the comparisons are analog. Thus, the bandwidths required for these computations must be several orders of magnitude higher than the actual analog signal. This extra bandwidth also increases the system noise. Increasing the number of bits, in order to increase the dynamic range, also increases the required computational bandwidth since there is less time available for each bit test.

In our previous discussion we allocated 15 μs for the 16-bit tests, thus giving about 930 ns per bit. When a bit is switched on by the SAR, a certain waiting time must elapse until the SAR can test for the sign of the comparator. If we assume that this settling process can be modeled as an RC time constant, then 12 time constants are required for the MSB bit to settle to an accuracy of ½ LSB. This translates into a combined 60-ns time constant for the SAR, D/A element, and comparator.

Even the best integrated circuits, with required bandwidths of more than 10 MHz, will generate noise levels over this band of more than 30 μV (rms). This implies that the peak signal must be dimensioned for at least 5 volts in order to produce the required 16 bits. Additional bits for increased dynamic range, or higher sampling rates, are difficult to incorporate because the noise-bandwidth tradeoff is close to the limit.

The next major advancement can only be obtained by using a parallel structure for the A/D conversion. Hypothetically, consider a converter made up of 2^{16} comparators, one for each quantization level. The bandwidth of the comparators could be reduced because the full 20 μs are available to make the decisions. This provides a factor of 5 decrease in the noise level; hence, an additional 2 bits of dynamic range are available.

Although this is more than a little impractical, it does

suggest that a hybrid technique could be used such that the higher order bits were converted with an algorithm that allocated a larger percentage of time to these decisions and hence a smaller noise bandwidth. The middle bit needs only half the time that the MSB needs to settle to the same accuracy. It would thus appear reasonable to modify the time-allocation procedure.

We must caution the audio profession from setting goals and standards which come too close to inherent limits. Moreover, most of the error types discussed in this paper are cumulative. This would imply an even stricter specification for each error type. See a later discussion on error budgets.

3.6 Output Sample-Hold

The output part of the conversion system almost always contains a sample-hold amplifier which is essentially identical to that of the input; however, its function is completely different. The output sampler is not sampling a continuous waveform; it is sampling the D/A converter which produces voltages at discrete times itself. In principle, no output sampler is required since the sampler input and output are identical. However, the D/A converter element generates erroneous and unpredicatable outputs during the transitions from one voltage to the next; the sampler removes these components. The sampler is allowed to acquire the D/A converter voltage only during the portion of the interval when the converter voltage is stable and accurate. During the transitions it is kept in the hold mode. The design criterion for the hold time is thus a function of the settling accuracy of the converter.

Although the sampler serves to avoid the introduction of noise and distortion components from the D/A element, it can also create its own distortion products. To understand this mechanism, we must observe that the output of the sampler is *continuously* connected to the low-pass filter, as shown in Fig. 25**a**. Thus any signal, regardless of its origin or type, appears at the system output after being low-pass filtered. This is not the case with the input sampler since the value acquired after sampling is the only value which is used by the A/D converter; all errors during the acquisition phase were irrelevant.

Specifically, we must consider the behavior of the output sampler when it switches from the hold phase (previous value) to the track phase (acquiring new value). If this transition is not a linear function of the difference between new and old values, nonlinear components will be created. Fig. 25 **b** and **c** shows two kinds of transitions (dashed lines) in comparison to the idealized transitions (solid lines). In example **c** the transition is proportional to the voltage difference of the current and final values, that is, the transition is a classical exponential step response. In example **b** the transition is always at a constant rate, often called slew-rate limited. For purposes of discussion, we may call the area enclosed by the ideal transition and the actual transition the error. A detailed error analysis of this process is available elsewhere [61], [64], but a brief summary is presented here for completeness.

Of the two types of error, only the slew-rate-limit process is interesting since it is a nonlinear mechanism. The exponential approach will be shown to be equivalent to a slight high-frequency deemphasis with no nonlinearity, that is, a filter. Fig. 26**a** shows the error term from the slew-rate process; the slope is constant and the height is proportional to the transition voltage. We can represent this as

$$v_n = x_n - x_{n-1} \tag{19}$$

where x_n is the nth voltage sample from the D/A converter element at time $t = nT$. Observe that the area is given by

$$A_n = (x_n - x_{n-1})^2 \operatorname{sgn}(x_n - x_{n-1}). \tag{20}$$

This type of signal can be modeled as a train of impulses having the same area function if the slew interval (percentage time in slew rate limit) is a small fraction of the total period. Once having made this approximation, we can construct a model of a process which would produce the same result (Fig. 26**c**).

The nonlinear process can generate harmonic distortion components which are higher than the Nyquist frequency. When sampled, new spectral energy is created at the beat frequency between the harmonics and the sampling frequency. Freeman [61] has computed the relative harmonic energy from this nonlinearity as being given by

$$a(n, f) = \frac{4 \sin^2 (\pi f/f_s)}{\pi n (n^2 - 4)} \frac{V^2}{V_{\max}} \frac{\tau}{T}, \tag{21}$$

where f is the signal frequency, n the harmonic number (odd only), $V_{\max}$ the maximum signal amplitude, τ/T the percentage time during which slew can occur for the maximum signal at the maximum frequency, and $a(n,f)$ the resulting harmonic amplitude.

Using the same analytic technique as previously, we

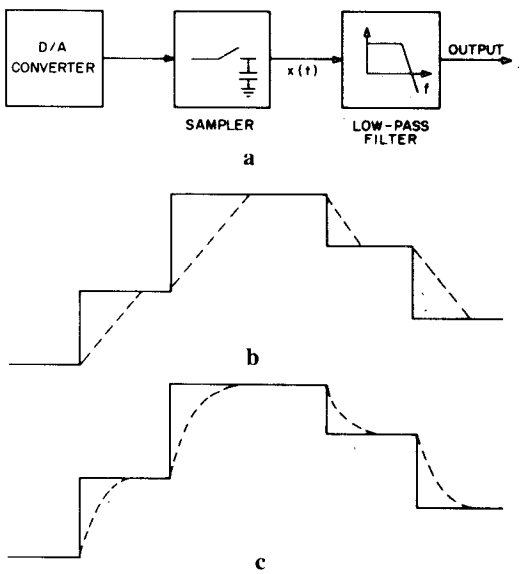


Fig. 25. **a.** Block diagram of output conversion system showing that the sampler output is continuously connected to low-pass filter. Idealized transitions at output of sampler (solid lines) in comparison to actual transitions (dashed lines). **b.** Slew-rate distortion. **c.** Exponential mechanism.

have reversed the sampling process and the nonlinearity. The input to the model may be thought of as an unsampled signal even though it is coming from the D/A. Thus, the harmonics indicated by Eq. (21) will appear at other frequencies after the sampling process.

There are several important aspects to this analysis. The distortion components are largest for input frequencies between one third of the Nyquist frequency and the upper band edge, and the effect is strongest for full amplitude signal. Because the denominator of Eq. (21) contains an n^3 term, we would expect that the third harmonic is the major term. Second harmonics can also occur if the negative and positive slew rates are different, but we will ignore this case. The worst case harmonic component is thus given by

$$a[\text{dB}]\ (n = 3; f = f_{\text{Nyquist}}; V = V_{\max}) = -21.4 + 20 \log_{10} (\tau/T). \quad (22)$$

A numeric example shows that the slew rate must occur for less than 10^{-4} of the sampling period for the harmonic to be 100 dB below the fundamental. Since this translates into a 2-ns transition in a 50-kHz sampling system, a 10-volt transition would imply an astronomical 5000-V/μs slew rate. This is clearly out of the range of possibility.

For this reason a pure exponential approach is used since the speed can be arbitrarily slow as long as the rate is proportional. Such a technique is sometimes referred to as an integrate-and-hold since the amplifier is integrating the difference between its current value and the input. Even if the final value is not accurately acquired, the error is still proportional: this is equivalent to filtering. We consider such a frequency effect to be benign, especially since the basic sample-hold process produces a significant attenuation at high frequencies, approximately 2.5 dB. This aperture mechanism is discussed in the next section.

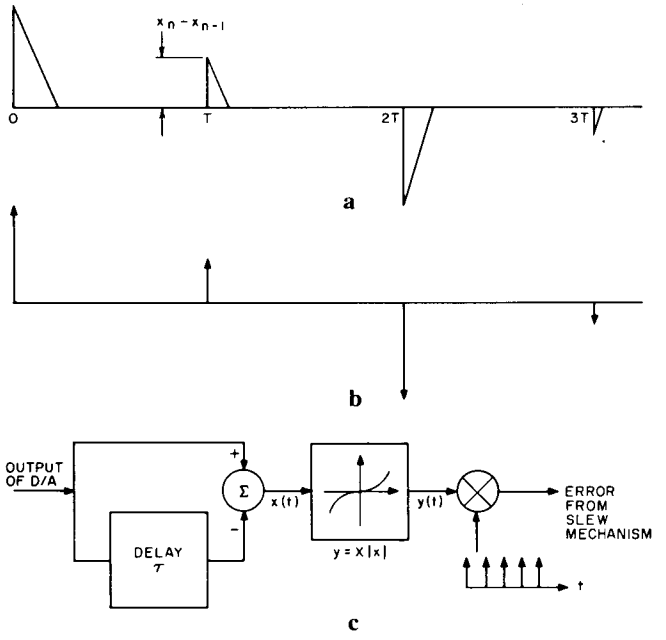


Fig. 26. **a.** Error signal resulting from slew-rate transitions in the output sampler. **b.** Equivalent approximation using impulses of the same area. **c.** Block diagram of a model which will generate the same error signal which is used for analysis purposes.

We should note that the 2-ns calculation gives us an indication of the sensitivity to any nonlinear influences. For example, if the integrate-and-hold amplifier has a small amount of nonlinearity as a result of rapid changes in the derivative, similar error components will result. Very-high-speed devices must be used even with integrating instead of sampling.

In all other respects, the output sampler can be quite primitive. An extreme amount of droop during the hold mode is acceptable since this represents only a dc shift at the output of the low-pass filter. Small amounts of control voltage leakage into the signal path are also acceptable. However, if this leakage is a result of a nonlinear capacitance in the FET switch of the integrate-hold amplifer, nonlinear signal components will again be created. It is not uncommon for an FET to have a hysteresis capacitance between the gate control and the source output.

3.7 Output Low-Pass Filter

This filter, in combination with the sample-hold amplifier, serves to transform the discrete series of voltages from the D/A into a continuous analog signal. From a spectral point of view, it removes those high frequencies which were created by the sampling process, as demonstrated in Fig. 2. A 1-kHz sine wave also created components at 49, 51, 99, 101 kHz, etc., when sampled at 50 kHz. It is these components which are to be removed by the filter. Since these components are all above the audible range, the filter specification cannot be derived from auditory phenomena. Rather, one must examine the equipment environment or the "measurement laboratory." These components may create audible interference when they combine with the bias frequency of a tape recorder in record mode, a local oscillator of a transmitter, or even the sampling process of another digitizing audio equipment.

Thus the output filter avoids degradation which *might* occur from the interconnection of the digital system to the next audio component. Moreover, if we elect to create a specification for attenuation outside the band, we must include the input low-pass action of the next equipment since it is in cascade. Since there is no current audio standard, we are left to create "reasonable" norms. We might assume that a total of 90 dB is desirable and that half should be allocated to the digital output and half to the input of the next system. A much more conservative approach is that of allocating all 90 dB to the output filter, except that such a filter is extremely difficult.

The spectral components are, however, audible in digital equipment which uses a much lower sampling frequency. For example, with a 30-kHz sampling rate, the components generated by frequencies above 10 kHz can create audible signals. Nevertheless, it is still difficult to create a specification for several reasons: 1) the auditory system is not very sensitive to these high-frequency components; 2) the main signal may create a high degree

of spectral masking if it is broadband with many musical overtones; 3) transients produce a high degree of temporal masking; and 4) the lower sampling frequency implies that the equipment is not archival main-channel quality. An attenuation of 40–60 dB would thus appear to be reasonable, although it is extremely difficult to detect these components even if a much more primitive filter is used. It is not possible to converge on a specification process, and we leave this topic to the design engineer.

The output filter needs to be designed such that the pass-band response has the required degree of flatness and that the stop-band attenuation reduces the components to a tolerably small level. We should note that the dominant energy in the stop-band region is near the sampling frequency since this corresponds to the low-frequency energy in the original signal. The spectral region of 0–5 kHz maps to 45–55 for a 50-kHz sampling rate. Thus a compromise specification might include an extra notch filter in this region in addition to the implicit notch created by the sample-hold amplifier. Sometimes an elliptic filter is chosen with a zero pair located in the region near the sampling frequency.

The output filter's pass-band response offers a special problem since the sample-hold amplifier has an excessive attenuation at higher signal frequencies (even if the integrate mode is minimal). We may consider the input to the sample-hold as being a hypothetical impulse train whose area is equal to the converted word. The frequency response of this kind of linear system is given by

$$H_1(f) = \frac{\sin(\pi f/f_s)}{\pi f/f_s} \tag{23}$$

where f_s is the sampling frequency. Because this response produces a significant attenuation in the pass band, a correction filter, often called aperture correction, must be included. One way to create a correction filter is to model the $H_1(f)$ function as if it were made up of poles and zeros. The correction filter would have the reciprocal characteristic to cancel the $H_1(f)$ function. The best approximation is a series of zeros on the jω axis except for the absence of one at the origin. This is the least interesting case since a correction filter would have to contain a series of poles on the axis.

In the following discussion we will assume that the compensation must cover the frequency region from 0 to 80% of the Nyquist frequency, and we will consider compensations composed of 1) a single zero, 2) a complex pole pair, and 3) a complex pole and zero pair. Moreover, the approximations are optimized in an equiripple sense in the pass band and an equiripple sense in the stop band when there are enough degrees of freedom to do so.

The simplest compensation filter is a single zero with a value given by

$$f_{\text{zero}} = -1.4550\,(f_s/\pi). \tag{24}$$

The resulting frequency response has a peak-to-peak pass band error of 0.2 dB which is probably at the limit of acceptability. Moreover, the zero must be embedded in the main output low-pass filter by combining it with an existing pole since a simple zero cannot exist alone. It would produce infinite gain at high frequencies. The major disadvantage of this compensation is that it completely removes the natural low-pass action of the $H_1(f)$ function.

A more sophisticated approach can incorporate the $H_1(f)$ function directly into the output low pass. A complex pole pair, when used for compensation, produces a composite frequency response shown in Fig. 27**b**. Observe that the pass-band error has been reduced to 0.0038 dB and that the stop band produces an attenuation of 21.67 dB. The pole location is given by

$$f_{\text{pole}} = -0.61845 \pm 2.1905\text{j}\,(f_s/\pi). \tag{25}$$

The low-pass property is not that useful since there is no attenuation in the region from the Nyquist frequency to the sampling frequency. If we use a zero pair also, the composite shown in Fig. 27**a** can be achieved. Although the pass-band ripple is worse, on the order of 0.1 dB, the stop-band behavior is very close to that of a real filter. There is 17 dB of attenuation at frequencies above the Nyquist frequency. The singularities are given by

$$f_{\text{pole}} = -0.24445 \pm 1.45906\text{j}\,(f_s/\pi)$$

$$f_{\text{zero}} = \pm 1.80578\text{j}\,(f_s/\pi). \tag{26}$$

The technique for creating these filter compensations is rather complicated and it involves extensive numeric calculations. There are no tables available. The method is briefly outlined here although the reader should see elsewhere for a full discussion [66]. First the original $H_1(f)$ function is taken to be equivalent to a filter made up of a limited number of poles and zeros, for example, a single pole and a complex zero pair. These are used to replace a similar pole and zero pair in a standard elliptic configuration. Since the equivalence is not exact, the remaining poles and zeros of the elliptic filter must be shifted somewhat to bring the overall characteristic back to being equiripple error.

Low-order compensations cannot create true equiripple because there are not enough degrees of freedom. A single pole pair offers two free parameters. The equiripple criterion in the pass band requires both degrees of freedom and the stop-band characteristics are unadjustable. In contrast, a compensation of a complex pole pair and a complex zero pair on the imaginary axis offers three degrees of freedom. Therefore one of the stop-band ripples could be balanced. The curves in Fig. 27**a** show that the first and second peak in the stop band is balanced in an equiripple sense. Much higher order compensation can be used and, in the limit, the $H_1(f)$ is directly embedded in the main low-pass filter configuration. The computation of the pole and zero values is extremely time consuming and requires extensive software. Only the low-order compensations can be done with a hand calculator.

In the preceding discussions we have treated the sample-hold as ideal with infinitely rapid transitions; the actual characteristics of an integrate-hold is somewhat

different and it has an additional effect given by

$$H_2(s) = \frac{(1 - e^{-a/T_{RC}} e^{-as})}{(1 - e^{-a/T_{RC}} e^{-Ts})} \quad \frac{1}{(sT_{RC} + 1)} \tag{27}$$

where T_{RC} is the integrator time constant, T the sampling period, a the time during which integration is taking place, and s the Laplace transform variable for $j\omega$. The last term in Eq. (27) originates from the integrator's natural low-pass action, whereas the first term is a result of the fact that the integration is taking place for less than the full period. A complete derivation is presented in the Appendix.

If we disregard phase effects, Eq. (27) can be written in terms of magnitude response:

$$|H_2(f)| = \sqrt{\frac{1 - K\cos(2\pi fa)}{1 - K\cos(2\pi fT)}} \left(\frac{1}{(2\pi fT_{RC})^2 + 1}\right)^{1/2} \tag{28}$$

where K is given by

$$K = \frac{-2\, e^{-a/T_{RC}}}{1 + e^{-2a/T_{RC}}}. \tag{29}$$

Because the hold phase begins before the integrator has reached its final value, the factor $(1 - e^{-a/T_{RC}})$ represents the fraction of the transition which is actually acquired. No distortion is produced by this mechanism since the factor is a linear constant: doubling the signal doubles the amount acquired.

This function $H_2(f)$ should be used with $H_1(f)$ when computing the compensation corrector. Thus each time the dimensions of the integrate system are changed, a new compensation process must be computed. However, we should observe that the integrator time constant offers another degree of freedom in the design of the compensation filter. With this being adjustable, the ripple error in the pass band can be reduced or the stop-band attenuation can be modified.

The function $H_2(f)$ can be made to produce more low-pass filtering either by increasing the time constant T_{RC} or by decreasing the integration time a. These are not actually equivalent because the limited integration time has a freqency response which is approximately periodic with the sampling frequency. In contrast, an increase in T_{RC} does produce more attenuation over the entire frequency spectrum.

Implementation of these low-pass filters is straightforward, and it is very similar to the input filter. However, unlike the input filter, each stage of the output must be able to follow the rapid changes in the output of the sample-hold amplifier. The slew rate analysis used with the sample-hold circuit also applies to the internal stages of the low-pass filter. Slew-rate limiting produces the same kind of aliasing components which were discussed previously.

As a final note, we might mention that the low-pass filter circuits are generally ineffective against very-high-frequency signals. Digital equipment can contain very high electromagnetic energy at frequencies beyond 100 MHz since the internal switching has transition times of 1 – 10 ns and the total logic current can be 10 A. Even if only 10% of the current is being switched at a given time, this represents significant radiation. Often this energy can escape through the low-pass filter because the circuits do not function at these frequencies. Also, RF can escape as common mode on any of the wires connected to the system, that is, input, output, control, power, etc., even the ventilation openings can create problems. Proper attention to grounding, shielding, and decoupling is mandatory.

3.8 Error Budget and Summary

In the previous discussions we have outlined the major sources of error in a 16-bit PCM conversion system operating at 50 kHz. If we actually expect the total system to work to this level of performance, we need to allocate a much smaller error criterion to each error mechanism. For discussion, assume that there are 10 major sources of error, to be chosen from the list below, and that all of these are equally important. To achieve our specification, the individual errors need to be reduced by a factor of at least 3 in order for the composite system to have the required performance. The list below is not unique and other groupings of error categories are possible [65]. At best, errors will accumulate statistically. This extra performance burden suggests that a true 16 PCM system having 96 dB dynamic range is almost beyond the range of possibility.

The engineer should note the following kinds of errors:

1) Slew-rate distortion in the input low pass for signals at at the upper end of the audio range and beyond

2) Insufficient filtering of very high-frequency-input signals

3) Noise generated by the input low-pass filter or sample-hold amplifier

4) Acquisition errors in the input sample-hold amplifier (nonlinear) because of settling time

5) Insufficient settling time in the successive approximation conversion

6) Errors in the quantization levels of the D/A converter element of the A/D system

7) Noise in the comparator or D/A element

8) Nonlinear aperture time of the input sample-hold switch

9) Clock jitter on the input or output sample-hold clock

10) Dielectric absorption on the sampling capacitors at the input or output sample-hold amplifier

11) Droop on the input sampler's hold phase

12) Low frequency nonlinearities in the analog circuits as a result of nonuniform heating of the input stage by high currents in the output

13) Power supply noise injection or ground coupling

14) Nonuniform quantization levels in the output D/A converter element

15) High-order derivative distortion in the output sampler (integrate-hold)

16) Noise in the output filter as a result of limited dynamic range of integrated circuit

17) Changes in characteristics as a function of temperature or aging.

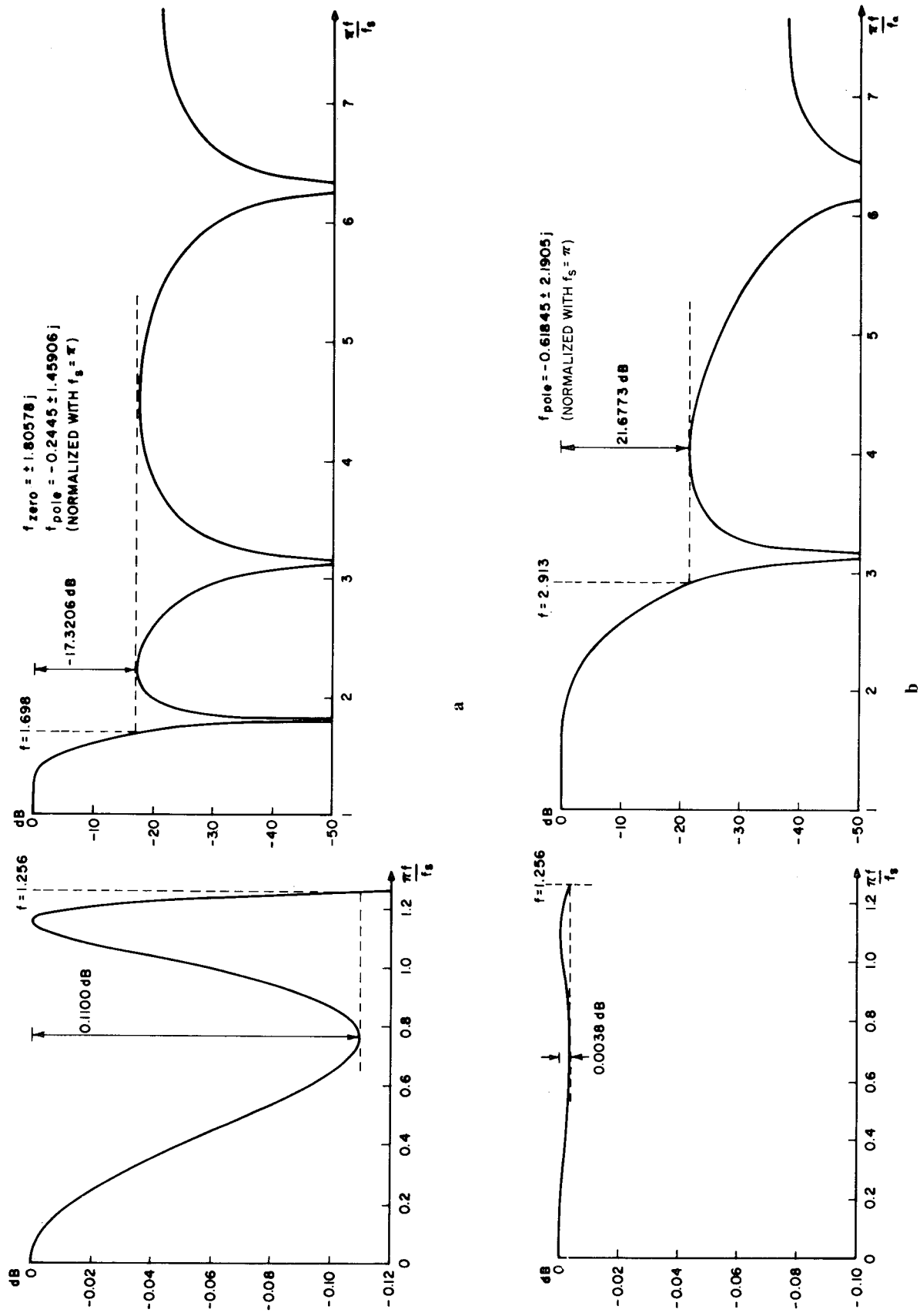


Fig. 27. Composite frequency response in pass band (left) and stop band (right) for the aperture effect and the correction filter. **a.** Complex pole pair with a complex zero pair on axis. **b.** Complex pole pair alone. Optimized for equiripple.

4. TESTING AND STANDARDS

The next stage in the advancement of digital audio will be the evolution of well defined testing and measurement methods. Unfortunately the profession has not yet advanced to this stage. Moreover, different conversion architectures require different kinds of measurements. For example, modulation noise is different when it originates from nonuniform quantization levels or from floating-point gain changes.

In general, there are two major areas of concern to the audio engineer. Nonlinearities any place in the total system are always worst in the region near one third of the sampling frequency and the perceptual manifestation is always clearcut. Second, modulation noise is almost an inherent part of the conversion system since the error mechanism changes as a function of the operating point. Both mechanisms are a very strong function of the program material. Thus the profession must first address itself to the nature of the testing signals.

The requirement of perfection places an impossible burden on the digital techniques since one can generally find an extreme signal type which produces perceptual degradation even though normal program is "perfectly" reproduced. Would a beat tone be unacceptable if the only signal which could produce it was a 16.6-kHz full-level signal? This may not be a reasonable test. Similarly, 10 dB of modulation noise may not be acceptable; but if the noise is already 90 dB below full level should we reject the system?

These tradeoffs are sufficiently complex that it may be at least 10 years before the standardization process will begin to converge. Most designers use one of the two criteria in the design of equipment: 1) it should be as good as current technology will allow; or 2) it must cost no more than a fixed limit because that is the market maximum. Observe that both procedures do not require a perceptual standard. The standard is irrelevant in the first case because the equipment cannot be made better, and it is irrelevant in the second case because improvement will usually raise the cost.

As the profession becomes more sophisticated, however, the users will indicate a willingness to trade off certain types of perceptual degradations for economic considerations. There will be such questions as: Is 6 dB improved SNR worth a 10% reduction in bandwidth?

It is the hope of the author that this paper will enhance this process. We hope that the user will begin to appreciate the consequences of his wishes, both technically and financially.

APPENDIX

The frequency response of the integrate-hold amplifier, which is used to prevent slew-rate limiting, can be derived by taking the Laplace transform of the impulse response. Although this derivation is straightforward, the mathematics is somewhat tedious and prone to error. A simpler derivation is illustrated in Fig. 28. We represent the system function of the integrate-hold amplifier as $H(s)$, the Laplace transform of $h(t)$.

To illustrate the process, consider the input to the amplifier as being a step rising from 0 to 1 at time $t = 0$. Such a signal is represented as a series of impulses of area 1 beginning at $t = 0$. We must use a sampled step function as input since the actual data are still in the sampled domain, that is, there is a discrete analog voltage corresponding to each discrete digital word. The output of the amplifier will be an exponential rising to the final value 1 with a time constant $T = RC$. This is shown as a dashed line in the figure. The exponential stops prematurely at $t = a$ since the amplifier is in the hold phase until the end of the sample period when $t = T$. It then returns to the integrate mode; the exponential rise continues until $t = T + a$. The resulting waveform shown at B (upper right) is the complete response to the pulse train at A (upper left).

The derivation continues by adding a judicious choice of additional linear system before and after the system function $H(s)$. On the second line of the figure, a differentiator is added at the output and the new output is shown at C. The differentiator has the system function given by

$$H_a(s) = s. \tag{30}$$

The resulting series of pulses can be made into a single pulse by subtracting a scaled and delayed replica of the signal at C to give the pulse at D. This system function is given by

$$H_b(s) = 1 - k\,e^{-sT} \tag{31}$$

where the factor k is given by

$$k = e^{-a/T_{RC}}. \tag{32}$$

This single pulse can be made into a simple exponential decay by a feedback loop which is shown on the fourth level of the figure. The delay in the loop results in the original pulse being continuously repeated at a frequency of $t = na$. Each time around the loop, the previous pulse is reduced in amplitude by the constant k. This has a system function given by

$$H_c(s) = \frac{1}{1 - k\,e^{-as}}. \tag{33}$$

Finally, the initial input can be represented as a single impulse at time $t = 0$ with a feedback loop having a delay of T seconds. This is shown on the fifth level. The system function is given by

$$H_d(s) = \frac{1}{1 - k\,e^{-sT}}. \tag{34}$$

The composite system made up of each of these additional linear system functions has the property that a single impulse (at point F) produces a simple exponential decay at the output (point E). The Laplace transform of the composite is equal to the Laplace transform of a simple exponential. This gives

$$\frac{1}{T_{RC}\,s + 1} = H(s)\,H_a(s)\,H_b(s)\,H_c(s)\,H_d(s). \tag{35}$$

Dividing both sides by the added systems gives

$$H(s) = \frac{(1 - e^{-sT})}{s} \cdot \frac{(1 - k\,e^{-as})}{(1 - k\,e^{-sT})} \cdot \left(\frac{1}{T_{RC}\,s + 1}\right). \quad (36)$$

The first term is simply the normal sample-hold system function which we defined as $H_1(s)$, and the remaining terms are therefore $H_2(s)$ which contains the effect from the integrate rather than sample operation:

$$H_2(s) = \frac{(1 - k\,e^{-as})}{(1 - k\,e^{-sT})} \cdot \left(\frac{1}{T_{RC}\,s + 1}\right). \quad (37)$$

ACKNOWLEDGMENT

The author must acknowledge the countless engineers who have made valuable contributions to the subjects presented in this paper. Whenever possible, they are

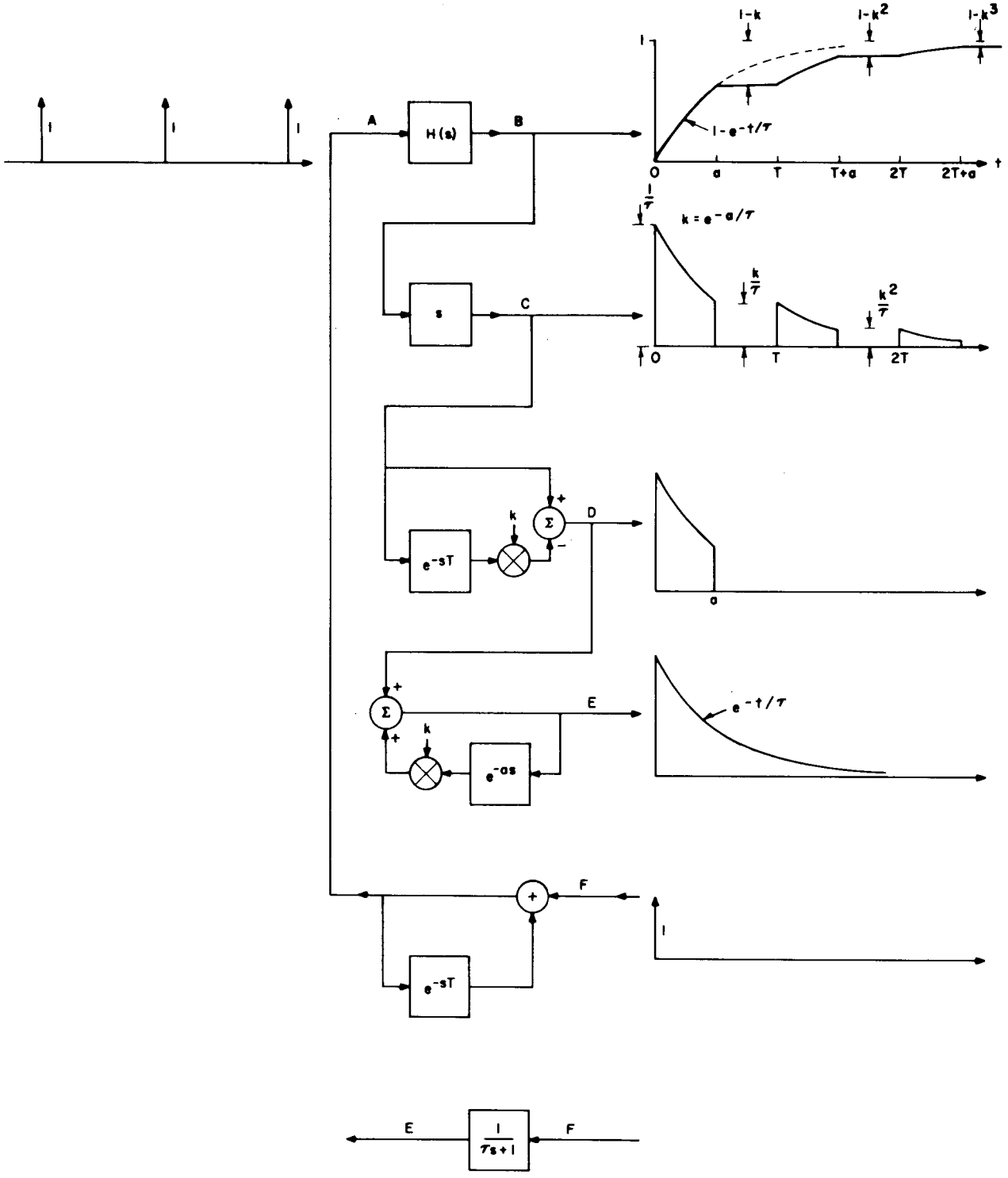


Fig. 28. Analytic block diagram for computing the frequency response of an integrate-hold amplifier. The basic system function $H(s)$ is combined with other functions until the composite can be represented as a simple 1-pole low pass.

referenced directly, but even those not specifically credited must be given an anonymous acknowledgment. Sometimes it has not been possible to separate the author's original contributions from those of his colleagues.

REFERENCES

[1] B. A. Blesser and F. F. Lee, "An Audio Delay System Using Digital Technology," *J. Audio Eng. Soc.*, vol. 19, pp. 393-397 (May 1971).

[2] B. A. Blesser, K. Baeder, and R. Zaorski, "A Real-Time Digital Computer for Simulating Audio Systems," *J. Audio Eng. Soc.*, vol. 23, pp. 698-707 (Nov. 1975).

[3] N. Sato, "PCM Recorder—A New Type of Audio Magnetic Tape Recorder," *J. Audio Eng. Soc.*, vol. 21, pp. 542-548 (Sept. 1973).

[4] H. Iwamura, H. Hayashi, A. Miyashita, and T. Anazawa, "Pulse-Code-Modulation Recording System," *J. Audio Eng. Soc.*, vol. 21, pp. 535-541 (Sept. 1973).

[5] J. P. Myers and A. Feinberg, "High-Quality Professional Recording Using New Digital Techniques," *J. Audio Eng. Soc.*, vol. 20, pp. 622-628 (Oct. 1972).

[6] M. Croll, D. Osborne, and D. Reid, "Digital Sound Signals: Multiplexing Six High-Quality Sound Channels for Transmission at a Bit Rate of 2.048 MB/s," BBC Research Eng. Div., Great Britain, Monograph 1973/42, 1973.

[7] D. Reid and M. Croll, "Digital Sound Signals: The Effect of Transmission Errors in a Near Instantaneous Digitally Companded System," BBC Research Eng. Div., Great Britain, Monograph 1974/24, 1974.

[8] T. Stockham, T. Cannon, and R. Ingebretsen, "Blind Deconvolution Through Digital Signal Processing," *Proc. IEEE*, vol. 63, pp. 678-692 (Apr. 1975).

[9] R. Willard, EMI Ltd., private comms. 1978.

[10] R. C. Cabot, "An Economical Digitally Controlled Audio Level Indicator," *J. Audio Eng. Soc. (Engineering Report)*, vol. 26, pp. 36-41 (Jan./Feb. 1978).

[11] B. Blesser and J. Kates, "Digital Processing in Audio Signals," in *Applications of Digital Signal Processing*, A. Oppenheim, Ed. (Prentice-Hall, Englewood Cliffs, NJ, 1978), chap. 2.

[12] T. Stockham, "A/D and D/A Converters: Their Effect on Digital Audio Fidelity," in *Digital Signal Processing*, L. Rabiner and C. Rader, Eds. (IEEE Press, New York, 1972).

[13] E. Baghdady, *Lectures on Communications Theory* (McGraw-Hill, New York, 1961), chap. 19.

[14] F. Gardner, *Phaselock Techniques* (Wiley, New York, 1966), chap. 5.

[15] B. Blesser, "An Investigation of Quantization Noise," *J. Audio Eng. Soc. (Project News/Engineering Briefs)*, vol. 22, pp. 20-22 (Jan./Feb. 1974).

[16] N. Jayant and L. Rabiner, "The Application of Dither to the Quantization of Speech Signals," *Bell Sys. Tech. J.*, vol. 51, pp. 1293-1304 (1972).

[17] W. Bennett, "Spectra of Quantized Signals," *Bell Sys. Tech. J.*, vol. 27, pp. 446-472 (1948).

[18] M. Croll, "Pulse Code Modulation for High Quality Sound Distribution: Quantizing Distortion at Very Low Signal Levels," BBC Research Eng. Div., Great Britain, Monograph 1970/18, 1970.

[19] L. Roberts, "Picture Coding Using Pseudo-Random Noise," *IRE Trans. Inform. Theory*, vol. 8, pp. 145-154 (Feb. 1962).

[20] L. Schuchman, "Dither Signals and Their Effect on Quantization," *IEEE Trans. Commun. Theory*, vol. COM-12, pp. 162-165 (Dec. 1962).

[21] D. Shorter and J. Chew, "Application of Pulse-Code Modulation to Sound Signal Distribution in a Broadcast Network," *Proc. IEE* (London), vol. 119, pp. 1442-1448 (1972).

[22] J. Candy, "A Use of Limit Cycle Oscillations to Obtain Robust Analog-to-Digital Conversion," *IEEE Hrans. Commun.*, vol. COM-22, pp. 298-305 (1974).

[23] H. Taub and D. Shilling, *Digital Integrated Electronics* (McGraw-Hill, New York, 1977).

[24] D. Hoeschele, *Analog-to-Digital Digital-to-Analog Conversion Techniques* (Wiley, New York, 1968).

[25] D. Sheingold, *Analog-Digital Conversion Handbook* (Analog Devices, Norwood, MA, 1972).

[26] F. F. Lee and D. Lipshutz, "Floating-Point Encoding for Transcription of High-Fidelity Audio Signals," *J. Audio Eng. Soc.*, vol. 25, pp. 266-272 (May 1977).

[27] B. Blesser and F. Ives, "A Reexamination of the S/N Question for Systems with Time-Varying Gain or Frequency Response," *J. Audio Eng. Soc.*, vol. 20, pp. 638-641 (Oct. 1972).

[28] J. Kriz, "A 16 bit A-D-A Conversion System for High-Fidelity Audio Research," *IEEE Trans. Acoust., Speech, Sig. Proc.*, vol. ASSP-23, pp. 146-149 (1975).

[29] D. Osborne and M. Croll, "Digital Sound Signals: Bit-Rate Reduction Using an Experimental Digital Compander," BBC Research Eng. Div., Great Britain, Monograph 1974/41, 1974.

[30] K. Euler, M. Schichte, and E. Pfrenger, "A PCM Single-Channel Codec in LSI Technology with a 13-Segment Characteristic," in *Proc. IEEE, Int. Zuerich Seminar on Digital Commun.* (Mar. 1974), pp. B2, 1-4.

[31] H. Hessenmueller, "The Transmission mf Broadcast Programs in a Digital Integrated Network," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 17-20 (1973).

[32] R. Karwoski, "Predictive Coding for Greater Accuracy in Successive Approximation A/D Converters," presented to the 57th Convention of the Audio Engineering Society, Los Angeles, May 1977, preprint 1228.

[33] S. Qureshi and G. Fornery, "A 9.6/16 kb/s Speech Digitizer," in *Proc. IEEE Int. Conf. on Commun.*, (June 16, 1975), pp. (30) 31-36.

[34] B. Atal and M. Schroeder, "Adaptive Predictive Coding of Speech Signals," *Bell Sys. Tech. J.*, vol. 49, pp. 1973-1986 (1970).

[35] H. Spang and P. Schultheiss, "Reduction of Quantization Noise by Use of Feedback," *IRE Trans. Commun.*, vol. CS-10, pp. 373-380 (Dec. 1962).

[36] J. O'Neal, Jr., "Differential PCM for Speech and Data Signals," *IEEE Trans. Commun.*, vol. COM-20, pp. 900-912 (Oct. 1972).

[37] D. Lindorff, *Theory of Sampled Data Control Systems* (Wiley, New York, 1965).

[38] R. Steele, *Delta Modulation Systems* (Wiley, New York, 1975).

[39] J. Abate, "Linear and Adaptive Delta Modulation," *Proc. IEEE*, vol. 55, pp. 298-308 (Mar. 1967).

[40] H. Schindler, "Delta Modulation," *IEEE Spectrum*, vol. 7, pp. 69-78 (Oct. 1970).

[41] H. Levitt, G. McGonegal, and L. Cherry, "Perception of Slope Overload Distortion in Delta Modulated Speech Signals," *IEEE Trans. Audio Electroacoust.*, vol. AU-18, pp. 240-247 (Sept. 1970).

[42] R. Rouquette, "An Audio Delta Modulator Design and Evaluation," S.M. Dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1975.

[43] R. DeFreites, "A System for Home Reverberation," presented at the 54th convention of the Audio Engineering Society, Los Angeles, May 4-7, 1976.

[44] F. DeJager, "Delta Modulation, a Method for PCM Transmission Using a 1-bit Code," *Philips Res. Rep.*, vol. 7, pp. 442-466 (1952).

[45] Y. Ching, "Idle Channel Noise Characteristic of Delta Modulation with Step Imbalance and Quantizer Hysteresis," in *Proc. IEEE Int. Conf. on Commun.* (June 17-19, 1974), pp. 13A, 1-6.

[46] C. Dalton, "Delta Modulation for Sound-Signal Distribution: A General Survey," *BBC Eng. Rev.*, pp. 4-14 (July 1972).

[47] P. Nielsen, "On the Stability of Double Integration Delta Modulators," *IEEE Trans. Commun. Tech.*, vol. COM-19, pp. 364-366 (1971).

[48] C. Song, J. Garodnick, and D. Schilling, "A Variable Step Size Robust Delta Modulator," *IEEE Trans. Commun.*, vol. COM-19, pp. 1033-1044 (1971).

[49] N. Jayant, "Adaptive Delta Modulation with a One-Bit Memory," *Bell Sys. Tech. J.*, vol. 49, pp. 321-342 (1970).

[50] L. Zetterberg and J. Uddenfeldt, "Adaptive Delta Modulation with Delayed Decision," *IEEE Trans. Commun.*, vol. COM-22, pp. 1195-1198 (1974).

[51] N. Jayant, "Digital Coding of Speech Waveforms: PCM, DPCM, and DM Quantizers," *Proc. IEEE*, vol. 62, pp. 611-634 (1964).

[52] P. Cummisky, N. Jayant, and J. Flanagan, "Adaptive Quantization in Differential PCM Coding of Speech, *Bell Sys. Tech. J.*, vol. 52, pp. 115-118 (1973).

[53] R. Wegel and C. Lane, "The Auditory Masking of One Pure Tone by Another and Its Probable Relation to the Dynamics of the Inner Ear," *Phys. Rev.*, vol. 23, ser. 2, pp. 266-289 (1924).

[54] C. Robinson and I. Pollack, "Interactions Between Forward and Backward Masking: A Measure of the Integration Period of the Auditory System," *J. Acoust. Soc. Am.*, vol. 53, pp. 1313-1316 (1973).

[55] L. Elliot, "Backward and Forward Masking of Probe Tones of Different Frequencies," *J. Acoust. Soc. Am.*, vol. 34, pp. 1116-1117 (1962).

[56] A. Zverev, *Handbook of Filter Synthesis* (Wiley, New York, 1967).

[57] D. Humpherys, "Rational Function Approximation of Polynomials with Equiripple Error" Ph.D. Dissertation, Dep. of Elec. Eng., University of Illinois, 1963.

[58] V. Kuleshov and D. Nelsen, "A First-Passage Time Statistics of a First-Order Markov Process with a Linear Ramp," *Res. Lab. Elec. Quart. Progress Report*, vol. 97, pp. 181-184 (Apr. 1970).

[59] E. Belger, "On Measuring Frequency Variations," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 79-80 (1972).

[60] W. Manson, "Digital Sound Signals: Subjective Effect of Timing Jitter," BBC Research Eng. Div., Great Britain, Monograph 1974/11, 1974.

[61] D. Freeman, "Slewing Distortion in Digital-to-Analog Conversion," *J. Audio Eng. Soc.*, vol. 25, pp. 178-183 (Apr. 1977).

[62] G. S. Moschytz, "2nd Order Pole Zero Selection for nth Order Minimum Sensitivity Networds," *IEEE Trans. Cir. Theory*, vol. CT-17, pp. 527-534 (1970).

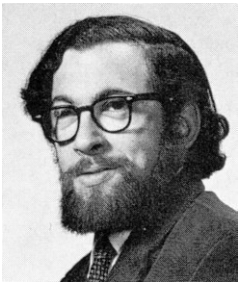
[63] C. M. Tsai, "A Design Technique for Testing A/D and D/A Converters," M.S. Dissertation, University of Utah, Department of Computer Science, May 1973.

[64] R. P. Talambiras, "Digital-To-Analog Converters: Some Problems in Producing High-Fidelity Signals," *Computer Design*, Jan. 1976, pp. 63-69.

[65] R. P. Talambiras, "Some Considerations in the Design of Wide-Dynamic Range Audio Digitizing Systems," presented to the 57th Convention of the Audio Engineering Society, Los Angeles, May 10-13, 1977, preprint 1226.

[66] G. Temes and D. Callahan, "Computer-Aided Network Optimization the State of the Art," *Proc. IEEE*, vol. 55, pp. 1832-1863, (Nov. 1967).

THE AUTHOR



Barry A. Blesser is currently devoting full time to his new consulting company which specializes in providing high-level engineering research and development services. He is especially active in the fields of professional audio, character recognition, and electronic system design (analog and digital). His most notable achievements have been the development of an all-electronic reverberation system and a real-time hand-print recognition system.

For the previous 10 years, Dr. Blesser pursued a more academic career at MIT as an associate professor of electrical engineering and computer science. He taught several graduate and undergraduate courses including advanced topics in instrumentation. In addition, he was a staff scientist in the Research Laboratory of Electronics where he maintained a research program in perceptual and cognitive processes.

Dr. Blesser is a Fellow of the Audio Engineering Society and a Senior Member of the IEEE.