Kenzo Akagiri, et. Al. “Sony Systems.”

Sony Systems

Kenzo Akagiri
Sony Corporation
(Tokyo)

M.Katakura
Sony Corporation
(Kanagawa)

H. Yamauchi
Sony Corporation
(Kanagawa)

E. Saito
Sony Corporation
(Kanagawa)

M. Kohut
Sony Corporation
(California)

Masayuki Nishiguchi
Sony Corporation
(Tokyo)

K. Tsutsui
Sony Corporation
(Tokyo)

43.1 Introduction
43.2 Oversampling AD and DA Conversion Principle
Concept • Actual Converters
References
43.4 The SDDS System for Digitizing Film Sound
Film Format • Playback System for Digital Sound • The SDDS Error Correction Technique • Features of the SDDS System
43.5 Switched Predictive Coding of Audio Signals for the CD-I and CD-ROM XA Format
Abstract • Coder Scheme • Applications
References
43.7 ATRAC (Adaptive Transform Acoustic Coding) and ATRAC 2
ATRAC • ATRAC2
References

43.1 Introduction

Kenzo Akagiri

In digital signal processing, manipulating of the signal is defined as an essentially mathematical procedure, while the AD and DA converters, the front end and the final stage devices of the processing, include analog factor/limitation. Therefore, the performance of the devices determines the degradation from the theoretical performance defined by the format of the system.

Until the 1970s, AD and DA converters with around 16-bit resolution, which were fabricated by module or hybrid technology, were very expensive devices for industry applications. At the beginning of the 1980s, the CD (compact disk) player, the first mass-production digital audio product, was introduced, and required low cost and monolithic type DA converters with 16-bit resolution. The two-step dual slope method [1] and the DEM (Dynamic Element Matching) [2] method were used in the first generation DA converters for CD players. These were methods which relieved the accuracy and matching requirements of the elements to guarantee conversion accuracy by circuit technology. Introducing new ideas on circuit and trimming, like segment decode and laser trimming of the thin film fabricated on monolithic silicon die, for example, classical circuit topologies using binary weighted current source were also used. For AD conversion at same generation, successive approximation topology and the two-step dual slope method were also used.

In the mid-1980s, introductions of the oversampling and the noise shaping technology to the AD and DA converters for audio applications were investigated [3]. The converters using the technologies are the most popular devices for recent audio applications, especially as DA converters.

43.2 Oversampling AD and DA Conversion Principle

M. Katakura

43.2.1 Concept

The concept of the oversampling AD and DA conversion, DS or SD modulation, was known in the 1950s; however, the device technology to fabricate actual devices was impracticable until the 1980s [4].

The oversampling AD and DA conversion is characterized by the following three technologies.

1. oversampling
2. noise shaping
3. fewer bit quantizer (converters used one bit quantizer called the DS or SD type)

It is well known that the quantization noise shown in the next equation is determined by only quantization step D and distributed in bandwidth limited by Nyquist frequency (2/fs), and the spectrum is almost similar to white noise when the step size is smaller than the signal level.

$$V_n = \Delta/\sqrt{12} \tag{43.1}$$

As shown in Fig. 43.1, the oversampling expands a capacity of the quantization noise cavity on the frequency axis and reduces the noise density in the audio band, and the noise shaping moves it to out of the band. Figure 43.2 is first-order noise shaping to show the principle of the noise shaping, in which the quantizer is represented by the adder fed an input $U(n)$ and a quantization noise $Q(n)$. $Y(n)$ and $U(n)$, the output and input signals of the quantizer, respectively, are given as follows:

$$Y_{(n)} = U_{(n)} + Q_{(n)} \tag{43.2}$$

$$U_{(n)} = X_{(n)} + Q_{(n-1)} \tag{43.3}$$

As a result, the output $Y(n)$ is

$$Y_{(n)} = X_{(n)} + \left\{Q_{(n)} - Q_{(n-1)}\right\} \tag{43.4}$$

The quantization noise in output $Y(n)$, which is a differentiation of the original quantization noise $Q(n)$ and $Q(n-1)$ shifted a time step, has high frequency boosted spectrum. Equation (43.4) is written as follows using z

$$Y_{(z)} = X_{(z)} + Q_{(z)}\left(1 - Z^{-1}\right) \tag{43.5}$$

The oversampling conversion using one bit quantizer is called DS or SD AD/DA converters. Regarding one bit quantizer, a mismatch of the elements does not affect differential error; in other

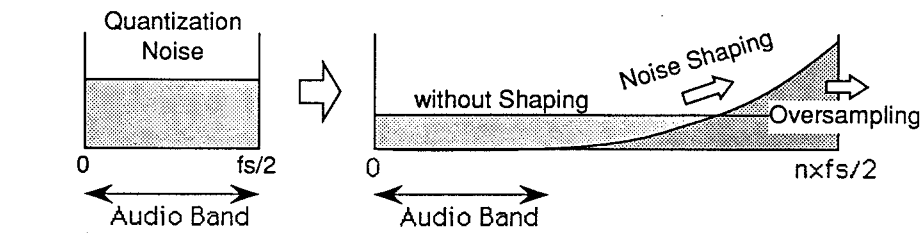


FIGURE 43.1: Quantization noise of the oversampling conversion.

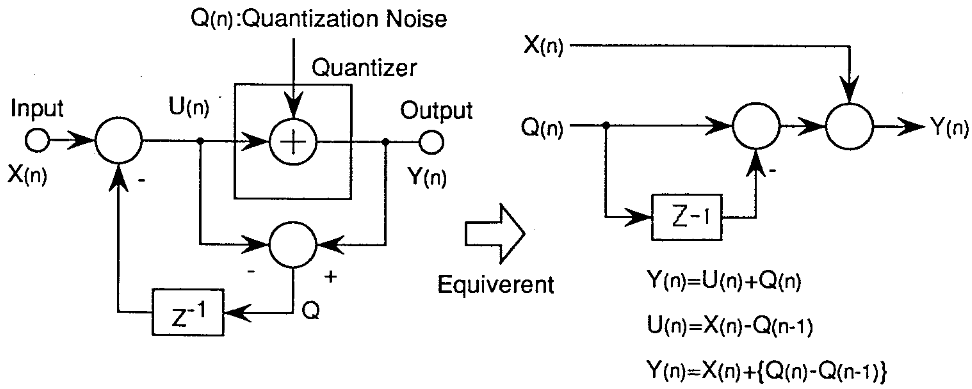


FIGURE 43.2: First-order noise shaping.

words, it has no non-linear error. Assume output swing of the quantizer is ± D, quantization noise $Q(z)$ is white noise, and the magnitude $|Q(Wt)|$ is D2/3, which corresponds to four times in power of Eq. (43.1) since the step size is twice that. Define q which is $2p \cdot f$ max $/fs$, where f max and fs are the audio bandwidth and the sampling frequency, respectively, then the in-band noise in Eq. (43.5) becomes

$$\begin{aligned}
\bar{N}^2 &= \left|Q_{(\omega T)}\right|^2 \frac{1}{2\pi}\int_{-\theta}^{\theta}\left|H_{(\omega T)}\right|^2 d_{(\omega T)} \\
&= \frac{\Delta^2}{3}\frac{1}{2\pi}\int_{-\theta}^{\theta}\left|1-e^{-j\omega T}\right|^2 d_{(\omega T)} \\
&= \frac{\Delta^2}{3}\frac{2}{\pi}(\theta - \sin\theta) \\
&= \frac{\Delta^2}{9\pi}\theta^3
\end{aligned} \tag{43.6}$$

The oversampling conversion has the following remarkable advantages compared with traditional methods.

1. It is easy to realize "good" one bit converters without superior device accuracy and matching.
2. Analog anti-aliasing filters with sharp cutoff characteristics are unnecessary due to oversampling.

Using the oversampling converting technology, requirements for analog parts are relaxed; however, they require large scale digital circuits because interpolation filters in front of the DA conversion, which increase sampling frequency of the input digital signal, and decimation filters after the AD conversion, which reject quantization noise in high frequency and reduce sampling frequency, are required.

Figure 43.3 shows the block diagram of the DA converter including an interpolation filter. Though the scheme of the noise shaper is different from that of Fig. 43.2, the function is equivalent. Figure 43.4 shows the block diagram of the AD converter including a decimation filter. Note that the AD converter is almost the same as with the DA converters regarding the noise shapers; however, the details of the hardware are different depending on whether the block handles analog or digital signal. For example, to handle digital signals the delay units and the adders should use latches and digital adders; on the other hand, to handle analog signals delay units and adders using switched capacitor topology should be used. In the DS type, the quantizer is just reduction data length to one bit for the DA converter, and is a comparator for the AD converter by the same rule.

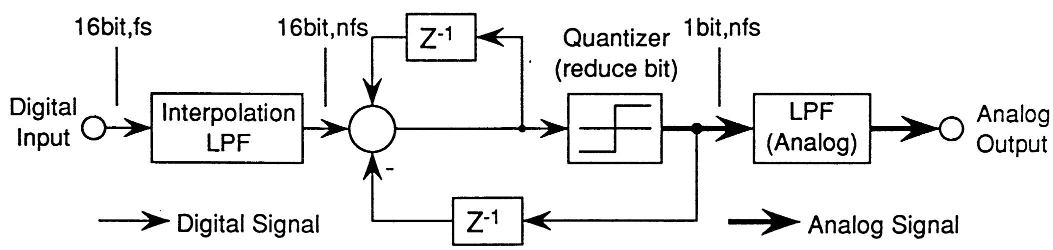


FIGURE 43.3: Oversampling DA converter.

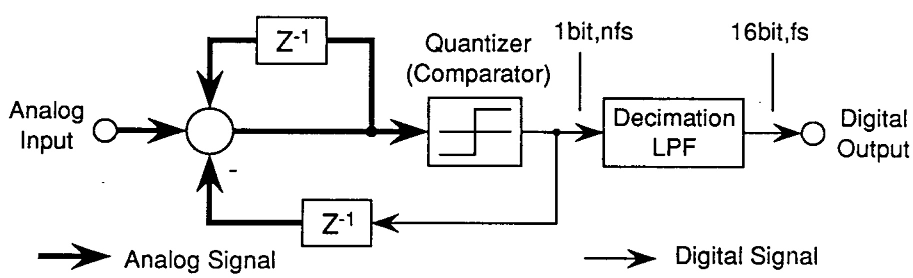


FIGURE 43.4: Oversampling AD converter.

43.2.2 Actual Converters

To achieve resolution of 16 bits or more for digital audio applications, the first-order noise shaping is not acceptable because it requires an extra high oversampling ratio, and the following technologies are actually adopted.

- High-order noise shaping

- Multi-stage (feedforward) noise shaping
- Interpolative conversion

1. High-order noise shaping

Figure 43.5 shows quantization noise spectrum for order of the noise shaping. The third-order noise shaping achieves 16-bit dynamic range using less than an oversampling ratio of 100. Figure 43.6 shows a third-order noise shaping for example of the high order. Order of the noise shaping used is 2 to 5 for audio applications.

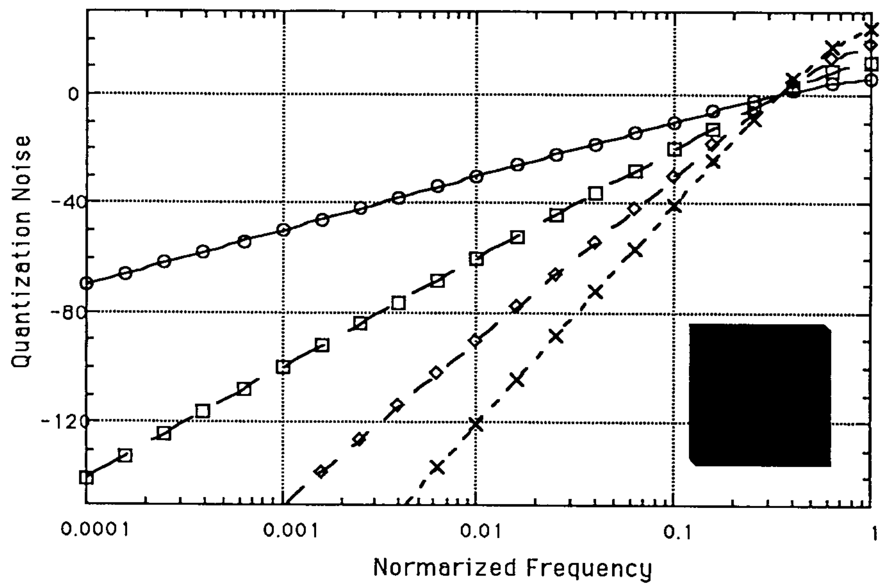


FIGURE 43.5: Quantization noise vs. order of noise shaping.

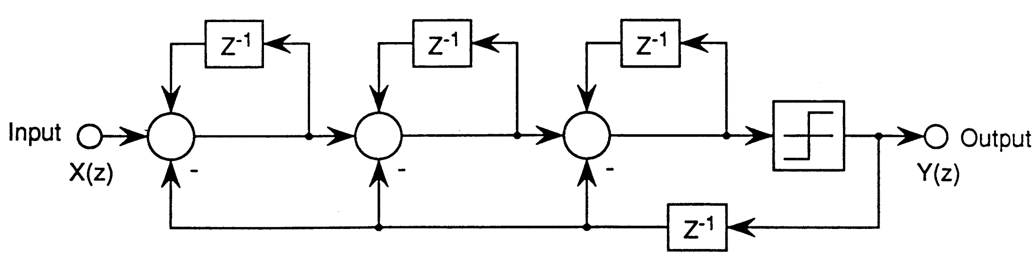


FIGURE 43.6: Third-order noise shaping.

In Fig. 43.6 output $Y(z)$ is given

$$Y_{(z)} = X_{(z)} + Q_{(z)}\left(1 - Z^{-1}\right)^3 \tag{43.7}$$

The high-order noise shaping has a stability problem because the phase shift of the open loop in more than a third-order noise shaping exceeds 180°. In order to guarantee the stability, an amplitude limiter at the integrator outputs is used, and modification of the loop transfer function is done, although it degrades the noise shaping performance slightly.

2. Multi-stage (feedforward) noise shaping [5]

Multi-stage (feedforward) noise shaping (called MASH) achieves high-order noise shaping transfer functions using not high-order feedback but feedforward, and is shown in Fig. 43.7. Though two-stage (two-order) is shown in Fig. 43.7, three-stage (three-order) is usually used for audio applications.

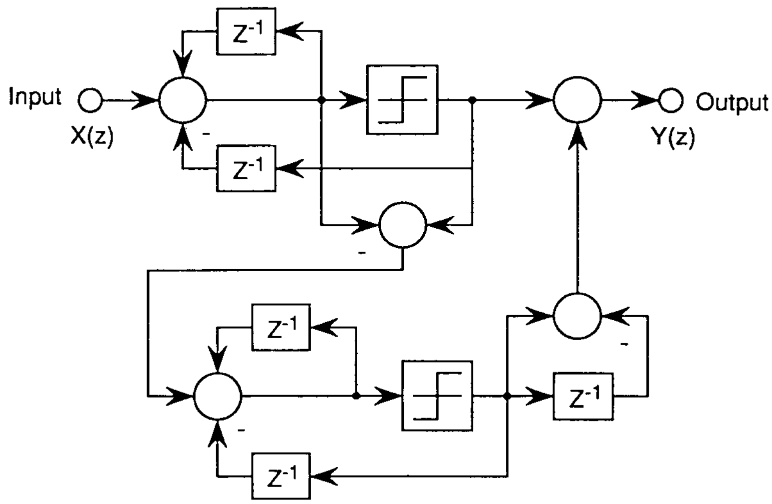


FIGURE 43.7: Multi-stage noise shaping.

3. Interpolative converters [6]

This is a method which uses a few bit resolution converters instead of one bit. The method reduces the oversamplimg ratio and order of the noise shaping to guarantee specified dynamic range and improve the loop stability. Since absolute value of the quantization noise becomes small, it is relatively easy to guarantee noise level; however, linearity of large signal conditions affects the linearity error of the AD/DA converters used in the noise shaping loop.

Oversampling conversion has become a major technique in digital audio application, and one of the distinctions is that it does not inherently zero cross distort. For recent device technology, it is not so difficult to guarantee 18-bit accuracy. Thus far, the available maximum dynamic range is slightly less than 20 bit (120 dB) without noise weighting (wide band) due to analog limitation. On the other hand, converters with 20-bit or more resolution have been reported [7] and are expected to improve sound quality in very small signal levels from the standpoint of hearing.

References

[1] Kayanuma, A. et al., An integrated 16-bit A/D converter for PCM audio systems, *ISSCC Dig. Tech. Papers*, pp. 56-57, Feb., 1981.

[2] Plassche, R. J. et al., A monolithic 14-bit D/A converter, *IEEE J. Solid State Circuits*, SC-14:552-556, 1979.
[3] Naus, P. J. A. et al., A CMOS stereo 16-bit D/A converter for digital audio, *IEEE J. Solid State Circuits*, SC-22:390-395, June, 1987.
[4] Hauser, M. W., Overview of oversampling A/D Converters, Audio Engineering Society Preprint #2973, 1990.
[5] Matsuya, Y. et al., A 16-bit oversampling A-to-D conversion technology using triple-integration noise shaping, *IEEE J. Solid State Circuits*, SC-22:921-929, Dec., 1987.
[6] Schouwenaars, H. J. et al., An oversampling multibit CMOS D/A converter for digital audio with 115 dB dynamic range, *IEEE J. Solid State Circuits*, SC-26:1775-1780, Dec., 1991.
[7] Maruyama, Y. et al., A 20-bit stereo oversampling D-to-A converter, *IEEE Trans. Consumer Electron.*, 39:274-276, Aug., 1993.

43.4 The SDDS System for Digitizing Film Sound

H. Yamauchi, E. Saito, and M. Kohut

43.4.1 Film Format

There are three basic concepts for developing the SDDS format. They can

1. Provide sound quality similar to CD sound quality. We adapt ATRAC (Adaptive TRansform Acoustic Coding) to obtain good sound quality equivalent to that of CDs. ATRAC is the compression method used in the mini disc (MD) which has been in sale since 1992. ATRAC enables one record digital sound data by compressing about 1/5 of the original sound.

2. Provide enough numbers of sound channels with good surround effects. We have eight discrete channel systems and six channels to the screen in the front and two channels in the rear as surround speakers shown in Fig. 43.8. We have discrete channel systems, making a good channel separation which provides superior surround effects even in a large theater with no sound defects.

3. Be compatible with the current widespread analogue sound system. There are limited spaces between the sprockets, picture frame, and in the external portion of the sprocket hole where the digital sound could be recorded because the analogue sound track is left as usual. As in the cinema scope format, it may be difficult to obtain enough space between picture frames. Because the signal for recording and playback would become intermittent between sprockets, special techniques would be required to process such signals. As shown in Fig. 43.9, we therefore establish track P and track S on a film external portion where continuous recordings are possible and where space can be obtained in the digital sound recording region on the SDDS format.

Data bits are recorded on the film with black and white dot patterns. The size of a bit is decided to overcome the effects caused by film scratch and is able to correct errors. In order to obtain the certainty of reading data, we set a guard band area to the horizontal and track direction.

Now, the method to record digital sound data on these two tracks is to separate eight channels and record four channels each in track P and in track S. A redundant data is also recorded about 18 frames later on the opposite track. By this method, it makes it possible to obtain the equivalent data from track S if any error occurs on track P and the correction is unable to be made, or vice versa. This is called the "Digital Backup System".

Figure 43.10 shows the block structure for the SDDS format. A data compression block of the ATRAC system has 512 bit sound data per film block. A vertical sync region is set at the head of the film block. A film block ID is recorded in this region to reproduce the sound data and picture frame

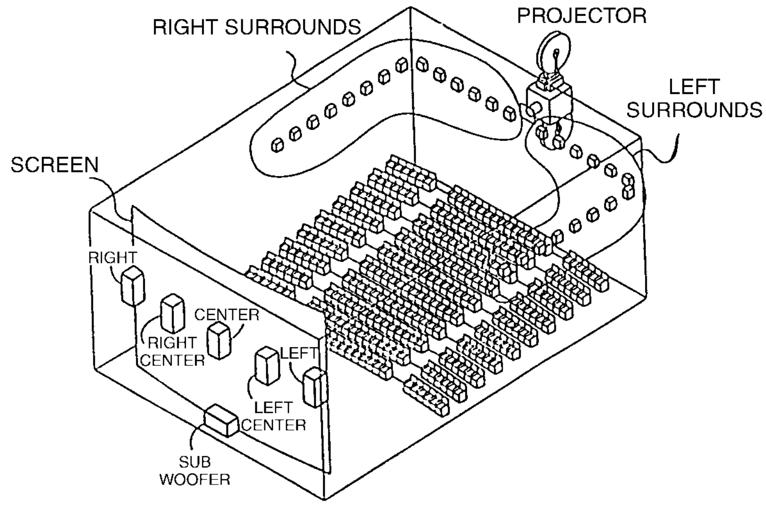


FIGURE 43.8: Speaker arrangement in theater.

with the right timing and to prevent the "lip sync" offset from discordance; for example, the time accordance between an actor's/actress' lip movement and his/her voice. Also, a horizontal sync is set on the left-hand side of the film block and is referred to correctly detect the head of the data in reading with the line sensor.

43.4.2 Playback System for Digital Sound

The digital playback sound system for the SDDS system consists of a reader unit, DFP-R2000, and a decoder unit, DFP-D2000 as shown in Fig. 43.11. The reader unit is set between the supply reel and the projector.

The principle of digital sound reading for the reader unit DFP-R2000 is described in Fig. 43.12. The LED light source is derived from the optical fiber and it scans the data portion recorded on track P and track S of the film. Transparent lights through the film give an image formation on the line sensor through the lens. These optical systems are designed to have the appropriate structures which can hardly be affected by scratches on the film. The output of a sensor signal is transmitted to the decoder after the signal processing such as the wave form equalization is made.

The block diagram of the decoder unit DFP-D2000 is shown in Fig. 43.13. The unit consists of EQ, DEC, DSP, and APR blocks.

In the EQ, signals become digital signals after being equalized. Then the digital signals are transmitted to the DEC together with the regenerated clock signal.

In the DEC, jitters elimination and lip sync control are done by the time base collector circuit, and errors caused by scratches and dust on the film are corrected by the strong error correction algorithm. Also in the DEC, signals for track P and track S which have been compressed by the ATRAC system are decoded. This data is transmitted to the DSP as a linear PCM signal.

In the DSP, the sound field of the theater is adjusted and concealment modes are controlled. A CPU is installed in the DSP to control the entire decoder, and control the front panel display and reception and transmission of external control data.

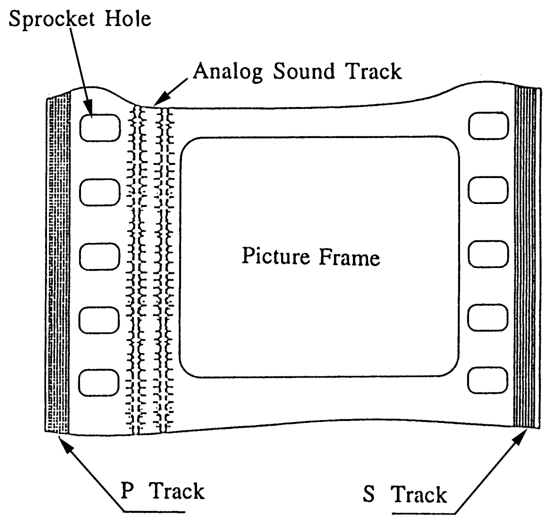


FIGURE 43.9: SDDS track designation.

Finally in the APR, 10 channels of digital filter including monitors, D/A converter, and line amplifier are installed. Also, it is possible to directly bypass an analogue input signal by relay as necessary. This bypass is prepared to cope with analogue sound if digital sound would not play back.

43.4.3 The SDDS Error Correction Technique

The SDDS system adapts the "Reed Solomon" code for error correction. An error correction technique is essential for maintaining high sound quality and high picture quality for digital recording and playback systems, such as CD, MD, digital VTR, etc. Such C1 parity + C2 parity data necessary for error correction are added and recorded in advance to cope with cases when the correct data are not able to be obtained. It enables recovery of the correct data by using this additional data even if a reading error occurs.

If the error rate is 10^{-4} (1 bit for every 10,000 bits), the error rate for C1 parity after correction would normally be 10^{-11}. In other words, an error would occur only once every 1.3 years if a film were showed 24 hours a day. Errors will be extremely close to "zero" by using C2 parity erasure correction. A strong error correction capability is installed in the SDDS digital sound playback system against random errors.

Other errors besides random errors are

- errors caused by a scratch in the film running direction
- errors caused by dust on the film
- errors caused by splice points of films
- errors caused by defocusing during printing or playback

These are considered burst errors which occur consistently. Scratch errors in particular will increase more and more every time the film is shown. SDDS has the capability of dealing with such burst errors. Therefore, in spite of the scratch on the film width direction, error correction towards the film length would be possible up to 1.27 mm and in spite of the scratch on the film running direction, error correction towards the film width would be possible up to 336 μ m.

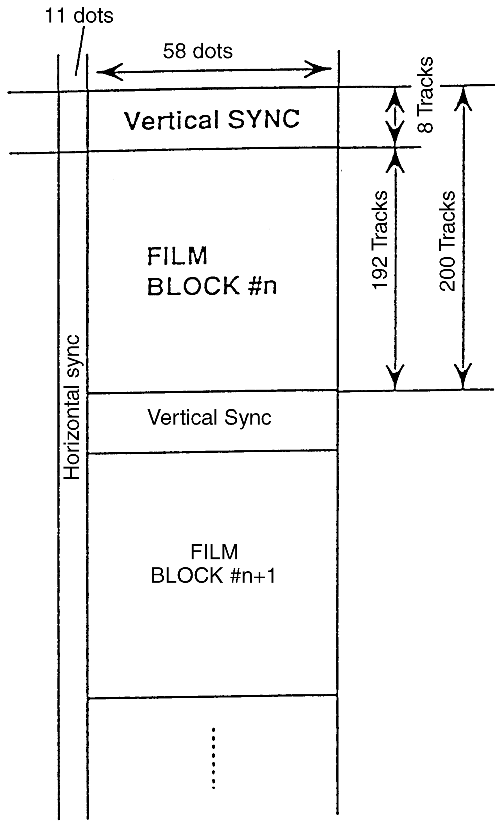


FIGURE 43.10: Data block configuration.

43.4.4 Features of the SDDS System

The specification characteristics for the SDDS player are shown in Table 43.1. It is not easy to obtain high fidelity in audio data compression compared to the linear recording system of CDs with regard to a sound quality. By adapting a system with high compression efficiency and making use of the human hearing characteristics, we were able to maintain a sound quality equivalent to CDs by adapting the ATRAC system which restrains deterioration to the minimum.

One of the biggest features of the SDDS is the adaption of a digital backup system. This is a countermeasure system to make up for the damage to the splicing parts of the digital data or the parts of data missing by using the opposite side of the track with a digital data recorded on the backup channel. By this system, it would be possible to obtain an equivalent quality. Next, when finally the film is worn out, the system switches over to an analogue playback signal.

This system also has a digital room EQ function. This supplies 28 bands of graphic EQ with 1/3 octave characteristics and a high and low pass filter. Moreover, a simple operation to control the sound field in the theater will become possible by using a graphic user interface panel of an external personal computer.

Such control usually took hours, but it can be completed in about 30 min with this SDDS player. The stability of its features, reproducibility, and reliability of digitizing is well appreciated.

Furthermore, the SDDS player carries a backup function and a reset function for setting parameters by using memories.

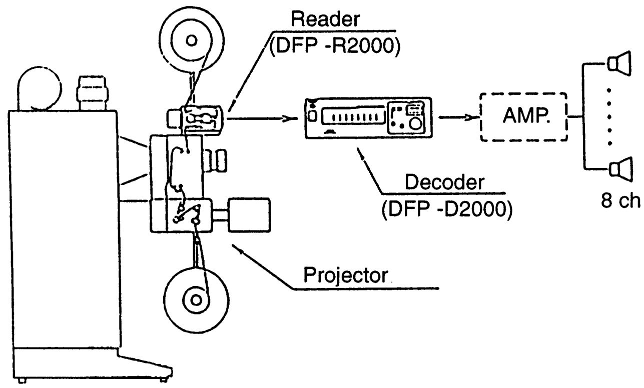


FIGURE 43.11: Playback system.

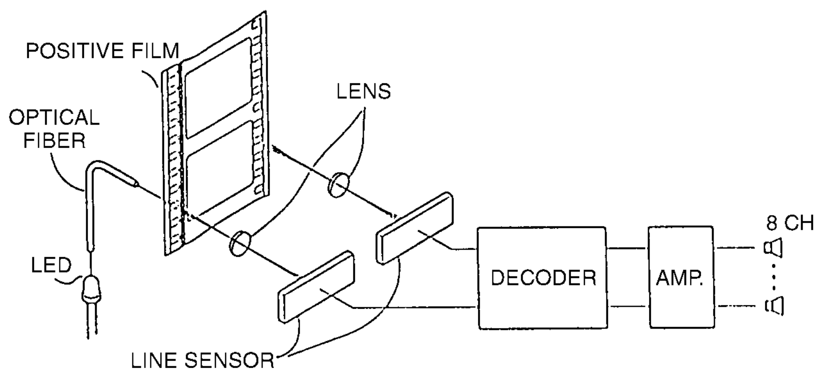


FIGURE 43.12: Optical reader concept.

43.5 Switched Predictive Coding of Audio Signals for the CD-I and CD-ROM XA Format

Masayuki Nishiguchi

43.5.1 Abstract

An audio bit rate reduction system for the CD-I and CD-ROM XA format based on switched predictive coding algorithm is described. The principal feature of the system is that the coder provides multiple prediction error filters, each of which has fixed coefficients. The prediction error filter that best matches the input signal is selected every 28 samples (1 block). A first-order and two kinds of second-order prediction error filters are used for signals in the low and middle frequencies, and the straight PCM is used for high-frequency signals. The system also uses near-instantaneous companding to expand the dynamic range. A noise-shaping filter is incorporated in the quantization stage, and its frequency response is varied to minimize the energy of the output noise. With a complexity of less than 8MIPS/channel, audio quality almost the same as CD audio can be achieved at 310 Kbps (8.2 bits/sample), near transparent audio can be achieved at 159 Kbps (4.2 bits/sample), and mid-fidelity audio can be achieved at 80 Kbps (4.2 bits/sample).

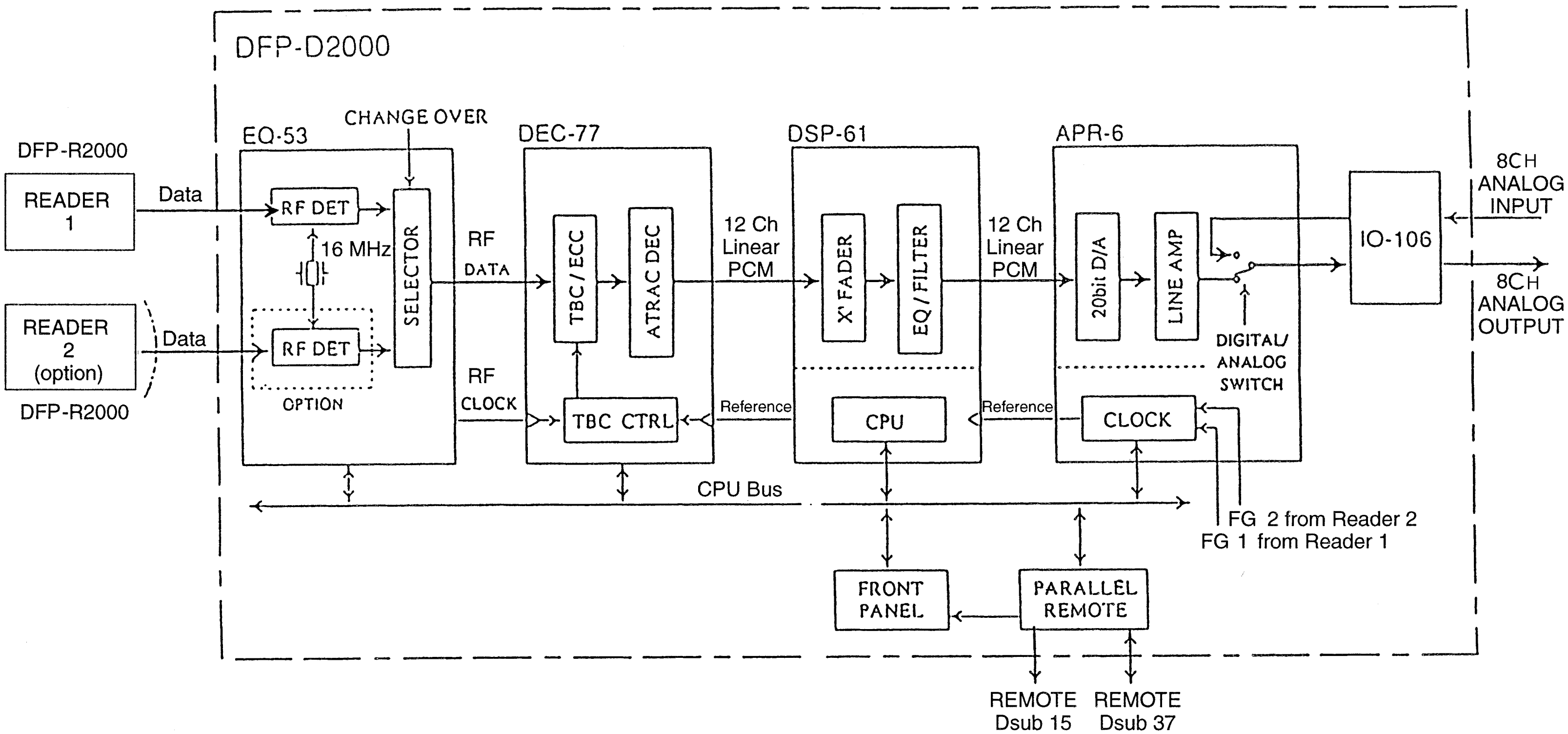


FIGURE 43.13: Overall block diagram.

TABLE 43.1 SDDS Player System Electrical Specifications

Item	Specification
Sampling frequency	44.1 KHz
Dynamic range	Over 90 dB
Channel	Max 8 cH
Frequency band	20 Hz - 20 KHz
	$+/-1.0$dB
K.F	$< 0.7\%$
Crosstalk	< -80 dB
Reference output level	-10 dB / Unbalanced
	$+4$ dB / Balanced
Head room	> 20 dB / Balanced

43.5.2 Coder Scheme

Figure 43.14 is a block diagram of the encoder and decoder system. The input signal, prediction error, quantization error, encoder output, decoder input, and decoder output are respectively expressed as $x(n)$, $d(n)$, $e(n)$, $\hat{d}(n)$, $\hat{d}'(n)$, and $\hat{x}'(n)$. The z-transforms of the signals are expressed as $X(z)$, $D(z)$, $E(z)$, $\hat{D}(z)$, $\hat{D}'(z)$, and $\hat{X}'(z)$. The encoder response can then be expressed as

$$\hat{D}(z) = G \cdot X(z) \cdot \{1 - P(z)\} + E(z) \cdot \{1 - R(z)\}\,, \tag{43.8}$$

and the decoder response as

$$\hat{X}'(z) = \frac{G^{-1} \cdot \hat{D}'(z)}{1 - P(z)}\,, \tag{43.9}$$

Assuming that there is no channel error, we can write $\hat{D}'(z) = \hat{D}(z)$. Using Eq. (43.8) and (43.9), we can write the decoder output in terms of the encoder input as

$$\hat{X}'(z) = X(z) + G^{-1} \cdot E(z) \cdot \frac{1 - R(z)}{1 - P(z)}\,. \tag{43.10}$$

where

$$P(z) = \sum_{k=1}^{P} \alpha_k \cdot z^{-k} \quad \text{and} \quad R(z) = \sum_{k=1}^{R} \beta_k \cdot z^{-k} \tag{43.11}$$

Here α_k and β_k are, respectively, the coefficients of predictor $P(z)$ and $R(z)$. Equation (43.10) shows the encoder-decoder performance characteristics of the system. It shows that the quantization error $E(z)$ is reduced by the extent of the noise-reduction effect G^{-1}. The distribution of the noise spectrum that appears at the decoder output is

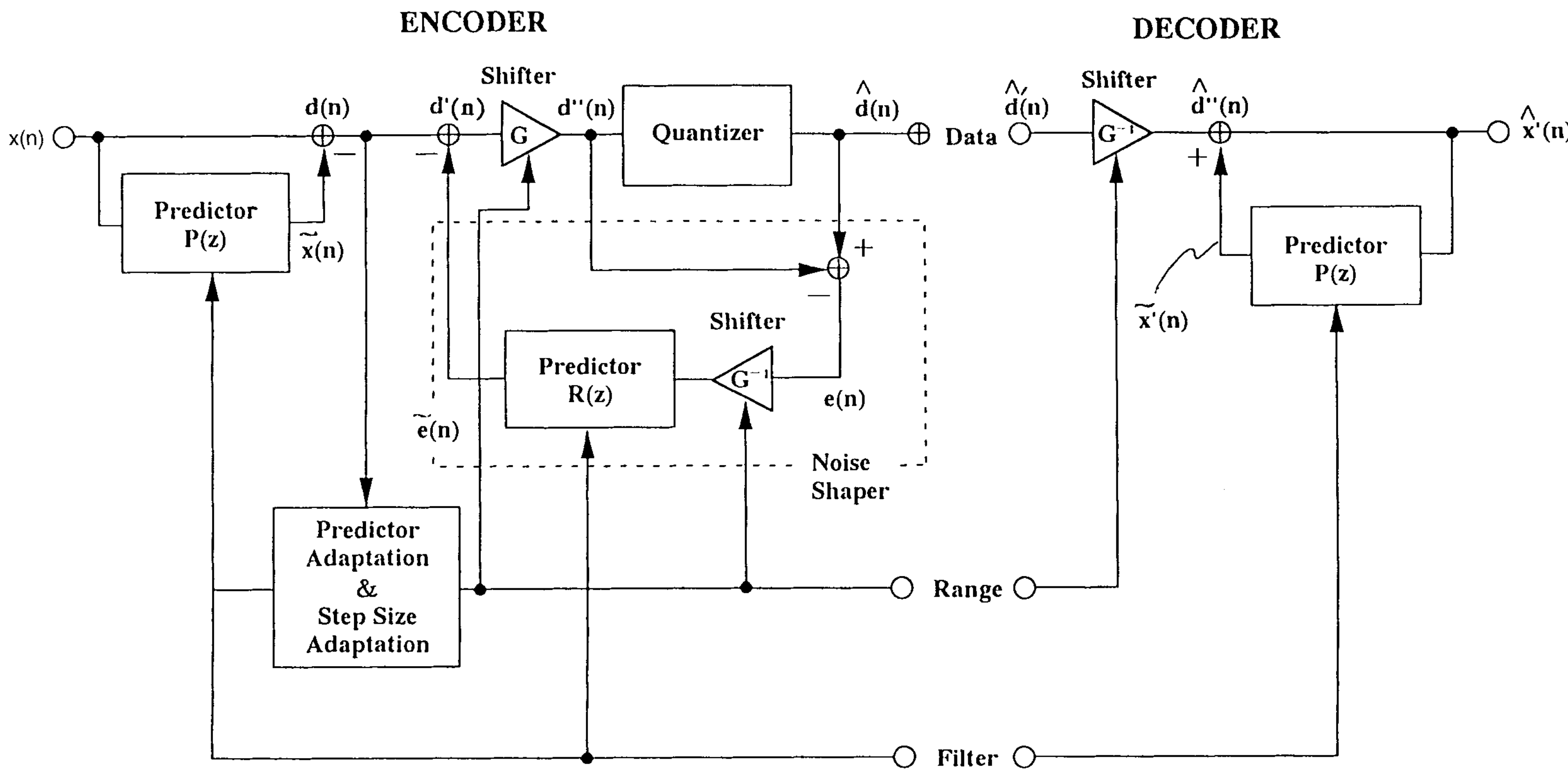
$$N(z) = E(z) \cdot \frac{1 - R(z)}{1 - P(z)}\,. \tag{43.12}$$

$R(z)$ can be varied according to the spectral shape of the input signal in order to have a maximum masking effect, but we have set $R(z) = P(z)$ to keep from coloring the quantization noise.

G can be regarded as the normalization factor for the peak prediction error (over 28 residual words) from the chosen prediction error filter. The value of G changes according to the frequency response of the prediction gain:

$$G \;\propto \frac{|X(z)|}{|D(z)|}\,. \tag{43.13}$$

This is also proportional to the inverse of the prediction error filter, $1/|1 - P(z)|$. So, in order to maximize G, it is necessary to change the frequency response of the prediction error filter $1 - P(z)$ according to the frequency distribution of the input signals.



$x(n)$: encoder input

$\tilde{x}(n)$: predicted value

$e(n)$: quantization error

$\tilde{e}(n)$: predicted value of quantization error

$d(n)$: prediction error

$\hat{d}(n)$: encoder output

$\hat{d}'(n)$: decoder input

$\tilde{x}'(n)$: predicted value

$\hat{x}'(n)$: decoder output

FIGURE 43.14: Block diagram of the bit rate reduction system.

Selection of the Optimum Filter

Several different strategies of selecting filters are possible in the CD-I/CD-ROM XA format, but the simplest way for the encoder to choose which predictor is most suitable is the following:

- The predictor adaptation section compares the peak value of the prediction errors (over 28 words) from each prediction error filter $1 - P(z)$ and selects the filter that generates the minimum peak.
- The group of prediction errors chosen is then gain controlled (normalized by its maximum value) and noise shaping is executed at the same time.

As a result, a high SNR is obtained by using a first-order and two kinds of second-order prediction error filters for signals with the low and middle frequencies and by using the straight PCM for high-frequency signals.

Coder Parameters

This system provides three bit rates for the CD-I/CD-ROM XA format, and data encoded at any bit rate can be decoded by a single decoder. The following sections explain how the parameters used in the decoder and the encoder change according to the level of sound quality. Table 43.2 lists the parameters for each level.

TABLE 43.2 The Parameters for Each Level

	Level A	Level B	Level C
Sampling frequency (KHz)	37.8	37.8	18.9
Residual word length (bits per sample)	8	4	4
Block length (Number of samples)	28	28	28
Range data (bits per block)	4	4	4
Range values	0-8	0-12	0-12
Filter data (bits per block)	1	2	2
Number of prediction error filters used	2	3	4
Average of bits used per sample (bits per sample)	8.18 $= (8 \times 28 + 4 + 1)/28$	4.21 $= (4 \times 28 + 4 + 1)/28$	4.21 $= (4 \times 28 + 4 + 1)/28$
Bit rate (Kbps)	309	159	80

Level A

We can obtain the highest quality audio sound with Level A, which uses only two prediction error filters. Either the straight PCM or the first-order differential PCM is selected. The transfer functions of the prediction error filters are as follows:

$$H(z) = 1 \tag{43.14}$$

and

$$H(z) = 1 - 0.975z^{-1}\,, \tag{43.15}$$

where $H(z) = 1 - P(z)$.

Level B

The bit rate at Level B is half as high as that at Level A. By using this level, we can obtain high-fidelity audio sound from most high-quality sources. This level uses three filters: the straight PCM, the first-order differential PCM, or the second-order differential PCM-1 is selected. The transfer functions of the first two filters are the same as in Level A, and that for the second-order differential PCM-1 mode is:

$$H(z) = 1 - 1.796875z^{-1} + 0.8125z^{-2} . \tag{43.16}$$

Level C

We can obtain mid-fidelity audio sound at Level C, and a monoaural audio program 16 hours long can be recorded on a single CD. Four filters are used for this level. The transfer function of the first three filters are the same as in Level B. The transfer function of the second-order differential PCM-2 mode, used only at this level, is

$$H(z) = 1 - 1.53125z^{-1} + 0.859375z^{-2} . \tag{43.17}$$

At all levels, the noise-shaping filter and the inverse-prediction-error filter in the decoder have the same coefficients as the prediction error filter in the encoder.

43.5.3 Applications

The simple structure and low complexity of this CD-I/CD-ROM XA audio compression algorithm make it suitable for applications with PCs, workstations, and video games.

References

[1] Nishiguchi, M., Akagiri, K. and Suzuki. T., A new audio bit-rate reduction system for the CD-I format, Preprint 81st AES Convention, Nov. 1986.

[2] Rabiner, L.R. and Schafer, R.W., *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, 1978.

[3] Oppenhein, A.V. and Schafer, R.W., *Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1975.

43.7 ATRAC (Adaptive Transform Acoustic Coding) and ATRAC 2

K. Tsutsui

43.7.1 ATRAC

ATRAC is a coding system designed to meet the following criteria for the MiniDisc system:

- Compression of 16-bit 44.1-kHz audio (705.6 kbps) into 146 kbps with minimal reduction in sound quality.
- Simple hardware implementation suitable for portable players and recorders.

Block diagrams of the encoder and decoder structures are shown in Figs. 43.15 and 43.16, respectively. The time-frequency analysis block of the encoder decomposes the input signal into spectral coefficients grouped into 52 block floating units (BFUs). The bit allocation block divides the available bits among the BFUs adaptively based on the psychoacoustics. The spectrum quantization block normalizes spectral coefficients with the scale factor given to each BFU, and then quantizes each

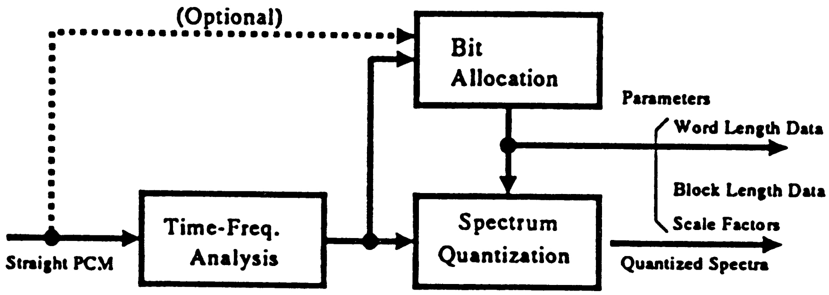


FIGURE 43.15: ATRAC encoder.

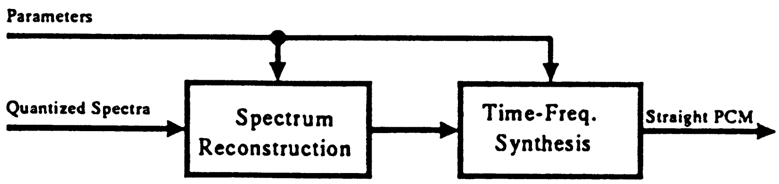


FIGURE 43.16: ATRAC decoder.

of them to the specified word length. These processes are performed in every sound unit, a block consisting of 512 samples per channel.

In order to generate the BFUs, the time-frequency analysis block first divides the input signal into three subbands. And then, each of these subbands is transformed into the frequency domain by modified discrete cosine transform (MDCT), producing a set of spectral coefficients. Finally, these spectral coefficients are nonuniformly grouped into BFUs. The subband decomposition is performed using cascaded 48-tap quadrature mirror filters (QMFs). The input signal is divided into upper and lower frequency bands by the first QMF, and then, the lower-frequency band is divided again by a second QMF. While the output samples of each filter are decimated by two, the aliasing caused by the subband decomposition is cancelled during reconstruction, due to the use of QMFs. MDCT block length is adaptively determined based on the signal characteristics in each band. There are two block-length modes: long mode (11.6 msec for $f_s = 44.1$ kHz) and short mode (1.45 ms in the high frequency band, 2.9 ms in the others). Normally, long mode is chosen, as this provides good frequency resolution. However, problems occur during attack portions of the signal since the quantization noise is spread over the entire block and the initial quantization noise is not masked by simultaneous masking. In order to prevent this degradation known as pre-echo, ATRAC switches to short mode when it detects an attack signal. In this case, as the noise before the attack exists only for a very short period of time, it is masked by backward masking. The window form is symmetric for both long and short modes, and the window form in the non-zero-nor-one region of the long mode is the same as that of the short mode. Although this window form is somewhat disadvantageous to the separability of the spectrum, it brings the following merits:

- The transform mode can be determined based only on the existence of an attack signal in the current sound unit, and hence, no extra buffer is required in the encoder.
- A smaller size of buffer memory is required to store the overlapped samples for the next sound unit in the encoder and decoder.

The mapping structure of ATRAC is summarized in Fig. 43.18.

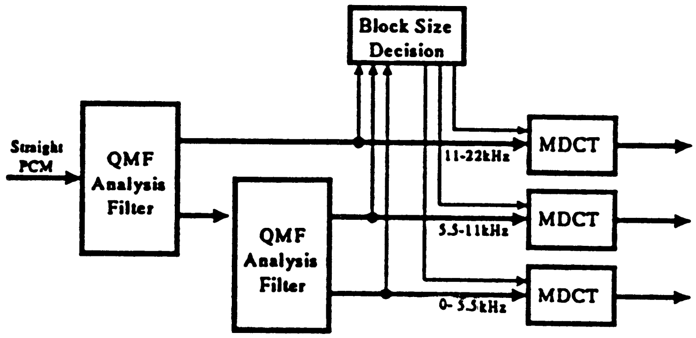


FIGURE 43.17: ATRAC time-frequency analysis.

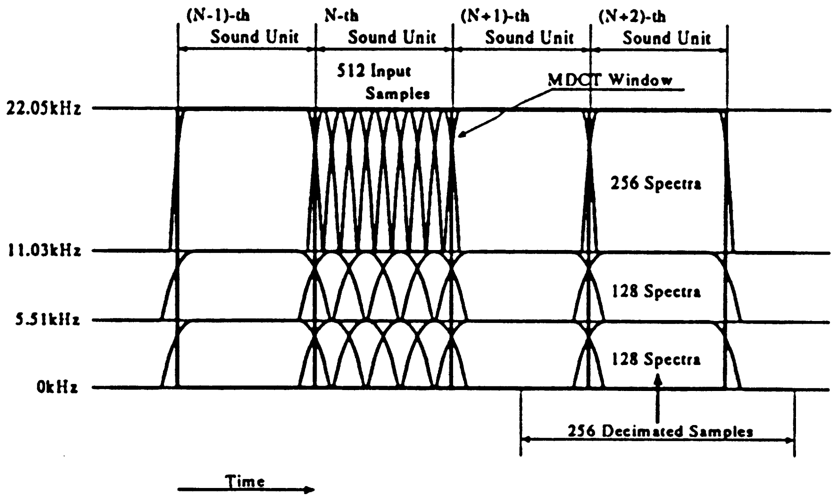


FIGURE 43.18: ATRAC mapping structure.

43.7.2 ATRAC2

The ATRAC2 system, taking advantage of the progress in LSI technologies, allows audio signals of 16 bits per sample with a sampling frequency of 44.1 kHz (705.6 kbps) to be compressed to 64 kbps, sacrificing almost no audio quality. It was designed focusing on efficient coding of tonal signals, as the human ear is very sensitive to distortions in such signals.

Block diagrams of the encoder and decoder structures are shown in Figs. 43.19 and 43.20. The encoder extracts psychoacoustically important tone components from the input signal spectra in order to encode them separately from the other less important spectrum data in an efficient way. A tone component is a group of consecutive spectral coefficients and is defined with several parameters including its location and width data. The remaining spectral coefficients are grouped into 32 non-uniform BFUs. Both the tone components and the remaining spectral coefficients may be encoded with Huffman coding, which is shown in Table 43.3 and for which simple decoding with a look-up table is practical due to its small size. Although the quantization step number is limited to 63, high S/N ratio can be obtained by repeatedly extracting tone components from the same frequency range.

The mapping structure of ATRAC2 is shown in Fig. 43.21. The frequency resolution is twice that

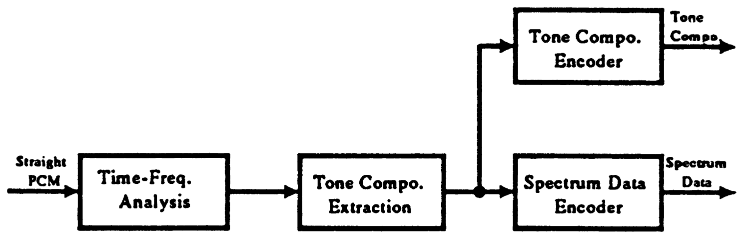


FIGURE 43.19: ATRAC2 encoder.

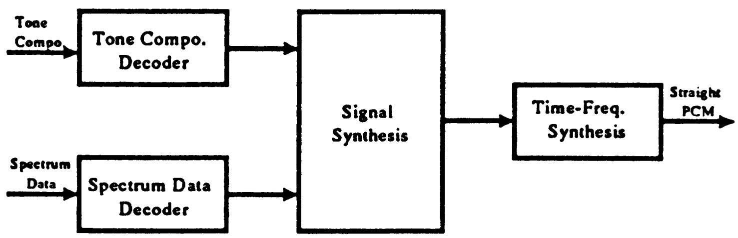


FIGURE 43.20: ATRAC2 decoder.

TABLE 43.3 Huffman Code Table

ID	Quantization step number	Dimension (spectr. num.)	Maximum code length	Look-up table size
0	1	—	—	—
1	3	2	5	32
2	5	1	3	8
3	7	1	4	16
4	9	1	5	32
5	15	1	6	64
6	31	1	7	128
7	63	1	8	256

Note: Total = 536

of ATRAC, and in order to secure the frequency separability, ATRAC2 performs a signal analysis using a combination of a 96-tap polyphase quadrature filter (PQF) and a fixed-length 50%-overlap MDCT whose forward and backward window forms are different from each other. ATRAC2 prevents pre-echo by amplifying the signal preceding an attack adaptively before transforming it into spectral coefficients in the encoder and restoring it to the original level after the inverse transform in the decoder. This technique, called gain control, simplifies the spectral structure of the system.

The subband decomposition realizes frequency scalability; decoders with smaller complexity can be constructed by simply decoding only lower-band data. Use of PQF lowers the computational complexity.

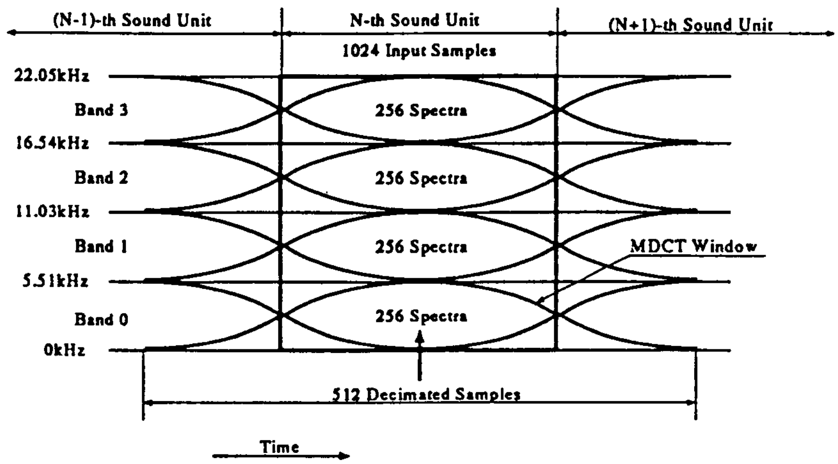


FIGURE 43.21: ATRAC2 mapping structure.

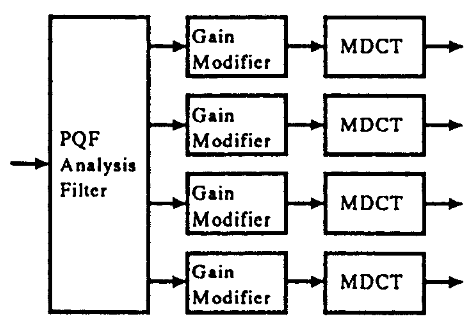


FIGURE 43.22: ATRAC2 time-frequency analysis.

References

[1] Kayanuma, A. et al., An integrated 16 bit A/D converter for PCM audio systems, *ISSCC Dig. Tech. Papers*, pp. 56-57, Feb. 1981.

[2] Plassche, R.J. et al., A monolithic 14 bit D/A converter., *IEEE J. Solid State Circuits*, SC-14, 552-556, 1979.

[3] Naus, P.J.A. et al., A CMOS stereo 16 bit D/A converter for digital audio, *IEEE J. Solid State Circuits*, SC-22, 390-395, June 1987.

[4] Hauser, M.W., Overview of oversampling A/D converters, an audio Engineering Society Preprint #2973, 1990.

[5] Matsuya, Y. et al., A 16-bit oversampling A to D conversion technology using triple-integration noise shaping, *IEEE J. Solid State Circuits*, SC-22, 921-929, Dec. 1987.

[6] Schouwenaars, H.J. et al., An oversampling multibit CMOS D/A converter for digital audio with 115dB dynamic range, *IEEE J. Solid State Circuits*, SC-26, 1775-1780, Dec. 1991.

[7] Maruyama, Y. et al., A 20-bit stereo oversampling D to A converter, *IEEE Trans. on Consumer Electronics*, 39, 274-276, Aug. 1993.