ATRAC: Adaptive Transform Acoustic Coding for MiniDisc

Kyoya Tsutsui
Hiroshi Suzuki
Osamu Shimoyoshi
Mito Sonohara
Kenzo Akagiri
Robert M. Heddle

Sony Corporate Research Laboratories
6-7-35 Kitashinagawa, Shinagawa-ku, Tokyo 141 Japan

Reprinted from the 93rd Audio Engineering Society Convention in San Fransisco, 1992 October 1-4

Abstract

ATRAC is an audio coding system based on psychoacoustic principles. The input signal is divided into three subbands which are then transformed into the frequency domain using a variable block length. Transform coefficients are grouped into nonuniform bands to reflect the human auditory system, and then quantized on the basis of dynamic sensitivity and masking characteristics. ATRAC compresses compact disc audio to approximately 1/5 of the original data rate with virtually no loss in sound quality.

1 Introduction

Recently, there has been an increasing consumer demand for a portable recordable high-quality digital audio media. The MiniDisc system was developed to meet this demand. The MiniDisc is based on a 64 mm optical or magneto-optical disc which has approximately 1/5 of the data storage capacity of a standard compact disc. Despite the reduced storage capacity, it was necessary that the MiniDisc maintain high sound quality and a playing time of 74 minutes. The ATRAC (Adaptive Transform Acoustic Coding) data compression system was therefor designed to meet the following criteria:

- Compression of 16-bit 44.1 kHz stereo audio into less than 1/5 of the original data rate with minimal reduction in sound quality.
- Simple and inexpensive hardware implementation suitable for portable players and recorders.

When digital audio data is compressed, there is normally a certain amount of quantization noise introduced into the signal. The goal of many audio coding systems [1-6] is to control the time-frequency distribution of this noise in such a way as to render it inaudible to the human ear. If this is completely successful, the reconstructed signal will be indistinguishable from the original.

In general, audio coders operate by decomposing the signal into a set of units, each corresponding to a certain range in time and frequency. Using this time-frequency distribution, the signal is analyzed according to psychoacoustic principles. This analysis indicates which units are critical and must be coded with high precision, and which units are less sensitive and can tolerate some quantization noise without degrading the perceived sound quality. Based on this information, the available bits are allocated to the time-frequency units. The spectral coefficients in each unit are then quantized using the allocated bits. In the decoder, the quantized spectra are reconstructed according to the bit allocation and then synthesized into an audio signal.

The ATRAC system operates as above, with several enhancements. ATRAC uses psychoacoustics not only in the bit allocation algorithm, but also in the time-frequency splitting. Using a combination of subband coding and transform coding techniques, the input signal is analyzed in nonuniform frequency divisions which emphasize the important low-frequency regions. In addition, ATRAC uses a transform block length which adapts to the input signal. This ensures efficient coding of stationary passages without sacrificing time resolution during transient passages.

This paper begins with a review of the relevant psychoacoustic principles. The ATRAC encoder is then described in terms of time-frequency splitting, quantization of spectral coefficients, and bit allocation. Finally, the ATRAC decoder is described.

2 Psychoacoustics

2.1 Equi-loudness Curves

The sensitivity of the ear varies with frequency. The ear is most sensitive to frequencies in the neighbourhood of 4 kHz; sound pressure levels which are just detectable at 4 kHz are not detectable at other frequencies. In general, two tones of equal power but different frequency will not sound equally loud. The perceived loudness of a sound may be expressed in sones, where 1 sone is defined as the loudness of a 40 dB tone at 1 kHz. Equi-loudness curves at several loudness levels are shown in Figure 1. The curve labeled "hearing threshold in quiet" indicates the minimum level (by definition, 0 sone) at which the ear can detect a tone at a given frequency.

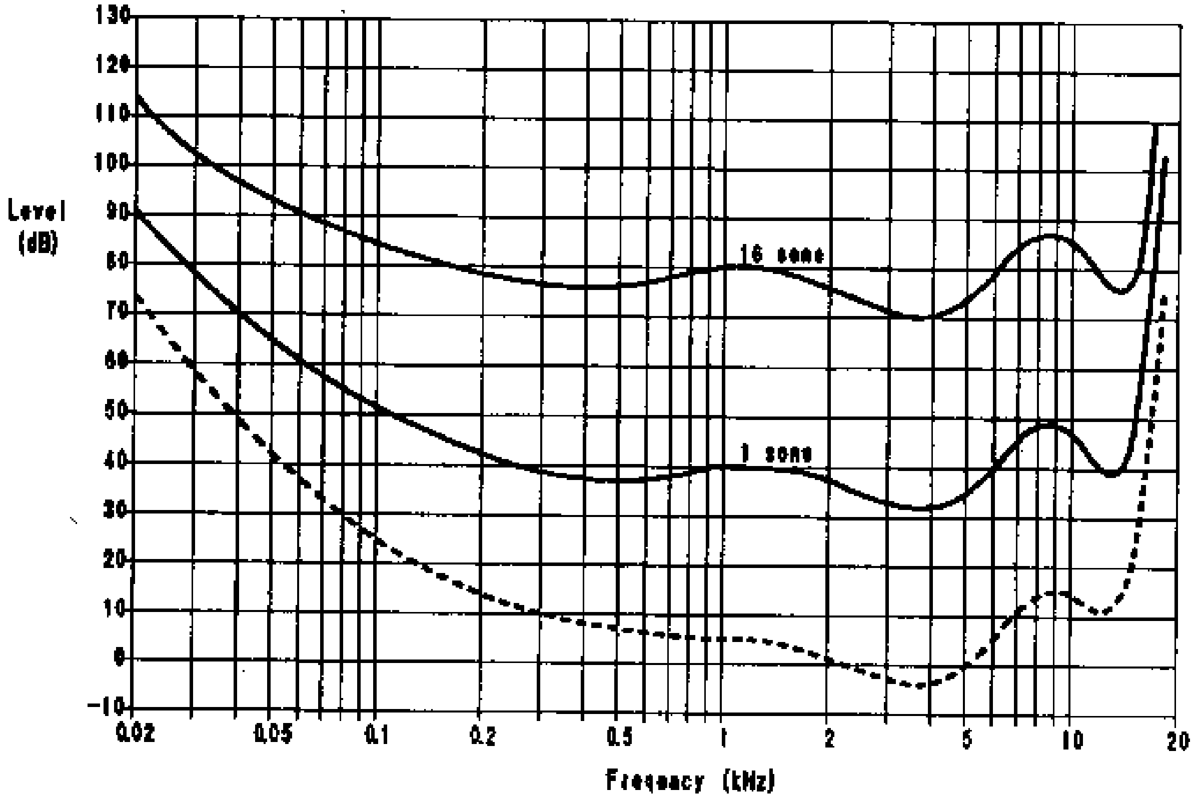


Figure 1: Equi-loudness curves (dotted line = hearing threshold in quiet) (adapted from Robinson-Dadson curves).

These curves indicate that the ear is more sensitive at some frequencies than it is at others. Distortion at insensitive frequencies will be less audible than at sensitive frequencies.

2.2 Masking

Masking [7] occurs when one sound is rendered inaudible by another. Simultaneous masking occurs when the two sounds occur at the same time, such as when a conversation (the masked signal) is rendered inaudible by a passing train (the masker). Backward masking occurs when the masked signal ends before the masker begins; forward masking occurs when the masked signal begins after the masker has ended.

Masking becomes stronger as the two sounds get closer together in both time and frequency. For example, simultaneous masking is stronger than either forward or backward masking because the sounds occur at the same time. Masking experiments are generally performed by using a narrow band of white noise as the masking signal, and measuring the just-audible level of a pure tone at various times and frequencies. Examples of simultaneous masking and temporal masking are shown in Figure 2 and Figure 3 respectively.

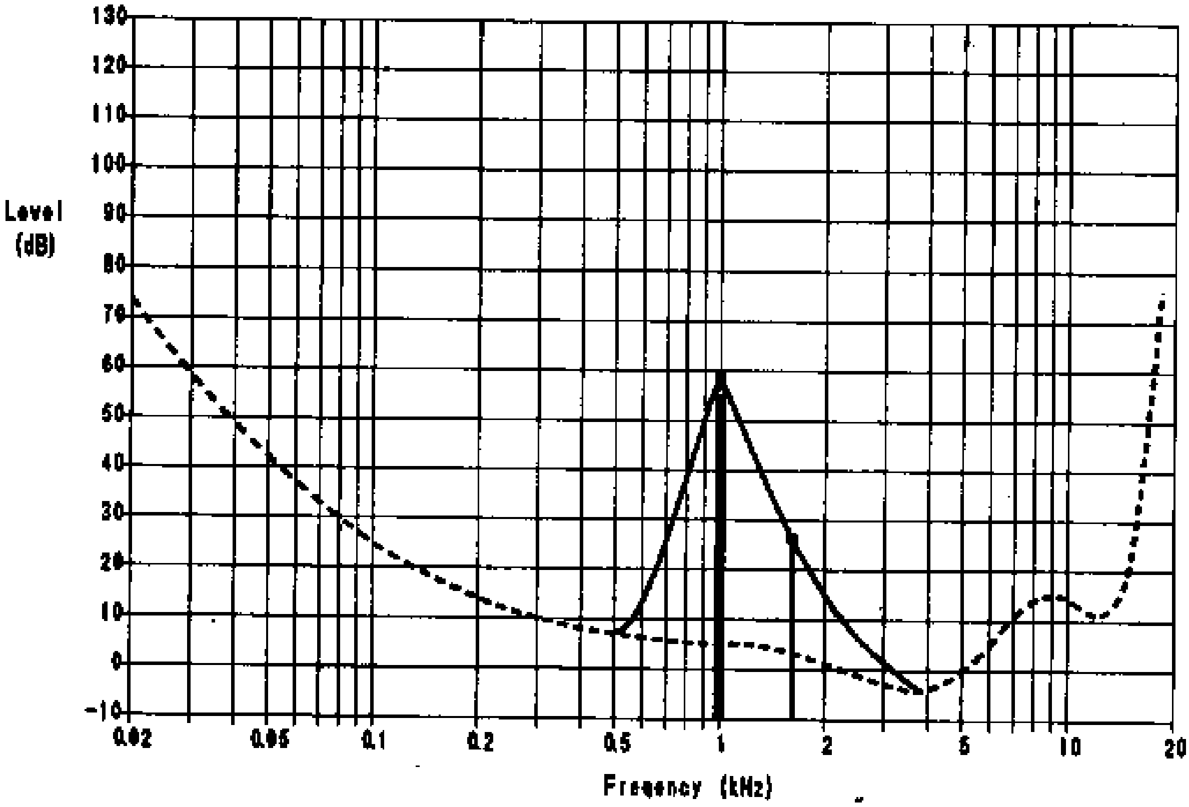


Figure 2: Simultaneous masking curve, including narrow-band noise and masked tone (adapted from [8]).

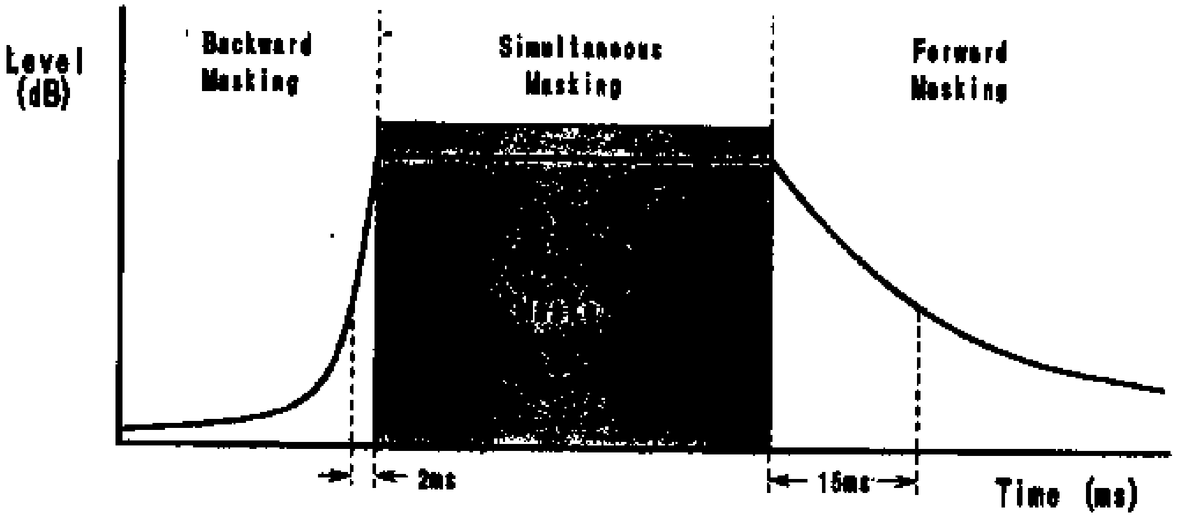


Figure 3: Example of temporal masking.

Important conclusions may be drawn from these graphs. First, simultaneous masking is more effective when the frequency of the masked signal is equal to or higher than that of the masker. Second, while forward masking is effective for a considerable time after the masker has stopped, backwards masking may only be effective for less than 2 or 3 ms before the onset of the masker.

2.3 Critical Bands

Critical bands [7] arose from the idea that the ear analyzes the audible frequency range using a set of subbands. The frequencies within a critical band are similar in terms of the ear's perception, and are processed separately from other critical bands. Critical bands arose naturally from experiments in human hearing and can also be derived from the distribution of sensory cells in the inner ear. Critical bands can be thought of as the frequency scale used by the ear [8].

The critical band scale is shown in Table 1. It is clear that the critical bands are much narrower at lower frequencies than at high frequencies; in fact, three quarters of the critical bands are located below 5 kHz. This indicates that the ear receives more information from the low frequencies and less from higher frequencies.

Table 1: Discrete critical bands [7]

Critical Band	Frequency (Hz)			Critical Band	Frequency (Hz)		
	Low	High	Width		Low	High	Width
0	0	100	100	13	2000	2320	320
1	100	200	100	14	2320	2700	380
2	200	300	100	15	2700	3150	450
3	300	400	100	16	3150	3700	550
4	400	510	110	17	3700	4400	700
5	510	630	120	18	4400	5300	900
6	630	770	140	19	5300	6400	1100
7	770	920	150	20	6400	7700	1300
8	920	1080	160	21	7700	9500	1800
9	1080	1270	190	22	9500	12000	2500
10	1270	1480	210	23	12000	15500	3500
11	1480	1720	240	24	15500	22050	6550
12	1720	2000	280				

3 The ATRAC Encoder

A block diagram of the encoder structure is shown in Figure 4. The encoder has three components. The analysis block decomposes the signal into spectral coefficients grouped into Block Floating units (BFU's). The bit allocation block divides the available bits between the BFU's, allocating fewer bits to insensitive units. The quantization block quantizes each spectral coefficient to the specified wordlength.

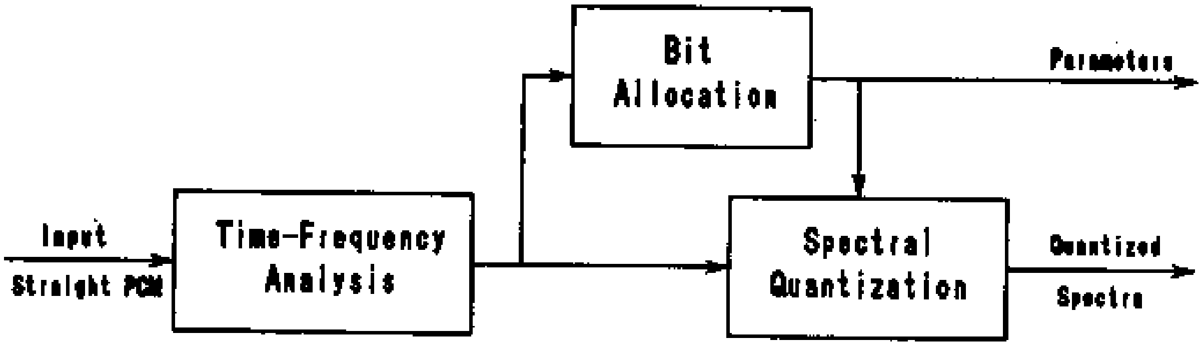


Figure 4: Block diagram of ATRAC encoder.

3.1 Time-Frequency Analysis

This block (Figure 6) generates the BFU's in three steps, combining techniques from subband coding and transform coding. First, the signal is broken down into three subbands: 0-5.5 kHz, 5.5-11 kHz, and 11-22 kHz. Each of these subbands is then transformed into the frequency domain, producing a set of spectral coefficients. Finally, these spectral coefficients are grouped nonuniformly into BFU's.

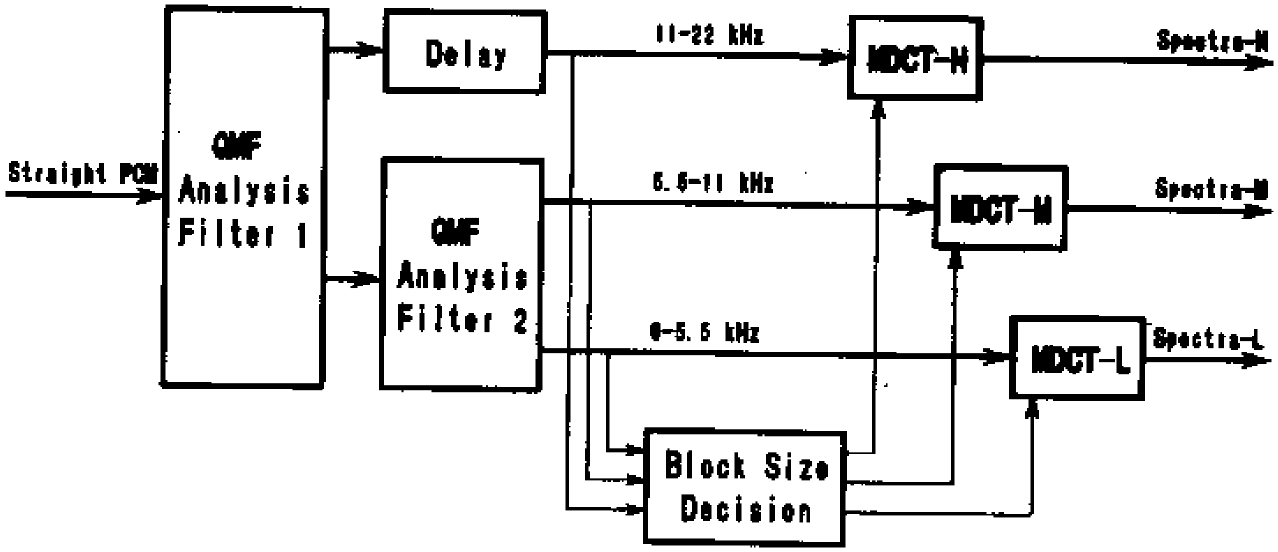
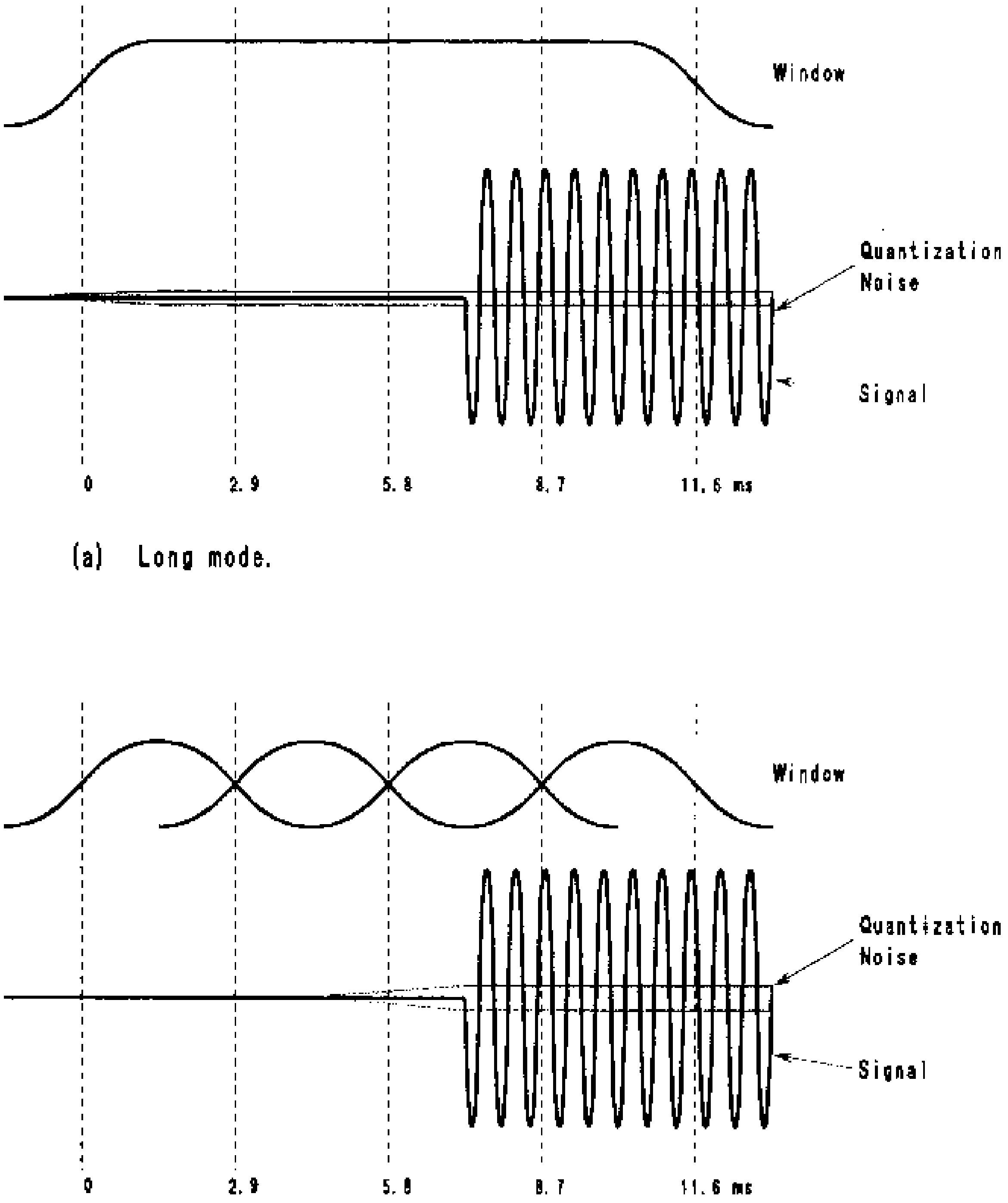


Figure 6: Time-frequency analysis structure.

The subband decomposition is performed using Quadrature Mirror Filters (QMF's) [0-10]. The input signal is divided into upper and lower frequency bands by the first QMF, and the lower frequency band is divided again by a second QMF. Use of QMF's ensures that time-domain aliasing caused by the subband decomposition will be cancelled during reconstruction.

Each of the three subbands is then transformed into the frequency domain using the Modified Discrete Cosine Transform (MDCT) [11-12]. The MDCT allows up to 50% overlap between time-domain windows, leading to improved frequency resolution while maintaining critical sampling. Instead of a fixed transform block length, however, ATRAC chooses the block length adaptively based on the signal characteristics in each band. There are two modes: long mode (11.6 ms) and short mode (1.45 ms in the high frequency band, 2.9 ms in the others). Normally long mode is used to provide good frequency resolution. However, problems may occur during attack portions of the signal. Specifically, the quantization noise is spread over the entire signal block, and the initial quantization noise is not masked (Figure 8a); this problem is called pre-echo. In order to prevent pre-echo, ATRAC switches to short mode (Figure 8b) when it detects an attack signal. In this case, because there is only a short segment of noise before the attack, the noise will be masked by backward masking (section 2.2). Backward masking is not effective for Long Mode because of its very short duration. Thus, ATRAC achieves efficient coding in stationary regions while responding quickly to transient passages.



(a) Long mode.

(b) Short mode.

Figure 8: Illustration of long and short MDCT block size modes. In this situation, short mode is preferable.

Note that short mode is not necessary for signal decay, because the quantization noise will be masked by forward masking which lasts much longer than backward masking. For maximum flexibility, the block size mode can be selected independently for each band.

The MDCT spectral coefficients are then grouped into BFU's. Each unit contains a fixed number of coefficients. In the case of long mode, the units reflect 11.6 ms of a narrow frequency band; in the case of short mode, each block reflects a shorter time but a wider frequency band (Figure 9). Note that the concentration of BFU's is greater at low frequencies than at high frequencies; this reflects the psychoacoustic characteristics of the human ear.

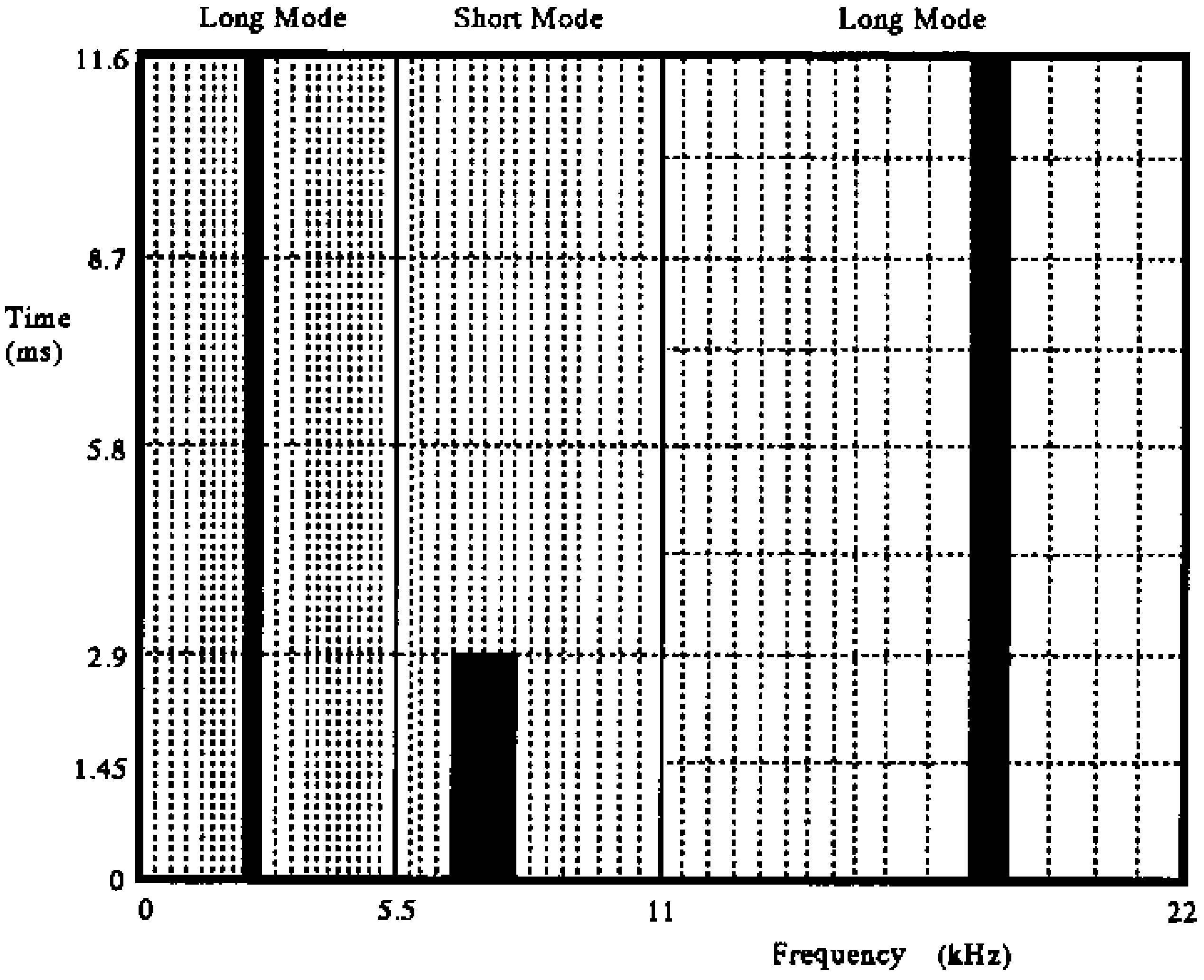


Figure 9: Example of nonuniform time-frequency divisions.

3.2 Spectral Quantization

The spectral values are quantized using two parameters: wordlength and scale factor. The scale factor defines the full-scale range of the quantization, and the wordlength defines the precision within that scale. Each BFU has the same wordlength and scale factor, reflecting the psychoacoustic similarity of the grouped frequencies.

The scale factor is chosen from a fixed list of possibilities, and reflects the magnitude of the spectral coefficients in each BFU. The wordlength is determined by the bit allocation algorithm (section 3.3).

For each sound frame (corresponding to 512 input points), the following information is stored in disc:

- MDCT block size mode (long or short).
- Wordlength data for each Block Floating unit.
- Scale factor code for each Block Floating unit.
- Quantized spectral coefficients.

In order to guarantee accurate reconstruction of the input signal, critical data such as the block size mode, wordlength and scale factor data may be stored redundantly. Information about quantities of redundant data is also stored on the disc.

3.3 Bit Allocation

The bit allocation algorithm divides the available data bits between the various BFU's. Units with a large number of bits will have little quantization noise; units with few or no bits will have significant quantities of noise. For good sound quality, the bit allocation algorithm must ensure that critical units have sufficient bits, and that the noise in non-critical units is not perceptually significant.

ATRAC does not specify a bit allocation algorithm; any appropriate algorithm may be used. The wordlength of each BFU is stored on the MiniDisc along with the quantized spectra, so the decoder is completely independent of the allocation algorithm. This provides for the evolutionary improvement of the encoder without changing the MiniDisc format or the decoder.

There are many possible algorithms, ranging from very simple to extraordinarily complex. For portable MiniDisc recorders, however, the possibilities are limited somewhat by the fact that they must be implemented on low-cost low-power compact hardware. Nevertheless, ATRAC is capable of good sound quality using even a simple bit allocation algorithm, provided it is soundly based on psychoacoustic principles. ATRAC's nonuniform adaptive time-frequency structure is already based on psychoacoustics, relieving the pressure on the bit allocation algorithm.

One suggested algorithm uses a combination of fixed and variable bits. The fixed bits emphasize the important low-frequency regions, allocating fewer bits to the BFU's in higher frequencies. The variable bits are allocated according to the logarithm of the spectral coefficients within each BFU. The total bit allocation b_{tot} is the weighted sum of the fixed bits $b_{fix}(k)$ and the variable bits $b_{var}(k)$. Thus, for each BFU k,

$$b_{tot}(k) = Tb_{var} + (1-T)b_{fix}$$

The weight T is a measure of the tonality of the signal, taking a value close to 1 for pure tones, and close to 0 for white noise. This means that the proportion of fixed and variable bits is itself variable. Thus, for pure tones, the available bits will be concentrated in a small number of BFU's. For more noise-like signals, the algorithm will emphasize the fixed bits in order to reduce the number of bits allocated to the insensitive high frequencies.

The above equation is not concerned with overall bit rate, and will in general allocate more bits than are available. In order to ensure a fixed data rate, an offset b_{off} (the same for all BFU's) is calculated. This value is subtracted from $b_{tot}(k)$ for each unit, giving the final bit allocation $b(k)$:

$$b(k) = \text{integer}\{b_{tot}(k)-b_{off}\}$$

If the subtraction generates a negative wordlength, that BFU is allocated 0 bits. This algorithm is illustrated in Figure 10.

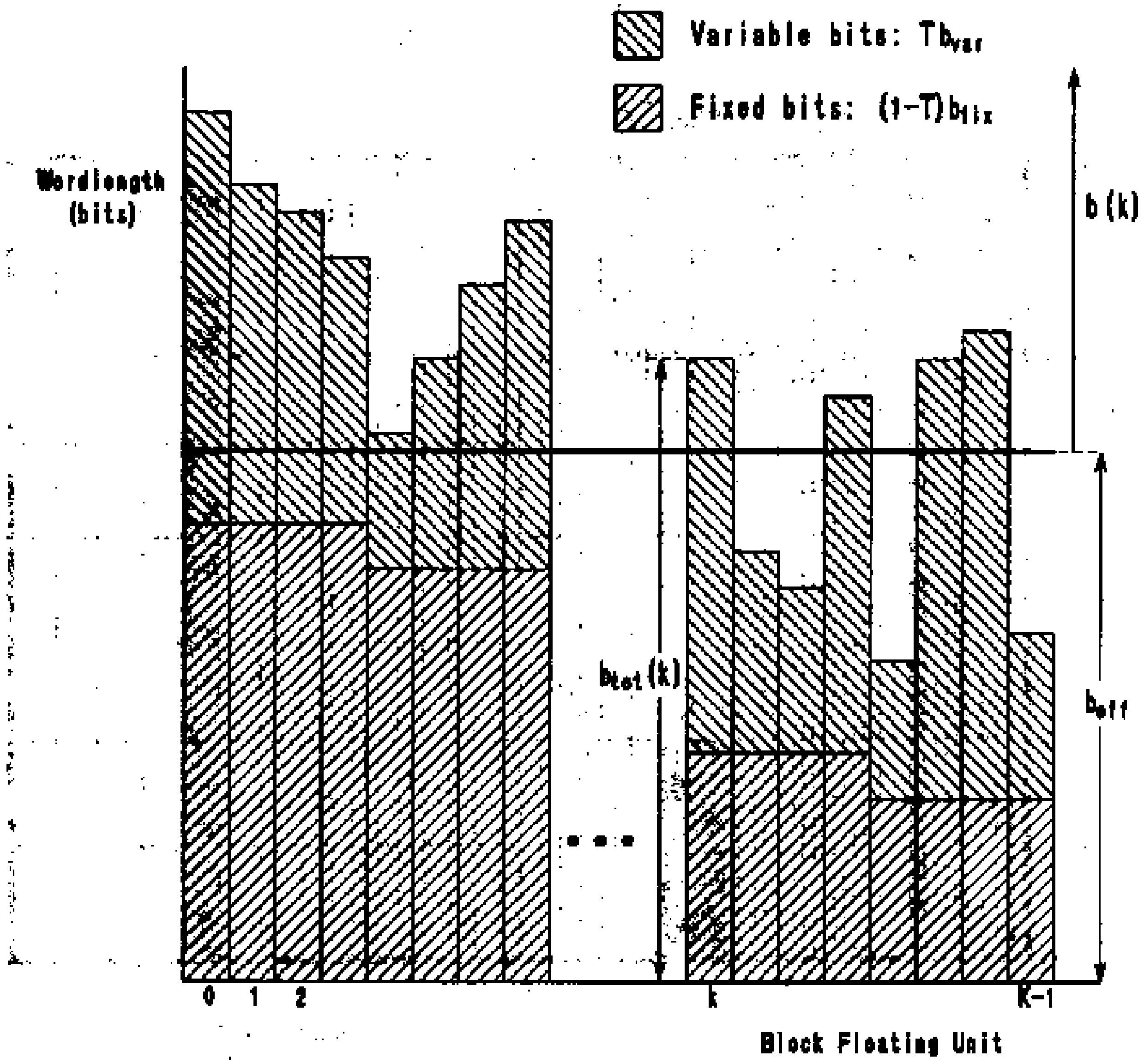


Figure 10: Example bit allocation algorithm showing final bit assignment b(k).

4 The ATRAC Decoder

A block diagram of the decoder structure is shown in Figure 5. The decoder first reconstructs the MDCT spectral coefficients from the quantized values, using the wordlength and scale factor parameters. These spectral coefficients are then used to reconstruct the original audio signal (Figure 7). The coefficients are first transformed back into the time domain by the inverse MDCT (IMDCT) using either long mode or short mode as specified in the parameters. Finally, the three time-domain signals are synthesized into the output signal by QMF synthesis filters.

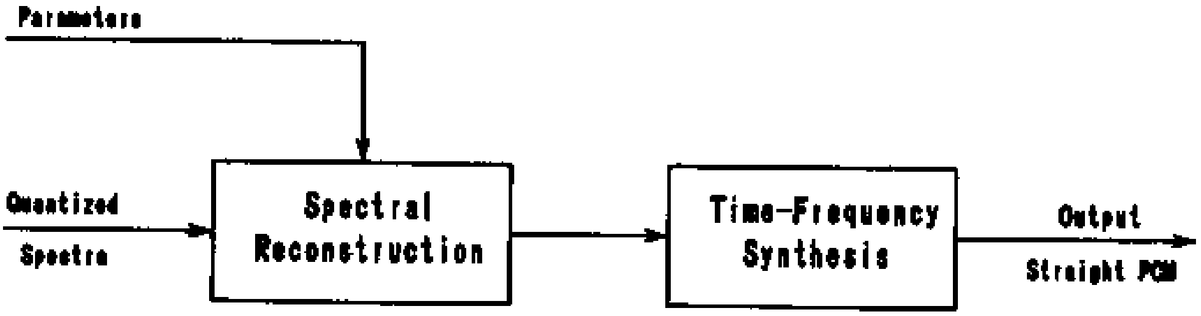


Figure 5: Block diagram of ATRAC decoder.

5 Conclusions

Through a combination of various techniques including psychoacoustics, subband coding and transform coding, ATRAC succeeds in coding digital audio with virtually no perceptual degradation in sound quality. Listening tests indicate that the difference between ATRAC sound and the original source is not perceptually annoying nor does it reduce the sound quality. Furthermore, the system is sufficiently compact to be installed in portable consumer products. Using ATRAC, the MiniDisc provides a practical solution for portable digital audio.

6 References

1. MPEG/AUDIO CA11172-3, 1992.
2. "ASPEC (Source: AT&T Bell Labs *et al.*)" Doc. No. 89/205, ISO-IEC JTC1/SC2/WG8 MPEG-AUDIO, Oct. 18, 1989.
3. R. Veldhuis, M. Breeuwer and R. van der Wall, "Subband coding of digital audio signals without loss of quality," *Proc. 1989 International Conference on Acoustics, Speech and Signal Processing,* Glasgow, pp. 2009-2012.
4. A. Sugiyama, F. Hazu, M. Iwadare and T. Nishitani, "Adaptive transform coding with an adaptive block size (ATCABS)," *Proc. 1990 International Conference on Acoustics, Speech and Signal Processing,* Albuquerque, pp. 1093-1096.
5. G. Davidson, L. Fielder and M. Antill, "High-quality audio transform coding at 128 kbits/s," *Proc. 1990 International Conference on Acoustics, Speech and Signal Processing,* Albuquerque, pp. 1117-1120.
6. G. Davidon, L. Fielder and M. Antill, "Low-complexity transform coder for satellite link applications," Audio Engineering Society 89th Convention preprint 2966, Sept. 1990.
7. J. S. Tobias, Ed., *Foundations of Modern Auditory Theory, Vol. 1,* Academic Press, New York, 1970.
8. E. Zwicker and U. T. Zwicker, "Audio engineering and psychoacoustics: Matching signals to the final receiver, the human auditory system." *J. Audio Engineering Society,* Vol. 39 No. 3, pp. 115-126, March 1991.
9. D. Estaban, and C. Galand, "Application of quadrature mirror filters to split band voice coding schemes," *Proc. 1977 IEEE International Conference on Acoustics, Speech and Signal Processing,* Hartford CT, pp. 191-195.
10. P. P. Vaidyanathan, "Quadrature mirror filter banks, M-band extensions and perfect-reconstruction techniques," *IEEE ASSP Magazine,* Vol. 4, pp. 4-20, July 1987.
11. J. Princen and A. Bradley. "Analysis/synthesis filter band design based on time-domain aliasing cancellation," *IEEE Trans. Acoustics, Speech and Signal Processing,* Vol. 34, pp. 1153-1161, 1986.
12. J. Princen, A. Johnson and A. Bradley, "Subband/transform coding using filter band designs based on time domain aliasing cancellation," *Proc. 1987 IEEE International conference on Acoustics, Speech and Signal Processing,* Dallas, pp. 2161-2164.

Translations:

- A Slovenian translation by Gasper Halipovich
- A A Russian translation has been provided by Best Sportscars Team.
- A Belorussian translation has been provided by Webhostingrating

- A Serbo-Croatian translation has been provided by Anja Skrba from Webhostinggeeks.com
- A Portugese translation has been provided by Artur Weber.

~ No further translations needed, thank you! ~

Return to the MiniDisc Community Page.