Special Article on Mobile Multimedia Signal Processing Technologies
Audio Coding Technology

The use of audio coding technology is spreading quickly on the Internet and in other areas to compress CD-quality audio signals down to around 100 kbit/s.

This article explains the fundamental technologies for compressing audio signals based on human psycho-acoustic models, the MPEG international standard and various de-facto standards, as well as applications in mobile communications.

Tomoyuki Ohya and Sanae Hotani

1. Introduction

Since the world's first portable analog cassette player appeared in 1979, listening to high-quality music outdoors (in town and in trains) has become extremely common in everyday life. Between the mid-1980s and the early 1990s, gadgets that use digital signal-processing technology appeared one after another, namely, portable CD and MD players. The market size of these gadgets is gigantic: more than 200 million CD and MD players have been shipped to date. In addition to this, the use of MP3 players and other memory-type players that have no moving mechanical parts or components are increasing at a rapid pace.

Owing to progress in digital audio technology, high-quality audio signals can be compressed into small volumes of data. This makes it possible to store extended pieces of music in small gadgets and enables more hours of playback by battery.

The CD requires a data rate of approximately 1.4 Mbit/s in order to digitize stereo audio signals. Adaptive Transform Acoustic Coding (ATRAC), which is used for MDs, can reproduce the same quality at about 300 kbit/s. Moving Pictures Experts Group-1 (MPEG-1), which was issued as an international standard in 1992, can do the same at roughly 192 kbits/s, whereas MPEG-2 Advanced Audio Coding (AAC), which became a standard in 1997, can compress audio data down to around 128 kbit/s. These technologies help expand the possibilities of communications; at DoCoMo, we are conducting monitoring experiments for music distribution services using compressed digital music data. On the Internet, streaming services, trial listening and distribution services are becoming widely available at an increasing pace.

This article explains the mechanism of digital audio compression technology and its application.

2. Basic Principles of Audio Coding

2.1 Psycho-acoustic Characteristics

Audio coding compresses data by taking advantage of our psycho-acoustic perception. Numerous experiments have proven that when there are two tones close in frequency, the stronger tone hides the occurrence of the weaker one, which is a phenomenon called acoustic masking (Figure 1).

Efficient data compression can be achieved by omitting data bits constituting sounds that are inaudible to the human ear, based on a detailed analysis of the input signal.

Another well-known fact is that the lowest level of sound

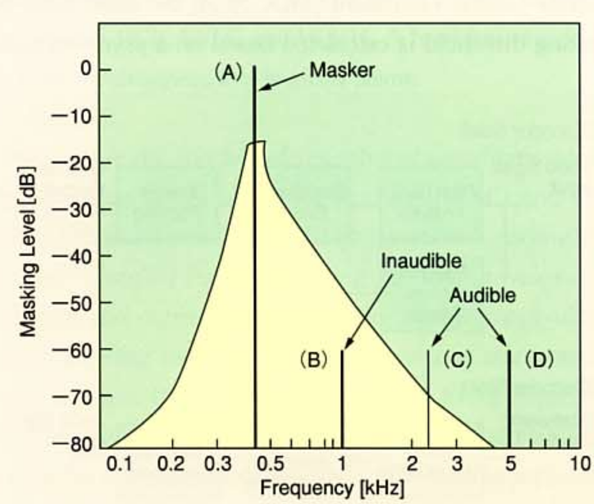

Figure 1 Acoustic Masking

that can be detected by the human ear depends on the frequency band, which is called the threshold of audibility. Efficient data compression can be achieved by calculating the level of noise that is inaudible to humans (the masking threshold) with reference to the threshold of audibility and the acoustic masking effect, and by ensuring that the quantization noise level is below the masking threshold.

2.2 Orthogonal Transform

Audio signal compression can be made more effective by taking advantage of its strong temporal correlation, by putting a number of audio signal samples into one frame. One of the techniques is to use prediction. In speech coding, the compression efficiency can be dramatically improved with the use of linear prediction analysis.

In audio coding, the technique often used to enhance the compression efficiency is to make the signals uncorrelated, by orthogonal transform. While it is widely known that the optimal orthogonal transform is Karhunen Loeve Transform (KLT), the quasi-optimal orthogonal transform called Discrete Cosine Transform (DCT) is normally used, due to the computational complexity of KLT. DCT technology can use fast calculation algorithms based on Fast Fourier Transform (FFT), and works well with the calculation of acoustic masking in the frequency domain, thereby contributing to higher audio-coding efficiency.

2.3 Basic Audio Coding Components

The basic structure of audio coding is as illustrated in Figure 2. On the encoder side, the frequency spectrum of the audio signal is split into subbands through an orthogonal transform method such as polyphase filter bank or Modified Discrete Cosine Transform (MDCT). At the same time, the masking threshold is calculated based on a psycho-acoustic model, and code bits are assigned to each frequency subband in an adaptive manner so that the distortion (quantization noise) is below the masking threshold. Each quantized frequency spectrum is then encoded by Huffman coding or another entropy coding, and in the frame-packing block, it is multiplexed with additional information, such as bit allocation information, to be transmitted through the channel.

On the decoder side, three modules are used to perform the coder's procedure in reverse and reproduce the audio signal, i.e., frame unpacking, decoding/inverse quantization and filter bank synthesis.

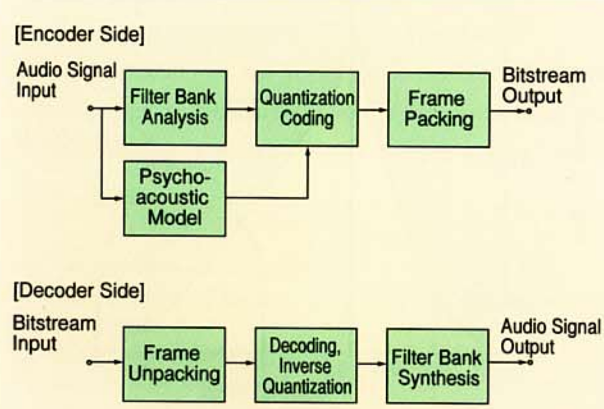

Figure 2 Basic Audio Coding Components

3. Audio Coding Standards

3.1 MPEG-1 Audio Coding Standard [1]

The International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) launched standardization activities in 1988 and finalized MPEG-1 in 1992 as an international standard. MPEG-1 supports monaural and stereo signals for input (sampling frequency: 48 kHz, 44.1 kHz or 32 kHz) and bitrates between 32 kbit/s and 448 kbit/s. MPEG-1 is mainly used for storage purposes and digital broadcasting. MPEG-1 consists of three coding modes (Layers I, II and III), in which the higher layer provides better quality and a higher compression ratio than the lower layer, in compensation for increased computational complexity and delay (Table 1).

MPEG-1 Layer III, which is referred to as MP3, is attracting a great deal of attention as a technology for music distribution over the Internet.

The filter bank analysis block uses a hybrid filter bank, which decomposes the input signal into 32 subbands, and transforms each subband into 18 spectral coefficients by MDCT, yielding a frequency resolution of 576 samples. Although a higher frequency resolution ensures a higher compression rate for stationary audio signals, temporal resolution must be raised to prevent pre-echoes for transient signals. Hence, in the case of transient parts, the temporal resolution is raised by transforming each subband into 6 spectral

Table 1 Key Parameters of MPEG-1

Mode	Bitrate required for CD-quality sound reproduction (stereo, bit/s)	Bitrate Range (bit/s)	Hardware Size
Layer I	256 k	32 k~448 k	Small
Layer II	192 k	32 k~384 k	Medium
LayerIII	192 k	32 k~320 k	Large

coefficients by MDCT, yielding a frequency resolution of 192 samples.

In regard to stereo signals, MPEG-1 supports two coding modes: Middle Side (MS) stereo coding and Intensity stereo coding. In MS stereo coding, one channel carries the sum signal (l+r) and the other the difference signal (l-r), both of which are encoded. In this case, the coding efficiency increases when the correlation between the channels is stronger. On the other hand, Intensity stereo coding involves the coding of the average spectrum of both channels and the power difference between them; the coding bitrate is reduced by taking advantage of the fact that the human ear is not sensitive to phase differences at high frequencies.

3.2 MPEG-2 Audio Coding Standard

MPEG-2, which is backward compatible with MPEG-1, offers the following features for audio coding.

(1) Multichannel...5 channels (front: 3, rear: 2) + Low Frequency Enhancement (LFE).

(2) Multilingual...Supports content coding for up to 7 lingual components.

(3) Low Sampling Frequency Enhancement...Supports 16 kHz, 22.05 kHz and 24 kHz.

MPEG-2 Backward Compatible (BC) [2] became a standard with these features in 1994. However, due to its focus on backward compatibility with MPEG-1, MPEG BC suffered some limitations in terms of quality. In order to overcome the quality limitations, MPEG-2 Advanced Audio Coding (AAC) [3] became a standard in 1997, to achieve multichannel broadcast-quality reproduction.

MPEG-2 AAC consists of three coding modes: Scalable Sampling Rate (SSR) Profile, Low Complexity Profile and Main Profile. It supports sampling frequencies between 8 kHz and 96 kHz, 1-48 channels, and bitrates from 8 kbit/s/ch to 576 kbit/s/ch.

The Filter Bank Analysis block uses 1024- or 128- sample MDCT, adapted to the input signal properties. The window shape used in the analysis can be switched between Kaiser-Bessel Derived (KBD) or the sine window, demonstrating flexibility in design.

In addition to these, a new technology was introduced into AAC to enhance performance, namely, Temporal Noise Shaping (TNS), which applies Linear Prediction Analysis to MDCT coefficients in the frequency domain, controls the shape of quantization noise in conformity with the shape of the spectrum, and concentrates the quantization noise in places that have a large amplitude along the time axis. It also predicts temporal variations by means of a backward adaptive predictor aimed at reducing the size of the required Huffman table.

3.3 MPEG-4 Audio Coding Standard [4]

Features supported by MPEG-4 include low bitrate speech and audio coding in the range of 2 kbit/s-64 kbit/s, structured audio like the Musical Instrument Data Interface (MIDI), and Text-to-Speech (TTS). MPEG-4 became an international standard in February 1999. Subsequently, efforts made to establish a standard with error robustness in a mobile communications environment and to add other new functions crystallized in the form of a revised version called the Amendment (AMD), released in February 2000 [5]. For audio coding, MPEG-4 is equipped with new tools like Transform Domain Weighted Interleave Vector Quantization (TwinVQ), Long Term Prediction (LTP) and Perceptual Noise Substitution (PNS) in addition to MPEG-2 AAC.

MPEG-4 has eight profiles for different applications. Among them, the Mobile Audio Internetworking (MAUI) profile, which consists of audio coding such as error robust AAC and TwinVQ, is expected to be useful in the mobile communications environment.

For audio services to be usable in mobile communications, technologies should be robust against errors caused in transmission channels (error-robustness tools). In order to achieve that, Error Protection (EP) tool, Huffman Codeword Reordering (HCR), Reversible Variable Length Code (RVLC), Virtual Codebook 11 (VCB11) and other technologies were standardized. Error-robustness tools can be divided into two categories: one relies on error detection and correction, and the other aims to reinforce the coded bitstream against errors. The former includes EP tool, whereas the latter includes HCR, RVLC and VCB11. A brief introduction of each error-robustness tool is given below.

(1) EP Tool

MPEG-4 Audio consists of a number of algorithms, including Harmonic Vector eXcitation Coding (HVXC) and Code Excited Linear Prediction (CELP) for speech coding, and AAC and TwinVQ for audio coding. EP tool provides error detection and correction functions for a wide range of purposes, meeting the requirements of various transmission environments. It aims to apply Unequal Error Protection (UEP) to audio coding, considering that UEP is already widely used for speech coding. Encoded audio data includes both error-sensitive and error-insensitive parameters. Each is classified according to its error sensitivity, and different error detection and correction methods are applied. This reduces the overall error-correction redundancies.

The key properties of EP tool are:

- Providing a set of error-correction/detection codes with wide and small-step scalability in performance and in redundancies;
- Providing UEP for both fixed and variable data lengths; and
- Reducing redundancies by sending frame-independent EP-tool configuration information at the beginning of the encoded bitstream.

(2) HCR

HCR is one of the error-robustness tools applied to AAC, and reorders the Huffman code in the AAC spectrum data. Huffman coding is used for encoding spectrum data, in which bit errors propagate due to variable length code properties, and the error sensitivity is dramatically increased. To reduce the error sensitivity, Huffman codewords are reordered every known section in order to prevent error propagation.

(3) RVLC

RVLC is one of the error-robustness tools applied to AAC. Error-robust RVLC is used when encoding a scale factor for scaling the spectrum. RVLC applied in this case is different from Huffman coding in that:

- It can decode in both directions; and
- It can detect errors in tree codes by reserving some unused nodes.

Bidirectional decoding can be used to minimize the error propagation by decoding from the opposite direction when any errors are detected.

(4) VCB11

VCB11 is one of the error-robustness tools applied to AAC. When encoding the spectrum data, information on the spectrum's maximum value is specified according to a codebook called Virtual Codebook 11, in order to prevent the occurrence of any annoying sound due to bit errors.

4. Other Audio Coding Technologies

The audio coding standards described in the previous chapter were defined by international standard bodies, and many organizations and companies were involved in the standardization process. Standards as such are generally available to the general public. In contrast to these standards, there are many de facto standards for audio coding, which are widely used in the marketplace, even though they are not approved by any official standard bodies. The detailed specifications of such de facto standards are often undisclosed. This chapter explains the technologies that are widely used and describes their key properties.

4.1 Adaptive TRansform Acoustic Coding (ATRAC) [6]

This is an audio coding technology developed by Sony for MDs, commercialized in 1992. The MD is an optical/magneto-optical disc measuring 64mm in diameter, which can record up to 74 minutes of stereo audio data. In regard to the sound quality, the distortion detected in MDs may be slightly greater than that in CDs.

ATRAC relies on a combination of subband coding and transform coding. It uses Quadrature Mirror Filter (QMF) to decompose input signals into three subbands, and transforms each subband by MDCT into the frequency domain. This allows transform coding with limited memory, leading to an economical LSI implementation.

The encoding process is performed on signals sampled at 44.1 kHz at a rate of every 512 samples, each of which is encoded into 424 bytes. Therefore, the coding bitrate is 424 (bytes) × 8 (bit/bytes) ÷ (512/44,100) = 292 (kbit/s).

In 1994, Sony developed an enhanced version, ATRAC2, and subsequently released ATRAC3 [7]. ATRAC3 is adopted as a Coder Decoder (CODEC) for memory sticks, and is becoming widely used in content delivery services combined with copyright protection technologies.

ATRAC3 performs adaptive bit allocation for quantization, and supports a wide range of coding bitrates. The difference from ATRAC in terms of technical specifications includes: expansion of the transform block length to 1024 samples; adoption of joint stereo coding; introduction of gain control to reduce pre-echoes; and the separate encoding of tonal components (Figure 3). At 132 kbit/s stereo, ATRAC3 achieves the same sound quality as ATRAC.

4.2 Dolby AC3

Dolby Audio Code Number 3 (AC3) is a pioneer of multi-channel audio coding technology that was developed by Dolby Laboratories Inc. before MPEG-2 AAC. Also known as Dolby Digital, it is widely used in DVDs, movies, laser discs (LDs) and digital TV broadcasting (in the U.S.). More than 5,000 movie titles have been released in the DVD format alone, and almost 60 million products using Dolby Digital have been shipped to date [8].

Dolby AC3 supports monaural to up to 5.1 channels (surround sound by 6 speakers, i.e., L: left, C: center, R: right,

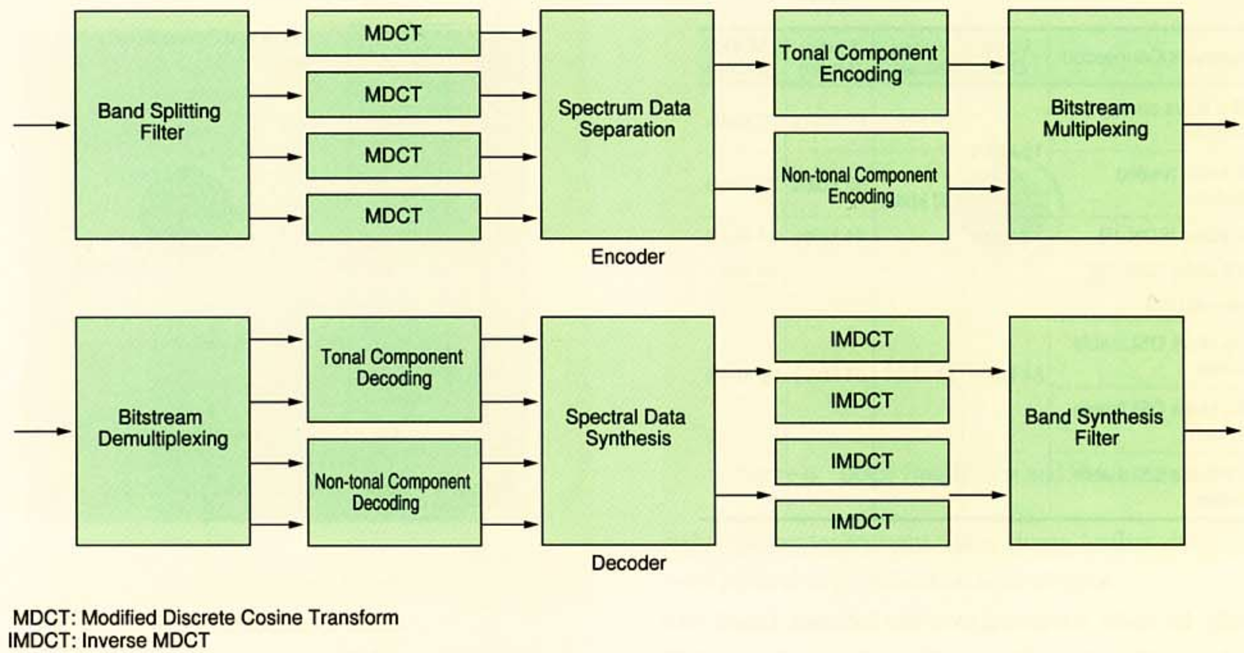

Figure 3 Block Diagram of ATRAC3

LS: left surround, RS: right surround and LFE). The typical coding bitrate is 192 kbit/s for stereo and 384 kbit/s for 5.1 channel surround (Table 2).

AC3 is the technology that served as the foundation for AAC. It uses orthogonal transform by MDCT for encoding, and adaptive bit allocation based on psycho-acoustic models. The window used for transform can be switched between long block mode (512 samples) and short block mode (256 samples). For multichannel coding, the same technique is used as that for Intensity stereo and MS stereo, which are applied to MPEG.

Unlike AAC, AC3 executes adaptive bit allocation based on an approximate psycho-acoustic model calculated from MDCT coefficients. On the decoder side, decoding can be performed even if the bit allocation data is not transmitted.

Table 2 Channel Configuration of AC-3

Channel Configuration (Number of speakers in front of listener/behind listener)	Note (The existence of LFE can be specified independently)
1+1	Dual Mono (e.g., bilingual broadcasts)
1/0	Center
2/0	Left, right
3/0	Left, center, right
2/1	Left, right, surround
3/1	Left, center, right, surround
2/2	Left, right, left surround, right surround
3/2	Left, center, right, left surround, right surround

LFE: Low Frequency Enhancement

4.3 Windows Media Audio (WMA) [9]

Windows Media Audio (WMA) is a speech and audio coding tool that comprises Windows Media Technologies, a group of audiovisual data processing tools developed by Microsoft Corp. It supports coding bitrates between 5 kbit/s and 160 kbit/s, and the quality is as described in Table 3. Although the compression ratio is not as high as other coding technologies like AAC, the decoder is included in the Windows Media Player, which is a part of the Windows operating system. The latest version, Windows Media Player 7, was downloaded by 10 million users within 6 weeks of release, and is spreading rapidly according to installation figures.

Table 3 Quality of Windows Media Audio (WMA)

Bitrate	Quality
28.8 kbit/s	FM Stereo Quality
64 kbit/s	Hi-Fi Audio Quality
160 kbit/s	CD-transparent Quality

4.4 Real Audio [10]

This is an audio CODEC included in Real Player, developed by Real Networks Inc. Real Audio has strong functions

Table 4 Typical Bitrate of Real Audio

Listener's Connection	Voice Only	Voice+ Music	Music (Mono)	Music (Stereo)
28.8 kbit/s analog modem	16 kbit/s	16 kbit/s	20 kbit/s	20 kbit/s
56 kbit/s analog modem		32 kbit/s	32 kbit/s	32 kbit/s
64 kbit/s ISDN 1B	32 kbit/s		44 kbit/s	44 kbit/s
112 kbit/s ISDN 2B	64 kbit/s	64 kbit/s	64 kbit/s	64 kbit/s
Internal LAN				96 kbit/s
256 kbit/s DSL/cable modem				
384 kbit/s DSL/cable modem				
512 kbit/s DSL/cable modem				

especially for audio streaming over the Internet. Based on a technology called Sure Stream, the same content can be delivered at the optimal bitrate to each user, even if they are connected at different speeds.

5. Multichannel Audio Coding and Virtual Audio Technology

Multichannel audio coding, referred to in Chapter 3 and Section 4.2, is designed primarily for theaters, home theaters and other listening environments in which many speakers are used. Although many commercial amplifiers and speakers are being produced for such environments, it is often difficult to find the right environment to accommodate them due to Japan's crowded housing situation, and even more difficult to apply them to a mobile communications environment. One solution to this is virtual audio technology, which simulates the ideal listening environment.

5.1 Sound Field Control Technology

What makes listening to music at a concert hall special is the reverberation from the hall. Researchers have found many techniques to reproduce the same effect in small listening rooms; sound field control using the speaker array is one of them.

As illustrated in Figure 4, many speakers are used to create a sound field around the listener's ears. In theory, a virtual sound source can be created anywhere by the array of speakers, which should present to the listener every detail of the sound characteristics, including the reverberations.

In practice, however, there are still many technical difficulties, such as high-speed processing to provide a stable inverse filter of the listening environment to each speaker. The technology still requires further research.

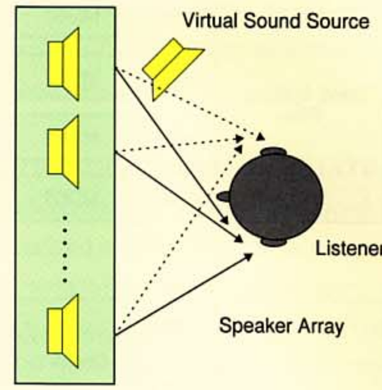

Figure 4 Sound Field Control by Speaker Array

Figure 5 Binaural Playback System

5.2 Binaural Playback Technology

When you listen to music using headphones, the sound image will be fixed at the center of your head. As this does not normally happen in a natural environment, you might feel uncomfortable and get tired easily. An alternative playback method using headphones is to accurately reproduce the sound in the same way as it would reach your ears in a "real environment," so that it sounds like you are listening to it without headphones (Figure 5). This technology is called binaural playback, and a dummy head with left and right ears is often used for recording the sounds, as shown in Photo 1.

5.3 Dolby Headphones [11]

Dolby Headphones were developed by Dolby Laboratories Inc., to enable listeners to enjoy 5.1-channel surround with headphones. The audio transmission functions in a theater

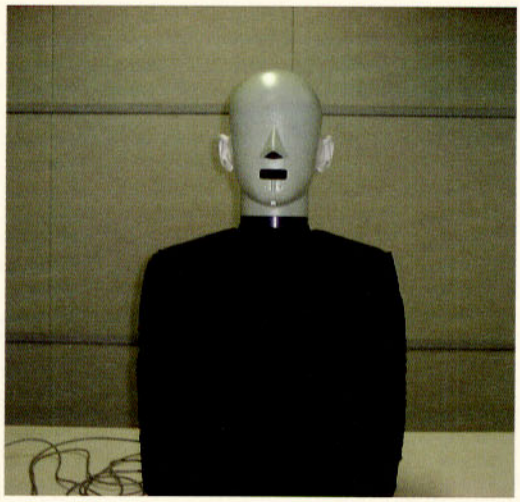

Photo 1 Dummy Head

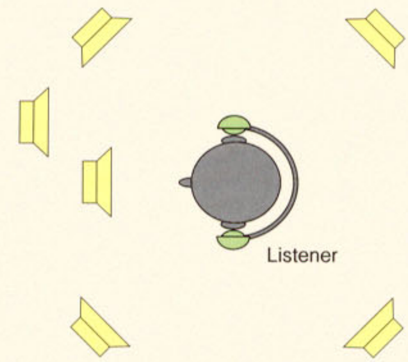

Figure 6 Dolby Headphone and Virtual Speakers

are built into the system in advance. Then, real-time convolution is applied with actual multichannel sound. Outputting the signals from the left and right speakers of the headphones generates a surround-sound audio experience, without relying on normal speakers (Figure 6). This example shows how progress in DSP and other signal processing technologies is making virtual audio reproduction possible, with the use of products available in the marketplace.

6. Conclusion

This article described various audio coding technologies and their applications. Since the appearance of CD and Digital Audio Tape (DAT), digital audio technology has made a negative impression on those who are wary of copyright protection, due to its ability to make an infinite number of copies without quality degradation. The unauthorized sharing of MP3 files over the Internet and the Napstar issue are recent examples. Nevertheless, we cannot deny that progress in audio coding technology has expanded the possibilities of the Internet and communications in general, including music content distribution.

The use of music information in the mobile communications environment has just begun. The possibilities for multichannel audio and virtual audio should expand, in line with the development of infrastructure for high-speed data communications as exemplified by IMT-2000, and described in Chapter 5. While this article did not address the distribution system or security issues, it should be noted that both are integral parts of audio coding technology. It is hoped that audio coding technology will undergo further development, hand in hand with application technologies.

References

[1] ISO/IEC 11172-3, "Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mb/s-Part 3: Audio", August, 1993.

[2] ISO/IEC 13818-3, "Generic Coding of Moving Pictures and Associated Audio Information-Part 3: Audio", May, 1995.

[3] ISO/IEC 13818-7, "Information Technology — Generic coding of moving Pictures and associated audio infomation, Part 7: Advanced Audio Coding (AAC)", December, 1997.

[4] ISO/IEC 14496-3: "Information Technology — coding of audio-visual objects — Part 3: Audio" 1999.

[5] ISO/IEC JTC1/SC29/WG11 N3058 "Text of ISO/IEC14496-3 FDAM1" December 1999.

[6] K.Tsutsui, H.Suzuki, O.Shimoyoshi, M.Sonehara, K.Akagiri and R.M. Heddle, "ATRAC: Adaptive Transform Acoustic Coding for MiniDisc," 93rd Audio Engineering Society Convention, 1992.

[7] http://www.world.sony.com/JP/ Electronics/ATRAC3/index.html

[8] http://www.dolby.com/

[9] http://www.asia.microsoft.com/japan/windows/windowsmedia/

[10] http://www.jp.real.com/

[11] http://www.dolby.co.jp/jp/AV/DH/

Glossary

AC: Advanced Audio Coding
AC3: Audio Coder Number3
AMD: Amendment
ATRAC: Adaptive Transform Acoustic Coding
BC: Backward Compatible
CELP: Code Excited Linear Prediction
CODEC: Corder Decorder
DAT: Digital Audio Tape
DCT: Discrete Cosine Transform
EP: Error Protection
FFT: Fast Fourier Transform
HCR: Huffman Code Reordering
HVXC: Harmonic Vector Excitation Coding
IEC: International Electrotechnical Commission
IMDCT: Inverse MDCT
ISO: International Organization for Standardization
KLT: Karhunen Loeve Transform
LFE: Low Frequency Enhancement
LSI: Large Scale Integration
LTP: Long Term Prediction
MAUI: Mobile Audio Internetworking
MDCT: Modified Discrete Cosine Transform
MIDI: Musical Instrument Data Interface
MPEG: Moving Picture Experts Group
MS: Middle Side
PNS: Perceptual Noise Substitution
QMF: Quandrature Mirror Filter
RVLC: Reversible Variable Length Code
SSR: Scalable Sampling Rate
TNS: Temporal Noise Shaping
TwinVQ: Transform Domain Weighted Interleave Vector Quantization
UEP: Unequal Error Protection
VCB11: Virtual Codebook11
WMA: Windows Media Audio