INTERNATIONAL JOURNAL OF TECHNOLOGY AND ACADEMIC RESEARCH

ISSN: 2224-6878 | Pakistan Volume 6, No. 10.

IF: 14.91 | 2026 http://waerasia.org/index.php/ijtar

DIGITAL AUDIO AND VIDEO TECHNOLOGIES: SIGNAL PROCESSING, COMPRESSION, AND APPLICATION IN MULTIMEDIA SYSTEMS

Orifjonova Mokhidilkhon Akiljon qizi
Assistant, Department of Telecommunication Engineering
Fergana State Technical University
Tel: +998908353969
E-mail: mokhidil.orifjonova@mail.ru
doi:https://doi.org/10.5281/zenodo.23361332

Abstract. This article analyzes the scientific and theoretical foundations of digital audio and video technologies, signal sampling and quantization, spectral analysis, audio compression based on the psychoacoustic model, and the operating principles of modern video codecs. It also examines spatial audio systems, video quality metrics (PSNR, SSIM), and the integration of audio-video streams with computer graphics. The results can be used in designing multimedia systems, developing virtual reality applications, and teaching related subjects.
Keywords: digital audio, sampling, Fourier transform, psychoacoustics, video codec, motion estimation, PSNR, SSIM, spatial audio, multimedia.

Introduction

Digital audio and video technologies have become an essential part of the modern information society. They are widely used in communications, education, mass media, film, computer games, and virtual environments. The transition from analog to digital signals has taken the storage, transmission, and processing of sound and images to a qualitatively new level. A digital signal can be copied without loss of quality, and it can be modified, compressed, and transmitted over networks by software methods.

However, the digital form has also created new problems. Uncompressed audio, and especially video, streams are very large, so efficient compression methods are necessary. On the other hand, taking into account the characteristics of the human auditory and visual systems makes it possible to find a balance between quality and size. This article examines these issues in a consistent manner.

1. Fundamentals of the Digital Audio Signal

Sound is a mechanical vibration in an elastic medium, which a microphone converts into a continuous electrical signal. To be processed on a computer, this signal is digitized in two stages. In sampling, signal values are taken at equal time intervals. In quantization, each value is approximated to one of a finite number of levels. Quantization error produces noise, and its level depends on the bit depth: each additional bit improves the signal-to-noise ratio by approximately 6 dB [2]. The choice of sampling frequency is based on the Kotelnikov–Shannon theorem: if the highest frequency of a signal is f_{max}, the sampling frequency must be greater than $2 \cdot f_{max}$ for the signal to be fully reconstructed [10]. Otherwise the spectra overlap and aliasing distortion occurs, which is why the signal is passed through a low-pass filter before digitization. Since the human ear perceives vibrations from approximately 20 Hz to 20 kHz, the audio CD format uses 44.1 kHz and 16 bits, while studio recordings widely use 48–192 kHz and 24 bits [1]. The Fourier transform is the main tool for analyzing digital audio signals. It represents a signal as a sum of sinusoids and determines its frequency content. This computationally intensive operation is carried out in practice with the Fast Fourier Transform (FFT) algorithm, which reduces

ISSN: 2224-6878 | Pakistan Volume 6, No. 10.

IF: 14.91 | 2026 http://waerasia.org/index.php/ijtar

computational complexity from N^2 to $N \cdot \log N$ [6, 9]. Filtering, equalization, noise suppression, reverberation, and sound effects all rely on spectral and filtering methods.

2. Audio Data Compression and Psychoacoustics

Audio compression methods are divided into two types: lossless and lossy. Lossless codecs (LAC, ALAC) restore the signal bit-for-bit and typically reduce file size by 40–60%. Lossy codecs (MP3, AAC, Opus) achieve much higher compression ratios because they exploit specific properties of human hearing.

The main concepts of the psychoacoustic model are the hearing threshold and the masking effect. The hearing threshold defines the quietest sound level that is audible at each frequency. Masking means that a loud sound can make quieter sounds at neighboring frequencies (and sometimes sounds in the brief moments before and after it) inaudible. The codec transforms the signal into the frequency domain (using the modified discrete cosine transform) and quantizes the inaudible components more coarsely or discards them. As a result, size is reduced by a factor of 8–12 while listening quality is in many cases almost preserved [11]. As the compression level increases, artifacts appear: “pre-echo,” cutoff of high frequencies, and an “underwater” sound. For this reason, codec parameters (bitrate, frequency range) are chosen according to the task. Speech codecs used for telephony and voice messages work at low bitrates, while music requires higher bitrates.

3. Spatial Audio Systems

Determining where a sound comes from in a virtual environment significantly increases realism. Humans determine the direction of a sound source from the interaural time difference (ITD) and the interaural level difference (ILD) between the signals arriving at the two ears, as well as from the influence of the shape of the ear and head. These effects are described by the head-related transfer function (HRTF) [8]. In practice, there are several approaches to spatial audio: channel-based formats (5.1, 7.1), ambisonics (decomposition into spherical harmonics), and object-based audio, in which each sound source is stored with its three-dimensional coordinates and the sound is rendered at that location according to the listener’s playback device. Binaural rendering for headphones is performed with HRTF filters. In modeling room acoustics, the reflection of sound from walls is calculated with geometric methods, namely ray tracing or the image-source method.

4. Video Signals and Video Codecs

Digital video is a sequence of frames, and each frame consists of a matrix of pixels. Color is usually represented by luminance (Y) and two color-difference components (Cb, Cr), because the eye is more sensitive to detail in luminance than in color. For this reason, the color components are stored at reduced resolution (for example, in the 4:2:0 format), which cuts the amount of data in half with almost no effect on quality.

Uncompressed Full HD video (1920×1080, 30 frames/s, 24 bits) produces a stream of about 1.5 Gbit/s, so it cannot be used in practice without video codecs. H.264/AVC, H.265/HEVC, and later standards remove two kinds of redundancy. Redundancy within a frame (spatial) is reduced with block-based prediction and the discrete cosine transform (DCT), while redundancy between frames (temporal) is reduced with motion estimation and compensation. In motion estimation, the current frame is divided into blocks, and for each block the most similar location is searched for in a previous frame. The encoder transmits only the motion vector and the residual difference. The residual is transformed with the DCT, the coefficients are quantized, and they are entropy coded (CAVLC or CABAC). Frames are divided into three types: independently coded I-frames, P-frames that rely on previous frames, and B-frames that use bidirectional prediction. These frames form a group of pictures (GOP).

ISSN: 2224-6878 | Pakistan Volume 6, No. 10.

IF: 14.91 | 2026 http://waerasia.org/index.php/ijtar

The HEVC standard extends block sizes compared with H.264 (coding tree units up to 64×64) and increases the number of prediction directions. As a result, it requires up to about 50% lower bitrate at similar quality [5]. JPEG, used for compressing still images, also relies on the DCT [12, 13], so video and image compression share a common theoretical foundation.

5. Audio-Video Quality Assessment

Quality assessment methods are divided into subjective and objective. The subjective method determines the mean opinion score (MOS) from a group of experienced and ordinary viewers. This method is reliable but expensive and time-consuming. Objective methods give faster and repeatable results. The simplest metric is PSNR, which is calculated from the mean squared error (MSE) between the original and reconstructed image using the formula $PSNR = 10 \cdot \lg(MAX^2/MSE)$ and is expressed in dB.

Although PSNR is simple to calculate, it does not always correctly reflect human perception. For this reason, the SSIM metric, which accounts for changes in image structure, was proposed [14], and in practice machine-learning-based methods such as VMAF are also used. For audio, assessment algorithms based on the psychoacoustic model are used. Quality assessment is also necessary for choosing codec parameters and for adapting bitrate in streaming.

6. Integration with Multimedia and Computer Graphics

In modern multimedia systems, audio, video, and graphics operate synchronized to a single time axis. In game engines and virtual and augmented reality applications, the 3D scene is rendered in real time, and the coordinates of sound sources are updated together with the scene geometry. When the user turns their head, both the image camera and the audio listener orientation change at the same time. If the time difference between audio and video (lip-sync) exceeds several tens of milliseconds, the viewer notices the mismatch, so streams are synchronized using timestamps [3]. The intersection of video and graphics also appears in areas such as virtual studios, frame interpolation, and scene reconstruction from video. In streaming services, video is prepared in versions with different resolutions and bitrates, and these are switched automatically depending on network speed (adaptive streaming). These processes place high demands on computing resources, which is why hardware encoding and decoding on graphics processors is widely used.

7. Challenges and Development Prospects

The development of audio and video technologies has also made a number of problems more pressing. First, the volume of high-resolution (4K, 8K), high-dynamic-range (HDR), and 360° video is growing rapidly, putting pressure on network bandwidth and storage costs. Second, the variety of codecs and formats for different devices and platforms makes compatibility harder to ensure. Third, new codecs are computationally heavier, so hardware support is needed for their widespread adoption [5].

From an educational perspective, it is important that the teaching of multimedia technologies combine theoretical foundations (signal theory, probability theory, linear algebra) with practical projects. Students can independently implement a simple codec (for example, DCT-based image compression or psychoacoustic filtering) and measure quality metrics. Such a hands-on approach helps them understand the internal logic of complex standards more deeply.

Conclusion

Digital audio and video technologies rest on a single mathematical apparatus based on signal sampling, spectral analysis, and the removal of redundancy. In audio, the psychoacoustic model, and in video, motion estimation and the DCT, provide efficient compression. Spatial audio and quality assessment methods increase the naturalness of multimedia systems for the user. Future research is expected to focus on machine-learning-based codecs, generative audio and video synthesis, and energy-efficient hardware solutions.

ISSN: 2224-6878 | Pakistan Volume 6, No. 10.

IF: 14.91 | 2026 http://waerasia.org/index.php/ijtar

References

1. Pohlmann K.C. Principles of Digital Audio. 6th ed. – New York: McGraw-Hill, 2010. – 784 p.
2. Zölzer U. Digital Audio Signal Processing. 2nd ed. – Chichester: John Wiley & Sons, 2008. – 336 p.
3. Richardson I.E.G. The H.264 Advanced Video Compression Standard. 2nd ed. – Chichester: John Wiley & Sons, 2010. – 346 p.
4. Wiegand T., Sullivan G.J., Bjøntegaard G., Luthra A. Overview of the H.264/AVC video coding standard // IEEE Transactions on Circuits and Systems for Video Technology. – 2003. – Vol. 13, No. 7. – P. 560–576.
5. Sullivan G.J., Ohm J.-R., Han W.-J., Wiegand T. Overview of the High Efficiency Video Coding (HEVC) standard // IEEE Transactions on Circuits and Systems for Video Technology. – 2012. – Vol. 22, No. 12. – P. 1649–1668.
6. Oppenheim A.V., Schafer R.W. Discrete-Time Signal Processing. 3rd ed. – Upper Saddle River: Pearson, 2010. – 1108 p.
7. Rao K.R., Yip P. Discrete Cosine Transform: Algorithms, Advantages, Applications. – San Diego: Academic Press, 1990. – 490 p.
8. Begault D.R. 3-D Sound for Virtual Reality and Multimedia. – Boston: Academic Press, 1994. – 293 p.