(4) Evolution of Coding Technologies for Mobile Multimedia

Minoru Etoh, Khosrow Lashkari,
Frank Bossen and Wai Chu

In this article we describe the latest trends in media coding technology, including recently developed standards for speech coding (AMR-WB), audio coding (AAC+), and video coding (H.264/AVC). We then present research activities at DoCoMo USA Labs and outline research directions for future coding technologies.

1. Introduction

Media coding over wireless networks is driven by two fundamental rules. One is the well-known Moore's Law, which states that the processing power of semiconductor devices doubles every 18 to 24 months. Moore's Law has held for the evolution of coders and decoders (hereinafter referred to as COder-DECoder (CODEC)): in the ten years since the adoption of the MPEG-2 standard, there have been significant advances in coding efficiency thanks to the increased availability of computational power. The other rule is not generally known as a law but can be stated here as follows: a large bandwidth gap (by one or two orders of magnitude) exists between wireless and wired networks. Because of this bandwidth gap, there is a demand for coding technologies that can achieve efficient and compact representations of media data over wireless networks. Improving the quality of transmitted media not only depends on wireless-access technologies but also on coding technologies. Given these circumstances, over the past three years, DoCoMo USA Labs has been developing compression technologies for speech, audio, and video in collaboration with DoCoMo Multimedia Laboratories.

This article covers recent technological progress, following up on a previous article [1] issued in 2001, and describes future research directions toward more advanced coding technologies. The current CODEC technologies were designed to conform to the restricted hardware architectures of the past. However, in line with the advances in semiconductor technologies forecast by Moore's Law, future CODEC technologies should be developed with a more open mind. With additional computational complexity*, we can further improve coding efficiency. CODEC evolution is a fundamental principle driving research at DoCoMo USA Labs.

2. Speech and Audio Coding

Speech and audio CODECs aim to represent speech and audio signals in a compact digital form. This objective implies re-creating a perceptually equivalent waveform, rather than an accurate copy. To do this, redundancies are removed from the signal. Even though their purpose is the same, the coding technologies for speech and audio signals are very different.

2.1 Speech Coding

Speech coding models speech sounds (i.e., the human sound-producing apparatus, including the glottis, mouth, and lips) by using a source-filter model. Speech-coding algorithms are highly specialized for speech signals and are generally unsuited for coding music. A large number of coding standards have been developed, and a chronological list of various speech and audio coding standards is given in [1]. **Figure 1** shows the data rates for various applications together with the standardized CODECs used in these applications and their Mean Opinion Score (MOS)s, which measures the perceptual quality of coded speech and audio. Generally speaking, speech-coding technologies can be divided into three main categories:

1) Coders such as Pulse Code Modulation (PCM), Differential Pulse Code Modulation (DPCM), and Adaptive Differential Pulse Code Modulation (ADPCM) [2].
2) Vocoders such as Linear Predictive Coding (LPC).
3) Hybrid coders such as Code Excited Linear Prediction (CELP) [3] and the Adaptive Multi Rate (AMR) CODEC.

The FS1015, FS1016, and Mixed Excitation Linear Prediction (MELP) standards were developed for secure communication over narrowband radio channels. The sound quality they provide is low. The ITU-T therefore designed the G.729 standard to provide speech quality adequate for cellular networks. The Regular Pulse Excitation with Long Term

* Computational complexity: a term used in computing and information technology studies. It expresses the amount of space and time needed for solving a problem with a Turing machine (i.e., modern computers).

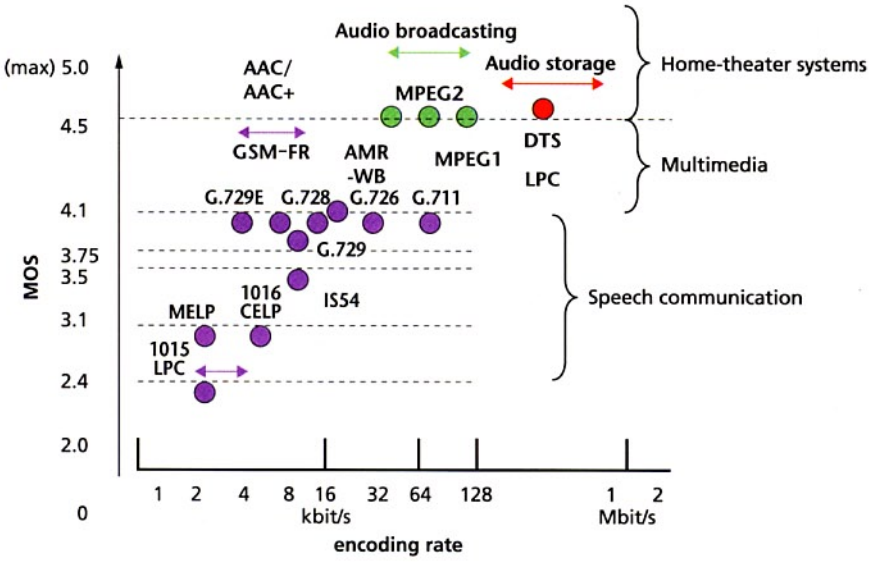

Figure 1 Position of various CODEC technologies in terms of MOS and encoding rate

Prediction (RPE-LTP) CODEC was developed by the European Telecommunications Standards Institute (ETSI) for use in the Global System for Mobile communications (GSM). The Adaptive Multi Rate-NarrowBand CODEC (AMR-NB) is suitable for applications in which bandwidth can fluctuate, such as Voice over IP (VoIP), and has been selected by the 3rd Generation Partnership Project (3GPP) as the mandatory speech CODEC for Wideband Code-Division Multiple Access systems (W-CDMA). In addition, the Adaptive MultiRate-WideBand (AMR-WB) CODEC provides high quality suitable for face-to-face speech in applications such as videoconferencing.

Figure 2 shows a conceptual diagram of the AMR-WB CODEC. It consists of nine source coders with bit rates ranging from 6.6 to 23.85 kbps. This coder is based on the CELP model, and it accomplishes efficient encoding by means of a sub-band filtering mechanism. It is similar to CELP coders of previous generations in many aspects. Nonetheless, it incorporates innovative techniques to efficiently achieve high-quality encoding. It is designed to work with wideband (7 kHz) speech sampled at 16 kHz. In the meantime, DoCoMo USA Labs is developing optimized encoders that reduce coding delay and increase the signal-to-noise ratio.

2.2 Audio Coding

A variety of musical sounds may be present in an audio signal for which no universal source model exists. Audio coding, therefore, does not attempt to model the sound source. Instead, it models the sound-perception mechanism, namely the human auditory system. MPEG-1 [6] audio coding consists of three different coding schemes—called Layers I, II, and III—and supports bit rates from 32 to 448 kbit/s. MPEG-1 Layer III (MP3) provides audio quality close to that of a compact disc (CD). MPEG-2 is an extension of MPEG-1 that provides lower sampling frequencies (16, 22.05, and 24 kHz) and multi-channel audio such as 5.1 surround sound. MPEG-2 AAC, a potential successor to MP3, is not backward compatible with MPEG-1 but achieves transparent stereo audio quality (i.e., it is indistinguishable from the uncompressed source) at 128 kbps. AAC can provide audio quality that exceeds CD quality, with the support of sampling rates up to 96 kHz. The most recent audio-coding standard, MPEG-4 High Efficiency AAC (hereinafter referred to as AAC+), provides high quality audio at low bit rates. While the perceived quality of most audio CODECs begins to break down below 128 kbps, AAC+ uses a technique known as Spectral-Band Replication (SBR) to achieve excellent stereo quality at 48 kbps and high quality at 32 kbps. In SBR, the full-band audio spectrum is divided into a low-pass section and a complementary high-pass section. The low-pass section of the spectrum is encoded with an AAC core, while the high-pass section of the audio spectrum is not coded directly. Instead, a small amount of information about this band is transmitted to enable a decoder to reconstruct the full-band audio spectrum. AAC+ attains high sound quality by taking advantage of the following two facts. First, the psycho-acoustic importance of the high frequencies in the audible frequency range is relatively low. Second, the lower and the higher frequencies of the audio spectrum are strongly correlated.

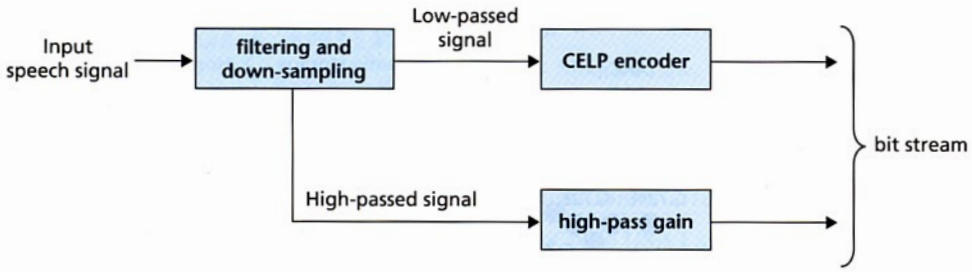

Figure 2 Conceptual diagram of the AMR-WB CODEC

2.3 Research Directions in Speech and Audio Coding

Current terminal devices must support two CODECs: one for speech (e.g., AMR) and one for audio (e.g., AAC). However, a single unified CODEC that can handle both speech and audio would be ideal. DoCoMo USA Labs—in collaboration with DoCoMo Multimedia Laboratories—has begun the development of a Unified Speech and Audio Codec (USAC). A prototype of this CODEC is based on the parametric audio-coding model used in MPEG-4: Harmonics, Individual Lines and Noise (HILN). A simplified block diagram of this coder is shown in **Figure 3**. The USAC relies on an information-source model that represents the input waveform as an integrated sound source. According to this model, an integrated speech and audio waveform is made up of the following three signal components:

1) harmonic tones described by their fundamental frequency and the spectral envelope of the amplitudes of their harmonics
2) single sinusoidal waves (also referred to as "individual lines") characterized by their frequency, amplitude, and phase
3) noise specified by its amplitude and spectral shape

The speech and audio signal is coded on the time axis—in terms of time intervals and frame units—using the parameters given by the above three components. The procedure is scalable, such that a single sound-source model can be tailored to form multiple models. As a result, the salient feature of this approach is that both speech signals and audio signals can be accurately represented.

Another goal of DoCoMo USA Labs is to improve the CODEC performance such as to produce three-dimensional, high-quality audio from the small loudspeakers fitted in wireless devices. We are working toward achieving these goals by using nonlinear signal-processing techniques.

3. Video Coding

3.1 New Video Codecs

The advances made in semiconductor technology and cellular networks have enabled video applications to be provided on mobile devices. To make the most out of the available resources and provide a realistic visual experience, advanced video-compression algorithms are needed. Several compression algorithms have been standardized over the last 15 years, mainly by two standardization bodies: MPEG, formed by the International Organization of Standardization and the International Electrotechnical Commission (ISO/IEC); and the Video Coding Experts Group (VCEG), formed by the ITU-T. The standards developed by these groups include H.261, H.263, MPEG-1, MPEG-2, and MPEG-4. While H.263 and MPEG-4 are currently used for mobile and videoconferencing applications, the quality they provide at low and medium bit rates is limited. MPEG and VCEG recently joined forces and created the Joint Video Team (JVT), which consequently developed H.264/AVC [7]. The standard was given final approval in May 2003, and it is expected to dramatically improve coding efficiency. Proprietary solutions such as Windows Media Video 9 [8] have also emerged and are aiming to become worldwide de-facto standards.

H.264/AVC is based around the same hybrid coding architecture featuring motion compensation and transformation used in previous standards (see the block diagram in **Figure 4**), which includes three frame types: I, P and B. Taken one by one, however, the techniques applied in this standard are very different from previous ones. One of the key factors that enable higher quality is the improvement of prediction. Prediction comes in two forms: spatial (i.e., within frame) and temporal (i.e., between frames). Whereas MPEG-4 uses a simple algorithm to predict coefficient values in either the horizontal or the vertical direction, H.264/AVC performs spatial prediction in the base-

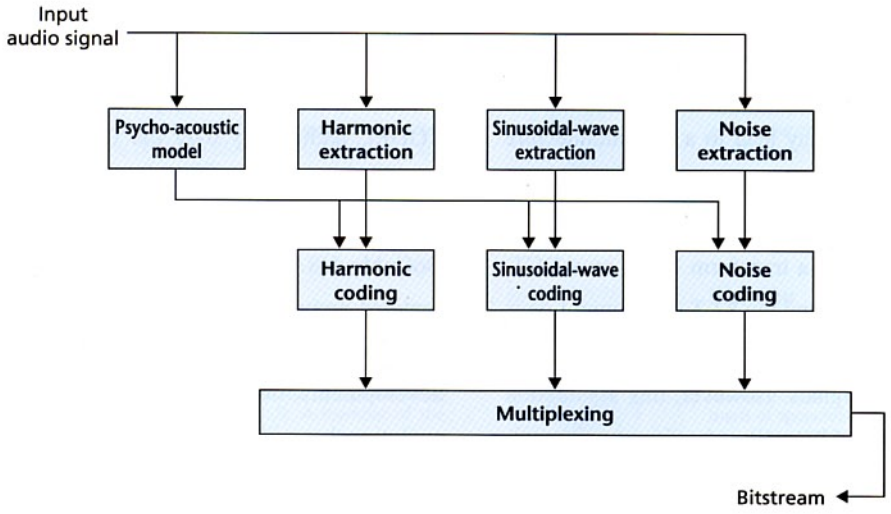

Figure 3 Basic concept of the unified speech and audio CODEC

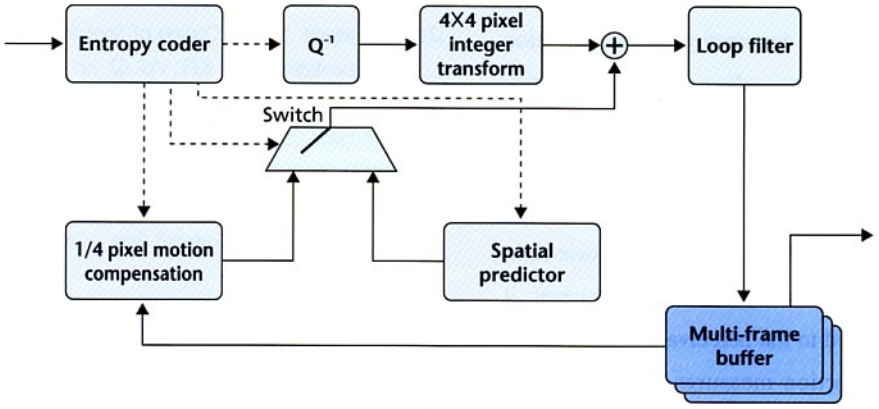

Figure 4 Block diagram of H.264/AVC decoder

band domain in one of nine orientations. Temporal prediction is also improved by using 1/4 pixel motion compensation (MPEG-4 Simple Profile features 1/2 pixel precision) and block sizes as small as 4×4 pixels that can capture motion boundaries more precisely (MPEG-4 Simple Profile uses block sizes down to 8×8 pixels). The ability to choose between multiple reference frames for temporal prediction also contributes to the improvement of coding efficiency. In addition, H.264/AVC also differs from other CODECs in that it performs an integer transformation approximating a Discrete Cosine Transform (DCT) in 4×4 pixel units. This departure from the floating-point 8×8 pixel transform found in previous standards removes any drift problem that might occur between the encoder and decoder and guarantees that all decoders reconstruct the same image. Another feature of H.264/AVC is that it incorporates a loop filter that removes blocking artifacts. Using this filter within the loop (as opposed to a post-process) also increases the accuracy of the motion-compensated prediction. Finally, H.264/AVC also uses advanced entropy-coding methods, including context-adaptive Huffman coding and context-adaptive binary arithmetic coding.

Windows Media Video 9 is a CODEC that shares many features with H.264/AVC, namely, I, P, and B frames, a loop-filter, 1/4 pixel motion compensation, and an integer transform. On the other hand, it differs in providing multiple sizes of block transforms and adaptive filters for motion compensation. In particular, the smallest block size for motion compensation is limited to 8×8 pixels and arithmetic coding is not supported. Despite these differences, the resemblance between both CODECs is significant and a matter of great interest.

3.2 Research Directions in Video Coding

While the current generation of video CODECs provides good quality, it is expected that the quality will have to be improved for future generations. These improvements may come in several forms:

1) Unlike in JPEG-2000, arithmetic coding was added on top of an existing algorithm in H.264/AVC. Designing a video CODEC under the assumption of arithmetic coding may provide further gains. This may lead to a new architecture that does not use macroblocks as the basic coding unit.
2) While H.264/AVC breaks with the conventional 8×8 DCT transform, it still depends on a transform based on a DCT. Transforms such as the DCT and the KLT (Karhunen-Loeve Transform) are designed under the assumption of a Gaussian distribution of signals. However, this assumption may not necessarily hold. Alternative transforms that maximize compression in a rate-distortion sense should be sought. The design of such transforms requires an optimization process that takes entropy coding into account.
3) Inter-frame prediction plays a key role in coding efficiency. Two factors may improve the accuracy of such prediction and thus produce more efficient coding: (i) better description of motion boundaries and (ii) adaptation of filters in the motion-compensation loop.
4) Encoder optimization typically relies on maximizing rate and distortion characteristics. The peak signal-to-noise ratio (PSNR) is the preferred measure for distortion. However, it is not always well correlated to the perceived visual quality. The use of alternate distortion measures (e.g., weighted PSNR) may lead to an increase in the perceived image quality.

Currently, DoCoMo USA Labs is undertaking the research focusing on above four improvements.

4. Conclusion

This article has given an overview of the advances in CODEC technology over the last few years, introduced the recently developed AMR-WB, AAC+, and H.264/AVC standards, and covered future CODEC developments directions. DoCoMo USA Labs is performing research and development centered around integrated speech and audio encoding as well as video CODECs based on arithmetic coding. As for future developments, our endeavor is to contribute to the CODECs evolution that will meet the demands of mobile multimedia.

References

[1] H. Nakano and M. Etoh, "An Overview of Signal-Processing Techniques for Mobile Multimedia", NTT DoCoMo Technical Journal Vol. 2, No. 4, pp. 4–9, Mar. 2001.
[2] N. S. Jayant: "Digital Coding of Speech Waveforms: PCM, DPCM and DM Quantizers," Proc. IEEE, Vol. 62, pp. 611–632, May 1974.
[3] ITU-T, Recommendation G. 729 Coding of speech at 8 kbit/s using Conjugate-Structure Algebraic-Code-Excited-Linear-Prediction (CS-ACELP), ITU-T, Geneva, Sep. 1998.
[4] K. Lashkari and T. Miki: "Joint Optimization of Model and Excitation in Parametric Speech Coders," Proc. ICASSP 2002, pp. 277–280, May 2002.
[5] W. Chu and T. Miki: "Optimization of Window and LSF Interpolation Factor for the ITU-T G.729 Speech Coding Standard," Proc. Eurospeech 2003, pp. 1061–1064, Sep. 2003.
[6] ISO/IEC 13818-3: "Coding of Moving Pictures and Associated Information-Part 3: Audio," May 1995.
[7] ITU-T Recommendation H.264 | ISO/IEC 14496-10, Geneva, May 2003.
[8] Microsoft, WMV9-an advanced video codec for 3GPP, Document S4(03)0613 submitted to 3GPP, Sep. 2003.

Abbreviations

AAC: Advanced Audio Coding
ADPCM: Adaptive Differential Pulse Code Modulation
AMR: Adaptive Multi Rate
AMR-NB: Adaptive Multi Rate-NarrowBand
AMR-WB: Adaptive Multi Rate-WideBand
AVC: Advanced Video Codec
CELP: Code Excited Linear Prediction
CODEC: COder DECoder
DCT: Discrete Cosine Transform
DPCM: Differential Pulse Code Modulation
ETSI: European Telecommunications Standards Institute
GSM: Global System for Mobile communications
HILN: Harmonics, Individual Lines and Noise
IEC: International Electrotechnical Commission
ISO: International Organization of Standardization
ITU-T: International Telecommunications Union, Telecommunication standardization sector
JPEG: Joint Pictures Experts Group
JVT: Joint Video Team
KLT: Karhunen-Loeve Transform
LPC: Linear Predictive Coding
MELP: Mixed Excitation Linear Prediction
MOS: Mean Opinion Score
MP3: MPEG-1 Audio Layer-3
MPEG: Moving Pictures Experts Group
PCM: Pulse Code Modulation
PSNR: Peak Signal to Noise Ratio
RPE-LTP: Regular Pulse Excitation-Long Term Prediction
SBR: Spectral Band Replication
SNR: Signal-to-Noise Ratio
USAC: Unified Speech and Audio Coding
VCEG: Video Coding Experts Group
VoIP: Voice over Internet Protocol
W-CDMA: Wideband Code Division Multiple Access
3GPP: 3rd Generation Partnership Project