Progress in ITU-T Video/Speech Coding Standardization

Shigeaki Sasaki†, *Hideaki Kimata, and Yusuke Hiwasaki*

Abstract

A number of recommendations related to video and speech coding, which form a crucial part of digital telecommunication, have been approved by ITU-T (International Telecommunication Union Telecommunication Standardization Sector). New work items, such as H264/SVC/MVC and G.711-WB, for enhancing the quality of telecommunication and introducing new services are currently being studied with aggressive schedules. This article gives an overview of the hot trends in video/speech coding at ITU-T.

1. Introduction

Along with the widespread usage of broadband networks based on Internet protocol (IP), IP telephony services over those networks, such as *Hikari Denwa*, are now becoming popular for both home and business use. Even if the networks have higher throughput, compression technologies for voices and images play important roles in transmitting data with high quality without delay. Moreover, it is necessary to provide standards for the video/speech coding to retain interoperability with legacy telecommunication systems. At ITU-T (International Telecommunication Union Telecommunication Standardization Sector), various Recommendations, e.g., H.262, H.263, G.711, and G.729, have already been published as standards for video/speech coding. The hierarchical structures of SG16 (Study Group 16), which is responsible for standardization related to multimedia service systems and terminals, WP3 (Working Party 3), which manages overall issues about media coding, and Questions 6, 9, 10, and 23, where work items are assigned, are shown in **Fig. 1**. A lot of experts in the media coding field participate in these groups and contribute to the standardization. Details of each activity in video and speech coding are given in the following sections.

† NTT Cyber Space Laboratories
Musashino-shi, 180-8585 Tokyo
Email: sasaki.shigeaki@lab.ntt.co.jp

2. Hot trends in speech coding standardization

The bandwidth of most speech coding standards in ITU-T, such as G.711 and G.729, was traditionally limited to the frequency bandwidth from 300 Hz to 3.4 kHz. However, the focus of such standardization has shifted towards standards that can provide higher quality with wider bandwidth, e.g., up to 7 or 14 kHz, while maintaining interoperability with legacy standards.

From the technological point of view, since the main focus of the conventional speech codecs was speech communication, those codecs, for example, CELP (code excited linear prediction), achieved high compression rate by using speech models based on the characteristics of human voices. However, the new wideband standards are required to encode both speech and music with high fidelity. Consequently, transform coding in the MDCT (modified discrete cosine transform) domain, which has generally been used only for audio codecs, such as AAC (advanced audio coding), is now popularly used in conjunction with the conventional speech coding technique.

SG meetings are usually held every eight to nine months; however, if there are urgent market demands for a new standard to be used in new services or new products, WP3 meetings and Expert meetings are organized once every two or three months to deal with the demands in a timely manner. Discussion also takes place by email to improve the efficiently of the procedure up to its approval. The current activities in

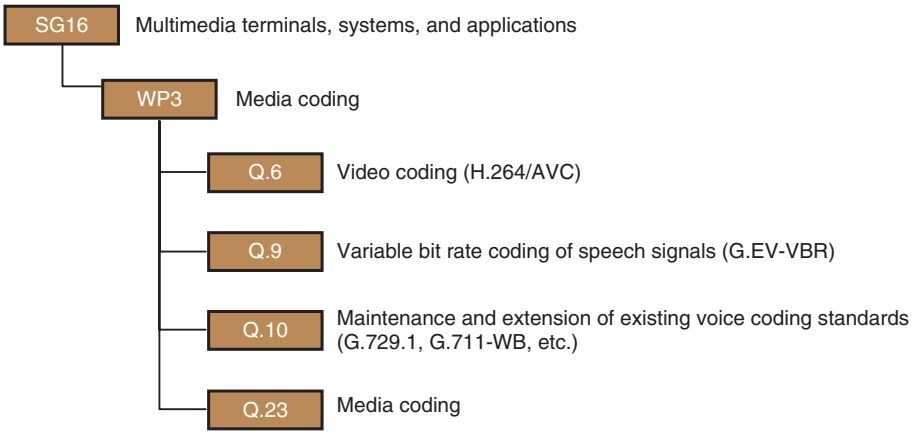

Fig. 1. Hierarchical structure of ITU-T SG16 related to video/speech coding.

each question are given below.

Q.9 Variable Bit-rate Coding of Speech Signals: In this question, a new work item for a speech coding standard, tentatively called G.EV-VBR, is being studied. The standard is designed to have a scalable bitstream and variable bit-rate and be applied to both the telephone band and 7-kHz wideband usage at 8–32 kbit/s. This standard is expected to enter service for 4G (fourth generation) wireless communications. It is scheduled to be approved by the end of 2008.

Q.10 Software Tools for Signal Processing Standardization Activities and Maintenance and Extension of Existing Voice Coding Standards: G.729.1, which is an extension of G.729 widely used for VoIP codecs and provides 7-kHz wideband speech and bitstream scalability, and G.722.1 Annex C, which is a 14-kHz-bandwidth annex of G.722.1, were standardized in this question during the current study period. Two work items are currently ongoing: (i) G.711 wideband extension (G.711-WB) to provide high-quality wideband speech interoperable with G.711 and (ii) G.722.1 full-band extension (G.722.1-FB), a 20-kHz bandwidth annex of G.722.1. The main features of the former are low frame delay and low computation, which are especially advantageous when mixing 7-kHz wideband signals. This item was launched in January 2007 based on a proposal by NTT. Five organizations worked closely together to integrate useful technologies into a candidate codec with NTT acting as the leader of this collaboration, and the candidate has been approved as a new ITU-T standard. The latter item is progressing under collaboration between Polycom and Ericsson and should be finalized by April 2008. The most recent topics are a 14-kHz-bandwidth/stereo extension to both G.711-WB and G.722 (G.711/722SWB) and a G.711 lossless compression (G.711LLC). In the expert meeting held in Nov. 2007, it was agreed to start standardization of these two work items.

Q.23 Media Coding: This question deals with study items that extend over several questions or have no relation to other existing questions. Currently, a 14-kHz-bandwidth/stereo extension that is to be applied commonly to G.EV-VBR and G.729.1 is being discussed.

The bit-rates and bandwidths of speech coding standards recently published or currently in progress are shown in **Fig. 2**.

In SG16, Q.8 regarding generic sound activity detection (GSAD), a more generalized voice activity detection (VAD) method, was newly established in April 2007. In the future, study items that have no relevance to interoperability, like VAD in Q.8, are expected to be organized as new questions if sufficient commercial merit is proved.

3. Hot trends in video coding standardization

Recommendations for video coding are also being published in ITU-T SG16. They are being studied in Q.6 VCEG (Video Coding Experts Group). Although the video codecs were initially intended for videoconferencing, they are now drawing more attention because they could play an important role in video content delivery, such as IPTV (Internet protocol television). This section describes hot trends in video coding, focusing on the recently standardized H.264 and its extensions (**Fig. 3**).

Traditionally, in video coding standards such as H.26x, the assumed network characteristics were

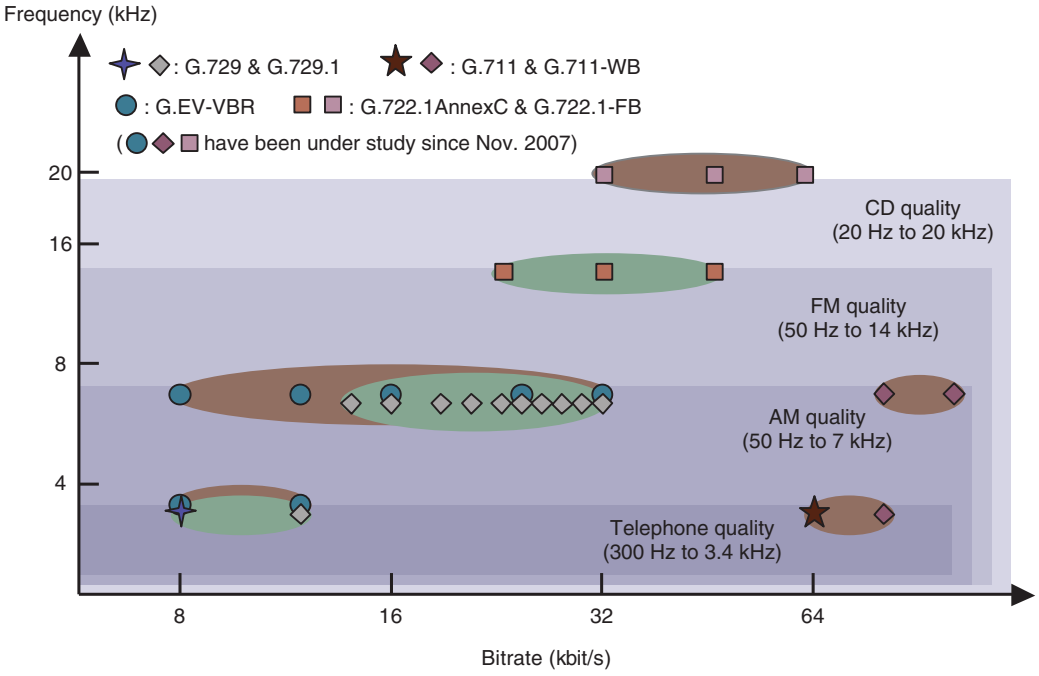

Fig. 2. Bitrates and bandwidths of current ITU-T speech coding standards.

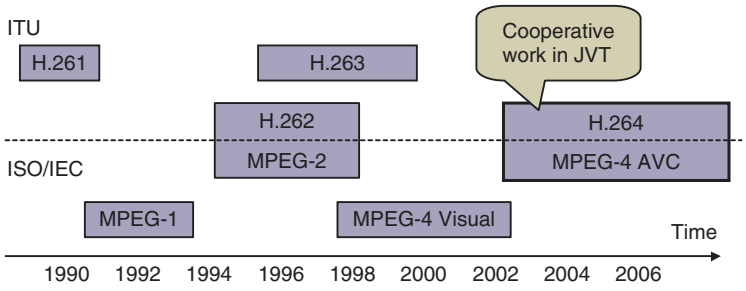

Fig. 3. History of video coding standards.

defined first and then, together with the terminal specification, the codecs were specified. In the case of H.261, the main target was ISDN (integrated services digital network), for H.262, it was the broadband ATM (asynchronous transfer mode) network, and for H.263, it was the PSTN (public switched telephone network) and the mobile network. SG16 has worked closely together with ISO (International Organization for Standardization), another international standardization body, and many Recommendations have been made common with ISO's MPEG series. H.262 was developed together with ISO and is identical to MPEG-2. MPEG-4 Visual is based on H.263 and has been enhanced to have high tolerance to bit errors. The standardization of H.264 was initiated and discussed only in ITU-T as H.26L, but a working group called JVT (Joint Video Team) was jointly established in 2001 by ITU-T and ISO/MPEG, and the work is now being moved forward by this group. As a result, H.264 is identical to the ISO standard MPEG-4 AVC (advanced video coding). The JVT meetings are held four times a year and attract about 100 contributions in every meeting. NTT has been active in JVT since its establishment and has made many contributions to all of the H.264-related standardizations, including the main body and its annexes, as described below.

H.264 is designed to be used on packet-based networks such as IP networks. The encoded data is packetized in an NAL (network adaptation layer) unit. The coding parameters are stored separately from the bitstream in another NAL and transmitted using the

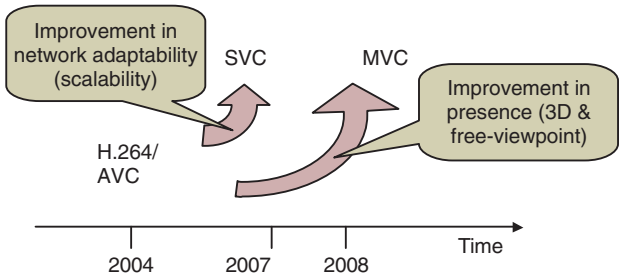

Fig. 4. H.264/AVC and its annexes.

protocol in order to enable parameter control. The introduction of the NAL also improves the compression efficiency. First of all, start codes are now unnecessary, and secondly, the bit-errors are no longer considered, so arithmetic coding can be applied. Moreover, the use of more complicated syntax is now possible and H.264 provides much greater flexibility for integrating various prediction methods. The first version of H.264 was completed in 2003 and its annexes are now being studied. Two representative annexes, SVC (scalable video coding) and MVC (multiview video coding), are described below and shown in **Fig. 4**.

(1) SVC

SVC was standardized as H.264 Annex G in 2007. It has spacial, temporal, and quality (signal-to-noise ratio) bitstream scalability. Although ordinary video standards also have scalability, the emphasis of this extension was to have single-loop decoding in order to achieve high compression efficiency without increasing the complexity of the decoding process. SVC was also designed for low-complexity distribution via streaming servers, where NAL packets are processed according to a newly specified Priority label for each packet.

(2) MVC

MVC is now being developed and the standardization is to be completed as H.264 Annex H in 2008. The scope is efficient compression method of multiview videos and the foreseen applications are three-dimensional and free-viewpoint videos. This technology will let viewers freely change their viewpoint in a scene. This new annex is designed to encode a huge amount of data in real time by using optimized parallel processing. It is also designed so that partial decoding can be achieved easily, thus enabling the reproduction of only the region of interest. There is a built-in mechanism for transmitting auxiliary parameters, so that camera parameters of the shot images can be transmitted as metadata.

Discussion about a post-H.264 study item has also been started. This item will be discussed along with its target network and the requirements of the system it is expected to be used for.

4. Future activities

Standardization of media coding with high quality and high usability is making progress in ITU-T. Although the current study period of SG16 is to be concluded in 2008 and its organization for the next period has yet to be discussed, the present work items will probably be continued into the next study period as planned.

Contributions to international standardization are gaining in importance as a means to increase the international competitiveness of Japanese technologies. This is certainly true for the standardization of media coding, and NTT will continue to contribute to the standardization work related to video/speech coding.

Shigeaki Sasaki
Research Engineer, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.
He received the B.E. degree in physics from Kyoto University, Kyoto, in 1991. He joined NTT in 1991 and has been engaged in research on wideband speech coding. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan and the Acoustical Society of Japan (ASJ).

Hideaki Kimata
Senior Research Engineer, Visual Media Communications Project, NTT Cyber Space Laboratories.
He received the B.E. and M.E. degrees in applied physics and the Ph.D. degree in electrical engineering from Nagoya University, Aichi, in 1993, 1995, and 2006, respectively. He joined NTT in 1995 and has been engaged in R&D of video coding and error-resilient video coding algorithms and visual communication systems. His research interests also include free viewpoint video coding, 3D video coding, and pre- and post-processing for video coding. He is a member of IEICE.

Yusuke Hiwasaki
Senior Research Engineer, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.
He received the B.E., M.E., and Ph.D. degrees from Keio University, Kanagawa, in 1993, 1995, and 2006, respectively. Since joining NTT in 1995, he has been engaged in research on low-bit-rate speech coding and voice-over-IP telephony. From 2001 to 2002, he was a guest researcher at the Royal Institute of Technology in Sweden. He is a member of IEEE, IEICE, and ASJ. He received the Technology Development Award from ASJ and the Best Paper Award from IEICE, both in 2006.