Current Progress of ITU-T Speech Coding Standardization

Shigeaki Sasaki†

Abstract

Since speech coding is a key technology required for telecommunication, several recommendations for it have been approved in ITU-T (International Telecommunication Union, Telecommunication Standardization Sector), and new standardization work, e.g., G.711.1 and G.729.1, which further enhance the speech bandwidth and have interoperability with the conventional telephony standard, is currently progressing for the new products and services demanded by the market.

1. Introduction

Broadband networks based on Internet protocol (IP) over wired networks, e.g., FLET'S HIKARI NEXT, have spread widely in Japan and wireless ones over cellular phone networks are also available. One of the services on those IP networks, VoIP (voice over IP), such as HIKARI DENWA, is also becoming popular for use in both homes and offices. Even though the access networks have become broadband, meaning that they have high capacity, speech coding technologies are still required in order to provide telecommunication services to many customers with high quality simultaneously because the network has limited throughput. Moreover, the standardization of speech coding methods guarantees the speech quality and interoperability of telecommunication services that use methods conforming to the standards.

2. Standardization in ITU-T

ITU-T (International Telecommunication Union, Telecommunication Standardization Sector) develops global standards for telecommunication and has standardized speech coding algorithms for telephony and voice communication. During the current study period, 2009–2012, Study Group 16 (SG16) is responsible for standardization of multimedia coding, systems, and applications. As shown in **Fig. 1**, WP3, one of its Working Parties, coordinates overall standards regarding multimedia coding and handles study items, which are assigned to Questions. A lot of experts in the media coding field participate in those groups and contribute to the standardization. Work items related to speech media are being or have been studied in the following Questions.

- Q.8 (Generic sound activity detection): Q.8 is developing a method of detecting a voice period and classifying signals present in that period, e.g., speech, music, or noise. It was responsible for the G.720.1 standard aimed at pre-processing for speech coding.
- Q.9 (Embedded variable bit rate coding of speech signals): This Question studied a coding algorithm that can provide different coding bitrates and produced G.718 as a result; it was then terminated in Oct. 2010.
- Q.10 (Speech and audio coding and related software tools): In the previous study period, 2005–2008, the main role of Q.10 was maintenance for existing coding standards. Its scope of work has been extended to take on most of the responsibilities regarding speech coding from this study period. Maintenance of G.191, which is a package of tools required for standardization work, is also an important role of Q.10.

SG meetings are usually held every eight or nine months, but interim meetings, such as WP3 meetings, are organized once every two or three months in order

† NTT Cyber Space Laboratories
Musashino-shi, 180-8585 Tokyo

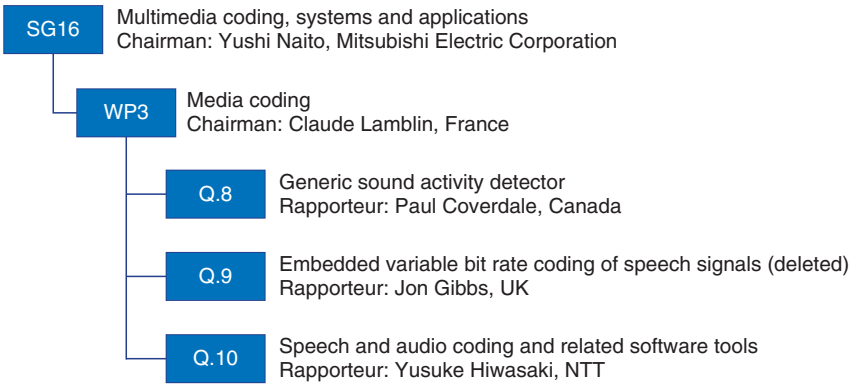

Fig. 1. SG16 structure related to speech coding.

to deal with market demands for new standards in a timely manner.

3. Hot trends in speech coding standard

Conventional standards for speech coding were designed to have speech bandwidth, e.g., telephone band or wideband, and bitrate to suit services for which the standard was expected to be used. However, since the last study period, before the start of standardization work, Requirements were defined to meet market needs for products and services. For the current market demands, the following features have been included in the recommendations.

- SWB and FB audio coding: Telecommunication systems with higher speech quality and higher presence, such as a telepresence system, need the capability to encode speech signals with bandwidths in the classes known as super-wideband (SWB, 50 Hz to 14 kHz) or full band (FB, 20 Hz to 20 kHz).
- Scalable audio coding: To give the bitstream a layered structure, both the bitrate and speech quality can be controlled according to the throughput of the transmission channel or some other conditions. Furthermore, in order to enhance the speech quality and the bandwidth of the conventional coding, enhancement bits can be layered onto the bitstream of the conventional structure. This structure also enables interoperability with the conventional coding through the extraction of the bits corresponding to those of the conventional structure.
- Lossless audio coding: Coding in which no information contained in the original signals is lost through compression and decompression. (In conventional speech coding, called lossy coding, distortion from the original is allowed in order to obtain a higher compression rate.)

Below, current recommendations with the above features are introduced in the order of their approval date.

(1) G.722.1 Annex* C

On the basis of a wideband speech standard, G.722.1, Polycom proposed its SWB extension as G.722.1 Annex C for videoconferencing and teleconferencing systems. G.722.1 Annex C was the first SWB speech coding standard in ITU-T and has very low computation for SWB coding. Its basic algorithm was common to G.722.1, but compatibility with G.722.1 was not provided.

(2) G.729.1

G.729.1 is a wideband scalable extension of G.729 (8 kbit/s), originally designed for VoIP. Since its bitstream contains that of G.729 as the core bitstream, terminals supporting G.729 can decode the core bitstream into speech signals by extracting it from the G.729.1 bitstream. Its bitrate can be controlled from 8 kbit/s to 32 kbit/s in 2-kbit/s steps. France Telecom, VoiceAge, Panasonic, and Siemens and others contributed its standardization.

(3) G.711.1

The proposal from NTT and four other organizations, ETRI, France Telecom, Huawei Technologies, and VoiceAge, was approved as G.711.1. The most important feature of this standard is interoperability

* Annex: An appendix recommendation that forms a part of the recommendation body. If related to the body, it might be made into a series as a part of the body.

with G.711, which is the most widely spread for speech coding. Scalable coding, for which the core layer consists of the G.711 bitstream, was introduced into it. 7-kHz wideband speech is obtained by decoding the whole bitstream and it is possible to extract only the core part and decode it by G.711. Because G.711.1 was based on G.711, which has a bitrate of 64 kbit/s, the bitrate of this new standard seems to be higher, up to 96 kbit/s, than that for speech compression, but it was targeted for use on the broadband network and no attention was paid to compression rate. Where legacy terminals that support only G.711 and new terminals that can handle G.711.1 co-exist, interoperability between the two standards is a great advantage. Moreover, its speech quality has been confirmed to be the highest among the wideband recommendations in ITU-T. G.711.1 is now used in the high-quality telephone service of HIKARI DENWA provided on NTT's broadband network FLET'S HIKARI NEXT.

(4) G.718

Using scalable coding, G.718 has also achieved both a bitrate that can range from 8 kbit/s to 32 kbit/s and wideband speech coding. The bitrate and basic structure of the bitstream are similar to those of G.729.1; however, the coding algorithm of the core layer at 8 kbit/s is a new design rather than the conventional one. Since there are no constraints with regard to interoperability with existing standards, a more efficient coding algorithm could be introduced. This recommendation was studied in the open consortium organized by Ericsson, Motorola, and Nokia and others.

(5) G.719

G.719 is the first FB speech coding standard in ITU-T. A joint proposal from Polycom and Ericsson, based on the algorithm of G.722.1 and G.722.1 Annex C, was approved as G.719. It has the advantage of much less computation like G.722.1 and G.722.1 Annex C.

(6) G.711.0

The concept of lossless audio coding was introduced for the first time in ITU-T in G.711.0. To be precise, this algorithm compresses the bitstream from G.711, not the input audio signals, without any degradation. The compression rate for lossless coding will be varied according to the redundancy of the input source to be encoded, but the rate has been confirmed to be 50% on average for a G.711 bitstream. Making new network devices, such as routers and gateways, that conform to this standard would save backbone network throughput that could be used for voice channels. A joint proposal from NTT, Huawei Technologies, and Cisco and others was approved as G.711.0.

(7) G.718-SWB/G.729.1-SWB

There are some common features between G.718 and G.729.1, such as the speech bandwidth, bitrate, and bitstream structure scalability. Their extensions for SWB capability were initially proposed independently, but additional layers for SWB enhancement, which are applicable to both standards, were developed considering their common features. As a result, the SWB extensions to G.718 and G.729.1 were approved as G.719 Annex B and G.729.1 Annex E, respectively.

(8) G.711.1-SWB and G.722-SWB

SWB extensions to G.711.1 and G.722 were developed in the same way as for the standardization of SWB extensions to G.718 and G.729.1. There are some features in common between G.711.1 and G.722, e.g., wideband capability, low delay, and low computation. Two additional enhancement layers at 16 kbit/s each consist of two types of sublayers: a common sublayer for SWB expansion and a sublayer specific to each core layer for wideband enhancement. G.711.1-SWB, which is composed of G.711.1 and these enhancement layers, was approved as G.711.1 Annex D and G.722-SWB was approved as G.722 Annex B. This work item was developed by five organizations: NTT, ETRI, France Telecom, Huawei Technologies, and VoiceAge.

Specifications of the above recommendations—usage, bitrate, delay, and computational complexity—are listed in **Table 1**.

The standards for stereo telecommunication are now being studied as the latest work items. Specifically, stereo extensions to the above extensions, (7) and (8), are aimed at videoconferencing systems and telepresence systems. The former is planned to be approved in 2012, and the latter was scheduled to be completed in 2011.

These recommendations are intended to be implemented in suitable digital signal processing chips and their algorithms are described in fixed-point arithmetic. Once the fixed-point recommendations have been completed, floating-point recommendations, which are guaranteed to be compatible with the fixed-point ones, would usually be planned for implementation on personal computers and for assistance in understanding the algorithms. For example, a floating-point implementation of G.711.1 was standardized as G.711.1 Annex A.

Table 1. Speech coding recommendations approved after previous study period.

Name	G.722.1 Annex C	G.729.1	G.711.1	G.718	G.719	G.711.0	G.718 Annex B	G.729.1 Annex E	G.711.1 Annex D	G.722 Annex B
Approval	May 2005	Apr. 2006	Mar. 2008	Jun. 2008	Jun. 2008	Sep. 2009	Mar. 2010		Nov. 2010	
Audio bandwidth	50 Hz to 14 kHz	50 Hz to 7 kHz	50 Hz to 7 kHz	50 Hz to 7 kHz	20 Hz to 20 kHz	300 Hz to 3.4 kHz	50 Hz to 14 kHz		50 Hz to 14 kHz	
Bitrate (kbit/s)	24, 32, 48	8–32	64, 80, 96	8–32	32–128	Not fixed (50% on average)	36–48	36–64	96, 112, 128	64, 80, 96
Frame length (ms)	20	20	5	20	20	5–40	20	20	5	5
Algorithmic delay (ms)	40	48.9375	11.875	43.875	40	Same as frame length	49.625	55.6875	12.8125	12.3125
Computational complexity (WMOPS)	11	35.8	8.7	57	21	1.67	80	63	21.498	22.76
Technology	MLT	G.729, MDCT, BWE, SVQ	G.711, MDCT, Interleave VQ	ACELP, MDCT, AVQ	Adaptive MDCT, FLVQ	Lossless compression	G.718, MDCT, Sinusoidal coding	G.729.1, MDCT, Sinusoidal coding	G.711.1, MDCT, BWE, AVQ	G.722, MDCT, BWE, AVQ
Usage	Teleconfer-encing and telepresence systems	VoIP	NGN and VoIP	VoIP	Teleconfer-encing and telepresence systems	Routers on backbone network	Teleconferencing and telepresence systems		Teleconferencing and telepresence systems	
Other features		Interoperable with G.729	Interoperable with G.711			Lossless coding of G.711 bitstream	Interoperable with G.718	Interoperable with G.729 and G.729.1	Interoperable with G.711 and G.711.1	Interoperable with G.722

ACELP: algebraic code excited linear prediction
AVQ: algebraic vector quantization
BWE: bandwidth extension
FLVQ: fast lattice vector quantization
MDCT: modified discrete cosine transform
MLT: modulated lapped transform
NGN: Next Generation Network
SVQ: spherical vector quantization
WMOPS: weighted millions of operations per second

4. Related speech processing standardization

As mentioned above, assuring interoperability and compatibility in telecommunication is the primary purpose of an ITU-T standard while keeping communication quality appropriate could be another purpose even if it is unrelated to the primary purpose. For example, methods such as packet loss concealment (PLC) and voice activity detection (VAD), which have no impact on interoperability, are usually integrated into the coding algorithm in order to qualify the performance requirements, but those methods could be standardized separately from the main body of the recommendation as Appendices.

(1) PLC

When packets containing the speech bitstream are lost on the IP network, the packet loss could be audible as an interruption in the speech. PLC makes the break inaudible by reproducing the lost speech. The original speech coding standard was made when the network was assumed to have no packet loss, so it did not have any PLC; therefore, it was newly developed as an Appendix. G.722 Appendices III and IV approved in November 2006 are examples of appendices that were established much later than the main body.

(2) VAD

The classification of input signals as speech, silence, etc. and use of a coding method appropriate to the signal's characteristics, such as not transmitting any information during silent periods, enable very efficient coding to be achieved. The input detection and classification method is called VAD. In the coding standards with VAD, e.g., G.729 Annex B and G.722.2, the algorithms, which are specifically designed to meet requirements, and the processing using the information, e.g., silence compression, are described together. In Q.8, the concept of a generic sound activity detector (GSAD), independent of the coding algorithm and applicable to general use, was newly defined; the first recommendation, which was based on this concept, was approved as G.720.1 in January 2010. It enables a 10-ms period of input to be classified into one of four categories: speech, music, noise, or silence.

Besides the speech processing methods, some other

recommendations were developed in order to support the standardization work. G.191 provides lots of tools, which are required for evaluating whether an algorithm meets the requirements, e.g., level adjustment, sampling rate conversion, filtering, error pattern generation, and fixed-point arithmetic library. They have been updated regularly and the current version of the set of tools is STL2009 (Software Tool Library 2009).

5. Use of ITU-T speech coding standard

This section describes how to obtain and use ITU-T speech coding standards. Most of the recommendations can be downloaded from the ITU-T website [1] except for preliminary versions. Since the current recommendations consist of recommendation text and source code, it is easy to evaluate the performance, e.g., speech quality and interoperability.

The recommendations are freely available for the purpose of standardization work and performance evaluation: however, for business use, a license must be obtained from the licenser holding the intellectual property rights, such as patents and software copyrights, included in the recommendations, which usually involves a fee set by the licenser. As mentioned above, several organizations jointly contributed to making some standards, so there are often multiple licensers per recommendation. To avoid the trouble of contacting all of them independently, the licensers have established a patent pool that provides a one-stop service for licensees.

6. Future activities

To meet market needs, new work items will progress in ITU-T such as speech and audio coding for multichannel telecommunication, NGN (Next Generation Network), FMC (fixed mobile convergence), and other audio processing for improving the quality of telecommunication.

NTT has participated in ITU-T standardization, such as contributing to the development of G.711.1 and providing the rapporteur for Q.10, and will continue to make further contributions to the standardization work in order to provide better telecommunication services to its customers.

Reference

[1] ITU downloads. http://www.itu.int/rec/T-REC/en

Shigeaki Sasaki
Senior Research Engineer, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.
He received the B.E. degree in physics from Kyoto University in 1991. Since joining NTT in 1991, he has been engaged in the research field of wideband speech coding. He received the Achievement Award from the Institute of Electronics, Information and Communication Engineers (IEICE) in 2009 and the Maejima Award from the Teishin Association in 2010. He is a member of IEICE and the Acoustical Society of Japan.