Global Standard for Wideband Speech Coding: ITU-T G.711.1 (G.711 wideband extension)

Shigeaki Sasaki†, *Takeshi Mori, Yusuke Hiwasaki, and Hitoshi Ohmuro*

Abstract

NTT has developed a scalable wideband speech coding method, whose core layer is ITU-T standard G.711, and has also acted as a leader of its standardization in ITU-T (International Telecommunication Union, Telecommunication Standardization Sector) since January 2007. As a result of the standardization activities, a new wideband speech coding method, proposed by NTT and four other organizations, was approved as ITU-T G.711.1 in March 2008. This article introduces the background and the process up to its approval and then describes the concepts, technical specifications, and quality evaluation of G.711.1.

1. Background

Along with the spread of broadband access networks based on optical fiber or ADSL (asymmetric digital subscriber line), IP-based telephony services using those networks, such as Hikari Denwa, are being used more and more in homes (IP: Internet protocol). In enterprises, many of the legacy private branch exchanges (PBXs) are also being replaced by IP-PBXs or VoIP (voice over IP) gateways installed in company intranets having a bandwidth of 100 Mbit/s or more. As a result, the speech communication services provided on IP networks are now becoming popular. The new generation of such services requires new speech coding algorithms designed with emphasis on factors such as wider audio frequency bandwidth, lower delay, and lower complexity rather than bitrate efficiency.

Today, the majority of fixed-line digital telecommunication terminals and VoIP terminals using the real-time transport protocol (RTP) over IP networks are capable of handling ITU-T (International Telecommunication Union, Telecommunication Standardization Sector) Recommendation G.711[*1] [1]. NTT Cyber Space Laboratories (referred to as NTT Labs. below) focused on the fact that G.711 is the most widely used codec. In 2005, it developed a scalable wideband speech coding method[*2] called UEM-CLIP (mu-law embedded coder for low-delay IP communication) [2], whose core codec is G.711. (The advantage of scalable coding is described in Section 3.) Since then, we have tried to promote wideband speech communication by implementing this codec in NTT's voice conference terminal MB-1000 and the high-quality IP telephones used in the NGN (Next Generation Network) field trial.

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

*1 G.711 is a speech coding method first standardized in ITU-T. The 8-kHz-sampled speech signal is encoded at 64 kbit/s using log-compressed pulse code modulation.

*2 Scalable speech coding makes the encoded bitstream structure layered so that output speech can be reproduced from even part of the bitstream.

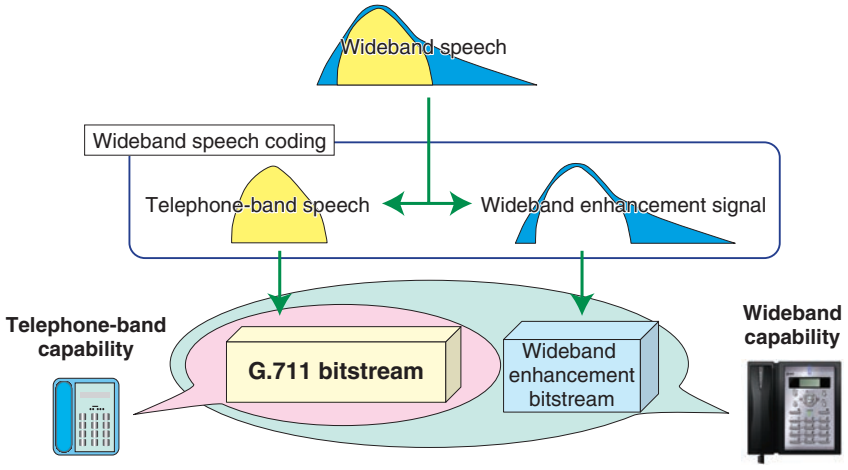

Fig. 1. Bitstream structure scalable with G.711.

2. Process up to the approval of G.711.1

To provide wideband speech communication in commercial NGN services, NTT Labs. proposed beginning standardization of wideband speech coding scalable with G.711 at the ITU-T meeting held in January 2008. It was recognized that such a codec would have advantages in situations where new wideband terminals and legacy G.711 terminals co-exist and the start of standardization was agreed. Afterward, NTT acted as a leader of the task and held two responsible positions: moderator in charge of coordination and editor in charge of drafting the recommendation. NTT and four other organizations—ETRI (Korea), France Telecom (France), Huawei Technologies (China), and VoiceAge (Canada)—jointly proposed a candidate algorithm, which combined UEM-CLIP and technologies of the other four organizations. Characterization tests, which were conducted in ITU-T using subjective listening, confirmed that this algorithm met all the requirements in terms of listening quality. As a result, at the ITU-T SG16 WP3[*3] meeting held in February 2008, consent was given for the candidate codec to progress to the alternative approval process (AAP)[*4], which is the final approval process for ITU-T standardization. This AAP was completed in March 2008. That is, the proposed candidate was approved as a new ITU-T standard, G.711.1.

*3 ITU-T SG16 WP3: Study Group 16 (SG16) is responsible for standardization related to multimedia terminals, systems, and applications. Working Party 3 (WP3) manages overall issues about media coding.

*4 AAP: The alternative approval process is applied for technical standards. The last call for comments is conducted online after consent has been granted at the ITU-T SG meeting. This can reduce the period until approval to about two months.

3. Concepts of G.711.1

3.1 Wideband speech capability

Since the frequency bandwidth of the telephone-band speech coded by G.711 is limited to the range from 300 Hz to 3.4 kHz, it has enough quality to handle conversations, but it loses the clearness and naturalness of human voices slightly. The first priority for the concepts required for G.711.1 was the ability to handle wideband speech (50 Hz to 7 kHz), which can transmit properties that are lost in the telephone band. It must also transmit both music and environmental sounds with a high level of listening quality.

3.2 Bitstream scalability with G.711

Several wideband speech coding standards have already been established, but their concepts do not include scalability with G.711. If the bitstream of the G.711 core layer and that of the enhancement layer are multiplexed for bandwidth expansion, wideband speech can be obtained using all parts of the bitstream, while telephone-band speech can be obtained from the G.711 part (**Fig. 1**). This bitstream structure brings two advantages, as described below.

3.2.1 Transcoding between G.711 and G.711.1

Until wideband speech terminals completely replace the telephone-band ones, both types of terminals will continue to coexist. The codec to be used

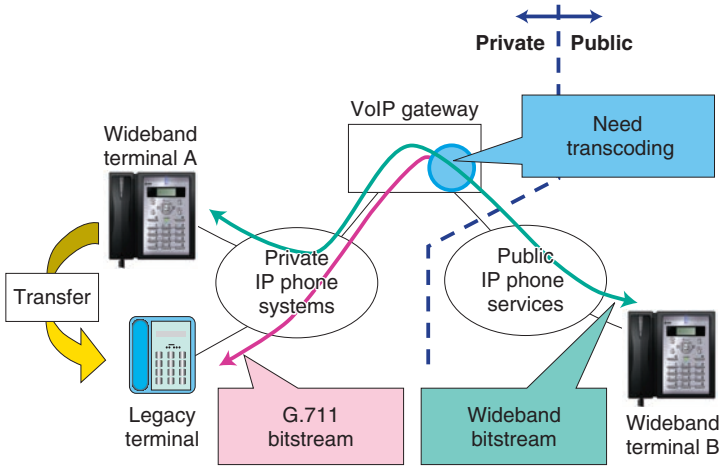

Fig. 2. Example of transcoding in call transference.

during a communication session is usually negotiated between the terminals when the call is set up. In the coexistence situation, in call transfers from wideband terminals to legacy terminals, transcoding will be needed after the negotiation has been established (**Fig. 2**). Such transcoding is generally performed in the following manner: decoding of the bitstream from the wideband terminal and re-encoding of the decoded speech signals into the G.711 bitstream for the legacy terminals. For transcoding between different types of bitstreams at a media gateway on a network, much more computation would be expected for the wideband case. Moreover, quality degradation caused by the transcoding cannot be ignored. However, the introduction of a bitstream structure scalable with the G.711 bitstream, as described here, enables transcoding to be completed just by extracting the G.711 bitstream. This needs hardly any additional computation and there will never be any quality degradation caused by re-encoding.

3.2.2 Wideband speech mixing

In multipoint conferencing, the mixing process must be performed at a signal mixer. In general, the signal mixer decodes all bitstreams from all locations, merges the decoded signals into one speech signal, re-encodes the merged signal into the bitstream, and then transmits it to all locations. However, to avoid echoes generated by sending back the signal received from each location, the mixed signal for each location must be prepared individually by removing that location's decoded signal from the full signal and then re-encoding it. That is, for multipoint-conferencing between N different points, it is necessary to complete all the following processes in the mixing server simultaneously: N decoding processes that produce N speech signals from N locations, N mixing processes that prepare N mixed signals for N locations, and N re-encoding processes that generate N bitstreams for N locations. Since the existing wideband speech codecs require much more computation than G.711 ones, they require much greater facilities in terms of both cost and scale to provide multipoint conferencing systems and services than the usual wideband codecs. This problem regarding the wideband speech mixing can be solved by introducing Partial Mixing [2], which was developed by NTT Labs. As shown in **Fig. 3**, partial mixing is performed in the following way. Since the amount of computation needed for decoding and re-encoding G.711 is much less, the G.711 core bitstreams are merged using the conventional mixing. The conventional mixing is not applied to the enhancement layer. The current location of the speaker is detected by analyzing the G.711 decoded signals and then the enhancement bitstream, received from the speaker site, is used for all locations. Each G.711 bitstream obtained by conventional mixing and the enhancement bitstream from the speaker's location are multiplexed and transmitted to every location. Partial mixing enables wideband speech mixing by just adding the slight computation required for detecting the speaker's location to that for conventional G.711 mixing.

3.3 Packet loss concealment

Packet loss concealment is necessary for the real-time voice communication provided on an IP network. On a best-effort network, packets might not

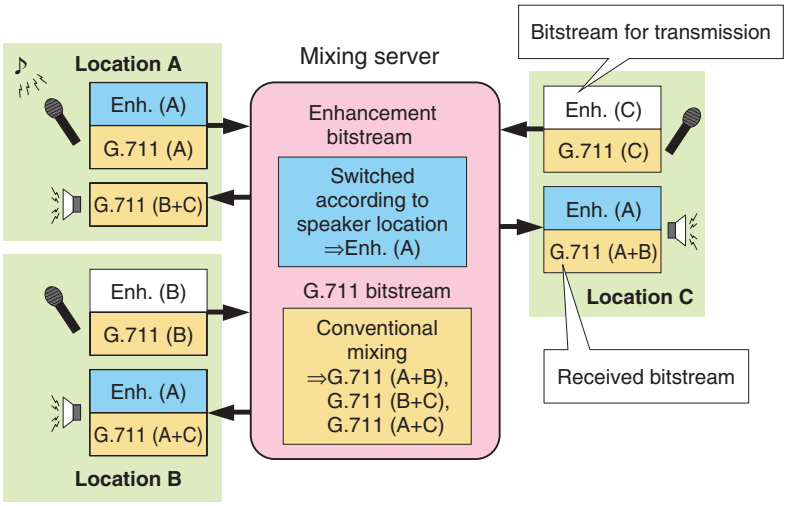

Fig. 3. Basic idea of partial mixing.

reach the receiver side with the proper timing. In that case, the speech would be interrupted. Conventional VoIP systems usually have jitter buffers to minimize the delay variation in packet arrival. However, making the jitter buffer longer would mean increasing the speech delay, so the buffer length cannot be lengthened without limit. G.711.1 has a mechanism for producing speech signals without interruption under conditions where several packet losses might occur successively.

3.4 Low delay and low computation

The lower the speech delay arising from the speech coding algorithm itself, the better the conversational quality. The frame length of G.711.1 has been confirmed to be 5 ms, so the speech packet length is set to the minimum of 5 ms or to a multiple of 5 ms. The algorithmic delay including the frame length is set to 11.875 ms.

Lower cost is also one of the most important factors for promoting the spread of wideband voice communication services. The algorithm of G.711.1 is designed to process speech signals with as little computation as possible so that it can be installed on inexpensive digital signal processors. The complexity of G.711.1, which is estimated using fixed-point simulation software, is 8.70 WMOPS (weighted million operations per second)[*5] in the worst case. This is comparable to that for ITU-T G.722, which is the first wideband speech coding method ever standardized in ITU-T.

*5 WMOPS: An abbreviation of weighted million operations per second. It indicates how many operations, simulating digital signal processor operations, are performed per second.

4. Technical descriptions of G.711.1

A high-level block diagram of G.711.1 is shown in **Fig. 4**. The input signal, which is sampled at 16 kHz, is processed frame-by-frame with a frame length of 5 ms. The input is split into lower-band and higher-band signals by an analysis quadrature mirror filter. The lower-band signal is encoded with an embedded lower-band pulse code modulation (PCM) encoder, which generates a G.711-compatible core bitstream at 64 kbit/s and a lower-band enhancement bitstream at 16 kbit/s. The higher-band signal is transformed into the modified discrete cosine transform (MDCT) domain and the frequency domain coefficients are encoded by the higher-band MDCT encoder, which generates a higher-band enhancement bitstream at 16 kbit/s. The lower-band enhancement bitstream improves the speech quality of the lower-band (50 Hz to 4 kHz) and the higher-band enhancement layer adds wideband capability (4 to 7 kHz). These three bitstreams are multiplexed and transmitted to the decoder. The bitstream received at the decoder is demultiplexed into three bitstreams. The G.711 bitstream and the lower-band enhancement bitstream are handed to the lower-band embedded PCM decoders. The higher-band enhancement bitstream is given to the higher-band MDCT decoder, and the decoded signal in the frequency domain is subsequently fed to an inverse MDCT (IMDCT), and the higher-band signal in the time domain is obtained. The lower- and

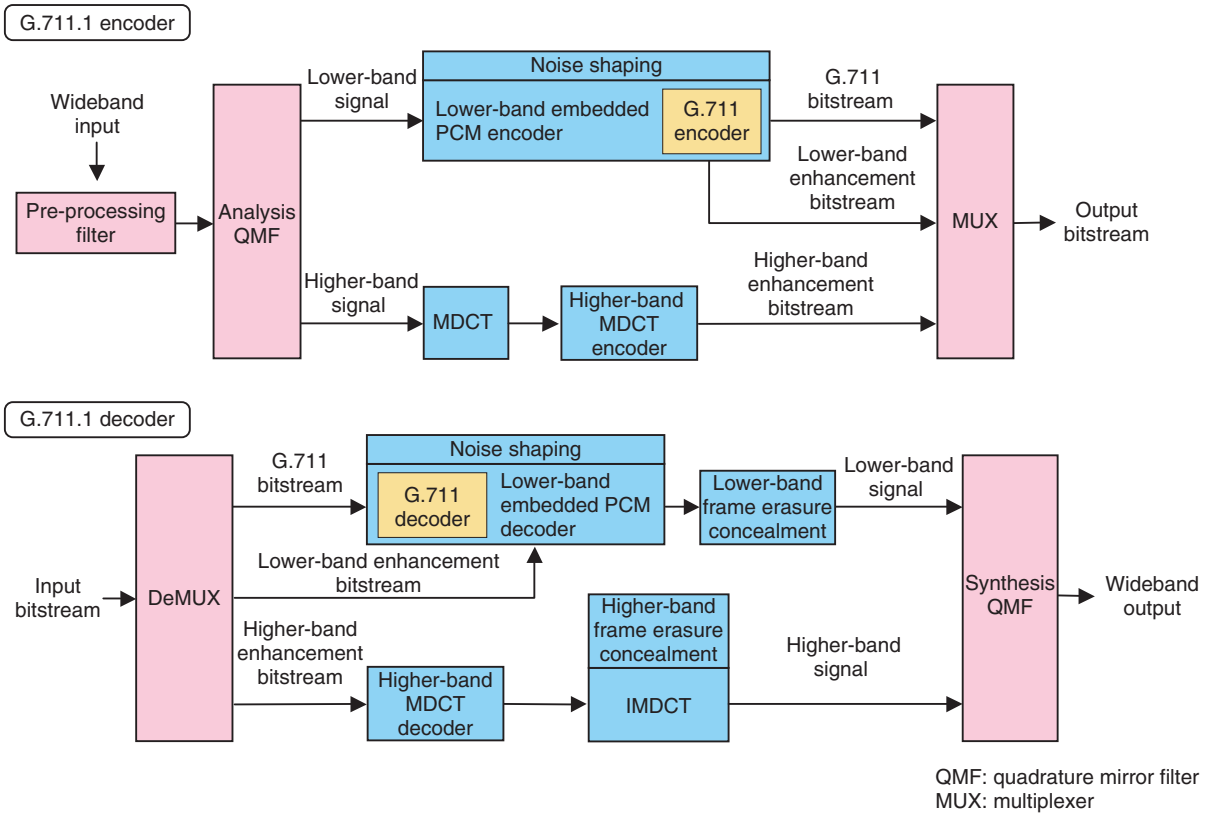

Fig. 4. High-level block diagram of G.711.1.

Table 1. Bitstream combination for each mode.

Mode	Sampling rate (kHz)	G.711 bitstream (64 kbit/s)	Lower-band enhancement bitstream (16 kbit/s)	Higher-band enhancement bitstream (16 kbit/s)	Bitrate (kbit/s)
R1	8	×	—	—	64
R2a	8	×	×	—	80
R2b	16	×	—	×	80
R3	16	×	×	×	96

higher-band signals are combined using a synthesis quadrature mirror filter to generate a wideband output signal. Thus, one of the concepts mentioned above, bitstream scalability with G.711, can be obtained. The four modes, which differ in sampling rate and bitrate, and respective combinations of the three bitstreams are given in **Table 1**. The structure described here is basically equal to that of UEMCLIP.

The main technologies introduced in G.711.1 are listed below.

- Noise shaping using linear prediction is applied to the lower-band embedded PCM encoder. It can perceptually suppress the G.711 quantization noise in encoding and also improves the listening quality of the speech decoded at legacy telephone terminals.
- The higher-band MDCT encoder efficiently compresses the input MDCT coefficients at 16 kbit/s using interleaved conjugate-structured vector quantization (compression rate of the higher-band MDCT encoder: 1/8). The preselection method and table optimization can greatly decrease the amount of computation required for vector quantization.

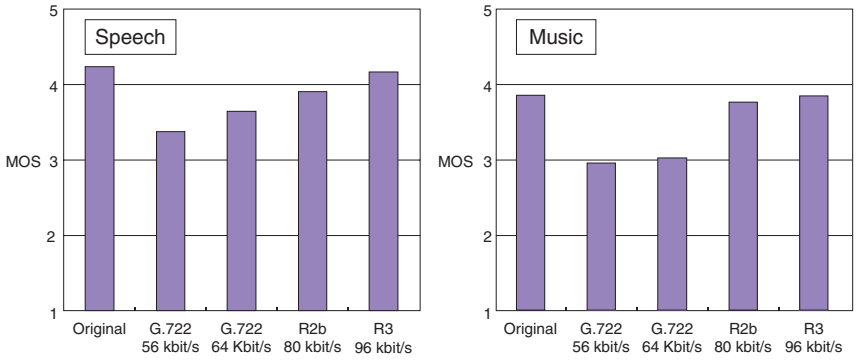

Fig. 5. Summary of G.711.1 subjective test results.

- If packet loss is detected at the decoder, the missing speech frame is reconstructed in the following way. The previously decoded signal is analyzed to estimate parameters such as signal class (e.g., voiced or unvoiced) and pitch and a replacement for the lost frame is synthesized using those parameters. Even if packet losses occur successively, the quality degradation caused by the lost frame is hardly perceived.

5. Quality evaluation

Before approving a proposed speech codec as a Recommendation, ITU-T usually conducts subjective listening tests to check whether it meets the requirements. The subjective tests of G.711.1 were performed for speech, music, speech with background noise, and mixed speech. In each condition, the speech quality was defined as the requirement. A summary of the test results obtained by NTT Labs. in conformance with the same procedure as that used in the ITU-T testing is given in **Fig. 5**. The estimated mean opinion scores (MOSs) in mode R2b at 80 kbit/s and mode R3 at 96 kbit/s were higher than those of G.722 at 64 kbit/s for speech and music.

6. Future deployment

To provide high-quality speech communication services using G.711.1 to our customers, NTT Labs. is cooperating with NTT operating companies and implementing G.711.1 in VoIP terminals, voice conferencing terminals, and software telephones on personal computers (PC softphones). In parallel with G.711.1 implementation, we will propose an RTP payload format for G.711.1 in IETF (Internet Engineering Task Force), so that it is available on the Internet and launch a framework for one-stop licensing of the patents, included those for G.711.1.

ITU-T has just started the standardization of speech coding with bandwidth capability of more than 7 kHz and stereo functionality while retaining scalability with G.711. NTT Labs. will develop speech coding technologies applicable to this standardization and contribute to it.

References

[1] ITU-T, Geneva, Switzerland, "ITU-T G.711—Pulse code modulation (PCM) of voice frequencies," Nov. 1988.

[2] Y. Hiwasaki, H. Ohmuro, T. Mori, S. Kurihara, and A. Kataoka, "A G.711 Embedded Wideband Speech Coding for VoIP Conferences," IEICE Trans. Inf. & Syst., Vol. E89-D, No. 9, pp. 2542–2551, Sept. 2006.

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, 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).

Takeshi Mori
Research Engineer, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.

He received the B.E. and M.E. degrees in computer science from Tokyo Institute of Technology, Tokyo, and the Ph.D. degree in engineering from Tsukuba University, Ibaraki, in 1994, 1996, and 2007, respectively. Since joining NTT in 1996, he has been engaged in research on speech and audio coding and the development of VoIP applications. He is a member of IEEE, IEICE, and ASJ.

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, 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.

Hitoshi Ohmuro
Senior Research Engineer, Supervisor, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.

He received the B.E. and M.E. degrees in electrical engineering from Nagoya University, Aichi, in 1988 and 1990, respectively. He joined NTT in 1990. He has been engaged in research on highly efficient speech coding and the development of VoIP applications. He is a member of IEEE and ASJ.