ITU-T G.711.1: Extending G.711 to Higher-Quality Wideband Speech

Yusuke Hiwasaki and Hitoshi Ohmuro, NTT Corporation

ABSTRACT

In March 2008 the ITU-T approved a new wideband speech codec called ITU-T G.711.1. This Recommendation extends G.711, the most widely deployed speech codec, to 7 kHz audio bandwidth and is optimized for voice over IP applications. The most important feature of this codec is that the G.711.1 bitstream can be transcoded into a G.711 bitstream by simple truncation. G.711.1 operates at 64, 80, and 96 kb/s, and is designed to achieve very short delay and low complexity. ITU-T evaluation results show that the codec fulfils all the requirements defined in the terms of reference. This article presents the codec requirements and design constraints, describes how standardization was conducted, and reports on the codec performance and its initial deployment.

INTRODUCTION

In the early years of voice communications, transmission bandwidth was somewhat limited, and the main technological focus at the time was to transmit voice at the best achievable quality given the bandwidth constraint. This led to using speech limited to a frequency range of 300 Hz to 3.4 kHz, today called *narrowband* speech. However, with the exponential growth of transmission bandwidth for both wired and wireless communications, broadband connections are now more widely available. For fixed lines, the trend is to transport all information and services, including voice, video, and other data, in packet-based networks. One advantage of a packet network such as an Internet Protocol (IP)-based one is that it can adapt to various bit rates, and this means that we are now free from having to use constant bit rates and band-limited audio. This means that the new generation of terminals can support richer services and functionalities. Consequently, speech coding algorithms can be designed with emphasis on factors such as low delay, low complexity, and, in particular, wider audio frequency bandwidth. We have seen the emergence of coders, such as International Telecommunication Union — Telecommunication Standardization Sector (ITU-T) G.722, G.722.1, AMR-WB (also known as ITU-T G.722.2), G.729.1, and G.718.[1] Those coders are for encoding conversational speech signals in the frequency range of 50 Hz to 7 kHz, called wideband speech, which is equivalent to audio signals conveyed in AM radio broadcasts. One of the most popular applications of voice over IP (VoIP) is remote audio-visual conferences, where hands-free terminals are often used. In that case, intelligibility becomes more important than when using handsets because participants usually sit around a terminal at a certain distance from a loudspeaker. This is where wideband speech coders, which can reproduce speech at high fidelity and intelligibility, are particularly favored.

Today, the majority of fixed-line digital telecommunications terminals are equipped with ITU-T G.711 (log-compressed pulse code modulation [PCM]) capability. In fact, for communication using Real-Time Transport Protocol (RTP) over IP networks, G.711 support is mandatory. Until wideband speech terminals completely replace narrowband ones, these two types of terminals will continue to coexist, meaning that the wideband ones must be capable of interoperating with those that carry only G.711. In an ordinary telecommunications scenario, the codec used during a session is negotiated between the terminals as the call is set up. However, there are some cases where this may not be possible, for example, in call transfers and multipoint conferencing. Therefore, transcoding[2] between different types of bitstream must be performed by a bitstream translator at a gateway or a signal mixer at a multi-point control unit (MCU). This is problematic when those devices must accommodate a large number of lines because transcoding usually requires a high computational complexity and will likely introduce quality degradations and additional delay. This can be an obstacle in the increased use of wideband voice-communication services.

To overcome this obstacle, ITU-T has standardized a new speech-coding algorithm, G.711.1 [1]. This codec, approved in March 2008, is an extension of ITU-T G.711. It was initially studied under the name G.711-WB (wideband extension). The coding algorithm is designed to provide a low-delay, low-complexity, and high-quality wideband speech addressing transcoding problems with legacy (narrowband) terminals with a bitrate trade-off. The main feature of this extension is to give G.711 wideband scalability.[3] It aims to achieve high-quality speech services over broadband networks, particularly for IP phone and

[1] *These ITU-T Recommendations can be freely downloaded: http://itu.int/rec/T-REC-G/*

[2] *Transcoding is the conversion processing of one encoding format to another.*

[3] *Scalability enables the best quality of service to be provided as the system load varies.*

0163-6804/09/$25.00 © 2009 IEEE

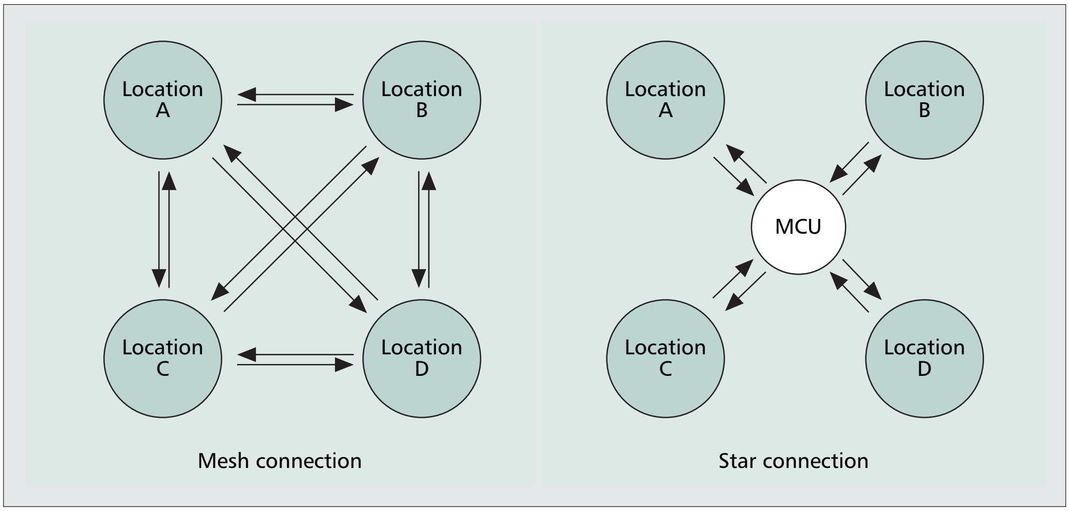

Figure 1. *Configurations for connecting remote conferences.*

In the partial mixing method, only the core bitstream (usually the lower band) is decoded and mixed, and the enhancement layers are not decoded, hence the name partial. Instead, one active endpoint is selected from all the endpoints, and its enhancement layers are redistributed to other endpoints.

multipoint speech conferencing, while enabling seamless interoperability with conventional terminals and systems equipped only with G.711.

In the next section one of the key applications of G.711.1, partial mixing, is described, and then the design constraints of G.711.1 are detailed in the section after. The succeeding section describes how the standardization progressed, and then a brief overview of the codec algorithm is presented. The characterized results of G.711.1, the speech quality of the codec, are presented next. Finally, new extensions to G.711.1 and the status of the codec deployment are discussed.

Partial Mixing

One of the primary applications of G.711.1 is audio conferencing, and *partial mixing* [2] is a solution to contain its growing complexity. When considering such conferences, there are two possible configurations, as shown in Fig. 1: one is a mesh connection where all endpoints are connected to all others, and the other is a star connection centered on a multipoint control unit (MCU). Mesh connections, such as those used in Skype, do not require a server but are restricted to conferences with a few endpoints. In large-scale n-point conferences, each endpoint needs to transmit and receive $2(n – 1)$ media bitstreams (counting both inbound and outbound streams). This is undesirable because $n – 1$ decoding computation would operate simultaneously, whereas only one decoder is required for star connections. In addition, each endpoint needs to have a sufficiently large transmission bandwidth, whereas in star connections only two (inbound and outbound) media bitstreams are required. Thus, mesh connections are suitable for conferences involving a few endpoints. However, in star connections the computation now occurs in the MCU, where mixing must be performed.[4] Here, the MCU has to decode all the bitstreams from endpoints, mix the obtained signals, and then re-encode the mixed signal. It should be noted that usually when calculating a mixed signal for an endpoint, the MCU would have to subtract the signal originating from that point. This means that n encoders and n decoders must operate at the MCU, in addition to the summation of the signals and the transmitting and receiving of $2n$ media streams. The number of mixed endpoints n might be limited to m ($m < n$, e.g., 3), selected on the active channels, but this would still require considerable computational complexity. Another issue contributing significantly to complexity is transcoding. Terminals with different coding capabilities require transcoding because various coding algorithms are used in VoIP systems. Usually, this capability is implemented by decoding a bitstream to a linear PCM signal and then re-encoding that with another encoder. For interconnection between wideband and narrowband coders, transcoding requires another intermediate step, down-/up-sampling, and this would further increase the required computational complexity at MCUs. Another less significant downside of transcoding is the accumulation of algorithmic delays caused by re-encoding and maybe down-/up-sampling, and this can be problematic when using codecs working on a certain frame length.

These problems can be overcome by taking advantage of a subband scalable bitstream structure because a signal can be reconstructed by decoding only part of the bitstream. In the partial mixing method, only the core bitstream (usually the lower band) is decoded and mixed, and the enhancement layers are not decoded, hence the name *partial*. Instead, one active endpoint is selected from all the endpoints, and its enhancement layers are redistributed to other endpoints. To implement this hybrid approach, which combines redistribution and mixing, the mixer must judge which endpoint to select by detecting voice activities and/or detecting the endpoint with the largest signal power.

Figure 2 illustrates how a partial mixer works. There are three locations connected to an MCU; assume that location A is talking, and locations B and C are listening in this instance. The partial mixer performs conventional core (G.711) layer mixing, but for enhancement layers, it detects the speaker location (A) and selects a set of enhancement layers for retransmission. This means that for location B, the received core layer is a mixture of locations A and C (Core[A + C]), but for enhancement layers, only A is received (Enh[A]), and for location C, the core layer is a mixture of

[4] Note that there is another type of MCU, a forwarding bridge, that also sends media bitstreams in a star configuration, but multiplexes media bitstreams. To endpoints this is similar to a mesh configuration, and the computational issue is still present.

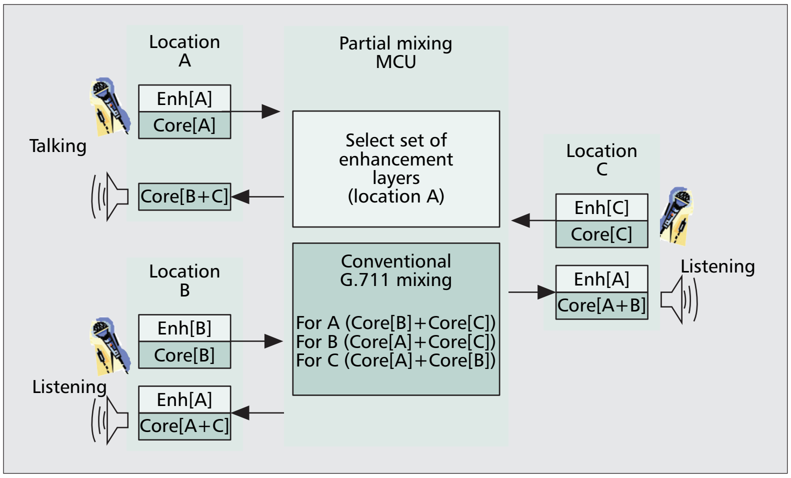

Figure 2. *Partial mixing.*

Mode	Layer 0	Layer 1	Layer 2	Bit rate (kb/s)
R1	X	—	—	64
R2a	X	X	—	80
R2b	X	—	X	80
R3	X	X	X	96

Table 1. *Sub-bitstream combination for each mode.*

locations A and B (Core[A + B]), with enhancement layers again from location A (Enh[A]). For location A, the received core layer is a mixture of locations B and C (Core[B + C]).

This method can considerably reduce the mixing complexity required for wideband codecs, and this advantage is more significant when used on a core coding scheme that requires very low complexity: G.711. Because the core layer is continuously mixed, there will be no disruptions in the reproduced speech due to switching of the bitstreams, only to bandwidth changes. In a conferencing scenario this is a good compromise because there is usually only one talker at a time. However, in the design of the codec, the quality degradation by the switching effect of the enhancement layers must be kept as low as possible.

Requirements and Design Constraints

As described in the introductory section, we have seen the emergence of numerous wideband speech coders. However, all of them, except G.729.1, lack one feature: interoperability with narrowband codecs. This is crucial in a scenario when transcoding is required, and this commonly occurs when speech is conveyed across different networks. It should also be noted that in modern wideband codecs, there is a compromise between computational complexity and encoding delay to achieve high quality at lower bit rates. However, we have seen increased use of broadband, such as fiber-to-the-x (FTTx) and high-speed wireless connections. This means that bit rate requirements can now be relaxed, and other design aspects such as quality, complexity, and encoding delay should be emphasized. The G.711 bit rate by default is 64 kb/s; compared to that of other media such as video, this is still a small value; hence, allocating a few more bits to enhance its quality would be feasible. This means that the design concept of G.711.1 was considerably different from previous ones.

At the start of the G.711.1 study, the design constraints of the coder were set as follows:

- The coder must be compatible with G.711 using an embedded approach.
- There are two enhancement layers:a lower-band enhancement layer to reduce G.711 quantization noise and a higher-band enhancement layer to add wideband capability.
- There must be a short frame length (sub-multiples of 5 ms) to achieve low delay.
- Low computational complexity and memory requirements must be used to fit existing hardware capabilities (also enabling energy-efficient designs).
- For speech signal mixing in multipoint conferences, complexity similar to G.711 must be achieved (i.e., no increase in complexity). It is preferable not to use interframe predictions to enable enhancement layer switching in MCUs for pseudo-wideband mixing and partial mixing, as detailed in the previous section.
- For robustness against packet losses, it is preferable to depend not too heavily on interframe predictions.
- All quality requirements are intended to be better than G.722 in wideband audio.

With three sub-bitstreams constructed from the core (layer 0 at 64 kb/s) and two enhancement layers (layers 1 and 2, both at 16 kb/s), four bitstream combinations can be constructed that correspond to four modes: R1, R2a, R2b, and R3. The first two modes operate at an 8 kHz sampling frequency (i.e., narrowband) and the last two at 16 kHz (i.e., wideband). Table 1 gives all modes and respective sub-bitstream combinations.

Standardization Process

The standardization of ITU-T G.711.1 was conducted under Question 10 of ITU-T Study Group 16, "Multimedia Terminals, Systems, and Applications" (Q.10/16). The work was launched in January 2007 with a proposal from Nippon Telegraph and Telephone (NTT, Japan) specifying the targeted market and its design constraints, as described in the previous sections: standardizing low-complexity, low-delay, 7 kHz bandwidth speech, with an embedded scalable structure on top of G.711. It was recognized that such a codec would have advantages, especially in the transitional stage between narrowband and wideband communications.

The terms of reference (ToR) and time schedule were finalized and approved in March and June 2007, respectively. All quality evaluation tests plans were formulated by Question 7 of ITU-T Study Group 12 (Q.7/12), "Performance and Quality of Service," and conducted accordingly, and the results were scrutinized by those

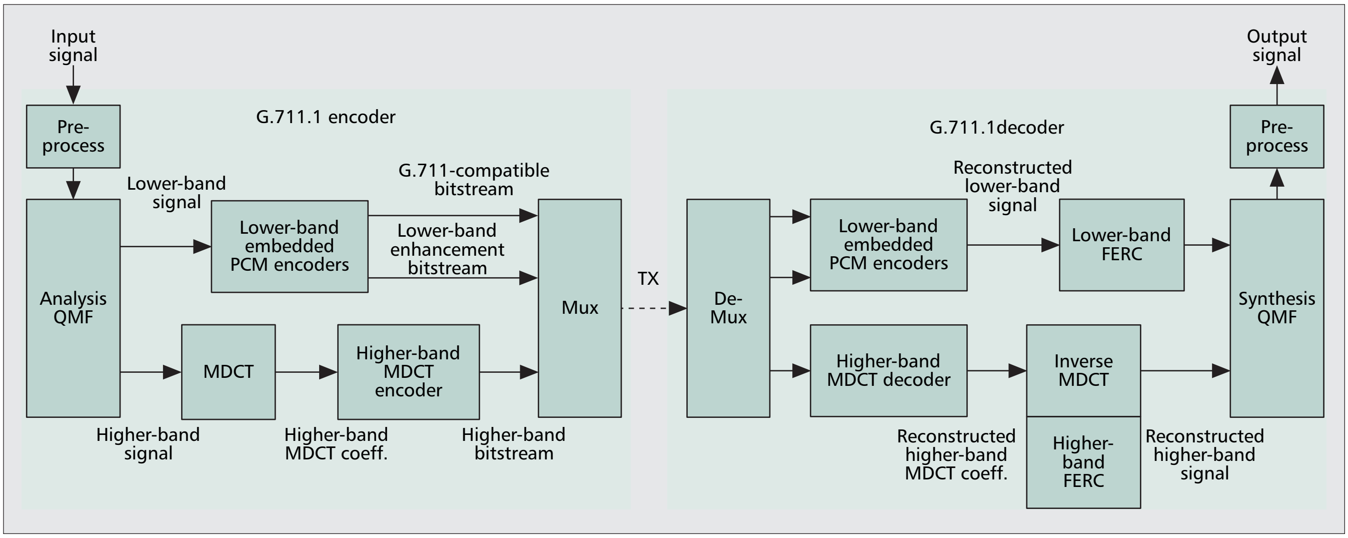

Figure 3. *High-level block diagram of G.711.1.*

experts. ETRI (Korea), France Télécom (France), Huawei Technologies (China), VoiceAge (Canada), and NTT participated in the qualification phase, which took place mid-2007. This was followed by the optimization and characterization phase, where all five organizations constructively collaborated to create a unified algorithm [3] implemented in fixed-point arithmetic. This means that the selection phase was skipped, and the candidate codec algorithm went straight to the characterization phase. At the February 2008 meeting of ITU-T Study Group 16 Working Party 3 (WP3/16), it was confirmed that the codec passed all the requirements set in the ToR. Finalized specifications, including text and ANSI-C source code utilizing basic operators [4] that simulates digital signal processing (DSP) instructions, entered the approval process. In March 2008, only 14 months after the launch of the standardization process, ITU-T Recommendation G.711.1 was formally approved.

Overview of Codec Algorithm

As an outcome of the standardization, the G.711.1 algorithm specification was defined as follows. The codec operates on 16 kHz sampled speech at a 5 ms frame length. A high-level block diagram of the codec is shown in Fig. 3. At the encoder, the input signal is preprocessed with a high-pass filter to remove low-frequency (0–50 Hz) components and then split into lower-band and higher-band signals using a 32-tap analysis quadrature mirror filterbank (QMF). The lower-band signal is encoded with an embedded lower-band PCM encoder that generates a G.711-compatible core bitstream (layer 0) at 64 kb/s and a lower-band enhancement (layer 1) bitstream at 16 kb/s. Note that the core codec is based on the legacy ITU-T G.711 standard, and that both μ-law and A-law schemes are supported. However, to achieve the best perceivable quality, the quantization noise of layer 0 (the G.711-compatible core) is shaped with a perceptual filter [5]. The higher-band signal is transformed into the frequency domain using modified discrete cosine transform (MDCT), and the MDCT coefficients are encoded by the higher-band encoder, which generates a higher-band enhancement (layer 2) bitstream at 16 kb/s. The transform length of MDCT in the higher band is 10 ms with a shift length of 5 ms. All bitstreams are multiplexed as a scalable bitstream. At the decoder, the whole bitstream is demultiplexed. Both layer 0 and 1 bitstreams are passed to the lower-band embedded PCM decoder. The layer 2 bitstream is fed to the higher-band MDCT decoder; the decoded signal is processed by inverse MDCT to generate the higher-band signal in the time domain. To improve the quality when speech frames do not arrive at the decoder due to packet losses, frame erasure concealment (FERC) algorithms are applied to the lower- and higher-band signals separately. The decoded lower- and higher-band signals are combined using a synthesis quadrature mirror filter (QMF) to generate a wideband signal. Noise gate processing is applied to the QMF output to reduce low-level background noise. As the decoder output, 16 or 8 kHz sampled speech is reproduced.

The codec has a very simple structure to achieve high-quality speech with low complexity and is deliberately designed without any interframe predictions. This is to increase the robustness against frame erasures and avoid annoying artifacts when enhancement layers are switched, which is required for the partial mixing in wideband MCU operations. This also contributes to reducing MCU complexity.

An optional post-filter designed to reduce the lower-band quantization noise at the decoder is also available as an Appendix to G.711.1. It enhances the quality of a 64 kb/s bitstream when communicating with a legacy G.711 encoder. More details of the algorithm can be found in [1, 3, 5].

Characterization of G.711.1

Complexity and Delay — The complexity of the codec, which is estimated using the basic operator set in the ITU-T Software Tool Library v. 2.2, was found to be 8.70 weighted million operations per second(WMOPS) in the worst case. This meets the ToR objective (“less than 10 WMOPS”), and compared to another wideband extension of a narrowband codec, G.729.1

In general, ITU-T standardizes speech codecs by first using fixed-point instructions, suitable for DSP platforms and to guarantee audio quality. However, it would also be beneficial to have an implementation on floating-point arithmetic because almost all CPUs for PCs have built-in floating point co-processors.

(35.8 WMOPS), this value is considerably lower. The memory size of the codec is 3.04 kWords RAM, 2.21 kWords table ROM, and around 1940 operators for program ROM. All figures also met the memory requirements in the ToR, meaning that the codec leaves only a small footprint in any implementation, which is advantageous for smooth switching from a narrowband to a wideband environment.

The tally of analysis and synthesis delays of the split-band QMF is 1.875 ms, and the delay due to the MDCT analysis for layer 2 is 5 ms. The overall algorithmic delay adds up to 11.875 ms (190 samples at 16 kHz), including the frame length (5 ms). Due to its short delay, the codec is suitable for any conversational applications.

Speech Quality — Before the acceptance of the codec, a set of formal subjective assessments was performed to fully characterize the G.711.1 performance. Several experiments were run, twice each in a different language using 32 naïve and native listeners. The languages tested were chosen from among North American English (NA English), Chinese, French, Japanese, and Korean. Four kinds of input signals were considered: clean speech, music, noisy speech with four types of background noise at various signal-to-noise ratios (SNRs): background music at 25 dB SNR; office noise at 20 dB SNR; babble noise at 30 dB; and interfering talker at 15 dB SNR, as well as mixed speech for MCU operation. Both µ-law and A-law were tested. For all conditions, the absolute category rating (ACR) was used, except for the conditions with background noise, which were tested with the degradation category rating (DCR). These methods are described in the P.800-series Recommendations. It should be noted that for testing mixed speech conditions, partial mixing was used for G.711.1 and was tested against full conventional mixing of the reference coder.

Table 2 gives a subset of the mean opinion scores (MOS) of the tested conditions limited to –26 dBov input signals. In this table, *CuT mode* designates G.711.1 (the coder under test [CuT]), *Reference* means the reference condition of the requirement/objective, Y_{CuT} and Y_{Ref} are the MOS of the coder and reference coders, respectively, and *Req*. and *Obj*. in the R/O column indicate a requirement or an objective, respectively. *FER* stands for bitstream frame erasure rate. The judgments were made based on the statistical comparison between the MOS of the codec candidate and the reference codecs, by means of a simple paired *t*-test at a 95 percent significance level. The codec met all requirements and all objectives, except in the objective condition of R3 high-level input (–16 dBov) in French, which is not included in Table 2. The shaded rows are the results for partial mixing, and these results show that partial mixing can provide reasonable quality better than the defined reference conditions.

Alongside the G.711.1 characterization, ITU-T Study Group 12 approved a revised version of G.113 Appendix IV [6] in April 2009, which defines the codec impairment factors for calculating the E-model value for VoIP scenarios. The E-model is a model that predicts the subjective effect of combinations of impairments to help network planners design networks, and the codec is a factor in this model that also takes packet losses into account. After extensive formal listening tests, separate from the one performed during G.711.1 characterization, the impairment factor for G.711.1 was found to be 1 for R2b and 0 for R3 mode, meaning that the G.711.1 R3 mode is literally transparent in this context.

Standardized Extensions and Functionalities

This section summarizes the developments after the G.711.1 main body and Appendix I were approved.

In general, ITU-T standardizes speech codecs by first using fixed-point instructions, suitable for DSP platforms and to guarantee audio quality. However, it would also be beneficial to have an implementation on floating-point arithmetic because almost all CPUs for PCs have built-in floating-point co-processors. This would be the best match for applications such as software phones. The standardization of such implementation was launched in January 2008 and was approved as G.711.1 Annex A in November 2008. This provided an opportunity for the coder to be implemented on a broader range of platforms. This alternative implementation is fully interoperable with the fixed-point main body.

To enable the transport of a G.711.1 bitstream over RTP, a Request dor Comments (RFC) specifying the payload format of the G.711.1 bitstream together with usage of the session description protocol (SDP) was prepared at the Internet Engineering Task Force (IETF). This became RFC 5391 [7] in November 2008. On this basis, the Internet Assigned Numbers Authority (IANA) registered the media types PCMU-WB and PCMA-WB for µ-law and A-law, respectively.

Once the transport over IP was defined, the next step in standardization was to facilitate G.711.1 use for H.323 "Packet-Based Multimedia Communications Systems" commonly used in VoIP applications. To do so, the generic capability of G.711.1 was defined in Annex B. This Annex is referenced from H.245, "Control Protocol for Multimedia Communication" Appendix VIII, which defines the call control of H.323. With this, G.711.1 can be made available in H.323-based audio-visual conferencing systems.

Of course, there are needs for communications using an even wider frequency range. Many important applications will benefit from an audio bandwidth wider than 7 kHz. ITU-T SG 16 is now studying a work item that extends G.711.1 to superwideband. Superwideband is a frequency range that reaches 14 kHz, which is equivalent to the audio bandwidth conveyed in FM radio broadcasting. This extension is being studied in parallel to the G.722 superwideband extension, and it is very likely that those extensions will be common to both core codecs. This work item was launched in October 2008, and the experts finalized the ToR in January 2009. This common extension to G.722 and G.711.1 is expected to be completed in the second quarter of 2010.

Table 3 is a summary of existing G.711.1-related standards.

CuT mode*	Reference	Exp	Condition	Language lab A	Y_{CuT} lab A	Y_{Ref} lab A	Language lab B	Y_{CuT} lab B	Y_{Ref} lab B	R/O
R1	G.711 A-law	Exp1a	Clean Speech	Korean	4.41	3.16	NA English	4.05	2.91	Req.
			3% Random FER	Korean	4.26	3.07	NA English	3.92	2.80	Req.
		Exp2a	Music		3.86	3.77		3.47	3.30	Req.
		Exp3	Background music	Japanese	4.77	4.56	Korean	4.58	4.35	Req.
			Office noise	Japanese	4.82	4.77	Korean	4.68	4.68	Req.
			Babble noise	Japanese	4.74	4.68	Korean	4.61	4.48	Req.
			Interfering talker	Japanese	4.64	4.52	Korean	4.62	4.43	Req.
R2a	16 bit PCM	Exp1a	Clean Speech	Korean	4.45	4.11	NA English	4.40	4.38	Obj.
	G.711 A-law		3% Random FER	Korean	4.35	3.07	NA English	4.23	2.80	Req.
	16 bit PCM	Exp2a	Music		3.85	3.90		3.46	3.43	Obj.
		Exp3	Background music	Japanese	4.80	4.82	Korean	4.80	4.81	Obj.
			Office noise	Japanese	4.83	4.83	Korean	4.73	4.73	Obj.
			Babble noise	Japanese	4.76	4.78	Korean	4.80	4.80	Obj.
			Interfering talker	Korean	4.72	4.73	NA English	4.82	4.85	Obj.
	G.726	**Exp5a**	**Mixed speech**	**Korean**	**4.45**	**2.96**	**NA English**	**4.12**	**2.44**	**Req.**
R2b	G.722 56 k	Exp1b	Clean Speech	French	4.10	3.70	Chinese	4.03	3.36	Req.
			3% Random FER**	French	4.08	3.17	Chinese	3.88	2.52	Req.
		Exp2b	Music		4.06	3.45		3.66	2.99	Req.
		Exp4	Background music	French	4.53	4.25	NA English	4.38	3.73	Req.
			Office noise	French	4.64	4.46	NA English	4.48	3.85	Req.
			Babble noise	French	4.71	4.41	NA English	4.56	3.79	Req.
			Interfering talker	French	4.61	4.48	NA English	4.68	3.89	Req.
	G.722 48 k	**Exp5b**	**Mixed speech**	**French**	**4.09**	**3.09**	**Chinese**	**4.02**	**2.66**	**Obj.**
R3	G.722 64 k	Exp1b	Clean Speech	French	4.41	3.73	Chinese	4.23	3.31	Req.
			3% Random FER**	French	4.31	3.20	Chinese	4.06	2.59	Req.
		Exp2b	Music		3.91	3.56		3.75	3.07	Req.
		Exp4	Background music	French	4.71	4.42	NA English	4.61	3.77	Req.
			Office noise	French	4.68	4.51	NA English	4.51	3.98	Req.
			Babble noise	French	4.79	4.52	NA English	4.62	3.82	Req.
			Interfering talker	French	4.78	4.56	NA English	4.69	4.01	Req.
	G.722 48 k	**Exp5b**	**Mixed speech**	**French**	**4.28**	**3.09**	**Chinese**	**4.23**	**2.66**	**Req.**

*CuT core in all conditions was G.711 A-law. **Random FER for reference G.722 was set to 1%.

Table 2. *Characterization test results.*

For transmission planning purposes, the codec is considered to add no speech quality degradation. This codec has already been deployed in the Japanese market, and subsequently, more products are expected to roll out.

	Approved	Description
G.711.1	Mar. 2008	G.711.1 main body with simulation ANSI-C source code in fixed-point arithmetic
G.711.1 Annex A	June 2008	Reference floating-point implementation of G.711.1, in ANSI-C
G.711.1 Annex B	Mar. 2009	Definition of H.245 usage, enabling implementation on H.323 systems
G.711.1 Appendix I	Mar. 2008	Post-filter for decoding bitstream by legacy G.711 encoders
RFC 5391	Nov. 2008	RTP payload type definition

Table 3. *G.711.1-related standards (as of July 2009).*

Deployment of the Codec

Due to the rapid completion of standardization processes at ITU-T and IETF, the codec was put on the market at a very fast rate. As anticipated, the initial market in which G.711.1 was introduced was Japan.

After one year of field trials, NTT East and West launched a multimedia converged network called *FLET'S Hikari Next*. This is a managed IP-based network with quality of service (QoS) mechanisms, which guarantees a certain level of performance for data flows. In early 2009 G.711.1 was put into service as its main wideband codec, and the first G.711.1-aware wideband handset was rolled out in February 2009.

Numerous other products are under preparation with G.711.1 capabilities, and a number of DSP implementers have already indicated the availability of the codec on various DSP platforms. For example, Israeli-based handset manufacturer AudioCodes has announced that a new wideband handset product line will soon roll out, capable of communicating using G.711.1. OKI Electric Industry of Japan recently announced a a carrier-grade large-scale wideband-speech transcoding unit that can handle G.711.1. It would be worth mentioning that the licensing of the codec can be obtained via a patent pool.

Conclusion

A new wideband speech codec, G.711.1, was standardized at ITU-T, with a novel design constraint different from conventional wideband codecs. The standardization was driven by a market need to move from narrowband telephony to a wider frequency bandwidth. This codec aims to achieve high-quality speech communication with an upward compatibility to G.711 while keeping complexity and delay as low as possible, but relaxes constraints on bit rate as a trade-off. As a result, the complexity and overall encoding/decoding delay were kept to very small values: 8.7 WMOPS and 11.285 ms, respectively. Extensive subjective evaluation tests showed that the codec performs better than the reference codec G.722 and achieves transparent quality. For transmission planning purposes, the codec is considered to add no speech quality degradation. This codec has already been deployed in the Japanese market, and subsequently more products are expected to roll out.

Acknowledgments

The authors would like to thank the experts of ITU-T Q.10/16 and Q.7/12, and the Rapporteurs of those two questions, Ms. Claude Lamblin, Ms. Catherine Quinquis, and Mr. Paolo Usai. Thanks also go to numerous colleagues from ETRI, Huawei Technologies, France Télécom, NTT, and VoiceAge: Do Young Kim, Jongmo Seong, Mi Suk Lee, Hyun Woo Kim, Balázs Kövesi, Stéphane Ragot, Jean-Luc Garcia, Claude Marro, Alexandre Guérin, Lei Miao, Jianfeng Xu, Shinji Hayashi, Akira Takahashi, Shigeaki Sasaki, Takeshi Mori, Vladimir Malenovsky, Jimmy Lapierre, Roch Lefebvre, and Redwan Salami, who have devoted considerable time to working collaboratively to finalize the standard. Finally, yet important, we are very grateful to Simão Campos, ITU-T SG16 counselor.

References

[1] ITU-T Rec. G.711.1, "Wideband Embedded Extension for G.711 Pulse Code Modulation," Mar. 2008.
[2] Y. Hiwasaki *et al.*, "A G.711 Embedded Wideband Speech Coding for VoIP Conferences," *IEICE Trans. Info. & Sys.*, vol. E89-D, no. 9, Sept. 2006, pp. 2542–51.
[3] Y. Hiwasaki *et al.*, "G.711.1: A Wideband Extension to ITU-T G.711," *Proc. EUSIPCO '08*, Lausanne, Aug. 2008.
[4] ITU-T Rec. G.191 STL-2005 Manual, "ITU-T Software Tool Library 2005 User's Manual," 2005.
[5] J. Lapierre *et al.*, "Noise Shaping in an ITU-T G.711-Interoperable Embedded Codec," *Proc EUSIPCO '08*, Lausanne, Switzerland, Aug. 2008.
[6] ITU-T Rec. G.113 Amendment 1, "Transmission Impairments Due to Speech Processing: Revised Appendix IV — Provisional Planning Values for the Wideband Equipment Impairment Factor and the Wideband Packet Loss Robustness Factor," Mar. 2009.
[7] RFC 5391, "RTP Payload Format for ITU-T Recommendation G.711.1," Nov. 2008.

Biographies

Yusuke Hiwasaki [M'96] (hiwasaki.yusuke@lab.ntt.co.jp) received his B.E., M.E., and Ph.D. degrees from Keio University, Yokohama, Japan, in 1993, 1995, and 2006, respectively. Since joining NTT Human Interface Laboratories (now Cyber Space Laboratories) in 1995, he has been engaged in research on low-bit-rate speech coding and VIP telephony. From 2001 to 2002 he was a guest researcher at Royal Institute of Technology in Sweden. Since 2007 he has been active in standardization of speech coding in ITU-T SG16 and is the editor of ITU-T Recommendation G.711.1. In 2009 he was appointed as Associate Rapporteur of ITU-T SG16 Q.10 (Speech and Audio Coding and Related Software Tools).

Hitoshi Ohmuro [M'93] is a senior research engineer, supervisor at the Speech, Acoustics and Language Laboratory at NTT Cyber Space Laboratories. He received his B.E. and M.E. degrees in electrical engineering from Nagoya University, Aichi, in 1988 and 1990, respectively. He has been engaged in research on highly efficient speech coding and the development of VoIP applications.