ITU-T CODERS FOR WIDEBAND, SUPERWIDEBAND, AND FULLBAND SPEECH COMMUNICATION

Richard V. Cox *Simao Ferraz de Campos Neto* *Claude Lamblin* *Mostafa Hashem Sherif*

This section of the magazine presents recent algorithms developed by the ITU to provide high quality coding beyond traditional narrowband telephony. Speech coders can be characterized by their *bit rate*, *quality*, *complexity*, and *delay*. Typical applications fall into one of two categories, one-way and two-way. The first includes storage applications such as telephone answering systems, streaming, multimedia delivery, and push-to-talk calls. The second includes real-time communications such as two person phone calls and conference calls. In this latter category, if the delay is too large — exceeding 300 ms round-trip — humans have difficulty communicating, while for storage and playback operations delay is not a factor. The complexity of a speech coder is one of the main contributing factors to its cost and energy usage. Complexity is most often measured in terms of memory usage (both RAM and ROM) and the number of instructions executed per second. All applications are sensitive to cost, and many are sensitive to energy usage as well. The desired bit rate is determined by channel capacity or storage capacity, depending on the application.

For decades the International Telecommunication Union (ITU) standardized speech coders having "telephone bandwidth," meaning an audio bandwidth of 300–3400 Hz with a sampling rate of 8 kHz. In the world of the public switched telephone network, or PSTN, this is what everyone has become accustomed to hearing on their telephones, be they wired or wireless. In the analog age this was the compromise point between the cost of transmission and the need for speech intelligibility. Quality was measured subjectively as a comparison, for both clean and noisy background conditions, with the performance of the first speech coding algorithm, G.711. If the fidelity of a new coder is judged comparable, it is called "toll quality." Each of the subsequent lower-bit-rate algorithms the ITU-T has approved aimed to achieve this quality for telephone bandwidth. Today, telephone bandwidth is referred to as "narrowband" for reasons explained below.

The first extension of the audio bandwidth, standardized about two decades ago, was ITU-T Recommendation G.722 having 50–7000 Hz bandwidth and a sampling rate of 16 kHz. This was dubbed "wideband" and was intended for video teleconference applications. Wideband improves sound reproduction without reaching the natural quality of face-to-face conversations or the high quality of professionally recorded speech. This is why recent coding algorithms have been developed for so-called superwideband (50–14,000 Hz) and fullband (20–20,000 Hz) to take advantage of the growth in available digital telecommunications bandwidth, particularly over the Internet. Nowadays, we need no longer be confined to narrowband communication; as a result, considerable effort has been expended on more natural teleconferencing for speech as well as video over Internet Protocol (IP).

One may reasonably ask why there are so many different ITU speech coding algorithms. The short answer is that today's applications go beyond speech, and different applications have different requirements. Generally speaking, if a higher bit rate is available, greater quality and lower delay can be supported. However, when bit rate is at a premium, different applications have made different trade-offs between bit rate, quality, complexity, and delay.

Sometimes having a coder with multiple bit rates is best. In today's applications, it is useful to have speech and audio coders that can adapt their bit rates on the fly according to the available channel capacity so that the best possible quality can be achieved at that instant in time. It is also possible to save channel bandwidth by reducing the coder bit rate during silence periods. In addition, any new coder requires new interworking equipment to interoperate with legacy coders. The cost of this new equipment can be considerably reduced if the bitstream of the new coder includes the bitstream of an earlier coder embedded in it. An embedded coder is a special kind of multiple bit rate coder with a bitstream structured into layers. In such a coder, the encoder generally operates at the highest rate, but groups of bits (layers) could be discarded by any component of the communication chain by simple truncation of the bitstream to reduce the bit rate. A layered bitstream

ITU-T number	Name	Rates (kb/s)	Algorithmic delay (ms)	Comments
Narrowband (300–3400 Hz, 8 kHz sampling rate) per sample coders				
G.711	Pulse code modulation	64, 56	0.125	Two types of companded PCM, A-law and μ-law, 8 (or 7) bits/sample
G.726	Adaptive differential pulse code modulation (ADPCM)	40, 32, 24, 16	0.125	Created for digital circuit multiplex encoding (DCME) to share channels for multiple conversations on undersea cables and satellite links; rates can be changed on the fly according to network congestion; 32 kb/s is toll quality
G.727	ADPCM	40, 32, 24, 16	0.125	Created for packet circuit multiplex encoding to share channels for multiple conversations on undersea cables and satellite links; each low rate is embedded in the next higher rate so that bits can be dropped in case of congestion; 32 kb/s is toll quality
Narrowband (300–3400 Hz, 8 kHz sampling rate) block coders				
G.728	Low-delay code excited linear prediction (LD-CELP)	40, 16, 12.8, 9.6	1.25	Created for DCME applications, 40 kb/s rate added to accommodate up to 14.4 kb/s modem signals; 16 kb/s is toll quality
G.729	Conjugate structure algebraic codebook excited linear prediction (CS-ACELP)	11.8, 8, 6.4	15	10 ms block, 5 ms look-ahead, explicit transmission of spectral and pitch parameters; 8 kb/s is toll quality; multiple applications, particularly for VoIP; 11.8 kb/s rate added to accommodate richer audio signals than speech (e.g., music signals)
G.723.1	Hybrid MPC-MLQ and ACELP	6.3, 5.3	37.5	Initially designed for PSTN videotelephony, yet primarily used for VoIP applications; 6.3 kb/s rate is toll quality
Wideband (50–7000 Hz, 16 kHz sampling rate) Coders				
G.722	Sub-band ADPCM	48, 56, 64	3	Lower band encoded with 4, 5, or 6 bit ADPCM (each low rate is embedded in the next higher rate so that bits can be dropped in case of congestion), upper band encoded with 2 bit ADPCM; initially designed for audio and video-conferencing applications, nowadays increasingly used in wideband telephony services (e.g., VoIP)
G.722.1	Transform coder	24, 32	40	Mainly used in audio and video conferencing applications; its superwideband extension G.722.1 Annex C provides 14 kHz bandwidth at 24, 32 and 48 kb/s
G.722.2	Adaptive multirate wideband (AMR-WB)	6.6, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05, 23.85	25.9375	Joint standard with 3GPP (3GPP TS 26.171 and TS 26.190); speech sampled at 16 kHz and processed at 12.8 kHz; 12.65 kb/s rate considered equal to G.722 at 56 kb/s; higher rates provide better quality for adverse background noise environments

Table 1. *Past ITU-T speech coding recommendations.*

offers higher flexibility and easier adaptation to any service requirements and interconnected networks/terminals or any variation in network capacity. An embedded variable bit rate coder is more succinctly referred to as a *scalable coder*. Table 1 is a list of previous ITU-T Recommendations in order to provide context as to why each new one was created. Many of these coders have multiple bit rates; some of them are embedded (or scalable) in bit rates.

G.711 is the basic coder, which some may consider as embedded because of "bit-robbing" for signaling in older North American networks. G.727 is an embedded coder whose adaptation mechanism and quantizer were constructed based on the 2 b/sample coder. Additional quantization levels were positioned between those of this fundamental quantizer. The encoder always operated at a high rate, 32 kb/s, but bits could be discarded due to net-

work congestion to reduce the bit rate to 24 or 16 kb/s. The lower band of G.722 is a narrowband algorithm. To convert from wideband to narrowband, the upper 16 kb/s can be discarded. G.728 was designed for much lower delay than G.729, but as a result has much higher complexity and therefore cost. Finally, G.723.1 has the most delay of the ITU-T telephone bandwidth coders and the lowest bit rate for toll quality performance.

Some of the new coders described in this section of the magazine reflect similar kinds of trade-offs. We begin with G.711.1. As its name implies, it is derived from G.711, which is the most ubiquitous, and oldest, of all the ITU-T speech coding Recommendations. Specifically, it extends G.711 to greater quality and audio bandwidth. G.711 is embedded inside the new coder's bitstreams of 80 or 96 kb/s. The additional bit rates were used to improve narrowband quality and to extend the bandwidth to wideband, 50–7000 Hz. A common superwideband (14 kHz audio bandwidth, 32 kHz sampling rate) extension of G.711.1 and G.722 is expected for mid-2010. Work on G.729.1 actually preceded that of G.711.1 with similar goals. G.729.1 is an embedded wideband extension of G.729 (8 kb/s CS-ACELP), the much-used algorithm of VoIP infrastructures. It can operate at 12 different bit rates from 32 down to 8 kb/s with wideband quality starting at 14 kb/s while keeping bitstream interoperability with G.729. Its fine bit rate granularity provides high flexibility to smoothly and finely improve the quality by increasing the bit rate with best possible network efficiency. G.718 is also an embedded coder, with a wideband core layer at 8 kb/s and four additional layers that increase quality for wideband speech and audio. Unlike G.729.1 and G.711.1, the G.718 core layer is not interoperable with any of the widely deployed coders. However, G.718 incorporates an alternate coding mode, which is bitstream interoperable with the 12.65 kb/s mode of G.722.2. A further extension to superwideband and stereo jointly for both G.729.1 and G.718 is currently under development.

In contrast, G.719 has a totally new structure that does not embed any older coder. It has multiple bit rates from 32 up to 128 kb/s, an audio bandwidth of 20–20,000 Hz, and a sampling rate of 48 kHz. Its low delay and complexity distinguish it from the majority of existing fullband audio coders, and make it particularly suitable for conversational applications such as videoconferencing

A word or two should be added about the standardization of speech and audio coding in the ITU. The first step is to identify the terms of reference (ToR), defining the application space for the new coder (or extension of an existing coder), the feature and performance requirements it must meet, and the objectives that are additional nonmandatory goals to be met in terms of features and performance. The requirements and objectives are established by polling the views of the various interested parties. A timetable is developed to allow competition where several phases of testing are planned. For full-blown codec development, three testing phases are generally planned: qualification, selection, and characterization. In the qualification phase, a potential candidate codec is tested against a subset of requirements to ascertain whether it can proceed to the next step. In the selection phase, the qualified candidates are more extensively evaluated by international laboratories using the subjective methods described below. This selection phase may be transformed into an optimization/characterization phase if the qualified candidates agree to collaborate toward a single solution or if only one candidate passes the qualification phase. The decision to start the approval process — or *consent* — depends on a thorough examination of the performance of the candidates (quality test results analysis, detailed algorithmic description, complexity figures). The best candidate providing the required features and performance is adopted through ITU-T's "Alternative Approval Procedure" (AAP) and then formally approved as an ITU-T Recommendation. After the approval, a characterization phase may follow to determine the codec performance in other application scenarios outside the original terms of reference. Through this phase, ITU-T is able to respond quickly to customer demands and cope with deployments that go beyond the originally intended applications. For instance, packet loss concealment (PLC) procedures were added to G.711 and G.722 — initially intended for PSTN and ISDN, respectively — to cope with packet losses over IP and other packet networks. Other functionalities could be wider audio bandwidth, stereo rendering capability, performance for specific signals or special languages, and more.

For narrowband and wideband audio coding, the subjective quality assessment of speech codecs follows the test methodologies of the ITU-T P.800 series of Recommendations. For wider audio bandwidths and richer audio signals (e.g., music), ITU-R testing methodologies such as BS.1116, BS.1285, and BS.1534 are considered. The choice of methodology is a function of the signal background, bandwidth, and so on; in all cases, testing deploys a large number of native listeners using at least two languages per experimental design. This ensures that test results reflect the quality opinion of the average user of the communication services identified in the ToR for the range of operating conditions required.

All recent ITU speech and audio codecs are specified through ANSI-C compilable source code. This serves the dual purpose of describing the implementation of the algorithm as well as a test model, particularly to verify the compliance of a given implementation. Since the 1990s, ITU-T coders come in two flavors, a 16-bit fixed-point arithmetic implementation suitable for implementation in digital signal processors, and a floating-point one that can be efficiently implemented in general-purpose CPUs with floating-point arithmetic units. All Recommendations (including C code and test vectors) are freely downloadable from the ITU Web site after their publication. Intellectual property right claims filed with ITU for any of its Recommendations can be checked online at http://itu.int/ipr.

In summary, the ITU-T audio/speech codecs portfolio is quite significant, covering a wide range of audio bandwidths and bit rates, offering different trade-offs to address various applications with different requirements (quality,

bit rates, complexity, robustness, delay). Conversational applications are still the primary applications; yet there has been an evolution from circuit-switched voice applications (e.g., PSTN and circuit multiplication equipment) to packet-based multimedia (notably IP).

The holy grail of speech coding is the quest for a single ("universal") sound coder suitable for any kind of future service requirements and interconnected networks that will provide flexibility in bit rates and bandwidths (full bandwidth down to lower bandwidth). Only the future will tell us whether this goal can be achieved.

BIOGRAPHIES

RICHARD V. COX [F] received his Ph.D. in electrical engineering from Princeton University. In 1979 he joined the Acoustics Research Department of Bell Laboratories. He conducted research in the areas of speech coding, digital signal processing, analog voice privacy, audio coding, and real-time implementations. He is well known for his work in speech coding standards. He collaborated on the low-delay CELP algorithm that became ITU-T Recommendation G.728 in 1992. He managed the ITU effort that resulted in the creation of ITU-T Recommendation G.723.1 in 1995. In 1987 he was promoted to supervisor of the Digital Principles Research Group. In 1992 he was appointed department head of the Speech Coding Research Department of AT&T Bell Labs. In 1996 he joined AT&T Labs as division manager of the Speech Processing Software and Technology Research Department. In August 2000 he was appointed Speech and Image Processing Services Research vice-president. In this capacity he had responsibility for all of AT&T's research in speech, audio, image, video, and multimedia processing research. In 1999 he was awarded the AT&T Science and Technology Medal. In 2008 he retired from AT&T. He is currently a senior staff scientist in the Human Language Technology Center of Excellence at Johns Hopkins University. He is a Past President of the IEEE Signal Processing Society and Past Member of the IEEE Board of Directors.

SIMÃO FERRAZ DE CAMPOS NETO [SM] joined the secretariat of the ITU Standardization Sector in 2002, and is the Counsellor for ITU-T Study Group 16 (for standardization work on multimedia services, protocols, systems, terminals, and media coding). He has organized several workshops (e.g. Multimedia in NGN, Telecoms for Disaster Relief, RFID, Standardization in E-health; SIIT 2005) and was the editor of the first version of the ITU-T Security Manual. Prior to joining ITU in 2002, he worked for eight years as a scientist in COMSAT Laboratories performing standards representation and quality assessment for digital voice coding systems, and before that he was a researcher at Telebras's R&D Center (CPqD). He has authored several academic papers and position papers, served on the review committees of several IEEE-sponsored conferences, and organized the first ITU-T Kaleidoscope Conference. He is a graduate of the State University of Campinas, Brazil (B.Sc. 1986 and M.Sc. 1993).

CLAUDE LAMBLIN graduated from the Ecole Nationale Supérieure des Télécommunications de Bretagne in 1983, and received her Ph.D. in electrical engineering from Sherbrooke University, Canada, in 1988. In 1983 she joined France Telecom Research Center (CNET) and studied speech and audio coding. Since 1989 she has taken an active part in the standardization process of speech coding algorithms in ETSI and ITU-T. She is a senior research engineer in the Speech and Sound Technologies & Processing Laboratory of France Telecom Research and Development Center, where she has managed R&D projects on compression and multidimensional representation of multimedia content. From 2002 to 2008 she was the Rapporteur of ITU-T Question 10/16 that dealt with the maintenance and extension of existing voice coding standards and software tools for signal processing standardization activities. Since 2005 she has been a Vice Chair of SG16 in charge of the Media Coding Working Party (WP3/16). She is a Senior Member of SEE (La société de l'électricité, de l'électronique, et des technologies de l'information et de la communication, http://www.see.asso.fr). In 2003 she received the Blondel Medal for her studies on algebraic vector quantization and its applications to audio compression standards.

MOSTAFA HASHEM SHERIF has been with AT&T in various capacities since 1983. He has a Ph.D. from the University of California, Los Angeles, an M.S. in the management of technology from Stevens Institute of Technology, New Jersey, and is a certified project manager from the Project Management Institute (PMI). Among the books he has co-authored are *Protocols for Secure Electronic Commerce* (2nd ed., CRC Press, 2003), *Paiements électroniques sécurisés, Presses polytechniques et universitaires romandes* (2006), and *Managing Projects in Telecommunication Services* (Wiley, 2006). He is a co-editor of two books on the management of technology published by Elsevier Science and World Scientific Publications in 2006 and 2008, respectively, and is the editor of the forthcoming *Handbook of Enterprise Integration* (3rd ed., Auerbarch). He is also a standards editor for *IEEE Communications Magazine*, an associate editor of the *International Journal of IT Standards & Standardization Research*, and a member of the editorial board of the *International Journal of Marketing*.