ABSTRACT

The International Telecommunications Union (ITU) has recently standardized three speech coders which are applicable to low-bit-rate multimedia communications. ITU Rec. G.729 8 kb/s CS-ACELP has a 15 ms algorithmic codec delay and provides network-quality speech. It was originally designed for wireless applications, but is applicable to multimedia communications as well. Annex A of Rec. G.729 is a reduced-complexity version of the CS-ACELP coder. It was designed explicitly for simultaneous voice and data applications that are prevalent in low-bit-rate multimedia communications. These two coders use the same bitstream format and can interoperate. The ITU Rec. G.723.1 6.3 and 5.3 kb/s speech coder for multimedia communications was designed originally for low-bit-rate videophones. Its frame size of 30 ms and one-way algorithmic codec delay of 37.5 ms allow for a further reduction in bit rate compared to the G.729 coder. In applications where low delay is important, the delay of G.723.1 may be too large. However, if the delay is acceptable, G.723.1 provides a lower-complexity alternative to G.729 at the expense of a slight degradation in quality. This article describes the attributes of speech coders such as bit rate, complexity, delay, and quality. Then it discusses the basic concepts of the three new ITU coders by comparing their specific attributes. The second part of this article describes the standardization process for each of these coders.

Low Bit-Rate Speech Coders for Multimedia Communication

Richard V. Cox, AT&T Laboratories Research

Peter Kroon, Bell Laboratories, Lucent Technologies

Speech coding refers to the process of reducing the bit rate of digital speech representations for transmission or storage, while maintaining a speech quality that is acceptable for the application. Multimedia refers to having a variety of media presented either simultaneously or sequentially. Thus, speech coding for multimedia automatically implies that the speech-coding bitstream will be sharing the communication channel with other signals. Some of the applications for such a speech coder include simultaneous voice and video, as in either a videophone or a stored video presentation, digital simultaneous voice and data (DSVD), where the data might be shared files which two or more parties are discussing or creating, or simultaneous voice and fax, in which a copy of a document is transmitted from one party to one or several others. Given this large number of applications, it might be difficult to select an appropriate speech coder. Although it would be convenient to have a "one size fits all" type of speech coder, it is often more economical to tailor the coder to the application. For example, sometimes the dominant factor is cost, sometimes quality.

An enormous number of new speech coders have recently been standardized. In the 1995–1996 time period three new international standards (International Telecommunications Union — ITU — G.729, G.729A, and G.723.1) and three new regional standards (enhanced full-rate coders for European and North American mobile systems) emerged. As a result, making an appropriate choice can be difficult. In addition, the reader might ask what the reason is for this abundance of coders. In the remainder of this article we will try to provide better insight into these matters.

We start this article with a discussion of coder attributes. These attributes can be used to make trade-offs during a coder selection process. This section is followed by a short description of three new international (ITU) standards. We briefly review the coding paradigm common to all these coders.

A large section of this article will discuss the history of the standardization process for these coders. Not only will this provide a better understanding of why so many standards exist; it will also provide better insight in how the specification came about and how it should be interpreted.

SPEECH CODER ATTRIBUTES

Speech quality as produced by a speech coder is a function of bit rate, complexity, delay, and bandwidth. Hence, when considering speech coders it is important to review all these attributes. It is important to realize that there is a strong interaction between all these attributes and that they can be traded off against each other. For example, low-bit-rate coders tend to have more delay than higher-bit-rate coders. They may also require higher complexity to implement and often have lower quality than the higher-bit-rate coders. In the remainder of this article we limit ourselves to telephone bandwidth speech (200–3400 Hz) sampled at 8 kHz. Additional factors that influence the selection of a given speech coder are availability and licensing conditions, or it could be the way the standard is specified. Some standards are only described as an algorithmic description, while others are defined by bit-exact American National Standards Institute ANSI-C code.

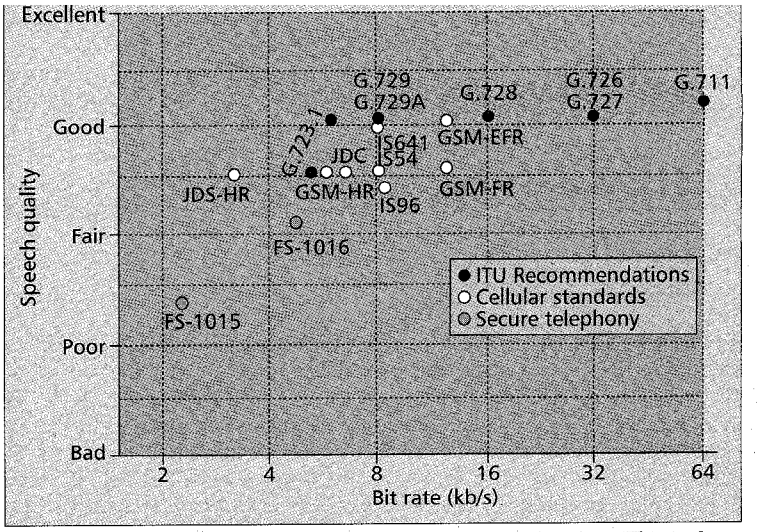
BIT RATE

Bit rate is the simplest attribute to understand, but is less straightforward than one might imagine. Since the speech coder is sharing the [illegible] th other data, the peak bit rate shc [illegible] as not to use a disproportionate share of the channel. Most speech coders operate at a fixed bit rate regardless of the input signal characteristics. Since multimedia speech coders share the channel with other forms of data, it is better to make the coder variable-rate. For simultaneous voice and data applications, a

0163-6804/96/$05.00 © 1996 IEEE

good compromise is to create a silence compression scheme as part of the coding standard. A common solution is to use a fixed rate for active speech and a low rate for background noise.

Silence compression consists of two main algorithms. The first is a voice activity detector (VAD), which determines if the input signal is speech or some sort of background noise. If the signal is declared speech, it is coded at the full fixed bit rate. If the signal is declared noise, it is coded at a lower bit rate. Sometimes no bits are transmitted at all. The second algorithm, comfort noise generation (CNG), is invoked at the receiver to reconstruct the main characteristics of the background noise. (The name "comfort noise" is used because listeners prefer low-level noise to dead silence.) Obviously, the performance of the VAD is critical to the overall speech quality. If speech is declared too often, the potential gains of silence compression are not realized. However, for loud background noises it may be difficult to distinguish between speech and noise. If the VAD fails to recognize the onset of speech, the beginning of the speech may be cut off. This is referred to as *front-end clipping* and seriously impairs the intelligibility of the coded speech. The comfort noise generation scheme must be designed in such a way that the encoder and decoder stay synchronized, even if there are no bits transmitted during some interval. This allows for smooth transitions between active and nonactive speech intervals.



■ **Figure 1.** *Quality comparisons of different standard coders. This is the performance for speech without background noise and without channel errors. Global System for Mobile Telecommunications (GSM), JDC, and ISXX are regional cellular standards in Europe, Japan, and North America, respectively. FS is a U.S. federal standard, and G.XXX are ITU standards. This figure is based on results from a number of subjective tests. It is presented to give approximate comparisons of relative quality. To the authors' knowledge, no single test has included all these coders.*

DELAY

The delay of a speech coding system usually consists of three major components. Most low-bit-rate speech coders process a frame of speech data at a time. The speech parameters are updated and transmitted for every frame. In addition, to analyze the data properly it is sometimes necessary to analyze data beyond the frame boundary. This is referred to as *look-ahead*. Hence, before the speech can be analyzed it is necessary to buffer a frame (plus look-ahead) worth of data. The resulting delay is referred to as *algorithmic delay*. This is the only delay component that cannot be reduced by changing the implementation; all other delay components depend on the implementation. Since they are unavoidable for practical systems, they need to be considered when analyzing the delay budget.

The second major contribution comes from the time it takes the encoder to analyze the speech and the decoder to reconstruct the speech. This is referred to as *processing delay*. It depends on the speed of the hardware used to implement the coder. The sum of the algorithmic and processing delays is called the *one-way codec delay*.

The third component is the communication delay, which is the time it takes for an entire frame of data to be transmitted from the encoder to the decoder. The total of these three delays is the *one-way system delay*. Maximum values of 400 ms for the one-way system delay can be tolerated if there are no echoes. However, new testing methodologies revealed that for ease of communication it is preferable if the one-way delay is below 200 ms. If there are echoes, the maximum tolerable one-way delay is only 25 ms! Hence, the use of echo cancellation devices is often necessary.

In many applications, such as teleconferencing, it is necessary to bridge several callers so that each person can hear all the others. For speech coders, this means decoding each bitstream, summing the decoded signals, and then re-encoding the sum signal. This process not only doubles the delay, it also reduces the speech quality due to the multiple encodings. In a bridged system the maximum tolerable one-way delay is 100 ms because the bridging will double the one-way system delay to 200 ms.

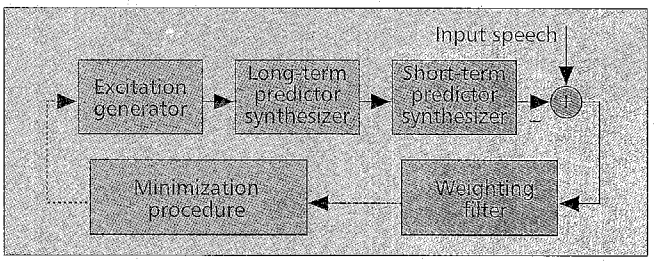
COMPLEXITY

Speech coders are often implemented on (or share) special-purpose hardware, such as digital signal processor (DSP) chips. Their attributes can be described as computing speed in millions of instructions per second (MIPS), random access memory (RAM), and read only memory (ROM). In creating a speech coder for any application, the system designer must make a choice about how much of these resources to allocate to the speech coder. Currently speech coders requiring less than 15 MIPS are thought of as low-complexity; those requiring 30 MIPS or more are considered high-complexity. Having to put more RAM and ROM on a chip results in a more expensive chip (greater chip size means lower yield).

From the system designer's point of view, more complexity results in higher costs and greater power usage; for portable applications, greater power usage means reduced time between battery recharges or using larger batteries, which means more expense and weight. Thus, complexity is an important factor. At the same time, it is reasonable to view the speech coder's share of the power and expense budget of the overall system of which it is to be a part. If that share is small (on the order of 10 percent), picking the best-quality coder for a given bit rate probably makes the best sense, since the difference in cost or power is unlikely to increase the percentage greatly. If the share is large, much more consideration must be given.

QUALITY

Of the four attributes, quality has the most dimensions. One of the most important is how well the coded speech sounds for ideal conditions (clean speech, no transmission errors, only one encoding). Figure 1 shows a typical picture which relates the performance of various standard coders. This picture represents the quality for a single encoding using speech



■ **Figure 2.** *Block diagram of LPAS coder.*

without background noise. Note that in the real world these ideal conditions are often not met. For many applications there are large amounts of background noise (car noise, street noise, office noises like typing or phones ringing, air conditioning noise, music in the background, etc.). How well does the coder perform under these adverse conditions? What happens when there are channel errors during transmission? These errors could be individual bit errors or the loss of entire frames. Are the errors detected or undetected? If undetected, the coder must perform even more robustly than when it is informed that entire frames are in error. How good does the coder sound when the speech is encoded and decoded twice, as in a bridging application? How well does it sound when tandemed with another standard speech coder that is likely to be in use for a given application? How well does the speech sound for a wide variety of speakers, including those with high or low pitch? What about two speakers talking at the same time? What about speech in other languages?

All of these are questions the standards bodies try to answer during the testing phase. The goal is to characterize the coder so that its performance is well known for its intended applications. The level of performance for some of these questions may be critical to the coder's intended application, in which case the standards body is likely to set requirements for those questions. In other cases there is a desired goal, but reaching it is not essential; thus, the intended performance level is an objective rather than a requirement.

Linear Prediction Analysis by-Synthesis Coding

Almost all recent international and regional speech coding standards belong to a class of linear prediction analysis-by-synthesis (LPAS) coders. This class of coders includes ITU Recommendations G.723.1, G.728, and G.729 and all the current digital cellular standards in:

- Europe: Global System for Mobile Telecommunications (GSM), full-rate, half-rate, and enhanced full-rate
- North America: full-rate and enhanced full-rate for time-division multiple access (TDMA) and code-division multiple access (CDMA) systems
- Japan: full-rate and half-rate
- Federal Standard 1016: a 4.8 kb/s speech coder for secure telephony

This section gives a brief description of what is in an LPAS coder and some of the details on what distinguishes G.729 and G.723.1.

The decoded speech is produced by filtering the signal produced by the excitation generator through both a long-term (LT) predictor synthesis filter and a short-term (ST) predictor synthesis filter. The excitation signal is found by minimizing the mean-squared error over a block of samples. The error signal is the difference between the original and decoded signal. It is weighted by filtering it through a weighting filter. Both short- and long-term predictors are adapted over time. Since the analysis procedure (encoder) includes the synthesis procedure (decoder), the description of the encoder defines the decoder. The short-term synthesis filter models the short-term correlations (spectral envelope) in the speech signal. This is an all-pole filter with an order between 8 and 16. The predictor coefficients are determined from the speech signal using linear prediction (LP) techniques as described in [1]. The coefficients of the short-term predictor are adapted in time, with rates varying from 30 to as high as 400 times/s.

The long-term predictor filter models the long-term correlations (spectral fine structure) in the speech signal. Its

About the ITU

The International Telecommunications Union (ITU) is a body within the United Nations Economic, Scientific and Cultural Organization (UNESCO). ITU headquarters are located in Geneva, Switzerland. Before 1993 the ITU consisted of two main bodies: the International Consultative Committee for Telephone and Telegraph (CCITT), which recommended telecommunications standards, and the International Consultative Committee for Radio (CCIR), which recommended radio standards. In 1993 the ITU was reorganized. Some portions of the CCIR became part of a new body that also contained all of the CCITT. The new body was called the ITU — Telecommunications Standardization Sector, and its acronym became the ITU-T. The remainder of the CCIR is now in a body called the ITU — Radio Standardization Sector, and its acronym is ITU-R. (Ironically, the ITU does not make standards. These standardization sectors produce documents that are formally known as Recommendations. They represent agreement within a segment of the telecommunications industry on a particular topic, but there is no force of law to mandate using them; hence the term "recommendations.")

Within each of these large bodies are smaller groups that focus on specific topics. In the ITU-T these are called Study Groups. Study Group 15 (SG15) is charged with making recommendations related to speech and video processing, such as speech coding or videotelephony. Study Group 14 (SG14) makes recommendations for modems, such as V.34 and V.32. Study Group 12 (SG12) is charged with studying matters related to network performance, such as speech quality. The Speech Quality Experts Group (SQEG) within SG12 designs and conducts the subjective testing experiments used to determine the performance of proposed ITU speech coder recommendations.

All ITU speech coding recommendations begin with the designation G.7. The specific speech coding recommendations that are currently active are G.711 64 kb/s pulse code modulation (PCM) (both A-law and μ-law), G.722 wideband speech coder operating at 64, 56 or 48 kb/s, G.726 adaptive differential PCM (ADPCM) operating at 40, 32, 24 or 16 kb/s, G.727 embedded ADPCM operating at 40, 32, 24 or 16 kb/s, G.728 16 kb/s low-delay code-excited linear prediction coder (LD-CELP), G.729 8 kb/s conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP), and G.723.1 low-bit-rate speech coder for multimedia communications operating at 6.3 and 5.3 kb/s. This last coder was designated with an extension (.1) because all the numbers in the G.711 and G.721 series have been used already. G.721 and G.723 were ADPCM coder recommendations that were folded into G.726. However, it was decided that it would be bad policy to reuse a previously used Recommendation number. Instead, G.723.1 was the designation.

ITU recommendations are often covered by patents. Those organizations holding patents agree in writing to charge fair and reasonable royalty rates and not to discriminate against any licensees. They may even agree to waive royalties altogether. The ITU does not administer intellectual property rights, but does have lists of organizations claiming to hold patents for each of the recommendations. Organizations wishing to manufacture or use ITU recommendations should negotiate licensing terms with those organizations holding the relevant patents.

parameters are a delay and a gain coefficient. For periodic signals, the delay corresponds to the pitch period (or possibly an integral number of pitch periods). The delay is random for nonperiodic signals. Typically, the long-term predictor coefficients are adapted at rates varying from 100 to 200 times/s.

A commonly used alternative structure for the pitch filter is the *adaptive codebook*. In this approach, the long-term synthesis filter is replaced by a codebook that contains the previous excitation at different delays. The resulting vectors are searched, and the one that provides the best match is selected. In addition, an optimal scaling factor can be determined for the selected vector. This representation simplifies the determination of the excitation for delays smaller than the length of the excitation frames.

To achieve a low overall bit rate, the average number of bits per sample for each frame of excitation samples has to be small. The *multipulse excitation coder* [3] represents the excitation as a sequence of pulses located at nonuniformly spaced intervals. The excitation analysis procedure has to determine both amplitudes and positions of the pulses. Finding these parameters all at once is a difficult problem and simpler procedures, such as determining locations and amplitudes one pulse at a time, are used. The number of pulses required for an acceptable speech quality varies from 4–6 pulses/5 ms. For each pulse, both amplitude and location have to be transmitted, requiring about 7–8 b/pulse.

Code-excited linear predictive (CELP) coders [4], which are the most common realization of the LPAS paradigm, use another approach to reduce the number of bits per sample. Here, both encoder and decoder store the same collection of C possible sequences of length L in a codebook. The excitation for each frame is described completely by the index to an appropriate vector in the codebook. In the configuration of Fig. 2, this index is found by an exhaustive search over all possible codebook vectors, selecting the one that produces the smallest error between the original and decoded signals. To simplify the search it is common to use a gain-shape codebook in which the gain is searched and quantized separately. The index requires $(\log_2 C)/L$ bits/sample (e.g., 0.2–2 bits/sample), and the gain requires 2 to 5 bits for each codebook vector. Further simplifications can be obtained by populating the codebook vectors with a multipulse structure. By using only a few nonzero unit pulses in each codebook vector, efficient search procedures can be derived. This partitioning of the excitation space is known as an algebraic codebook, so the excitation method is known as *algebraic codebook-excited linear prediction* (ACELP) [5].

Error Weighting Filter

In the paradigm shown in Fig. 2, the coder parameters are selected such that the error energy between the reference and reconstructed signal is minimized. Minimizing a mean-squared error results in a quantization noise that tends toward having equal energy at all frequencies of the input signal. By using the properties of the human auditory system, one can try to reduce the perceived amount of noise. Frequency-masking experiments have shown that greater amounts of quantization noise are undetectable by the auditory system in the frequency bands where the speech signal has high energy [2]. To make use of this masking effect, the quantization noise has to be properly distributed among the different frequency bands. Spectral shaping of the noise can be achieved by minimizing a weighted error in the block diagram of Fig. 2. To shape the noise as a function of the spectral peaks in the speech signal, it makes sense to derive the error-weighting filter from the short-term predictor filter. In most speech coders, this filter is indeed derived from the LP filter coefficients. Noise shaping increases the mean squared error between the original and reconstructed speech, resulting in a reduction in segmental signal-to-noise ratio (SNR).

Parameter	G.729	G.729A	G.723.1	G.723.1
Bit rate (kb/s)	8	8	6.3	5.3
Frame size (ms)	10	10	30	30
Subframe size (ms)	5	5	7.5	7.5
Algorithmic delay (ms)	15	15	37.5	37.5
MIPS (fixed-point DSP)*	20	10.5	14.6	16
RAM (16-bit words)	2.7 k	2 k	2.2 k	2.2 k

*The MIPS numbers were taken from implementations on different high-end fixed-point DSPs, and are the lowest values known to the authors.

■ **Table 1.** *Comparison of coder parameters for the three new ITU Recommendations.*

Adaptive Postfilter

Despite the error-weighting filter, it is not always possible to mask the noise in speech caused by the quantization of the excitation signal. Especially in the low-energy frequency regions, this quantization noise can dominate the speech signal. Using a separate post-processing technique after reconstruction by the decoder, the perceived noise can be further reduced. This operation, referred to as *postfiltering*, trades off spectral distortion in the speech versus suppression of the quantization noise by emphasizing the spectral peaks and attenuating the spectral valleys.

To have more flexibility in the shape of the postfilter, it is generally implemented as a combination short-term/long-term filter. The short-term postfilter modifies the spectral envelope. This spectral envelope is usually based on the transmitted short-term predictor coefficients, but can also be derived from the reconstructed signal. The parameters for the long-term postfilter are either derived from the transmitted long-term predictor coefficients or computed from the reconstructed speech.

Comparison between G.729 and G.723.1

Table 1 shows the different parameters of each coder. The principal difference between the G.729 and G.723.1 excitation signals is their partitioning of the excitation space. Both assume that all pulses have the same amplitudes and that the sign information will be transmitted. G.729 has excitation frames of 5 ms and allows four pulses to be selected. The 40-sample frame is partitioned into four subsets. (The first three subsets have eight possible locations for pulses, the fourth has 16.) One pulse must be chosen from each subset. This is referred to as an algebraic codebook, and the excitation method is ACELP [5]. G.723.1 has excitation frames of 7.5 ms, and also uses a four-pulse ACELP excitation codebook for the 5.3 kb/s mode. For the 6.3 kb/s rate, the frame positions are grouped into even-numbered and odd-numbered subsets. A sequential multipulse search is used for a fixed number of pulses from the even subset (either five or six, depending on whether the frame itself is odd- or even-numbered). A similar search is repeated for the odd-numbered subset; then the set resulting in the lowest total distortion is selected for the excitation. This is known as multipulse excitation with a maximum likelihood quantizer (MP-MLQ).

At the decoder, the LPC information, adaptive codebook

information, and fixed codebook information are demultiplexed and then used to reconstruct the output signal. An adaptive postfilter is used to reduce the perceptible noise. For G.723.1, the long-term postfilter is applied to the excitation signal before passing it through the LPC synthesis filter and the short-term postfilter.

The DSVD version of G.729 (G.729A) is bitstream-compatible with G.729. This means that a signal analyzed with the DSVD coder can be reconstructed with the G.729 decoder, and vice versa. The dominant complexity reduction in the DSVD was obtained by simplifying the codebook search for both the fixed and adaptive codebooks; in addition, the postfilter has been simplified. In all, the complexity is reduced by about 50 percent, at the expense of a slight degradation in performance for some operating conditions.

A History of the Three New ITU-T Standards

In this section we give a historical overview of the standardization process for the three ITU coders. We start with a description of how the requirements are set. This is illustrated by the specifics for each of the speech coder attributes: bit rate, delay, complexity, and quality.

Setting Requirements

In most standardization procedures, it is common to specify the requirements using the Terms of Reference (ToR). This document not only contains a schedule, but also specifies the performance requirements and objectives. Besides the quality, it also specifies the other coder attributes such as bit rate, delay, and complexity.

Bit Rate

For G.729, the ToR requirement was that the speech coder should operate at 8 kb/s. This rate was selected in part because it fit the range of first-generation digital cellular standards, from 6.7 kb/s in Japan to 7.95 kb/s in the United States to 13 kb/s in Europe. In addition, it was a natural division by two of speech coder bit rates already planned or standardized by the ITU (64, 32, and 16). For G.723.1, the ToR requirement was that the speech coder should operate below 9.6 kb/s. As it turned out, all the coders tested ranged between 5.0 and 6.8 kb/s. In the later development of G.723.1, a second lower rate was added for flexibility. For the DSVD coder, the ToR requirements for bit rate were derived from the amount of speech data that could be carried over a 14.4 kb/s modem. The bit rates of the five candidate coders submitted for this standard were all near 8 kb/s. None of the three coders had a silence compression scheme as part of the main body of the recommendation. Subsequent work created silence compres-

Subjective Testing of Speech Quality

When a speech coder is required to match a certain level of subjective quality, how is it tested? That is the challenge of subjective testing. Within SG12 several methodologies have been accepted as producing meaningful results, while other, newer methods are still being studied.

The most frequently used test is the *absolute category rating* (ACR) test. Subjects listen to about 8–10 s of speech material and are asked to rate the quality of what they heard. Usually a five-point scale is used to represent the quality ratings, such as

- excellent (5)
- good (4)
- fair (3)
- poor (2)
- bad (1).

By assigning the corresponding numerical values to each rating, a mean opinion score (MOS) can be computed for each coder by averaging these scores. Typically a test must include a balanced selection of material; that is, there should be an equal number of male and female utterances, and the quality of the test material should cover the range of possible quality ratings so that listeners will use all categories in their voting. This is often done by incorporating well-established anchor conditions, such as other standard coders or the source material corrupted by speech-correlated noise. The latter is referred to as modulated noise reference units (MNRUs), and can be used as a means to normalize scores. This is done by a curve fitting procedure between MNRU conditions (expressed in dB SNR and MOS scores). Once this curve is obtained, all other MOS results can be expressed in equivalent MNRU SNR or dBQ values.

In determining whether one condition in a test is better or worse than another, it is necessary to do a statistical analysis to determine if MOS differences are significant. A commonly used procedure is Tukey's HSD which groups coders into statistically equivalent performance groups.

The ACR test works well when the votes are spread over the entire range of responses. However, if the input signal quality is poor to begin with, coding will not make it better. Thus, an ACR test for background noise conditions will be unlikely to find significant differences among any coders in the test if they all receive low scores. One type of test that SQEG has used for this test scenario is the *degradation category rating* (DCR) test, in which the listeners hear the original passage first and then the processed (coded) version second. They are asked to rate how much degradation they heard in the second stimuli compared with the first. Possible ratings are:

- No perceptible distortion (5)
- Perceptible but not annoying (4)
- Mildly annoying (3)
- Annoying (2)
- Very annoying (1)

A degradation mean opinion score (DMOS) can be computed from the ratings using the numbers in parenthesis. However, the DCR test has not achieved as much acceptance as the ACR test because listeners seem to equate difference with degradation.

A newer test is the comparative category rating test (CCR) in which the listener hears the original and the processed speech in a random order. They are asked to rate the second condition on a seven-point scale; three of the points suggest it sounds better than the first condition, and three suggest it sounds worse; they can also be rated equal. A comparison mean opinion score (CMOS) can be computed by normalizing the scores with reference to the unprocessed original. This test was used for the first time in the G.729 characterization phase testing. Like the DCR test, it appears to be very sensitive to small differences between stimuli.

One comparison that is difficult to make is between different coders on different tests. Even when each test used similar anchors and the scores are normalized, it is still difficult to make reliable comparisons between coders that were not included in both tests.

When a coder is determined to give robust performance under many different input and channel conditions, it is considered to be toll- or wireline-quality. A good definition would be that either any differences between the performance of 32 kb/s G.726 ADPCM and a toll-quality coder are not statistically significant or the toll-quality coder is better. Thus, the toll-quality designation was not applied to G.723.1 or the DSVD coder because they were not widely enough tested. However, G.729 was widely tested and did meet this statistical definition for all conditions using ACR testing. This is also true of 16 kb/s G.728 and 32 kb/s G.727. Nevertheless, it should be kept in mind that all these tests reflect only a limited aspect of normal everyday usage, and especially for lower-bit-rate coders (8 kb/s and below) the user experience might not be as favorable as with high-bit-rate coders.

sion schemes for both G.723.1 and G.729, which are included as annexes to each recommendation.

DELAY

Delay is a major differentiation between G.723.1 and G.729. The ToR requirement for delay for G.729 was discussed for over a year. Initially it was a maximum one-way codec delay of 10 ms. Later, the frame size was allowed to grow to 16 ms. Ultimately, at the request of the Speech Quality Experts Group (SQEG) (see sidebar), the frame-size was specified to be 10 ms. G.729 has a 5 ms look-ahead. Assuming 10 ms processing delay and 10 ms transmission delay, the one-way system delay of G.729 is 35 ms.

The principal application of G.723.1 is low-bit-rate videophones, which typically operate at 5 frames/s or fewer. This rate equates to a video frame period of 200 ms. The Experts Group felt that a one-way delay of 100 ms for the speech coder was needed so that if bridging were used, the one-way delay would only increase to 200 ms. Working backwards from the 100 ms value, a maximum frame size of 32 ms was set. The final version of G.723.1 has a look-ahead of 7.5 ms, making the one-way system delay 97.5 ms.

In deliberating on the delay requirements for a DSVD coder, Study Group 14 (SG14) was cognizant of the delay inherent in V.34 modems. These modems often have one-way delays greater than 35 ms. If bridging were used, the modem delay would be greater than 70 ms. In rejecting G.723.1 for use with simultaneous voice and data applications, SG14 was well aware that the combined one-way delay for a single encoding could be 135 ms or greater using G.723.1. SG14 and SG15 agreed on a one-way codec delay maximum of 40 ms.

Parameter	Requirements	Objectives
Quality (without bit errors)	Not worse than 32 kb/s G.726	
Quality (with bit errors)		
Random bit errors BER < 10^{-3}	Not worse than G.726	Equivalent to 32 kb/s G.726
Detected frame erasures	No more than 0.5 MOS	
Random and bursty 3%	degradation from 32 kb/s ADPCM without errors	As small as possible
Undetected burst errors		For further study
Level dependency	Not worse than 32 kb/s G.726	As low as possible
Talker dependency	Not worse than 32 kb/s G.726	
Capability to transmit music		No annoying effects generated
Tandeming capability for speech	Two asynchronous codings with a total distortion of ≤ 4 asynchronous 32 kb/s G.726	3 asynchronous codings with a total distortion ≤ 4 Asynchronous 32 kb/s G.726
Tandeming with other ITU standards	≤ 4 asynchronous 32 kb/s G.726	Synchronous tandeming property
Tandeming with regional DMR standards	For further study	
Idle channel noise		
•Weighted	For further study	Not worse than 32 kb/s G.726
•Single-frequency	Not worse than 32 kb/s G.726	
Capability to transmit signaling/information tones	DTMF, CCITT Nos. 5,6 and 7, CCITT R2, Q.35, Q.23, V.25	Distortion as low as possible

■ **Table 2.** *Speech quality performance requirements and objectives for G.729.*

COMPLEXITY

In formulating the requirements for G.729, the trade-offs discussed all involved delay and complexity. As indicated above, the ITU — Radiocommunications Standards Sector (ITU-R) was concerned about creating a coder that would be too complex with too high a delay. Ultimately, they accepted a delay target that allowed a significant reduction in complexity compared with the G.728 coder. The MIPS are reduced to around 17. However, the amount of RAM required is 3000 words, 50 percent more than G.728. Much of this extra memory usage is due to the use of larger frames.

G.723.1 is of lower complexity than G.729 (14.6 at 5.3 kb/s and 16 at 6.3 kb/s), and uses 2200 words of RAM. However, even this complexity was considered a little too high by SG14. Their requirements for the DSVD coder were 10 MIPS, 2000 words of RAM, and 10,000 words of ROM.

QUALITY

Table 2 is the list of performance requirements and objectives for the coder that became G.729. This table does not include requirements unrelated to speech quality, such as bit rate, delay, and complexity, which are also discussed in the ToR. The first requirement is that for error-free conditions, a single encoding should be rated not worse than 32 kb/s G.726. In separate testing of G.729, G.723.1, and the DSVD version of G.729, all three coders met this requirement. The difficult part was measuring performance for the various background noise conditions. Both the initial coders submitted for G.729 failed some or all these conditions because the coded noise sounds different from the original. In degradation category rating (DCR) tests, subjects seem to equate different with worse. As a result, G.729 received lower scores than G.726 for DCR testing. However, if absolute category rating (ACR) tests were performed, the MOS scores of G.729 were never significantly worse than G.726 and were sometimes better. The testing of G.723.1 and G.729A was less extensive. The second requirement concerned speech quality with noisy channels. For 10^{-3} random bit error rate, the speech quality should again be no worse than G.726 under similar conditions. The requirement for frame erasure conditions was more difficult to determine. Ultimately the coder was tested with 1, 3, and 5 percent random and bursty frame erasures. The requirement was that the 3 percent case for both random and bursty conditions should be no worse than the MOS score of G.726 minus 0.5 points. This was still a difficult requirement to test, since the score for 3 percent frame erasures is greatly influenced by other conditions under test. Ultimately G.729 passed all these conditions. G.723.1 was not tested for random bit errors because this condition is unlikely to occur for high-speed modems; burst errors are more likely to occur, so G.723.1 was tested for 3 percent random burst errors and met this requirement. The same was also true of the DSVD version of G.729. The third set of requirements are concerned with speech that is input at 10 dB either higher or lower than the nominal input level. All three coders were tested for this condition and passed. To measure the fourth requirement, a large number of speakers are needed, including children as well as adults. G.729 was tested in four different languages with at least eight speakers in each language. G.723.1 and G.729A were

Activity	G.729	G.723.1	G.729A
Initial discussion of ToR	7/90	11/92	11/94
Finalization of ToR	11/91	9/93	2/95
Elapsed months	16	10	3
Submission of candidate coders	9/92	12/93	3/95
Host lab session for selection	2/93	2/94	6/95
Selection of candidate	2/95*	10/94	11/95
Elapsed months	39	13	9
First draft recommendation	6/95	1/95	11/95
Submitted for determination	2/95	2/95	11/95
Submitted for decision	11/95	11/95	5/96
Elapsed months	9	13	6
Total months for entire process	64	36	18

*No initial candidate met all requirements in the ToR, so an optimized coder was created, retested, and did meet all requirements. This process added nine months to the overall schedule.

■ **Table 3.** *Schedule of standardization activities for low-bit-rate speech coders.*

not tested as extensively. All coders met the requirement for the testing that was done.

All three coders pass music signals, but the quality of the music is poor. The reason for this is that LPAS coders rely on pitch prediction to achieve high coding efficiency. Most music signals lack a pitch structure, and all the coding burden has to be carried by the excitation and low-order LP predictor.

The next requirement concerns tandeming, with itself or other standard coders. In the cases of G.723.1 and the DSVD version of G.729, tandeming with other standards was not an issue. The requirement for self-tandeming is that two encodings should produce distortion no worse than four encodings with 32 kb/s G.726. All the coders meet this requirement. Only G.729 was tested in tandem with other coders. There were no requirements for tandeming with regional digital cellular standards. Idle channel noise and signaling tone tests were done for G.729. Only limited testing for dual tone multifrequency (DTMF) tones was done for G.723.1, and no tone tests were done for G.729A.

The overall performance of the three coders was similar. It seemed that the G.723.1 and G.729A coders were slightly less robust for background noises and tandem conditions. Their performance for clean speech and general robustness are sufficient that the ITU sees fit to recommend them for use in simultaneous voice and data applications, such as low-bit-rate multimedia communications.

Schedule

The ITU creates recommendations only because there is a need. The standard-making effort attempts to create a recommendation that will meet the requirements of the need. In this way performance requirements are determined. Depending on the urgency of the need, the schedule may be either short or long. If we review these three schedules, it seems that we can divide the process into three main parts: time spent determining the requirements and objectives (which is culminated by the completion of the ToR), time spent on submissions and testing (which is culminated by the selection of the coder), and time spent drafting the recommendation and following the procedures of the ITU required for ratification. Table 3 summarizes these times in months. The 10 months spent on G.723.1 is probably a reasonable time period. Urgent, highly focused requests such as DSVD can be handled more quickly, but will still depend on the date of the next SG15 meeting for formal approval. Under ideal conditions, a complete testing process such as that used for G.729 could probably be completed in 24 months between the completion of the ToR and the selection of the proposed coder. Smaller testing efforts can be accommodated in less time. Also, software testing of coders reduces the time needed for the proponents to prepare their coders and for the host laboratory to set up equipment. The amount of time spent on the final portion of the process depends on the schedule of meetings for SG15. Two full Study Group meetings are required, one to put the recommendation forward for determination and the second to put it forward for decision (balloting). For Annex A of G.729 (the DSVD coder), the two meetings are closer together (6 months rather than 9). This is just the luck of the calendar. This analysis suggests that from finalization of the ToR to putting the recommendation forward for ballot can take as short as 15 months for a derivative standard to as long as three years for the most ideal conditions and four or more years for two recent examples (G.728 and G.729).

Conclusion

The tremendous growth of multimedia communications has increased the demand for speech coders. The availability of new-generation speech coders capable of providing toll-quality speech at bit rates between 6.3 and 8 kb/s allows tailoring the coder to the application.

Each of the three new ITU Recommendations G.729, G.729 Annex A for DSVD, and G.723.1 has the potential to become the principal bearer of speech on the Internet and other networks. All three are low enough in complexity to be executed on a host processor, such as a PC, or implemented as part of a modem chip. When this happens and the coder becomes ubiquitous, it will enable speech to become another form of data available via electronic access. The number of potential applications is myriad, as are the economic possibilities arising from them. It is fun to speculate on which applications will eventually become the most widely used and most profitable.

However, the speech coder will have become a commodity item. Does that mean there will be no need for future coding standards? By no means. Anyone who has heard telephone bandwidth speech compared to live conversation knows that much of the richness and presence of the speaker is removed by the bandpass filtering. The ITU is already working on a next-generation low-bit rate, low-complexity, 7 kHz bandwidth speech coder. That is one way to improve quality further. Why stop at speech signals? As part of MPEG-4, ISO is standardizing a family of audio coders that will span the range from telephone bandwidth at 2 kb/s to CD-quality audio and beyond at rates up to 128 kb/s. In spite of the bandwidth explosion, it seems that the number of signals people would like to hear, see, or access is growing even more rapidly. Speech and audio compression will have a future as lively as their past.

References

[1] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs, NJ: Prentice Hall, 1978.

[2] B. S. Atal and M. R. Schroeder, "Predictive Coding of Speech Signals and Subjective Error Criteria," *IEEE Trans. ASSP*, vol. ASSP-27, June 1979, pp. 247–54.

[3] B. S. Atal and J. R. Remde, "A New Model of LPC Excitation for Producing Natural Sounding Speech at Low Bit Rates," *Proc. IEEE ICASSP*, Apr. 1982, pp. 614–17.

[4] M. R. Schroeder and B. S. Atal, "Code-Excited Linear Prediction (CELP): High Quality Speech at Very Low Bit Rates," *Proc. IEEE ICASSP*, 1985, pp. 937–40.

[5] J.-P. Adoul *et al.*, "Fast CELP Coding Based on Algebraic Codes," *Proc. IEEE ICASSP*, Apr. 1987, pp. 1957–60.

Additional Reading

In addition to the above references, ITU-T Recommendations G.729 and G.723.1 are available from the ITU (http://www.itu.ch). For further information on speech coding, several comprehensive papers are listed below, as well as one comprehensive book.

[1] A. Gersho, "Advances in Speech and Audio Compression," *Proc. IEEE*, vol. 82, June 1994 pp. 900–18.

[2] A. S. Spanias, "Speech Coding: A Tutorial Review," *Proc. IEEE*, vol. 82, no. 10, , Oct. 1994, pp. 1541–82.

[3] W. B. Kleijn and K. K. Paliwal (Eds.), *Speech Coding and Synthesis*, Elsevier, 1995.

Biographies

RICHARD V. COX [F] received the B.S. degree from Rutgers University, and the M.A. and Ph.D. from Princeton University, all in electrical engineering. He joined Bell Laboratories in 1979 and has worked in various aspects of speech and audio coding, speech privacy, digital signal processing, combined speech and channel coding for noisy channels, and real-time signal processing implementations. He has been active in the creation of speech coding standards for digital cellular telephony and the toll network. From 1993 to995 he served as editor of Recommendation G.723.1. In 1987 he was appointed supervisor of the Digital Principles Research Group and in 1992 he was appointed head of the Speech Coding Research Department. In AT&T Laboratories Research he is currently division manager of the Speech Processing Software and Technology Research Department with responsibility for speech and audio coding, text-to-speech synthesis, and human hearing research. Dr. Cox is active in the IEEE Signal Processing Society. He is a past chairman of the Speech Technical Committee, served on the AdCom and Board of Governors for six years and as Treasurer/Vice President Finance. He is currently Vice President-Elect for Publications.

PETER KROON [F] received the M.S. and Ph.D. degrees in electrical engineering from Delft University of Technology, Delft, The Netherlands. His Ph.D. work focused on time-domain techniques for toll-quality speech coding at rates below 16 kb/s. The regular-pulse excitation coding technique described in his thesis forms the basis of the current GSM cellular system. In 1986 he joined AT&T Bell Laboratories, Murray Hill, New Jersey, where he has been working on a variety of speech coding applications, including the definition of the 4.8 kb/s secure voice standard FS1016. For the last two years he has been involved in the design and definition of the new ITU-T 8 kb/s speech coding standard G.729. His current research activities are concentrated on high-quality speech coding algorithms suitable for second-generation digital cellular systems and applications such as voice store-and-forward systems. Dr. Kroon received the 1989 IEEE SP Award for authors under 30 years old for his paper on regular pulse coding. He just finished a four-year term as a member of the IEEE SP Society Speech Technical Committee and the IEEE SP Conference Board.