CODING OF SPEECH AND WIDEBAND AUDIO

Nikil S. Jayant, Victor B. Lawrence, and Dimitrios P. Prezas

Advances in coding algorithms and digital signal processing have led to sophisticated technologies for speech communication for a variety of applications, as well as to greater flexibilities in the design of ISDN terminals for integrated communication of speech, images, and data. For traditional telephony with a signal bandwidth of 3.2 kHz, the transmission rate for network-quality speech is now down to 16 kb/s. Robust communications-quality speech appropriate for cellular radio has been realized at 8 kb/s. Research attention is shifting toward 4 kb/s, focused on improving speaker identification and the naturalness of coded speech. For wideband audio with a signal bandwidth of 7 kHz, high-quality coding is now possible at 32 kb/s, which implies stereo teleconferencing or dual-language programming over a 64-kb/s channel. Transparent coding of 20-kHz audio has been demonstrated at 128 kb/s, with near-transparent performance at rates as low as 64 kb/s for some classes of signals.

Nikil S. Jayant *is head of the Signal Processing Research Department of AT&T Bell Laboratories, Murray Hill, New Jersey.* ***Victor B. Lawrence*** *is head of the Data Communications Research Department of AT&T Bell Laboratories, Middletown, New Jersey. The late* ***Dimitrios P. Prezas*** *was a supervisor in the Advanced Services Technology Department of AT&T Bell Laboratories (Indian Hill Park), Naperville, Illinois. Mr. Jayant is responsible for research in signal processing, including coding and communication of speech, image, and wideband audio. He joined the company in 1968 and has a Ph.D. in electrical communications engineering from the Indian Institute of Science (Bangalore, India). Mr. Lawrence is responsible for exploratory development of data communications equipment and services. He joined the company in 1974 and* (continued on page 41)

Introduction

This paper is a review of the technology for digital speech coding. First, we discuss traditional telephone speech with a bandwidth of about 3.2 kHz (kilohertz). Then, we turn our attention to higher grade wideband speech with a bandwidth of 7 kHz, and we briefly discuss wideband audio with a bandwidth of 20 kHz.

The bit rate in the digital representation of speech could vary from 2 to 128 kb/s (kilobits per second), depending on the application and on user expectations of signal quality. To describe the performance of a digital coding system, we use several parameters, such as:

- Processing delay
- Tolerance of transmission errors and multiple stages of coding and decoding

Panel 1. Acronyms and Terms

ADPCM	adaptive-differential pulse-code modulation
AM	amplitude modulation
APC	adaptive-predictive coder
APL	analog private line
ATM	automatic-teller machine
AUDIX	audio-information exchange
CADN	cellular access digital network
CCITT	International Telegraph and Telephone Consultative Committee
CD	compact disk
CELP	code-excited linear prediction
centrex	central exchange; a service provided by the local telephone company that permits any telephone extension within a company to call another extension within the company or dial directly to an outside line
codec	coder-decoder
CPE	customer-premises equipment
CTIA	Cellular Technology Industry Association (North America)
DAM	diagnostic acceptability measure; reflects acceptability of speech communication in a multidimensional sense
DCME	digital circuit-multiplication equipment
DDS	digital data service
DRT	diagnostic rhyme test; a measure of word intelligibility
DSI	digital speech interpolation
DSP	digital signal processor
FM	frequency modulation
FX	foreign exchange
G.711	CCITT standard for PCM at 64 kb/s
G.721	CCITT standard for ADPCM at 32 kb/s
G.722	CCITT standard for 7-kHz audio; a 64-kb/s algorithm for ISDN teleconferencing and loudspeaker telephony
G.723	CCITT standard for ADPCM at 24, 32, and 40 kb/s
G.727	draft CCITT standard for ADPCM at 16, 24, 32, and 40 kb/s
G.764	CCITT standard for packet speech transmission
GSM	Group Speciale Mobile (Europe); standards organization for digital cellular radio
HDTV	high-definition television
IACS	integrated access and cross-connect system
INMARSAT	international maritime satellite
ISDN	Integrated Services Digital Network
ISO	International Organization for Standardization
LPC	linear-predictive coding
LD-CELP	low-delay CELP
MFLOP	10^6 floating-point arithmetic operations
modem	modulator-demodulator
MOS	mean opinion score; used for evaluating the performance of coding algorithms
MSAT	mobile satellite
MSE	mean squared error
NSA	National Security Agency (U.S.)
PBX	private-branch exchange
PCM	pulse-code modulation
PSTN	public-switched telephone network
SELP	sum-excited linear prediction
SNR	signal-to-noise ratio
STU	secure telephone unit (STU-II or STU-III)
TASI	time-adaptive speech interpolation
vocoder	voice coder

- Ability to handle nonvoice signals, such as voiceband modem waveforms.

However, the most important descriptors of coder performance are the quality of the digitized speech at a target bit rate and the way the quality diminishes with decreasing bit rate.

Measuring Speech Quality. The measurement of speech quality has been a difficult and long-standing problem. In this paper, we use a subjective rating scale of 1 to 5, wherever possible, to quantify the level of digital speech quality. This is the so-called *mean opinion score*, or MOS scale,[1,2] that is widely used for evaluating coding algorithms for digital telephony. (Panel 1 defines acronyms and terms.)

A score of 4.0 on the MOS scale will signify *high quality*, or *near-transparent* coding. *Network quality* will imply high quality as a necessary condition, but not the only one. It also implies that the speech coder provides further capabilities demanded by the telecommunications network environment.

An MOS of 3.5 will denote *communications quality*. At this level, speech degradation is easily detectable, but not bad enough to impede natural communication.

Finally, *synthetic quality* will imply a signal that is

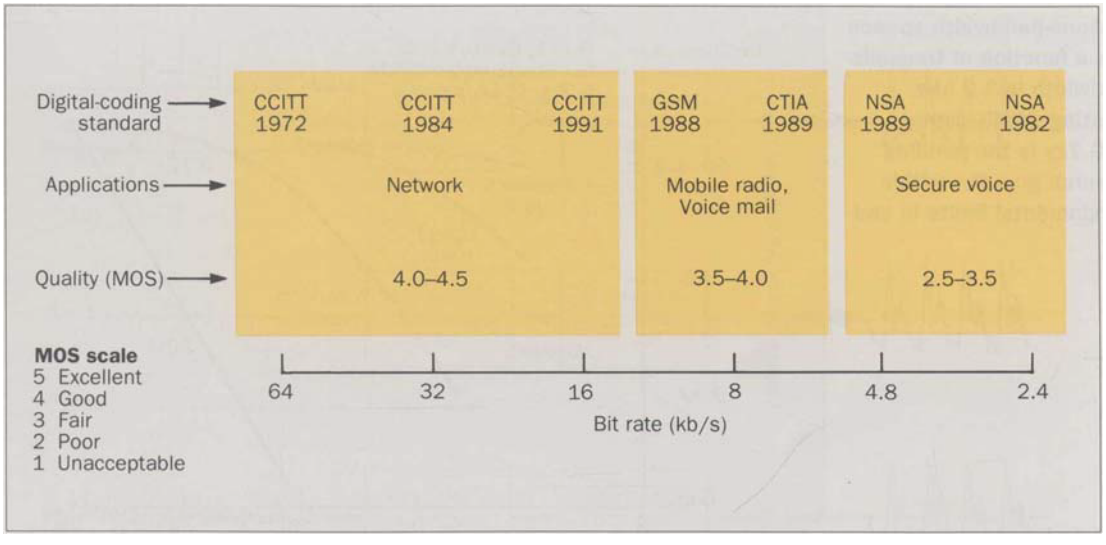

Figure 1. Digital telephony standards, typical applications, and ranges of speech quality (which is expressed using the five-point MOS scale). The frequency range of telephone speech is 200 to 3400 Hz; hence, the speech bandwidth is 3.2 kHz. An MOS score of 4.0 signifies high quality, or near-transparent coding. The coding standards define digital-coding algorithms at the particular bit rate; the 16-kb/s standard is likely to be a hybrid coding algorithm.

characterized by an inadequate level of naturalness and speaker recognizability, although it may have high intelligibility. These deficiencies are usually reflected by an MOS that does not exceed 3.0.

We will also use the five-point MOS scale in our discussion of coding algorithms for 7-kHz speech.

MOS measurements of speech quality are supplemented, especially in low-bit-rate speech technology, by scores of DRT (diagnostic rhyme test) and DAM (diagnostic acceptability measure). The DRT is a word-intelligibility measure, while the DAM reflects acceptability for speech communication in a broader multidimensional sense.[3,4]

Digital Coding of Telephone Speech

Figure 1 describes the current state of telephone speech coding in terms of standards activity, bit rate, typical application, and quality of decoded speech. We assume that the frequency range of the input signal is 200 to 3400 Hz (hertz), and that the quality of the output speech is measured on the five-point MOS scale.

The current goals in speech coding include the achievement of near-transparent or transparent quality at 8 kb/s, and robust telecommunications quality at 4.8 kb/s and lower. (By *robust*, we mean the performance is not degraded drastically across various speech signals, various speakers, and various transmission environments.)

Figure 2 presents a more quantitative description of speech quality as a function of bit rate. The historical progression is from right to left. In the figure, the characteristics depicted as solid curves refer to generic examples of coding algorithms. The dashed curve describes a research goal that is believed to be achievable in that it does not violate fundamental limits in coding capability.

Pulse-code modulation (PCM) is the simplest

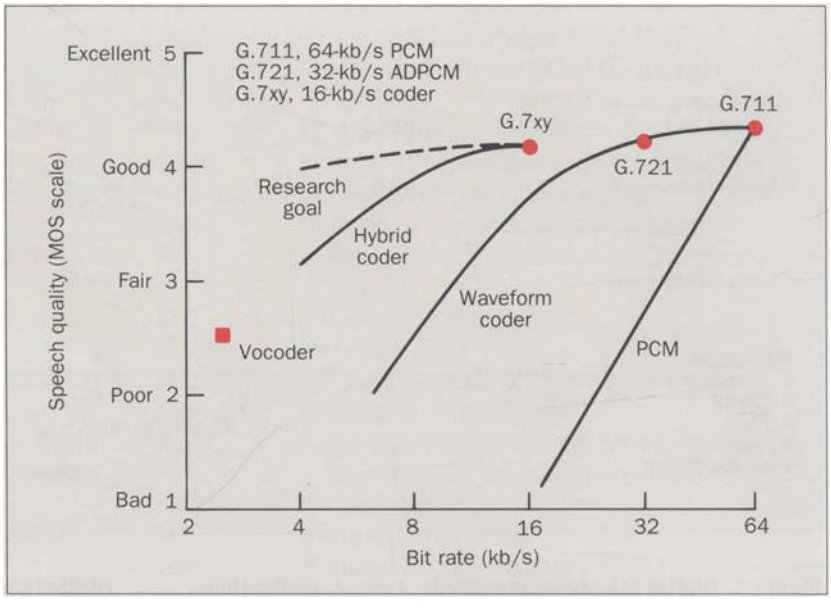

Figure 2. Quality of telephone-bandwidth speech (using the MOS scale) as a function of transmission rate. The signal bandwidth is 3.2 kHz. G.711 and G.721 are existing CCITT digital-coding standards, while G.7xy is the pending CCITT standard. The research goal fits within the constraints of the fundamental limits in coding capability.

coding system, a memoryless quantizer.

The waveform coder curve is that of a high-complexity algorithm, such as adaptive predictive coding. Waveform coding uses redundancy-removing operations to present a signal of lower energy to the amplitude quantizer, which results in a lower bit rate for a specified level of output-speech quality.

The vocoder point represents an algorithm that produces intelligible but synthetic-sounding speech at very low transmission rates by using a highly compact excitation-modulation model (Figure 3a). The synthetic quality in this system is accepted in applications where digital encryption and low transmission rate are of paramount importance. These could include commercial applications, such as banking, but the principal customers, by far, are government and defense agencies.

In Figure 2, the hybrid coder curve describes the performance of a class of algorithms that combine the high-quality potential of waveform coding with the compression efficiency of a model-based vocoder. Here, the idea is to use a time-varying excitation model that is much more sophisticated than that of a traditional vocoder.[5,6] This model uses waveform-coding principles to compute an excitation that minimizes distortion for every frame [say, 16 ms (milliseconds)] of input speech. (See Figures 3b and 3c.) Hybrid coders for 4.8, 8, and 16 kb/s are discussed later.

The solid dots in Figure 2 refer to coding algorithms that provide high quality at 64, 32, and 16 kb/s. Both PCM at 64 kb/s and adaptive-differential PCM (ADPCM) at 32 kb/s are CCITT standards, as defined in Figure 1, and are called G.711 and G.721, respectively. (CCITT is the International Telegraph and Telephone Consultative Committee.) These algorithms provide network-quality coding.

Currently, the CCITT is considering the definition of a high-quality speech standard at 16 kb/s. The technique is likely to be a hybrid coding algorithm.

Figure 2 shows that our current understanding of coding has not yielded high-quality speech at bit rates below about 8 kb/s—in particular, in the important neighborhood of 4 kb/s.

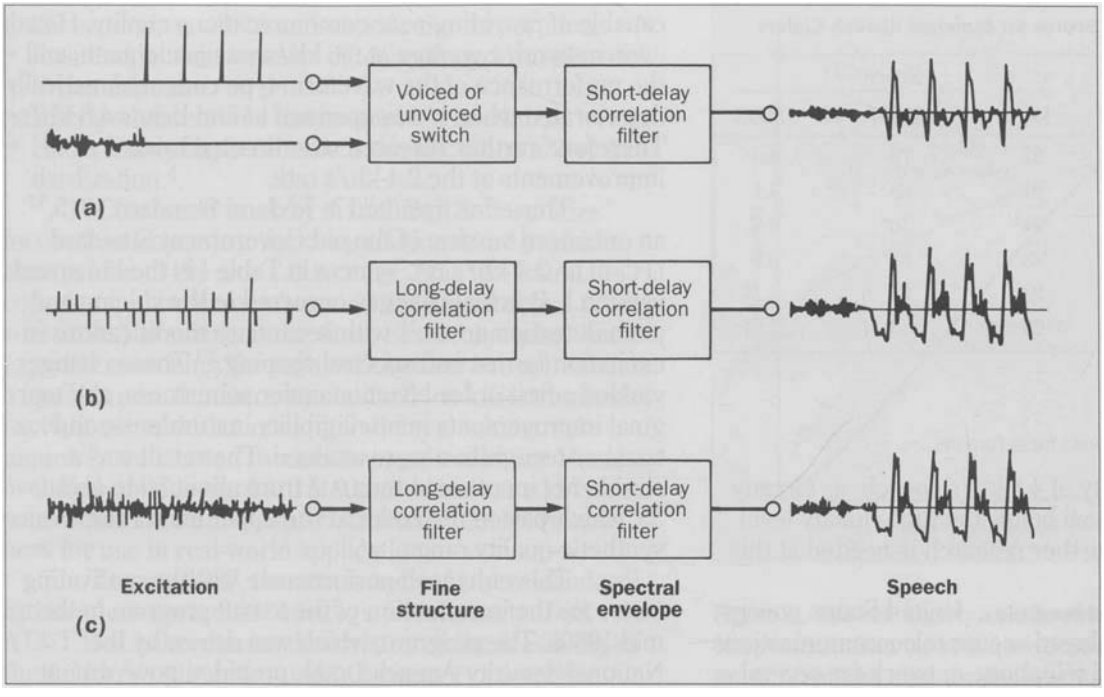

Figure 3. Models of speech excitation. (a) LPC vocoder and hybrid coders, whose performance is given in Figure 2. (b) Multipulse LPC. This coder uses a more sophisticated time-varying excitation model than a traditional vocoder. The excitation computed minimizes distortion for every frame (e.g., 16 ms) of input speech. (c) Codebook-excited LPC. This coder selects the best excitation vector from a codebook of possible vectors.

Digital Telephony at 4 kb/s

High-quality coding at 4 kb/s is a primary focus in speech research.[5-11] Speech coding at bit rates of 4 kb/s is important for:

- Enhancing secure telephony in government and military applications.[12] (See the next section.)
- Providing the central capability of future band-efficient systems for digital radio; for example, cellular channels with a user bandwidth of 5 kHz. These are wireless-access channels for moving vehicles in rural, suburban, or urban environments that are served by terrestrial base stations.
- Mobile-satellite (MSAT) communication applications for providing wireless access to moving vehicles in remote areas.
- INMARSAT (International Maritime Satellite) applications with a 6.4-kb/s target for the total transmission rate (i.e., the speech-coder bit rate plus overhead for channel-error protection).
- Storage of coded speech. When coded at 4 kb/s, one hour of spoken material could be stored on a single 16-Mb (megabit) memory chip.

Table I summarizes DRT, DAM, and MOS performances of various speech coders that range from 64 to 2.4 kb/s.[1,13,14] The numbers in this table are not the result of a single well-controlled experiment, but are accumulated from various independent sources. As such, the scores in the table should be used for a rough, overall indication of performance rather than for very fine comparisons and judgments.

The code-excited linear prediction (CELP) algorithms in Table I typify the hybrid coders depicted in Figure 2. Recent advances in CELP coding have produced significant improvements relative to 2.4-kb/s vocoding. This is reflected in a standardized, 4.8-kb/s algorithm (Figure 1) with high levels of DRT and DAM.

Table I. DRT, DAM and MOS Scores for Standard Speech Coders

Coder	Score		
	DRT	DAM	MOS
64-kb/s PCM	95	73	4.3
32-kb/s ADPCM	94	68	4.1
16-kb/s LD-CELP	94*	70*	4.0
8-kb/s CELP	93†	68*	3.9
4.8-kb/s CELP	93‡	64	3.2†
2.4-kbps LPC (vocoder)	90	53	2.5*

* estimates
† upper bound
‡ lower bound
NOTE: See Panel 1 for definitions for acronyms.

However, the MOS quality of 4.8-kb/s speech, as already noted in Figure 2, falls well below the high-quality level of 4.0, which suggests further research is needed at this important bit rate.

Secure Voice—A Case Study. United States government agencies have deployed secure telecommunications over the public-switched telephone network for several years. As modem technology advanced, highly effective digital encryption techniques became available for such applications. Low-bit-rate voice coding was required to take advantage of these sophisticated methods.

In the early 1980s, the Department of Defense introduced the Government Standard LPC voice-coding algorithm at 2.4 kb/s (see Figure 1).[15] (LPC is *linear predictive coding*.) This vocoder featured a simplified source-excitation model that provided fair speech intelligibility. However, the vocoder and its speech lacked naturalness and robustness, exhibited inconsistent performance across the speaker population, and allowed little, if any, speaker recognition. (That is, the listener usually could not identify the person who was speaking.)

This vocoder technology was incorporated in a limited number of bulky and expensive secure-telephone units (called the STU-II). These units also featured a 9.6-kb/s adaptive-predictive coder (APC)[16] that was capable of providing near-communications quality. However, network coverage at 9.6 kb/s was inadequate, and the performance of the waveform-type coder dramatically deteriorated when it was operated at and below 4.8 kb/s. Therefore, further research was directed toward improvements at the 2.4-kb/s rate.

This effort resulted in Federal Standard 1015,[17] an enhanced version of the old Government Standard LPC. (The 2.4-kb/s LPC system in Table I is the enhanced version.) Primary changes occurred in the voicing and pitch-detection areas,[18] with secondary modifications in excitation format and spectral shaping.[19] These changes yielded a first-order effect on coder robustness, and marginal improvements in intelligibility, naturalness, and speaker recognition were attained. The result was a sizable net increase in the DAM from about 50 to about 55, which placed the coder at the upper end of the synthetic-quality range.

This enhanced performance was the motivating factor for the introduction of the STU-III program in the mid-1980s. The program, which was driven by the National Security Agency (NSA), provided government support for the development of the next-generation secure-telephone units (called the STU-III) and fostered purchase of the units by various government agencies. Compact and cost-effective compared to their predecessors, the STU-IIIs featured the new 2.4-kb/s standard. AT&T's participation in this program resulted in the development of the Security-Plus terminal, which has been in production for the last three years.

The introduction of CELP coders[5] in the mid-1980s made communications quality feasible at 4.8 kb/s. At the same time, new modem technology permitted wide network coverage at this bit rate. Overall coder robustness, naturalness, and speaker recognition far exceeded those of a 2.4-kb/s system. After its introduction by the Acoustics Research Department at AT&T Bell Laboratories (a division of AT&T), CELP coding attracted the attention of numerous potential users, including the NSA. AT&T's ability to develop organiza-

tional synergies, within and outside the corporation:

- Resulted in swift transfer of technology from research to development
- Effectively identified the customer's needs
- Had a timely impact on the NSA's decisions about standardization.[4]

Continued work by AT&T Bell Laboratories focused on improving the computational and performance profile of the new CELP coder. This facilitated the coder's rapid implementation on the digital signal processors (DSPs) available at the time. Fast techniques that expedite searching through codebooks[9] brought about a tenfold reduction in computational load. Constrained excitation[11] and fractional pitch-delay tracking techniques[10] contributed to a net DAM gain of about 10 units over Federal Standard 1015. Algorithms that address source and channel noise[7,20] increased CELP's robustness for use in real-world applications.

During 1987, the NSA launched a new standardization effort toward the 4.8-kb/s rate. In early 1988, AT&T Bell Laboratories demonstrated the feasibility of the 4.8-kb/s CELP coder, using laboratory prototype hardware that reflected the computational capacity of the voice section of the Security-Plus terminal. In mid-1988, AT&T demonstrated, at the NSA, secure-call completion at 4.8 kb/s using CELP-modified Security-Plus terminals. The 4.8-kb/s standardization process, which had been accelerating, peaked by mid-1989 when the NSA issued the first predraft of Federal Standard 1016 and submitted it to the U.S. Office of Standards for approval.[12] (The 4.8-kb/s coder in Table I is this version of the standard.)

Toward the end of 1989, the NSA awarded contracts to vendors for incorporating into security-terminal production a 4.8-kb/s CELP coder that was compatible with Federal Standard 1016. Compatibility requirements permitted shorter codebooks and allowed for optional features to accommodate immediate implementations that used current DSP products. AT&T's early efforts and contributions in the CELP-coding area have facilitated the inclusion of new technology into a preexisting product.

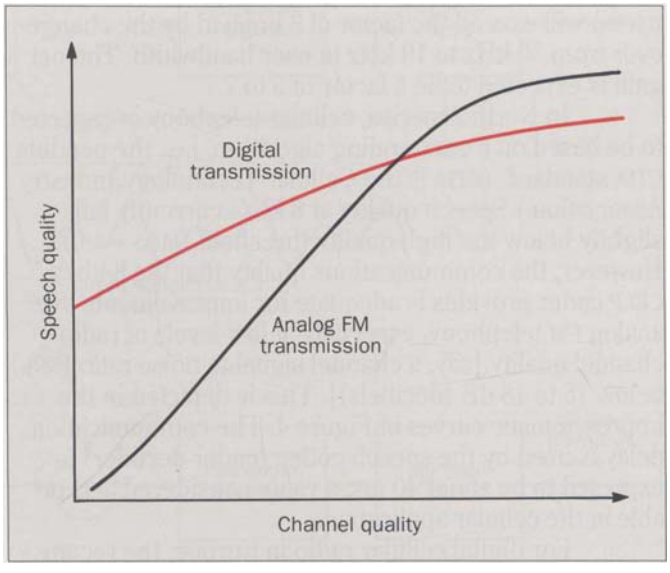

Figure 4. Speech quality versus channel quality in cellular telephony, based on semiquantitative estimates. In the first North American digital-cellular standard (i.e., the pending CTIA standard), speech coding is at 8 kb/s.

AT&T's new version of the Security-Plus terminal, the first of such products to feature a 4.8-kb/s CELP coder, has been in production since August 1990.

Digital Telephony at 8 kb/s

Systems for first-generation digital cellular radio use bit rates of about 8 kb/s for speech coding (Figure 1). In North America, the proposed system has a per-user channel bandwidth of 10 kHz and a total transmission rate of about 13 kb/s for speech coding and channel-error protection.[7,21,22] The system will eventually replace the current practice of analog FM speech that has a 30-kHz user bandwidth. The digital system provides greater robustness to channel noise and fading, as well as better reuse of individual carrier frequencies. As a result, the improvement in call capacity (number of

users) will exceed the factor of 3 implied by the change-over from 30 kHz to 10 kHz in user bandwidth. The net gain is expected to be a factor of 5 to 7.

In North America, cellular telephony is expected to be based on a CELP-coding algorithm, i.e., the pending CTIA standard. (CTIA is the Cellular Technology Industry Association.) Speech quality at 8 kb/s currently falls slightly below the high-quality threshold (MOS = 4.0). However, the communications quality that the 8-kb/s CELP coder provides is adequate for improvements over analog FM telephony, especially at low levels of radio-channel quality [say, a channel signal-to-noise ratio (SNR) below 15 to 18 dB (decibels)]. This is depicted in the impressionistic curves of Figure 4. The communication delay caused by the speech codec (coder-decoder) is expected to be about 40 ms, a value considered acceptable in the cellular application.

For digital cellular radio in Europe, the recommendation of the GSM (Group Speciale Mobile) is also a hybrid coder. It is a regular pulse-excitation algorithm with a bit rate of 13.2 kb/s (out of a total transmission rate of 22.8 kb/s) and a codec delay of 40 ms.[23,24] The coding technique is similar to a 9.6-kb/s multipulse excitation coder for the Skyphone® airline application. (Skyphone is a registered service mark of British Telecommunications, PLC.)

Network-Quality Speech Coding

For ubiquitous application in networks, a speech-coding algorithm has to satisfy several performance criteria, including:

- A level of speech quality that is high enough to withstand multiple stages of coding and decoding
- A processing delay that is low enough to withstand echoes and additional delay components in the network
- The ability to handle nonspeech signals in the telephone band.

PCM and Variable Bit-Rate ADPCM. Algorithms at 64 kb/s (G.711, PCM) and 32 kb/s (G.721, ADPCM) satisfy a broad class of network requirements and are international CCITT standards.[25-28] These standards are in widespread use in both public and private speech telecommunications.

The 32-kb/s standard, G.721, is relatively recent (i.e., 1984). An important application of this codec is in digital circuit-multiplication equipment (DCME). Here, the combination of 2:1 compression (from 64 kb/s to 32 kb/s) with exploitation of the so-called 2.5:1 TASI or DSI gain (i.e., the effect of silences in speech) provides effective circuit expansions of 5:1 over traditional telephone systems. (TASI is *time-adaptive speech interpolation*. The *silences* in natural speech occur when the talker pauses to breathe or collect his or her thoughts or stops speaking and waits for the other person to begin speaking, and when he or she is listening to the other speaker.)

The 32-kb/s algorithm has been extended to 24 and 40 kb/s in the G.723 standard. Also, embedded ADPCM[29] is a draft CCITT standard, G.727, at the 40, 32, 24, and 16-kb/s rates. G.727 can be used with G.764 for wideband packet network applications, such as in AT&T's integrated access and cross-connect system (IACS).[30]

Lower transmission rates such as 24 and 16 kb/s, if based on ADPCM, do not provide network-quality coding. However, they permit occasional bursts of heavy telephone traffic to be accommodated, without explicit coordination among all nodes in the path of the call. Adaptive algorithms can be used for postfiltering[31] at the receiver to enhance the lower speech quality of the 24- and 16-kb/s systems. However, the use of postfiltering adversely affects the performance of a system that has multiple stages of encoding and decoding.

The higher ADPCM rate of 40 kb/s (from the G.723 standard) provides the capability for transmitting 9.6-kb/s modem waveforms. The simplicity of the G.721 algorithm and related algorithms also makes them attractive for wireless-access applications that require very low transmitter power; for example, in terminals that are within or near a building that has indoor wireless communication.[32,33]

The 32-kb/s ADPCM algorithm of the G.721 standard is also robust to (i.e., it can tolerate) multiple stages

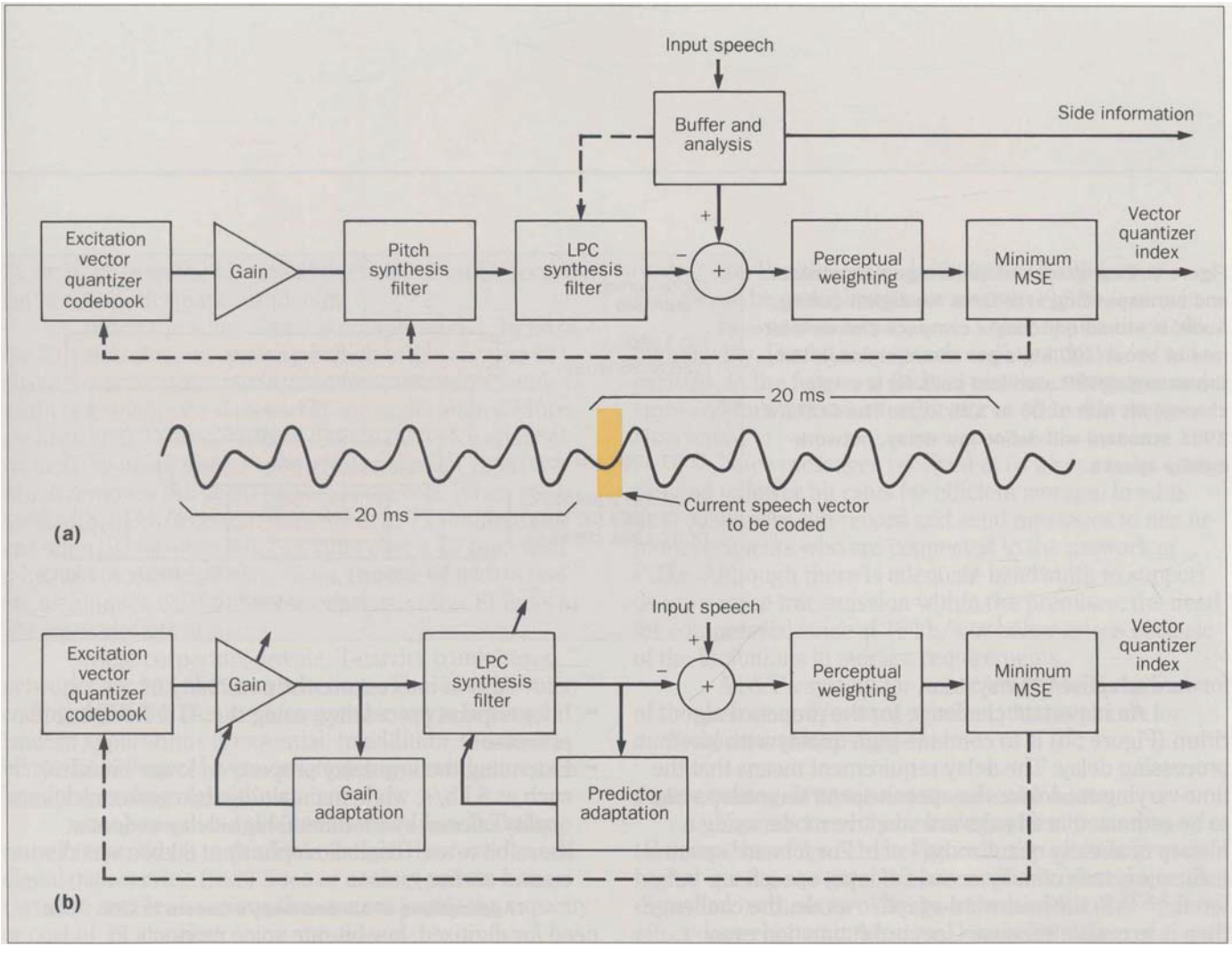

Figure 5. Code-excited linear prediction. The waveform's 20-ms segments are used to perform speech analysis to provide LPC filter coefficients. (a) Conventional or fully forward-adaptive CELP; uses the 20-ms segment on the right in the waveform. (b) Backward-adaptive system for low-delay CELP, the proposed CCITT standard at 16 kb/s; uses the left 20-ms segment of the waveform. In this system, the forward adaptation of the shape of the excitation signal is the only source of encoding delay.

of encoding and decoding, and more robust than 64-kb/s PCM to digital errors in transmission. At a bit-error rate of 1 in 1000, the degradation in speech quality for 32-kb/s ADPCM is graceful. At a bit-error rate of 1 in 100, speech intelligibility is good, although the quality is poor.

Low-Delay Speech Coding at 16 kb/s. Currently, the CCITT is considering the definition of a low-delay network-quality speech standard at 16 kb/s (Figure 1). Possible applications include DCME, ISDN transmission, packetized speech, cordless telephones, and speech for videophone service. (ISDN is the Integrated Services Digital Network.)

Figure 5b is the block diagram of a backward-adaptive CELP coder proposed by AT&T for this CCITT standard.[34,35] In this system, the only source of encoding delay is the forward adaptation of the shape of the excitation signal. This delay comes from selecting the best excitation vector from a rich codebook of possible excitation vectors, each a sample vector of length 5 (0.625 ms). Figure 5a is the more traditional, fully

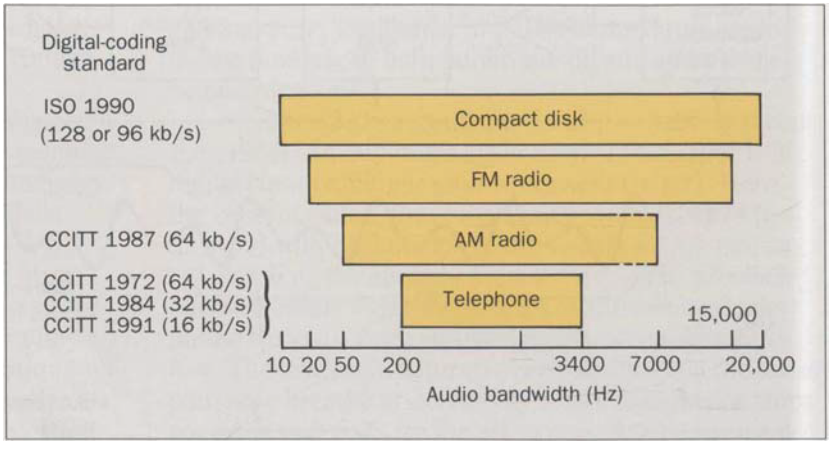

Figure 6. Four grades of audio-signal bandwidth and corresponding standards for digital coding. Audio is stored on today's compact disk at a bit rate of about 700 kb/s per sound channel, but the emerging ISO standard calls for a single-channel bit rate of 96 or 128 kb/s. The CCITT's 1991 standard will define low-delay, network-quality speech.

forward-adaptive CELP system.

An important challenge for the proposed algorithm (Figure 5b) is to combine high quality with low processing delay. The delay requirement means that the time-varying model for the speech-spectral envelope has to be estimated in a backward-adaptive mode, using a history of already quantized speech. For forward spectral estimation, tens of milliseconds of input speech are buffered.[5,6,22] In the backward-adaptive mode, the challenge then is to realize adequate spectral estimation even though quantization noise is present in the past speech samples used for backward spectral analysis.

The algorithm is complex, with the codebook search as the single, most demanding component. A 25-MFLOP processor with advanced memory capabilities (i.e., the AT&T WE® DSP32C processor) is available, which permits a single-chip (half-duplex) implementation of the coder. Currently, a full-duplex coder requires a two-chip implementation, but prospects for a single-chip implementation with nearly equal speech quality are good. (MFLOP stands for 10^6 floating-point arithmetic operations. In *full duplex* transmissions, data is transmitted and received simultaneously. With *half duplex*, data can be transmitted and received, but only in one direction at a time.)

Also of interest are the possibility of:

- Integer-point processing, using the AT&T WE DSP16A processor.
- Extending the low-delay property to lower bit rates, such as 8 kb/s, while maintaining the communications quality offered by traditional high-delay coders at those bit rates. (Digital telephony at 8 kb/s was discussed earlier.)

Applications of 16- and 8-kb/s Coders in CPE. The need for digitized, low-bit-rate voice products in customer-premises equipment (CPE) is expected to increase, as will the use of digital transmission facilities for integrated voice and data services. The use of low-bit-rate voice will also grow because of the demand for store-and-forward voice mail and for voice-security applications. Speech coders at 16 and 8 kb/s are prime candidates for CPE applications.

Intelligent T1 multiplexers. Today, several vendors are offering intelligent T1 or fractional-T1 multiplexers for large, corporate, T-carrier trunk-based networks. These networks are complete telecommunications systems that carry both voice and data traffic. For these networks, the economies of scale and the dynamic reallocation of bandwidth offer potential cost savings.

In many of these applications, users select 64, 32,

24, or 16 kb/s as the bit rate of the voice circuits, according to cost-performance tradeoffs.

When the voice signal is compressed to 16 kb/s, the T1 voice-channel capacity (which originally was 24 channels) is increased to 96 channels, and extra bandwidth is available for data and image applications. More sophisticated T1 multiplexers double the voice-channel capacity by using digital speech interpolation (DSI), which removes the silent pauses in speech. When DSI is used with 16-kb/s compressed speech, T1 multiplexers can offer 192 or more voice circuits over a T1 link, with minimal voice degradation. Soon, the use of 8-kb/s coding techniques will double the capacity of the T1 links to 384 voice circuits or more.

Most corporate, private, T-carrier trunk-based networks are PBX to PBX connections. Therefore, a voice-coding algorithm that performs well in asynchronous tandem applications is essential. In addition, the desirability of avoiding echo cancellation suggests the use of a low-delay coding algorithm.

Compressed voice over APL and DDS circuits. Today, some CPE products multiplex voice and data over leased, digital-data-service (DDS) lines or analog, private lines (APLs) to smaller locations that cannot justify the capacity or cost of T1 circuits.

The DDS systems can be configured to provide 56-kb/s service for multiple 8-kb/s voice channels and 16-kb/s data channels between PBX, centrex, or FX locations. These multiplexers are especially needed in international circuits. Because such circuits are expensive, users normally like to multiplex as many voice connections as possible onto a single circuit [for example, five voice channels plus one data channel; i.e., $(5 \times 8$ kb/s) + $(1 \times 16$ kb/s) = 56 kb/s]. For international applications that use satellite links, the delay and echo characteristics associated with the links make a low-delay, high-quality, compressed-voice algorithm highly desirable.

Another application for low-bit-rate speech is for automatic-teller machines (ATMs). In this application, a high-speed, 19.2-kb/s APL circuit connects each ATM to a central site. Besides voice, both data and still-frame images can be multiplexed onto a single 19.2-kb/s circuit.

Store-and-forward voice mail. Other applications of high-quality, low-bit-rate speech coding at 16 kb/s (and perhaps, in the future, at 8 kb/s) are the call-answer and store-and-forward voice mail features offered in many PBXs today.

Voice messages received at 64 kb/s can be compressed to lower bit rates for efficient storage. In addition, customers can record and send messages to one or more recipients who are connected to the network of PBXs. Although there is adequate bandwidth to support 64-kb/s voice transmission within the premises, the need for compressed voice at 16 kb/s or below arises because of the limitations in storage requirements.

AT&T's system for AUDIX Voice Mail is typical of these store-and-forward services. (AUDIX stands for *audio-information exchange.*)

Digital Coding of Wideband Speech and Audio

Figures 1 and 2 referred specifically to telephone-band speech. In Figure 2, the achievement of higher quality at a given bit rate implied reduced speech distortion, without any change of bandwidth. But what effect does changing the signal bandwidth have on speech quality and intelligibility?

Figure 6 defines four commonly understood grades of audio bandwidth. If the audio signal is speech instead of music, the perceived gains in quality are, perhaps, greatest when one progresses from the telephone level to the commentary, or AM-radio, level. The gains in quality are in terms of increased intelligibility, naturalness, and speaker recognition. Low-frequency enhancement (i.e., 50 to 200 Hz) contributes to increased naturalness and speaker presence, and high-frequency enhancement (i.e., 3400 to 7000 Hz) provides greater intelligibility and fricative differentiation (for example, *s* versus *f*).

In the rest of this section, we describe high-quality compression of wideband audio and ISDN

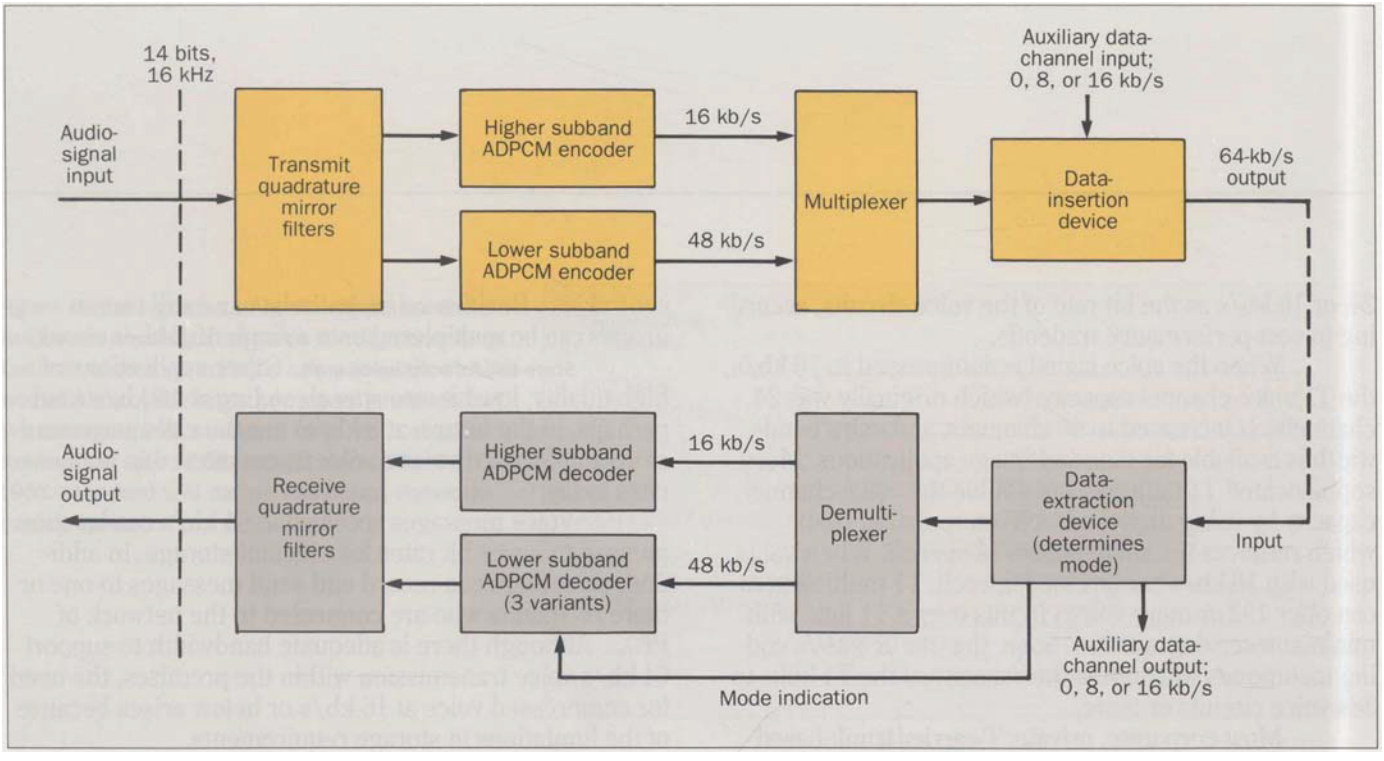

Figure 7. Block diagram of a two-band subband coder for 64-kb/s coding of 7-kHz audio,[37] the basis for the G.722 standard. The low- and high-frequency subbands are quantized using 6 and 2 bits per sample, respectively. The analysis and synthesis filters produce a communication delay of about 3 ms.

applications of digital audio. We also discuss a CCITT coding standard for 7-kHz audio and a 20-kHz audio standard that is being defined by the ISO (International Organization for Standardization).

The naturalness of wideband speech is a significant feature for extended telecommunications processes, such as audio teleconferencing and program broadcasting. Basic-rate ISDN provides a natural framework for a 64-kb/s algorithm to encode wideband audio for such applications. [Basic-rate ISDN provides two 64-kb/s circuit-switched channels (bearer channels for the customer's voice, data, or video) and one 16-kb/s packet-switched channel (data channel for the network's information).] The digital connectivity afforded by ISDN[36] has prompted a worldwide revisiting of audio-transmission quality. In particular, end-to-end digital connectivity has made possible the inclusion of low frequencies down to 50 Hz in the transmitted audio band.[37]

Coding of 7-kHz Audio. The CCITT standard for 7-kHz audio (G.722) is a 64-kb/s algorithm developed primarily for ISDN teleconferencing and loudspeaker telephony. Because of the 64-kb/s capability, a single "voice-grade" channel on a digital or analog, public-switched telephone network (PSTN) can transport a commentary-quality sound program over any distance and yield a broadcast-grade voice program at the receiving end.

The G.722-coding algorithm is based on a two-band subband coder, with ADPCM coding of each subband (Figure 7).[37,38] The low- and high-frequency subbands are quantized using 6 and 2 bits per sample, respectively. The filter banks that are used for analysis and synthesis produce a communication delay of about 3 ms. This delay turns out to be a desirable feature because of the expected interconnections of G.722 with narrowband links. For these interconnections, uncanceled echoes could pose a problem, if compounded by codec delay. (In isolation, digital wideband links do not have two-wire/four-wire hybrids and the resulting uncanceled echoes.)

The 64-kb/s algorithm can tolerate random error

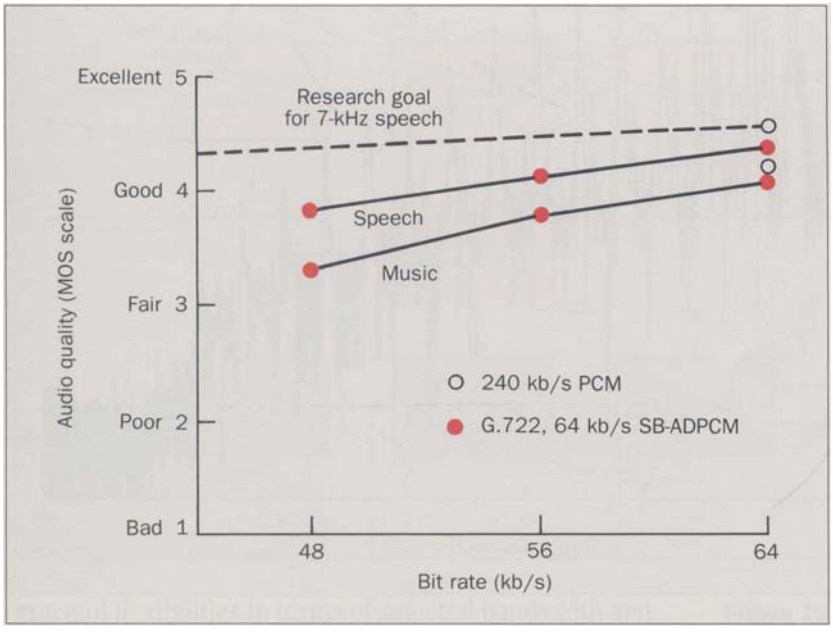

Figure 8. Quality of 7-kHz digital audio as a function of bit rate in the G.722 algorithm.[37] Signal bandwidth is 7 kHz. The G.722 points are for the two-band subband coder (Figure 7), which uses ADPCM coding for each subband. The PCM points are supplied for comparison and represent a 15-bit audio input sampled at 16 kHz. Again, the research goal is realistic.

rates of about 1 in 10,000 and four tandem stages of repeated encoding and decoding. The simplicity of the quantizing, predicting, and filtering (24-tap) algorithms permits a single-chip, fixed-point implementation on the DSP16A processor.

ISDN applications suggest that the audio-coding algorithms be operated at slightly lower bit rates. Here, the use of an embedded coding technique for ADPCM permits operation of the low-frequency subband at one of three quantizing rates (i.e., 6, 5, or 4 bits per sample), with graceful degradation of quality. The corresponding audio rates are 64, 56, and 48 kb/s. For the 56- and 48-kb/s rates, capacities of 8 and 16 kb/s are available for simultaneous data transmission over the 64-kb/s basic-rate channel.

Figure 8 shows audio quality on the MOS scale for speech and music material at rates of 64, 56, and 48 kb/s. For comparison, we also show the performance of 240-kb/s linear PCM (i.e., a 15-bit audio input that was sampled at 16 kHz). In another comparison that involves G.722, G.721, and 128-kb/s PCM (16 kHz × 8 bits per sample), the G.722 algorithm at 64 kb/s was shown to have an equivalent SNR gain of 13 dB over the G.721 algorithm.[38,39] Of this SNR gain, 6 dB can be attributed to increased input bandwidth.

Figure 8 also shows the current research goal for the coding of 7-kHz audio, a goal that is believed to be realistic. One implication of this goal is the possibility of coding 7-kHz audio at 32 kb/s with high quality (MOS = 4.0). This will permit the transmission of two bilingual or stereo wideband channels at 64 kb/s. For stereo, the use of cross-channel correlations can provide a further increase of capability. For example, a bandwidth greater than 7 kHz could be accommodated in the 64-kb/s system.

There are at least two approaches to the problem of high-quality coding of audio at 32 kb/s:

- Linear-prediction approach, exemplified by CELP

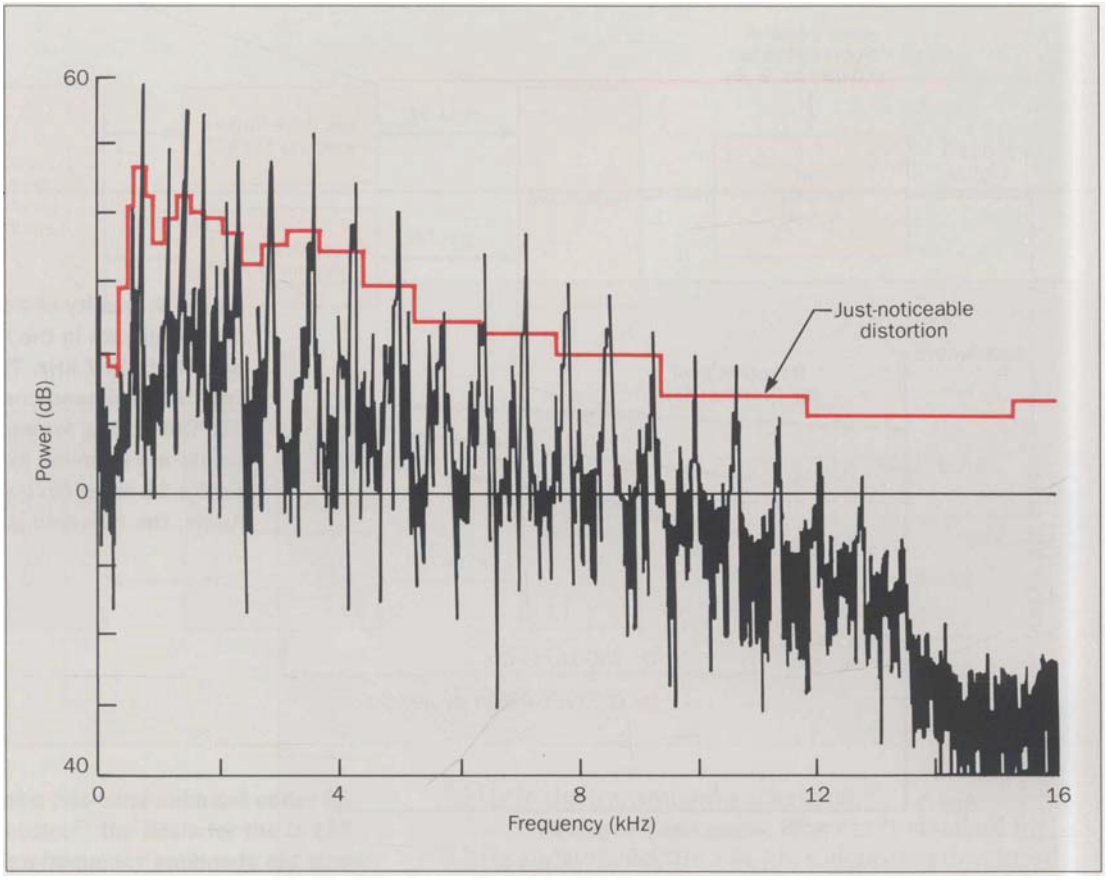

Figure 9. Threshold of just-noticeable distortion as a function of frequency for an illustrative audio signal (a trumpet).[40] Research on perception will play an increasing role by enhancing our understanding of how to mask noise, especially in the time domain.

- Frequency-domain technique of transform or subband coding.

For both, the attainment of high audio quality will depend on the use of perceptual tuning of the algorithm to provide effective shaping of the quantization noise. The CELP technique offers the additional possibility of low-delay coding through backward adaptation, as illustrated in Figure 5b.

Coding of 20-kHz Audio. Although a bandwidth of 7 kHz provides very natural reproduction of speech, 20 kHz is a well-accepted bandwidth standard for more general classes of audio, including vocal and instrumental music.

The ISO is committed to the standardization (in the 1990 to 1991 time frame) of a low-bit-rate coding algorithm for 20-kHz audio. Applications for low-bit-rate wideband speech include electronic publishing, travel and guidance, teleteaching, multilocation games, multimedia memoranda, and database storage. Another major application for 20-kHz digital audio is in advanced television systems, such as high-definition television (HDTV).

On current compact disks (CDs), audio is stored at a bit rate of about 700 kb/s per sound channel (i.e., 16-bit PCM coding of 44.1-kHz sampled signals). However, the emerging ISO standard calls for a single-channel bit rate of 96 or 128 kb/s. Production of high-quality audio at these very low rates calls for a new generation of coding algorithms. These algorithms will achieve coding gains by removing signal redundancy. But these gains must be augmented by the liberties permitted by the human auditory process, as predicted by sophisticated models of just-noticeable distortion (Figures 9 and 10).[40,41]

Future Trends

Sophisticated algorithms for coding will lead to transmission techniques that do not permit quantization noise to limit speech quality. In addition, the notion of enhancing speech quality by using greater input bandwidth will become more pervasive. Coding systems in the 8- to 64-kb/s range will thus provide

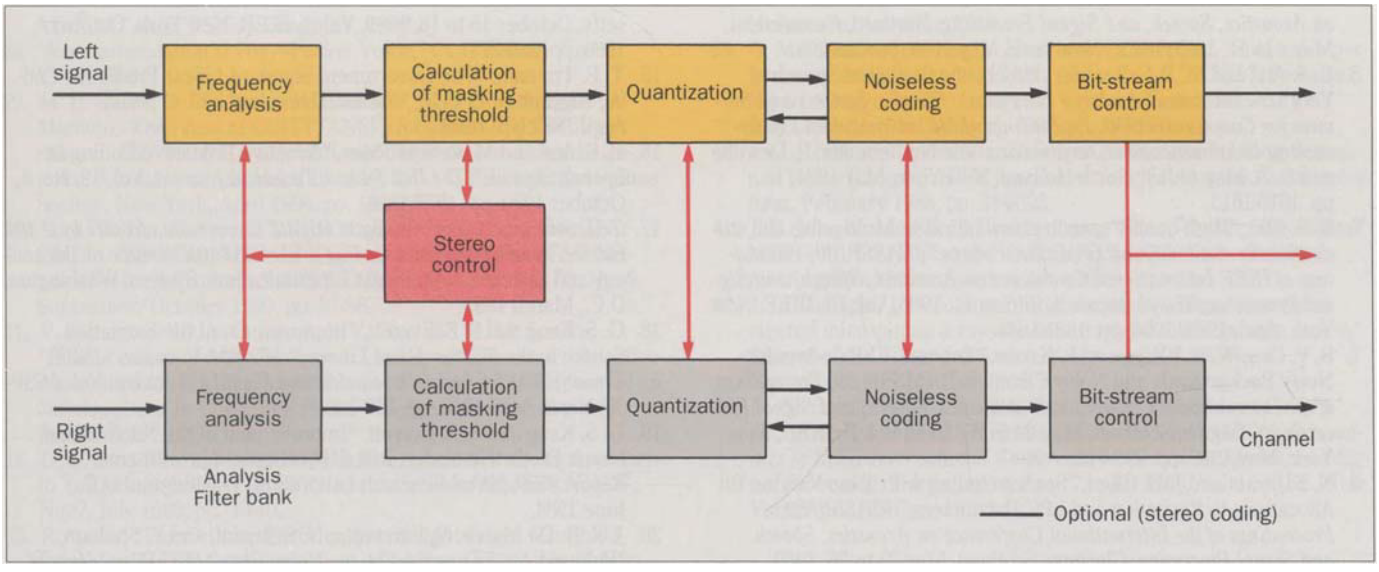

Figure 10. Perceptual coding of wideband audio; block diagram of a perceptual frequency-domain coder.[41] The dashed lines identify an option for using left-right channel correlations to increase efficiency in stereo coding. For high-quality audio at low bit rates, the liberties permitted by human perception must augment the coding gains achieved from removing signal redundancy.

graceful flexibilities in terms of selected bandwidth and special features, such as stereo separation in teleconferencing. Advances in coding will be supported by new technologies for wideband transducers, noise-canceling systems for audio pickup, and autodirective microphone arrays.[42,43]

As coding algorithms become increasingly efficient and approach fundamental capabilities, research on perception will play an increasing role by enhancing our understanding of noise masking, especially in the time domain.

Advances in signal-processor technology will continue to support increasingly complex algorithms for coding and decoding. The synergistic working of coding theory, perception science, and signal processing will bring sophisticated speech technology to the human listener in affordable forms.

Acknowledgment

We thank the following colleagues for reviewing an earlier version of this paper: B. S. Atal, K. H. Brandenburg, J.-H. Chen, R. V. Cox, J. D. Johnston, W. B. Kleijn, D. J. Krasinski, P. Noll, M. H. Sherif, and Y. Shoham.

References

1. W. R. Daumer, "Subjective Evaluation of Several Efficient Speech Coders," *IEEE Transactions on Communications*, Vol. 30, No. 4, April 1982, pp. 655-662.
2. N. S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*, Prentice Hall, Englewood Cliffs, New Jersey, 1984.
3. W. D. Voiers, "Diagnostic Evaluation of Speech Intelligibility," *Speech Intelligibility and Speaker Recognition*, M. E. Hawley (ed.), Dowden Hutchinson Ross, Stroudsburg, Pennsylvania, 1977.
4. W. D. Voiers, "Diagnostic Acceptability Measure for Speech Communication Systems," *ICASSP '77, IEEE International Conference*

on Acoustics, Speech, and Signal Processing, Hartford, Connecticut, May 9 to 11, 1977, IEEE, New York, May 1977, pp. 204-207.
5. B. S. Atal and M. R. Schroeder, "Stochastic Coding of Speech at Very Low Bit Rates," *Links for the Future: Science, Systems and Services for Communications, Proceedings of the International Conference on Communications*, Amsterdam, The Netherlands, P. Dewilde and C. A. May (eds.), North-Holland, New York, May 1984, pp. 1610-1613.
6. B. S. Atal, "High-quality speech at low bit rates: Multi-pulse and stochastically excited linear predictive coders," *ICASSP '86, Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Tokyo, Japan, April 7 to 11, 1986, Vol. III, IEEE, New York, April 1986, 1986, pp. 1681-1684.
7. R. V. Cox, W. B. Kleijn, and P. Kroon, "Robust CELP Coders for Noisy Backgrounds and Noisy Channels," *ICASSP '89, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Glasgow, Scotland, May 23 to 26, 1989, Vol. II, IEEE, New York, May 1989, pp. 739-742.
8. N. S. Jayant and J.-H. Chen, "Speech Coding with Time-Varying Bit Allocations to Excitation and LPC Parameters," *ICASSP '89, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Glasgow, Scotland, May 23 to 26, 1989, Vol. I, IEEE, New York, May 1989, pp. 65-68.
9. W. B. Kleijn, D. J. Krasinski, and R. H. Ketchum, "Improved Speech Quality and Efficient Vector Quantization in SELP," *ICASSP '88, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, New York, April 11 to 14, 1988, Vol. I, IEEE, New York, April 1988, pp. 155-158.
10. P. Kroon and B. S. Atal, "Pitch Predictors with High Temporal Resolution," *ICASSP '90, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Albuquerque, New Mexico, April 3 to 6, 1990, Vol. II, IEEE, New York, April 1990, pp. 661-664.
11. Y. Shoham, "Constrained Excitation CELP Coding," *IEEE Workshop on Speech Coding for Telecommunications*, Vancouver, British Columbia, Canada, September 1989, p. 65.
12. *Telecommunications: Analog to Digital Conversion of Radio Voice by 4800 bit/sec. Code Excited Linear Prediction (CELP)*, FED-STD-1016, Second Draft, Office of Technology and Standards, National Communications System, Washington, D.C., November 1989.
13. D. P. Kemp, R. A. Sueda, and T. E. Tremain, "An Evaluation of 4800 bps Voice Coders," *ICASSP '89, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Glasgow, Scotland, May 23 to 26, 1989, Vol. I, IEEE, New York, May 1989, pp. 200-203.
14. V. C. Welch, T. E. Tremain, and J. P. Campbell, "A Comparison of U.S. Government Standard Voice Coders," *MILCOM 89, Bridging the Gap: Interoperability, Survivability, Security*, Conference record, IEEE Military Communications Conference, Boston, Massachusetts, October 15 to 18, 1989, Vol. 1, IEEE, New York, October 1989, pp. 269-273.
15. T. E. Tremain, "The Government Standard Linear Predictive Coding Algorithm: LPC-10," *Speech Technology*, Vol. 1, No. 2, April 1982, pp. 40-49.
16. B. S. Atal and M. R. Schroeder, "Adaptive Predictive Coding of Speech Signals," *The Bell System Technical Journal*, Vol. 49, No. 8, October 1970, pp. 1973-1986.
17. *Telecommunications: Analog to Digital Conversion of Voice by 2,400 Bit/sec. Linear Predictive Coding*, FED-STD-1015, Office of Technology and Standards, National Communications System, Washington, D.C., March 1983.
18. G. S. Kang and S. S. Everett, "Improvement of the Excitation Source in the Narrow-Band Linear Prediction Vocoder," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-33, No. 2, April 1985, pp. 317-386.
19. G. S. Kang and S. S. Everett, "Improvement of the Narrowband Linear Predictive Coder, Part 2: Synthesis Improvements," Report 8799, Naval Research Laboratory, Washington, D.C., June 1984.
20. J. R. B. De Marca, N. Farvardin, N. S. Jayant, and Y. Shoham, "Robust Vector Quantization for Noisy Channels," *Proceedings of the M-SAT Conference*, Jet Propulsion Laboratories, Pasadena, California, May 1988, pp. 515-520.
21. E. S. K. Chien, D. J. Goodman and J. E. Russell, "Cellular Access Digital Network (CADN): Wireless Access to Networks of the Future," *IEEE Communications Magazine*, Vol. 25, No. 6, June 1987, pp. 22-27.
22. I. A. Gerson and M. A. Jasiuk, "Vector Sum Excited Linear Prediction (VSELP)," *IEEE Workshop on Speech Coding for Telecommunications*, Vancouver, British Columbia, Canada, September 1989, pp. 66-69.
23. P. Kroon, E. F. Deprettere, and R. J. Sluyter, "Regular-Pulse Excitation—A Novel Approach to Effective and Efficient Multipulse Coding of Speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-34, No. 5, October 1986, pp. 1054-1063.
24. P. Vary, K. Hellwig, R. Hofmann, R. J. Sluyter, C. Galand, and M. Rosso, "Speech Codec for the European Mobile Radio System," *ICASSP '88, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, New York, April 11 to 14, 1988, Vol. I, IEEE, New York, April 1988, pp. 227-230.
25. M. Taka and X. Maitre, "CCITT Standardizing Activities on Speech coding," *ICASSP '86, Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Tokyo, Japan, April 7 to 11, 1986, Vol. II, IEEE, New York, April 1986, pp. 817-820.
26. "G.721—32 kbits/s Adaptive Differential Pulse Code Modulation (ADPCM)," Report R57, Part II, CCITT Study Group XVIII, IXth Plenary Assembly, Melbourne, Australia, 1988.
27. "Recommendation G.727—5, 4, 3, 2 bit per Sample Embedded

ADPCM," CCITT Study Group XV, 1990.
28. "Recommendation G.764—Packet Voice," CCITT Study Group XVIII, 1990.
29. M. H. Sherif, D. O. Bowker, G. Bertocci, B. A. Orford, and G. A. Mariano, "Overview of CCITT/ANSI Embedded ADPCM Algorithms," *ICC '90, IEEE International Conference on Communications,*, Atlanta, Georgia, April 16 to 19, 1990, IEEE Communications Society, New York, April 1990, pp. 1014-1018.
30. M. K. Verma, D. Prezas, T. L. Russell, M. H. Sherif, and R. Thorkildsen, "Novel Applications of Speech Processing in AT&T Network Systems Products," *AT&T Technical Journal*, Vol. 69, No. 5, September/October 1990, pp. 77-86.
31. V. Ramamoorthy, N. S. Jayant, R. V. Cox, and M. M. Sondhi, "Enhancement of ADPCM Speech Coding with Backward-Adaptive Algorithms for Postfiltering and Noise Feedback," *IEEE Journal on Selected Areas in Communications*, Vol. 6, No. 2, February 1988, pp. 364-382.
32. D. C. Cox, "Portable Digital Radio Communications—An Approach to Tetherless Access," *IEEE Communications Magazine*, Vol. 27, No. 7, July 1989, pp. 30-40.
33. R. Steele, "The Cellular Environment of Lightweight Handheld Portables," *IEEE Communications Magazine*, Vol. 27, No. 7, July 1989, pp. 20-29.
34. J.-H. Chen, "A Robust Low Delay CELP Speech Coder at 16 kbps," *Communications Technology for the 1990s and Beyond*, Globecom '89, 8th IEEE Global Telecommunications Conference, Dallas, Texas, November 27 to 30, 1989, IEEE, New York, 1989, pp. 3411-3415.
35. J.-H. Chen, "High Quality 16 kbps speech coding with a one-way delay less than 2 ms," *ICASSP '90, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Albuquerque, New Mexico, April 3 to 6, 1990, Vol. I, IEEE, New York, April 1990, pp. 453-456.
36. T. Irmer, "An Idea Turns Into a Reality—CCITT Activities on the Way to ISDN," *IEEE Journal on Selected Areas in Communications*, May 1986, pp. 316-319.
37. P. Mermelstein, "G.722, A New CCITT Coding Standard for Digital Transmission of Wideband Audio Signals," *IEEE Communications Magazine*, Vol. 26, No. 1, January 1988, pp. 8-15.
38. "G.722—7 kHz Audio Coding Within 64 kbits/s," Report R57, Part II, CCITT Study Group XVIII, IXth Plenary Assembly, Melbourne, Australia, 1988.
39. G. Modena, A. Coleman, P. Usai, and P. Coverdale, "Subjective performance evaluation of the 7 kHz Audio Coder," *CSELT Technical Report* (Centro Studi e Laboratori Telecommunicazion, Turino, Italy), Vol. 15, No. 2, March 1987, pp. 171-176.
40. J. D. Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria," *IEEE Journal on Selected Areas in Communications*, February 1988, pp. 314-323.
41. J. D. Johnston and K. H. Brandenburg, "Sound Coding Algorithm," MPEG-891-148, Report of ISO-IEC/JTCI/SC2/WG8 committee meeting, Stockholm, Sweden, June 1989.
42. J. L. Flanagan, J. D. Johnston, R. Zahn, and G. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *Journal of the Acoustical Society of America*, Vol. 78, No. 5, November 1985, pp. 1508-1518.
43. M. M. Sondhi and G. W. Elko, "Adaptive Optimization of Microphone Arrays under a Nonlinear Constraint," *ICASSP '86, Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Tokyo, Japan, April 7 to 11, 1986, Vol. II, IEEE, New York, April 1986, pp. 19.9.1-19.9.4.

Biographies (continued)

has a B.Sc. in electrical engineering from London University (England), a D.I.C. from Imperial College of Science and Technology (London, England), and a Ph.D. in electrical engineering from London University. Mr. Prezas was responsible for development of speech coding and automatic speech recognition technologies for intelligent network applications and secure voice equipment. He joined the company in 1979 and had a B.S. in physics from the University of Athens, Greece, and an M.S. and Ph.D. in electrical engineering from the Illinois Institute of Technology in Chicago.

(Manuscript received June 14, 1990)