SPEECH CODING

Speech is the predominant means of communication between human beings, and since the invention of the telephone by Alexander Graham Bell in 1876, speech services have been the core of almost all telecommunication systems. The original analog methods of telephony had the disadvantage that the speech signal was corrupted by noise, crosstalk, and distortion. Long-haul transmissions that use repeaters to compensate for the loss in signal strength on transmission links also increase the associated noise and distortion. On the other hand, digital transmission is relatively immune to noise, crosstalk, and distortion, primarily because of its capability to faithfully regenerate digital signal at each repeater on the sole basis of binary decisions. Hence the end-to-end performance of a digital link essentially becomes independent of the length and operating frequency bands of the link. Hence, from a *transmission* point of view, digital transmission has been the preferred approach due to its higher immunity to noise.

The need for digital speech transmission has become extremely important from a *service provision* point of view as well. Modern requirements have introduced the need for robust, flexible, and secure services that can carry a multitude of signal types (such as voice, data, and video) without a fundamental change in infrastructure. Such a requirement could not have been easily met without the advent of digital transmission systems, which require speech to be coded digitally.

The term “speech coding” is often used for techniques that represent, or *code,* speech signals, either directly as a waveform or as a set of parameters by analyzing the speech signal. In either case, the codes are transmitted to the distant end, where the speech is reconstructed, or *synthesized,* using the received set of codes. A more generic term that is applicable to these techniques and that is often used interchangeably with “speech coding” is “voice coding.” This term is more generic in the sense that the coding techniques are equally applicable to any voice signal, whether or not it carries any intelligible information, as the term “speech” implies. Other terms that are commonly used are “speech compression” and “voice compression,” since the fundamental idea behind speech coding is to reduce (compress) the transmission rate (or equivalently the bandwidth) and/or reduce storage requirements. In this article “speech” and “voice” will be used interchangeably.

Digital Speech: An Introduction

Speech, which is a continuous-time, continuous-amplitude signal, is typically sampled at a fixed rate, and each sample is represented (or *quantized,* in speech-processing terminology) using a certain number of bits. Sampling is the process of converting a continuous-time signal such as speech to a discrete-time signal. The duration between samples—or, inversely, the sampling rate—that is used for speech is governed by Nyquist sampling theorem, which depends upon the speech spectral characteristics as described below. It is observed that speech energy falls off rapidly after 4 kHz and that the intelligibility and practically all of the naturalness and talker peculiarities are present in the speech spectrum below 3.5 kHz. Thus according to the Nyquist sampling theorem, if speech is filtered by a sharp-cutoff analog filter prior to sampling so that the maximum frequency component is 4

J. Webster (ed.), *Wiley Encyclopedia of Electrical and Electronics Engineering.* Copyright © 1999 John Wiley & Sons, Inc.

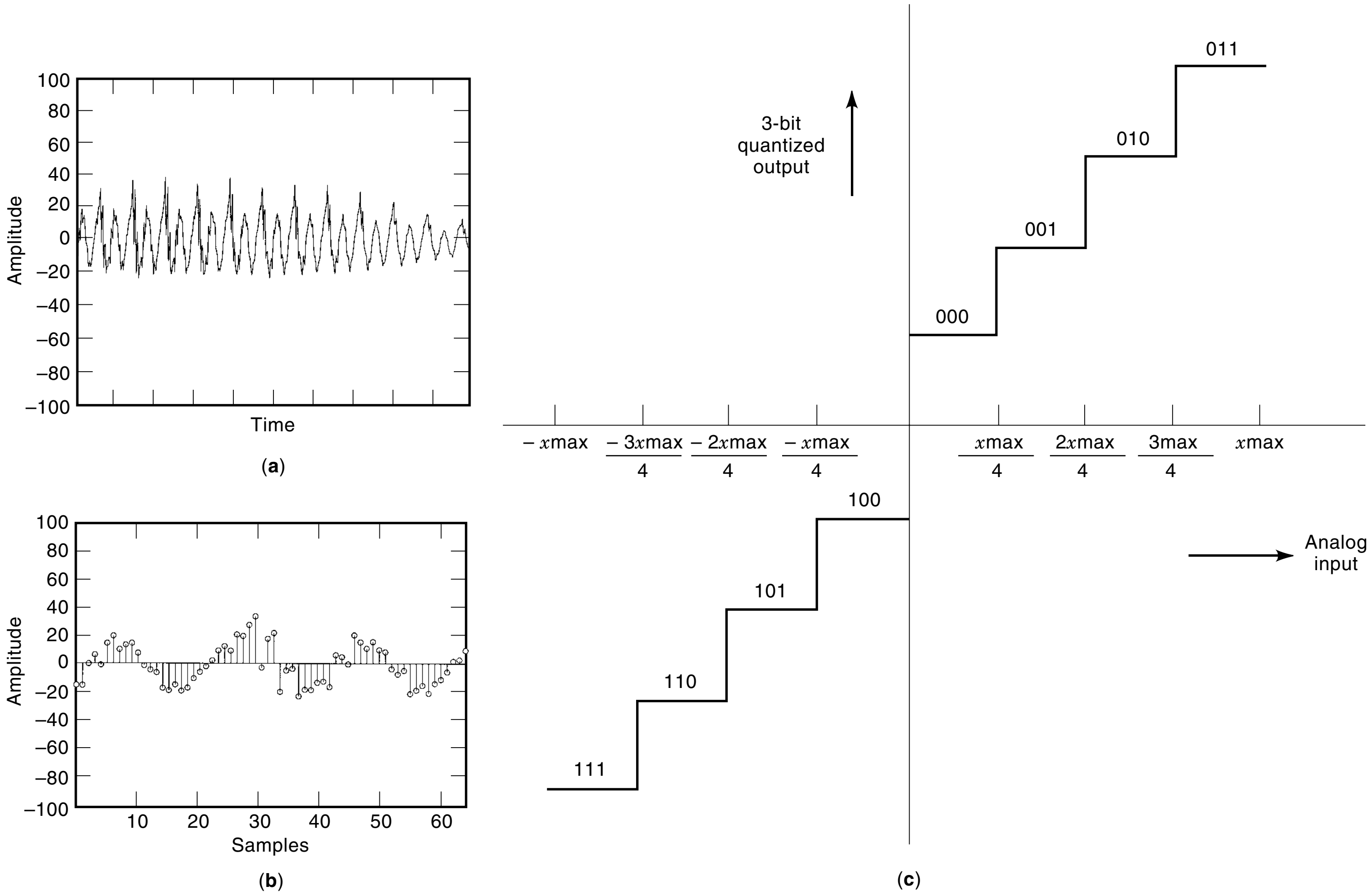
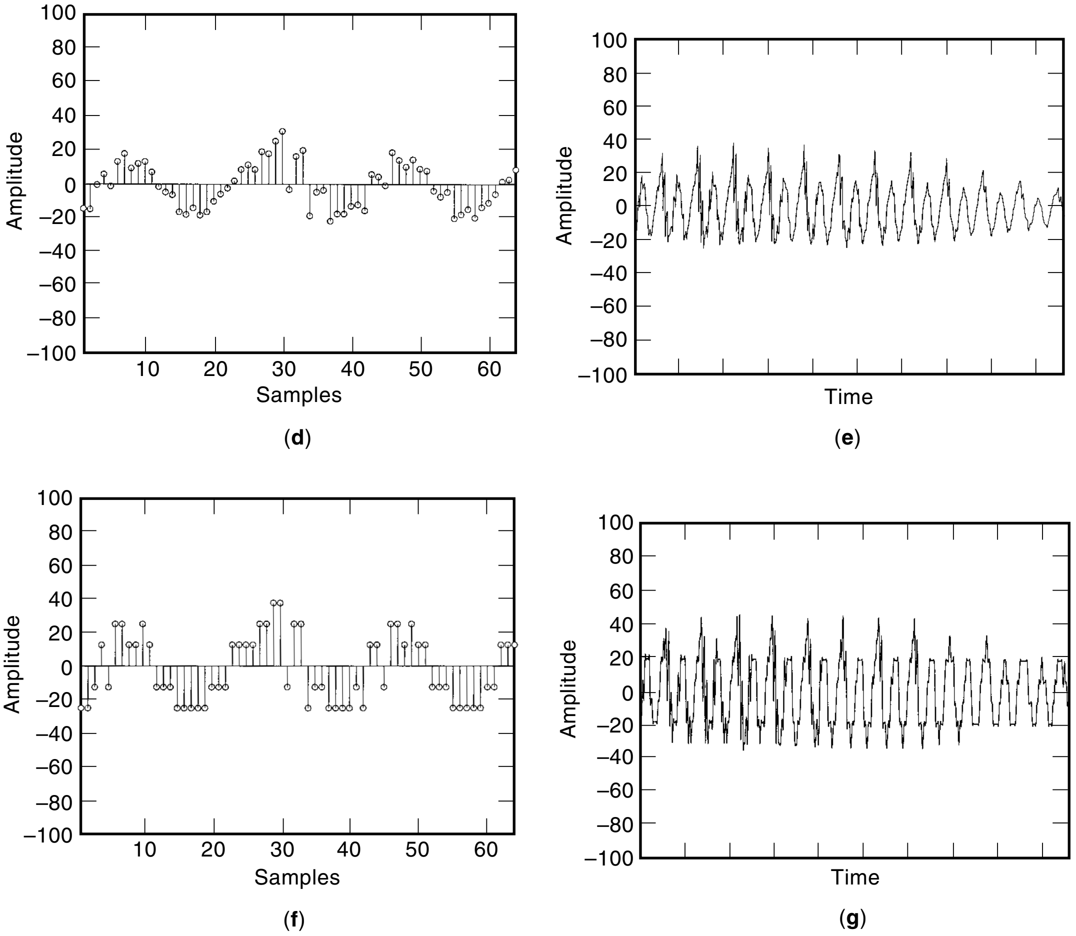


Figure 1. (a) Illustration of a continuous-time continuous-amplitude speech signal of duration about 60 ms. (b) Discrete-time (sampled) continuous-amplitude version of the speech waveform in (a); sampling rate 8000 samples/s; only first 64 samples shown for clarity. (c) Illustration of 3-bit quantization scheme. (d) Discrete-time, discrete-amplitude (sampled and quantized) version of samples in (b), using 8 bits (256 levels). (e) Reconstructed speech signal after 8-bit quantization. (f) Discrete-time, discrete-amplitude (sampled and quantized) version of samples in (b), using 3 bits (8 levels). (g) Reconstructed speech signal after 3-bit quantization.

kHz, then a sampling rate of 8 kHz (or 8000 samples per second) can be used without losing any information. Practically all speech coders used for telephony applications use the 8-kHz sampling rate. For specialized applications such as audio/videoconferencing, wideband speech coding techniques are used where sampling rates are as high as 20 kHz. Unless otherwise specified, it is assumed in this article that for digital speech coding techniques, a sampling rate of 8 kHz is used.

Quantization is the process whereby the discrete-time continuous-amplitude samples are converted to discrete-time discrete-amplitude samples, and the discrete amplitude signal is represented using a certain number of bits. Conversion from continuous-amplitude sample to a discrete-amplitude sample is performed simply by dividing the dynamic range of the signal into discrete levels and approximating the input sample to the level closest to it. It is obvious, then, that more bits are used, the better will be the representation and the lower will be the error due to quantization in the reconstructed signal. The process of sampling and quantization is illustrated in Fig. 1. Figure 1 also illustrates the effects of increased quantization distortion due to decreased number of bits. It will be shown in the subsection "Pulse Code Modulation" that the signal-to-quantization-noise ratio increases (or decreases) by 6 dB for every increase (or decrease) of one bit. If f_s is the sampling rate (in samples per second) and B is the number of bits per sample, then the channel capacity required to transmit digitally represented speech is $C = f_sB$. It can be shown using Shannon's channel capacity theorem that the larger the value of C, the larger will be the bandwidth required to transmit on a channel with a given signal-to-noise ratio. As described above, for a given speech bandwidth (4 kHz for telephony applications), the sampling frequency f_s is fixed at 8000 samples per second and the only other variable that controls the channel bandwidth required is B, the number of bits per sample.

For a given channel bandwidth it is highly desirable to make C (or B) as low as possible in order to accommodate bits from multiple users to be multiplexed on the same channel bandwidth. However, as noted above, B cannot be arbitrarily reduced, since the signal-to-quantization noise ratio (and hence the quality of the reconstructed signal) degrades sig-



(d) (e) (f) (g)

Figure 1. *Continued*

nificantly with every reduction by one bit. It is then obvious that the objective of digital speech coding is to satisfy two conflicting requirements: On the one hand it is required to maintain good speech quality, and on the other hand it is required to make the bit rate as low as possible. To achieve this objective, most low-bit-rate speech coding techniques rely on a parametric approach, whereby a block (usually called a *frame* in speech coding terminology) of speech samples is represented (coded) by a minimal set of parameters that will permit reconstructing speech with a desired quality. While the set of parameters that speech coders use varies from one technique to another, practically all modern speech coding techniques have relied heavily upon the vast amount of research that has been conducted over the past several decades in speech production, perception, analysis, and synthesis (1) for the choice of a given set of parameters. The parameters are then quantized and transmitted digitally to the remote decoder, where speech is reconstructed.

In this article, speech coders that use the digitization technique described above, where each sample is represented using a certain number of bits, will be referred to as *waveform coders,* since the objective there is to achieve a reconstructed signal whose waveform is as close to the original as possible. On the other hand, speech coders that parametrize speech signals purely by extracting parameters of an assumed model, without necessarily having the objective of reproducing a signal whose waveform looks like the original one, will be referred to as *parametric speech coders* or more simply as *vocoders.*

Speech coders have been categorized into various other ways in the literature. For example, they have been classified as high bit rate, medium bit rate, and low bit rate using bit rate as the criterion; wireline or toll quality, cellular quality, communications quality, intelligible quality, and synthetic quality using speech quality as the criterion; and time domain, frequency domain, quefrency domain, and time–frequency domain according to the domain in which speech processing is performed.

The Vocoder: A Historical Perspective

As described above, the concept of speech coding has been an area of study for several decades. The first vocoder apparatus was reported as early as 1939 by Homer Dudley in his paper titled “Remaking Speech” (2). The vocoder apparatus consisted of electrical circuits that first analyzed speech signals and extracted certain parameters, and then synthesized, or “remade,” them using those parameters. The parameters that were extracted to synthesize speech were based on the understanding gained from a significant amount of work in the areas of speech analysis and synthesis prior to Dudley’s automatic (and almost instantaneous) vocoder apparatus. These efforts led to the understanding that in order to produce intelligible speech, the analyzer had to extract information on the

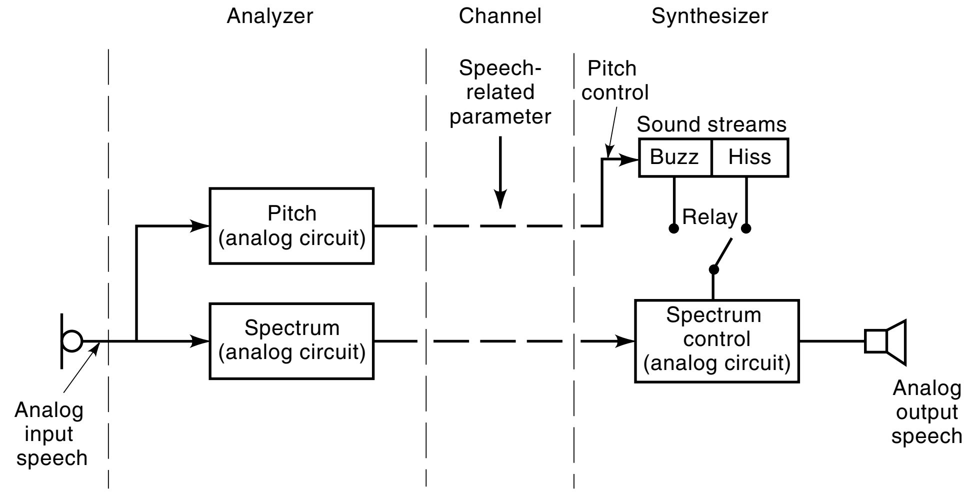


Figure 2. Homer Dudley's vocoder apparatus built in 1939 using analog electronic circuits.

pitch (determined by the fundamental frequency) of the talker, on the spectrum (the relative power in different frequency bands), and on the intensity (determined by the total sound power).

At the synthesizer, two streams of sounds were generated, based on pitch extraction by the analyzer. Properly controlled variations of these two streams generated intelligible synthetic speech. The first type of sound stream was generated when the analyzer determined a nonzero pitch in the talker's speech signal, and the synthesized speech signal was hence characterized by the fundamental frequency of the talker, the spectrum shape, and the intensity. The second type of sound stream was generated when the analyzer observed a zero pitch in talker's speech, and the synthesized speech signal was characterized by random frequency components (independent of the talker or speech material), the spectrum shape, and the intensity. Figure 2 illustrates the plan of the circuit (2) used by Homer Dudley for his vocoder apparatus used to remake speech. Long before then, serious efforts had been put into the development of manually operated speech synthesizers (or speaking machines)—as early as 1791 by Von Kempelen, and in 1922 an electrically operated vowel synthesizer) by Stewart. However, as noted above, these efforts led to *making* of speech, unlike Dudley's vocoder apparatus, which performed electrical analysis of human speech and in realtime performed electrical *remaking* (or synthesis) of speech.

A significant amount of research was simultaneously being conducted to understand the human speech production mechanism that would serve to benefit workers in multitude of disciplines in a variety of ways. For phoneticians and linguists these studies would provide a tool to describe in simple ways the acoustical features associated various phonemes in different languages. For physiologists, laryngologists, and physicists these studies would help detect, diagnose, and isolate problems related to organs involved in human speech production. For communication engineers these studies would help determine the essential features that need to be preserved in speech events in order to reconstruct intelligible speech. In this way, only the essential features of speech events need to be transmitted rather than the speech signal itself, thereby achieving significant bandwidth compression.

Speech Production Mechanism

The vocal system consisting of the vocal tract, the nasal tract, and the lungs is responsible for producing the various sounds in human beings (3). A schematic is illustrated in Fig. 3. The vocal tract begins at the opening between the vocal cords, or glottis, and ends at the lips. The nasal tract begins at the velum and ends at the nostrils. The lungs are the source of energy for the production of speech. Speech is simply the acoustic wave that is radiated from the vocal system when air is expelled from the lungs. Physiological features such as lengths and cross sections of the vocal tract and nasal tracts and the tensions in vocal cords distinguish different talkers for the same speech sound. The relative positions and cross sections within the vocal tract and nasal tract, as well as the positions of lips and velum, distinguish different speech sounds generated by the same talker.

Speech sounds are broadly classified into three different classes depending on their mode of excitation: voiced sounds, unvoiced sounds, and plosive sounds. Voiced sounds are produced by forcing air through the glottis with the tension of the vocal cords adjusted so that they vibrate in a relaxation oscillation, thereby producing quasiperiodic pulses of air, which excite the vocal tract. Unvoiced sounds, or fricatives, are generated by forming a constriction at some point in the vocal tract (usually towards the mouth end), and forcing air through the constriction at a high enough velocity to produce turbulence. This creates a broad spectrum noise source to ex-

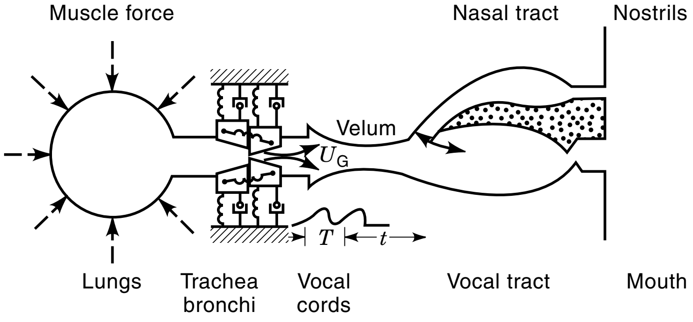


Figure 3. Schematic of the human speech production system.

cite the vocal tract. Plosive sounds results from making a complete closure at or near the lip region, building up pressure behind closure, and abruptly releasing it.

The vocal tract and nasal tract in Fig. 3 have been modeled as tubes of nonuniform cross-sectional area for the purposes of analysis (3). As the air expelled from the lungs propagates along these two tubes, the frequency spectrum of the resulting sound is determined by the resonant frequencies of the tubes. These resonant frequencies are popularly known as formant frequencies in speech processing. Different sounds are formed by varying the shape of the vocal tract, thereby yielding different formant frequencies for different sounds. Thus spectral properties of a speech signal vary with time as the vocal tract shape varies.

The time-varying spectral characteristics of a speech signal were first graphically displayed using a spectrograph (4), whereby speech energy at different frequencies as a function of time could be observed. The two-dimensional pattern (horizontal time axis and vertical frequency axis) produces dark patches in regions where the signal energy is high and light patches where it is low. Voiced regions are typically characterized by a dark striated appearance due to periodicity in the time waveform, whereas unvoiced regions are characterized by lighter and uniformly filled patches. Unlike unvoiced regions, plosives appear darker on the spectrogram, with a sharp transition from lighter bands, and unlike voiced regions, plosives typically do not exhibit the same amount of periodicity.

This article is organized as follows: Waveform coding techniques are described in the next section, and parametric coding techniques using linear prediction coding (LPC) the section following. Thereafter we discuss modeling the excitation of the human speech production mechanism. This will be used as the input to the LPC synthesis filter at the remote speech decoder. The next section deals with some important speech coding techniques that are not processed in time domain; rather, speech is transformed into a different domain and then analyzed to extract parameters of a given speech model. The section after that describes some important international and regional speech coding standards that are in use today, based on techniques described in the preceding sections. Then we provide a qualitative overview of the methods involved in assessment of speech coder performance. The effects of having multiple speech coding technologies as a result of having multiple links in an end-to-end connection are then discussed. Trends in speech coding follow, and conclusions are given in the final section.

WAVEFORM CODERS

As discussed in the introduction, waveform coders strive to encode speech in a manner that will permit reconstructing speech that is close to the original sampled speech waveform on a sample-by-sample basis. The first and most popular waveform coding technique, which is predominantly used in most national digital telephone networks, is pulse code modulation (PCM). Another waveform coding technique that is being increasingly used on international satellite links and in some large national networks is *adaptive differential PCM,* or ADPCM. Other waveform coding techniques (successors of PCM and predecessors of ADPCM), such as *adaptive delta modulation* (ADM) and *continuously variable slope delta* (CVSD) modulation, are used in some special networks such as those for military communication. These techniques are simply special cases of ADPCM. In the sequel, some fundamentals of PCM, ADM, and ADPCM techniques are described.

Pulse Code Modulation

Here speech samples are quantized in the speech encoder using B bits per sample, and these B bits are transmitted to the PCM decoder. The decoder uses the received bits to reconstruct speech as shown in Fig. 4. The choice of B depends on the available channel capacity. The choice of B determines the quantization step size and hence the signal-to-quantization-noise ratio as described below.

Let $s(n)$ represent a speech sample at time instant n, and $s(n)$ be its quantized value. Let Δ be the step size of the quantizer. The value of Δ satisfies the equation

$$2^B\Delta = 2S_{\max} \tag{1}$$

where $S_{\max}$ is the maximum amplitude of the speech signal fed into the quantizer. Then

$$\tilde{s}(n) = s(n) + e(n)$$

where $e(n) \in (-\Delta/2, \Delta/2]$. The signal-to-quantization-noise ratio $\mathrm{SNR_q}$ in decibels (dB) for such a configuration is defined by

$$\mathrm{SNR_q} = 10\log_{10}\frac{E\{s^2(n)\}}{E\{e^2(n)\}} \tag{2}$$

where $E\{\ \}$ denotes the mathematical expectation, or mean, or first moment of the random variable under consideration. In practice, $E\{\ \}$ is replaced by an unbiased estimate of the mean based upon short segments of speech. As discussed in the introduction, a speech signal can be considered as a nonstationary random process whose characteristics change slowly in time depending on the type of sound that is being produced and the talker that produces it. Therefore, in reality $E\{s^2(n)\}$ and hence $\mathrm{SNR_q}$ are time-varying quantities.

Assuming that $e(n)$ is uniformly distributed in $(-\Delta/2, \Delta/2]$, we have

$$E\{e^2(n)\} = \frac{\Delta^2}{12}$$

Substituting the value of $E\{e^2(n)\}$ in Eq. (2) above and substituting for Δ from Eq. (1), it can be shown that $\mathrm{SNR_q}$ can be written in the form

$$\mathrm{SNR_q} = 10\log_{10} E\{s^2(n)\} + 6B + f(S_{\max})$$

From the above analysis, it is clear that (i) the signal-to-quantization-noise ratio decreases by 6 dB when the number of bits per sample, B, is reduced by one bit, and (ii) for an assumed $S_{\max}$ and chosen B, the short-term $\mathrm{SNR_q}$ is a monotonic function of $E\{s^2(n)\}$, implying that $\mathrm{SNR_q}$ is poor for durations where speech signals have smaller amplitudes as compared to durations where speech signal has larger amplitudes. Item (ii) above is highly restrictive since studies

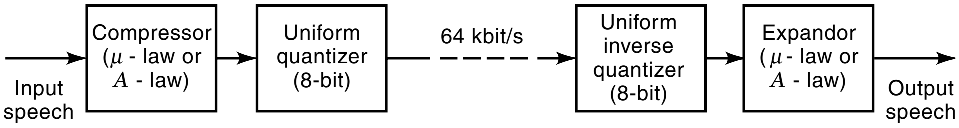

Input speech
Compressor (μ - law or A - law)
Uniform quantizer (8-bit)
64 kbit/s
Uniform inverse quantizer (8-bit)
Expandor (μ- law or A - law)
Output speech

Figure 4. Block diagram of a PCM coder.

have repeatedly shown that speech signal amplitudes are less than $S_{\max}/4$ for most of the time. Hence a quantization scheme is desirable whose step size is smaller for lower-amplitude speech signals and larger for larger-amplitude signals. Such a quantizer is called a nonuniform quantizer. In practice, nonuniform quantization is achieved by first transforming the signal amplitudes in a manner that will result in approximately a uniform distribution and then performing uniform quantization. At the decoder, an inverse transformation is applied to recover the original distribution. In this manner, SNR_q can be significantly increased, or alternatively, it is possible to use fewer bits with nonuniform quantization than with uniform quantization for obtaining the same SNR_q.

Two types of signal transformation that are used worldwide are called the μ law and the A law. The μ law, which is used in North American telephone networks, is defined as follows:

$$s_c(n) = S_{\max}\frac{\log(1+\mu|s(n)|/S_{\max})}{\log(1+\mu)}\mathrm{sign}[s(n)]$$

The A-law method of compressing speech signals, which is employed in European telephone networks, is defined by

$$s_c(n) = \begin{cases} \dfrac{As(n)}{1+\log A} & \text{for } 0 \le |s(n)| \le \dfrac{S_{\max}}{A} \\ S_{\max}\dfrac{1+\log(A|s(n)|/S_{\max})}{1+\log A}\mathrm{sign}[s(n)] & \text{for } \dfrac{S_{\max}}{A} < |s(n)| \le S_{\max} \end{cases}$$

In both cases $s_c(n)$ represents the compressed version of original speech $s(n)$. A mapping table corresponding to these laws is provided in International Telecommunications Union–Telecom Sector (ITU-T, formerly CCITT) Recommendation G.711 for PCM signals.

Due to the logarithmic curves in both A-law and μ-law transformations, which tend to *compress* larger-amplitude signals, and the inverse transformations (exponential in nature) which tend to *expand* larger-amplitude signals, the two functionalities together have been called *companding* and PCM in which speech signals have undergone companding is called companded PCM. It has been found that the signal quality of a 13-bit uncompanded PCM is equivalent to that of a 8-bit companded PCM (5). Today, μ-law or A-law companding is used in almost all telephone networks, and each speech sample is represented using 8 bits, so that the transmission bit rate of PCM signal is 64 kbit/s.

Adaptive Differential Pulse Code Modulation

PCM is extremely robust across a variety of speech signals, since it does not make any inherent assumptions about the time-varying spectral characteristics of the speech signal. However, the bit rate of 64 kbit/s can be prohibitively high, especially in bandwidth-scarce links such as satellite links. Hence there arose a need for a lower-bit-rate speech coding technique. This led to the development of speech coders that encoded differences between adjacent samples or differences between the current sample and a predicted value of the current sample (based on previous samples), broadly referred to as *differential PCM* or DPCM (64). The DPCM speech coder is based upon the observations that (i) adjacent speech samples are highly correlated and (ii) a correlated speech sample can be predicted whose associated prediction error has a small variance as derived below.

Let $s(n)$ be the speech sample at time instant n. For simplicity, let the prediction formulation be of the form

$$s(n) = \alpha s(n-1) + d(n)$$

Then it can be shown that mean square of the prediction error $d(n)$, or $E\{d^2(n)\}$, is minimum when

$$\alpha = \alpha^* = E\{s(n)s(n-1)\}/E\{s^2(n)\}$$

that is, when α is equal to the correlation coefficient [assuming $s(\ )$ to be zero-mean and identically distributed] between adjacent samples $s(n)$ and $s(n-1)$. The corresponding mean squared error between predicted value and actual value is then given by

$$E_{\min} = E\{s^2(n)\}(1-\alpha^{*2})$$

It is therefore easy to see that when the correlation a^* is large, the mean squared error between actual speech sample and predicted sample becomes small, and hence fewer bits are needed to represent $d(n)$.

The prediction formulation above is often referred to as a linear first-order predictor, since the current sample is being predicted from one previous sample and the relation between $s(n-1)$ and $s(n)$ is linear. In practice, for speech signals, the prediction formulation is usually is of higher order, of the form

$$s(n) = \sum_{i=1}^{N}\alpha_i s(n-i) + d(n) \tag{3}$$

It is important to note that, in general, it can be shown that the mean squared prediction error $E\{[s(n) - f(s(n-1), s(n-2),\ldots, s(n-N)]^2\}$ is minimum if

$$f(\ ) = f^*(\ ) = E\{s(n)\,|\,s(n-1), s(n-2), \ldots, s(n-N)\}$$

In practice however, $f^*(\ )$ is assumed to be linear in $s(n-1)$, $s(n-2),\ldots, s(n-N)$ as shown in Eq. (3), primarily because of the simplicity and analytical tractability that a linear formulation provides as compared to that of a nonlinear formulation as shown above. A secondary but important reason for formulation of linear predictors in most systems stems from the fact that $f^*(\ )$ in the equation above is indeed linear in $s(n-1)$, $s(n-2)$,. . ., $s(n-N)$ if the joint probability distribution of $s(n)$, $s(n-1)$,. . ., $s(n-N)$ is normal.

Thus, if α^* and $d(n)$ are transmitted and $s(0)$ is known, then it is possible for the decoder to reproduce $s(n)$ exactly for any $n > 0$. However, in practice $d(n)$ has to be quantized to $d_q(n)$ before transmission. Hence the decoder output $\tilde{s}(n)$ will not be equal to $s(n)$. In such an event the formulation at the decoder is of the form

$$\tilde{s}(n) = \alpha\tilde{s}(n-1) + d_q(n)$$

Such a formulation, however, leads to a situation where the difference between the reconstructed speech sample $\tilde{s}(n)$ at the output of the ADPCM decoder and the input speech sample $s(n)$ at the input of the ADPCM encoder is an accumulation of quantization errors $d(m) - d_q(m)$, $0 \le m \le n$. For the first-order predictor formulation it can be shown that

$$\begin{aligned} s(n) - \tilde{s}(n) &= \alpha[s(n-1) - \tilde{s}(n-1)] + d(n) - d_q(n) \\ &= \alpha^n[s(0) - \tilde{s}(0)] + \sum_{i=0}^{n-1} \alpha^i[d(n-i) - d_q(n-i)] \end{aligned}$$

To avoid the accumulative effect of quantization errors, the encoder replicates the operation of the decoder and estimates α based on $\tilde{s}(n-1)$—or, in the Nth-order predictor case of the equation above, estimates $\boldsymbol{\alpha} = [\alpha_1, \alpha_2,. . ., \alpha N]^t$ based on $\tilde{s}(n-i)$, $1 \le i \le N$—as shown in Fig. 5. Such a configuration ensures that the error between original sample and reconstructed sample is simply the quantization error associated with $d(n)$. For example, for the first-order predictor, at the encoder we have

$$s(n) = \sum_{i=1}^{N} \alpha_i \tilde{s}(n-i) + d(n)$$

and at the decoder

$$\tilde{s}(n) = \sum_{i=1}^{N} \alpha_i \tilde{s}(n-i) + d_q(n)$$

Therefore, $s(n) - \tilde{s}(n) = d(n) - d_q(n)$ is the quantization error associated with $d(n)$ alone and has no contributions from $d(n-1)$, $d(n-2)$, etc.

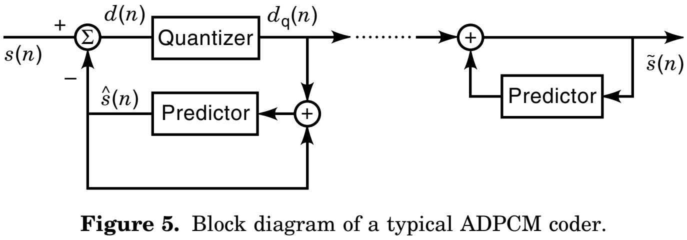


Figure 5. Block diagram of a typical ADPCM coder.

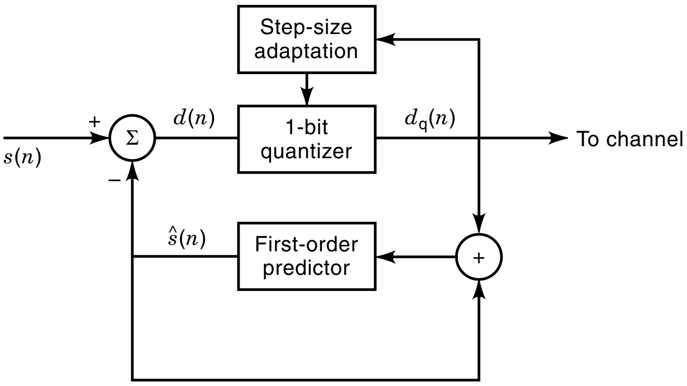


Figure 6. A simplified block diagram of an ADM coder; note the 1-bit quantizer.

It is noted that the prediction formulation in the equations above is essentially an all-pole formulation, since the current sample is predicted from previous output samples and not from previous inputs. However, many ADPCM speech coders employ a pole–zero prediction of the form

$$\tilde{s}(n) = \sum_{i=1}^{N_1} \alpha_i \tilde{s}(n-i) + \sum_{j=0}^{N_2} \beta_j d_q(n-j) \tag{4}$$

Such a formulation permits higher prediction gain (in other words, lower variance or dynamic range for prediction residual) for nasal sounds. Such a pole–zero adaptive predictor is employed in ADPCM-based ITU-T speech-coding standards G.726 and G.727 operating at bit rates of 40, 32, 24, and 16 kbit/s. These will be discussed in the section "Transform-Domain Speech Coding."

Adaptive Delta Modulation

The ADM technique is a special case of ADPCM, in the sense that α of the first-order prediction formulation is usually constrained to be equal to 1, and $d_q(n)$ is constrained to be equal to $\pm\Delta$, where δ is known at encoder and decoder. This implies that only one bit per sample needs to be transmitted to the remote decoder, depending on the sign of Δ. Although such a scheme is suboptimal as compared to ADPCM, the significant reduction in bit rate (equal to the sampling rate) has rendered it useful in some military applications with sacrifice in voice quality. A major difficulty comes from the inability of the model to track very large (larger than Δ) and very small (less than Δ) variations, leading to noticeable distortions in reconstructed speech samples. To improve voice quality at the output of the decoder, the scheme is made adaptive by adjusting the value of Δ to track the variations in the input speech signal. One such scheme is based on adjusting Δ in the following recursive manner

$$\Delta(n) = \Delta(n-1)K^{d_q(n)d_q(n-1)}$$

This is illustrated by the step-size adaptation box in Fig. 6. Such a scheme does not require any additional information at the decoder as compared to the nonadaptive approach, since the adaptation of Δ is based upon parameters known to the remote decoder.

One drawback of ADM is that transmission errors can cause degradation of speech quality that can last for a long time, especially when α is constrained to 1. To recover from transmission errors, it is necessary to introduce a leakage factor in both prediction and Δ adaptation. One such method is the CVSD method.

PARAMETRIC SPEECH CODERS

Parametric coders typically analyze the speech signal (in blocks or frames) and extract parameters that are deemed necessary to synthesize speech with a given quality objective under the constraint of a given bit rate. Among all the parametric coders that have been investigated and reported in the literature, the most widely used for speech analysis is LPC. Here the prediction model is similar to that used in ADPCM described above in the subsection "Adaptive Differentiate Pulse Code Modulation," but the order of the model and its objective are different from that used in ADPCM coders. In ADPCM coders, the prediction filter is used to reduce the variance of the difference between the actual and predicted signal, thereby reducing the number of bits necessary to represent and transmit the prediction residual. Parametric coders that will be described below use the predictive model to estimate the poles of the vocal tract transfer function and hence obtain the spectral envelope of the speech signal. An Lth order linear predictive model is of the form

$$\hat{s}(n) = \sum_{i=1}^{L} \alpha_i s(n-i) \tag{5}$$

where $\tilde{s}(n)$ is the predicted value of the current speech sample $s(n)$ based on previous speech samples. The prediction error $e(n)$ is defined as

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{i=1}^{L} \alpha_i s(n-i)$$

or in terms of the transfer function,

$$E(z) = S(z)A(z)$$

or

$$S(z) = \frac{E(z)}{A(z)} \tag{6}$$

where

$$A(z) = 1 - \sum_{k=1}^{L} \alpha_k z^{-k} \tag{7}$$

The coefficients α_i in Eq. (7) are known as LPC coefficients. The basic problem of linear prediction is determination, representation, and quantization of the LPC coefficients α_i. Equation (6) can be interpreted as the vocal tract being modeled as an all-pole system $1/A(z)$ whose input is $e(n)$ and output is the speech sample $s(n)$. Then $e(n)$ represents the excitation source to the vocal tract system as shown in Fig. 3. An excellent treatment on LPC analysis and synthesis can be found in Refs. 65, 66, and 67.

One of the earliest objectives of linear prediction coding was to model the speech spectral envelope with a prediction (all-pole) filter whose input $e(n)$ was a quasiperiodic sequence (with period equal to the pitch period) for voiced speech and random noise for unvoiced speech. This is illustrated in Fig. 7. The acoustic theory of speech production points to the important fact that in order to obtain intelligible speech, it is sufficient to identify five dominant resonant frequencies (poles) of the vocal tract function. Hence a large number of vocoders that use LPC to analyze speech signal use a 10th-order linear prediction [$L = 10$ in Eq. (5)] that will result in an all-pole filter with five complex poles corresponding to the five dominant resonant frequencies of the vocal tract. Furthermore, it is interesting to note that the model described in Fig. 7 is no different from the model Dudley's vocoder apparatus was based upon, except for the fact that digital computers were not available then to determine the coefficients of the linear predictive model; instead Dudley used an electric circuit to extract the spectral envelope of the speech signal.

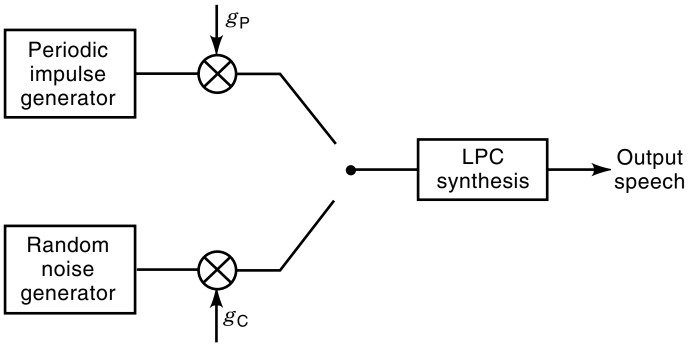


Figure 7. A two-stage excitation model that can produce intelligible-quality speech.

The LPC coefficients are determined by minimizing the mean squared prediction error over a short segment of speech waveform. Because of the time-varying nature of the speech signal, the LPC coefficients have to be determined from short segments of it, called *frames.* In most speech coders the LPC coefficients are computed approximately every 10 ms to 20 ms. The frequency with which LPC coefficients are determined (or updated) is based upon the observation of spectrograms of speech, which indicate that the spectral envelope changes slowly with time, and that for a duration of about 10 ms to 20 ms it can be assumed that the spectral envelope remains reasonably constant. Furthermore, the optimal choice of LPC coefficients is computationally expensive and hence computed only as often as necessary. In order to retain the smoothness with which the spectrum changes are registered on the spectrogram, most speech coders perform linear interpolation of LPC parameters in between LPC coefficient updates.

The mean squared prediction error E_n is given by

$$E_n = E\{[s(n) - \hat{s}(n)]^2\}$$

where $E\{\ \}$ denotes the expected value. However, because of the time-varying nature of speech, the mathematical expectation is replaced by a short-term average $\hat{E}_n$ defined as

$$\hat{E}_n = \frac{1}{m_{\text{high}} - m_{\text{low}}} \sum_{m=m_{\text{low}}}^{m_{\text{high}}} \left(s_n(m) - \sum_{i=1}^{L} \alpha_i s_n(m-i) \right)^2 \tag{8}$$

where $s_n(m) = s(n+m)$, $m_{\text{low}} \leq m \leq m_{\text{high}}$ is a segment of speech surrounding the speech sample $s(n)$ of interest, and the speech samples between $s_{m_{\text{low}}}$ and $s_{m_{\text{high}}}$ constitute a frame of speech.

Minimizing $\hat{E}_n$ with respect to α_i by setting $\partial\hat{E}_n/\partial\alpha_i = 0$, $i = 1, \ldots, L$, leads to a set of L simultaneous equations given by

$$\sum_{m=m_{\text{low}}}^{m_{\text{high}}} s_n(m-i)s_n(m) = \sum_{k=1}^{L}\alpha_k \sum_{m=m_{\text{low}}}^{m_{\text{high}}} s_n(m-i)s_n(m-k)$$
$$1 \le i \le L$$

or more compactly

$$\varphi_n(i,0) = \sum_{k=1}^{L}\alpha_k\varphi_n(i,k), \qquad 1 \le i \le L \tag{9}$$

where

$$\varphi_n(a,b) = \sum_{m=m_{\text{low}}}^{m_{\text{high}}} s_n(m-a)s_n(m-b) \tag{10}$$

Solution of this set of equations will yield the set of optimal LPC coefficients. There are two fundamental approaches, the autocorrelation method and the covariance method, that have been used to arrive at a solution. The basic difference in the two approaches is the choice of the limits m_{low} and m_{high}. They will be explained below.

Autocorrelation Method of Linear Predictive Coding Analysis

In this method, it is assumed that the waveform segment $s_n(m)$ is zero outside the interval $0 \le m \le N-1$. Such an assumption is equivalent to a windowing operation on the original speech samples, where the window $w(n)$ is represented as

$$w(n)\begin{cases} = 1, & 0 \le n \le N-1 \\ = 0 & \text{otherwise} \end{cases}$$

Such a window is called a rectangular window. It has the effect of spreading the spectrum of the speech segment, since it has high sidelobes. Some of the more common windows that have been used for speech coding are as follows:

Hamming window:

$$w(n) = \begin{cases} 0.54 - 0.46\cos\dfrac{2\pi n}{N-1}, & 0 \le n \le N-1 \\ 0 & \text{otherwise} \end{cases}$$

Hanning window:

$$w(n) = \begin{cases} 0.5 - 0.5\cos\dfrac{2\pi n}{N-1}, & 0 \le n \le N-1 \\ 0 & \text{otherwise} \end{cases}$$

Bartlett window:

$$w(n) = \begin{cases} \dfrac{2n}{N-1}, & 0 \le n \le \dfrac{N-1}{2} \\ 2 - \dfrac{2n}{N-1}, & \dfrac{N-1}{2} \le n \le N-1 \\ 0 & \text{otherwise} \end{cases}$$

Blackman window:

$$w(n) = \begin{cases} 0.42 - 0.5\cos\dfrac{2\pi n}{N-1} + 0.08\cos\dfrac{4\pi n}{N-1}, & 0 \le n \le N-1 \\ 0 & \text{otherwise} \end{cases}$$

The choice of the window shape is critical for handling sidelobe effects in the windowed speech in the spectral domain. The Blackman window has the least frequency leakage in terms of sidelobe contribution, but at the same time has the lowest frequency resolution. The performance of Hamming and Hanning windows is in between the two extremes of rectangular and Blackman windows, and makes them the most popular in speech coding.

Since $s_n(m)$ is zero for $0 \le m \le N-1$ when any of the above windows is used, it is easy to verify that the prediction residual $e(m)$ is nonzero over the interval $0 \le m \le N-1+L$, so that lower and upper limits for $\hat{E}_n$ in Eq. (8) have to be $m_{\text{low}} = 0$ and $m_{\text{high}} = N-1+L$. It is also noted that the prediction error $e(m)$ for $0 \le m \le L-1$ is likely to be large, since the equation implies that nonzero samples are being predicted from zero samples in this region. Similarly, $e(m)$ is likely to be large in the region $N \le m \le N+L-1$, since zero samples are being predicted from nonzero samples. Substituting the values of m_{low} and m_{high} in Eq. (10), we get

$$\varphi_n(a,b) = \sum_{m=0}^{N-1+L} s_n(m-a)s_n(m-b)$$

and since $s_n(m) = 0$ for $m < 0$ and $m > N-1$, we can rewrite $\varphi_n(a,b)$ as

$$\varphi_n(a,b) = \sum_{m=0}^{N1-(a-b)} s_n(m)s_n(m+a-b)$$

From the above equation, $\varphi_n(a,b)$ can be viewed as a short-term autocorrelation function of $s_n(m)$ evaluated at a lag of $a-b$, i.e.,

$$\varphi_n(a,b) = R_n(a-b)$$

where

$$R_n(\tau) = \sum_{m=0}^{N-1-\tau} s_n(m)s_n(m+\tau)$$

It can also be shown, under the assumption of $s_n(m) = 0$ for $m < 0$ and $m > N-1$, that $R_n(\tau)$ is an even function: $R_n(\tau) = R_n(-\tau)$. This property is exploited in the autocorrelation method of LPC analysis to reduce the computational complexity of the algorithm, and this is evidenced when the set of simultaneous equations is represented in the matrix form

$$\begin{bmatrix} \varphi_n(1,0) \\ \varphi_n(2,0) \\ \vdots \\ \varphi_n(L,0) \end{bmatrix} = \begin{bmatrix} \varphi_n(1,1) & \varphi_n(1,2) & \varphi_n(1,3) & \cdots & \varphi_n(1,L) \\ \varphi_n(2,1) & \varphi_n(2,2) & \varphi_n(2,3) & \cdots & \varphi_n(2,L) \\ \vdots & \vdots & \vdots & & \vdots \\ \varphi_n(L,1) & \varphi_n(L,2) & \varphi_n(L,3) & \cdots & \varphi_n(L,L) \end{bmatrix}$$
$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_L \end{bmatrix}$$

Substituting for φ_n in terms of R_n and using the symmetric property of R_n, we get

$$\begin{bmatrix} R_n(1) \\ R_n(2) \\ \vdots \\ R_n(L) \end{bmatrix} = \begin{bmatrix} R_n(0) & R_n(1) & R_n(2) & \cdots & R_n(L-1) \\ R_n(1) & R_n(0) & R_n(1) & \cdots & R_n(L-2) \\ \vdots & \vdots & \vdots & & \vdots \\ R_n(L-1) & R_n(L-2) & R_n(L-3) & \cdots & R_n(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_L \end{bmatrix}$$

From this it is easy to see that the $L \times L$ matrix has a Toeplitz structure and that the number of correlation computations necessary to solve the above equation is simply $L + 1$. Several efficient recursive procedures have been devised to solve it, the most efficient being Durbin's recursive procedure, which is outlined below.

The recursive procedure entails computing the following quantities recursively L times, where superscripts indicate the recursion number:

$$\begin{aligned} E_n^{(0)} &= R_n(0) \\ k_i &= \frac{R_n(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R_n(i-j)}{E^{(i-1)}}, \qquad 1 \le i \le L \\ \alpha_i^{(i)} &= k_i \\ \alpha_j^{(i)} &= \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)}, \qquad 1 \le j \le i-1 \\ E_n^{(i)} &= (1 - k_i^2) E_n^{(i-1)} \end{aligned} \tag{11}$$

The above equations are solved for $1 \le i \le L$, after which the solution to the matrix equation above is given by

$$\alpha_i = \alpha_i^{(L)}$$

It is noted that $\alpha_j^{(i)}$, $1 \le j \le i$, above actually are the LPC coefficients of an ith-order LPC model. This property has been exploited by some speech coders to extract lower-order LPC coefficients from a higher-order model as well as to make voiced–unvoiced decisions on speech segments.

Covariance Method of Linear Predictive Coding

In this approach, rather than windowing the speech segment, the duration over which the prediction error is computed is windowed. The limits on E_n for the covariance method are $0 \le m \le N - 1$, and hence $m_{\text{low}} = 0$ and $m_{\text{high}} = N - 1$, which leads to

$$\hat{E}_n = \frac{1}{N} \sum_{m=0}^{N-1} \left(s_n(m) - \sum_{i=1}^{L} \alpha_i s_n(m-i) \right)^2$$

$$\varphi_n(a, b) = \sum_{m=0}^{N-1} s_n(m-a) s_n(m-b)$$

It can be shown in this case that $\varphi_n(a, b) = \varphi_n(b, a)$; however, $\varphi_n(a, b) \neq f(a - b)$ and hence does not truly represent the autocorrelation function. Instead, $\varphi_n(a, b)$ represents cross-correlation between similar, but not identical sequences. Hence, in matrix form, the covariance method leads to

$$\begin{bmatrix} \varphi_n(1,0) \\ \varphi_n(2,0) \\ \vdots \\ \varphi_n(L,0) \end{bmatrix} = \begin{bmatrix} \varphi_n(1,1) & \varphi_n(1,2) & \varphi_n(1,3) & \cdots & \varphi_n(1,L) \\ \varphi_n(1,2) & \varphi_n(2,2) & \varphi_n(2,3) & \cdots & \varphi_n(2,L) \\ \vdots & \vdots & \vdots & & \vdots \\ \varphi_n(1,L) & \varphi_n(2,L) & \varphi_n(3,L) & \cdots & \varphi_n(L,L) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_L \end{bmatrix}$$

written compactly as

$$\Omega = \Phi \boldsymbol{\alpha}$$

where the $L \times L$ matrix Φ is symmetric, but not Toeplitz. The above matrix equation is solved using Cholesky decomposition, where Φ can be decomposed into a product of upper triangular (U), diagonal (D) and lower triangular (U^{t}) matrices:

$$\Omega = UDU^{\text{t}} \boldsymbol{\alpha}$$

Letting

$$DU^{\text{t}} \boldsymbol{\alpha} = \boldsymbol{\alpha}'$$

it is easy to see that since U is an upper triangular matrix, the solution for each element of the vector $\boldsymbol{\alpha}'$ can be obtained recursively. After obtaining α', the vector $\boldsymbol{\alpha}'$ is obtained in a similar manner by noting from above equation that $U^{\text{t}}\boldsymbol{\alpha} = D^{-1}\boldsymbol{\alpha}'$ and exploiting the fact that U^{t} is a lower triangular matrix.

It was seen from the autocorrelation method and the covariance method of LPC analysis that determination of LPC coefficient vector $\boldsymbol{\alpha}$ involved precomputing R_n or φ_n and then solving the matrix equation. However, there exists a third technique, called the *lattice* method, of LPC analysis that eliminates the precomputation of correlation values and obtains LPC coefficients directly from speech samples. This is described in Ref. 9.

Quantization and Alternative Representations of Linear Predictive Coding Parameters

In the above discussion, techniques to estimate the LPC coefficients were described. Since the ultimate goal is to reduce the bit rate, the LPC coefficients have to be quantized using as few bits as possible. The LPC coefficients may be scalar quantized or vector quantized. In scalar quantization, each LPC coefficient is quantized independently of the others in a manner similar to that in PCM, except that different coefficients may be represented using different numbers of bits, depending on their importance. In vector quantization (VQ), the entire vector of coefficients (or subvectors) is quantized jointly, and the quantization is based on finding an element (or vector) of a codebook that is close (with respect to a defined distortion measure) to the vector of LPC coefficients to be quantized. The index of the codebook is transmitted to the voice decoder. In such systems, both encoder and decoder are expected to have a copy of the same codebook. The codebook itself is generated from a large training set. VQ of LPC parameters has received considerable attention (and is perhaps the only alternative) for low-bit-rate speech coders operation at 4 kbit/s or below. Whereas the original VQ techniques involved having a single large codebook and an exhaustive search procedure, modern-day techniques typically have multiple smaller codebooks that are searched in multiple stages, thereby reducing search complexity and storage. A classical example can be found in Ref. 10.

Let $1/\hat{A}(z)$ represent the synthesis filter with quantized LPC coefficients, that is,

$$\frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{L} \alpha_1 \hat{z}^{-i}}$$

where $\hat{\alpha}_j$ represents the jth LPC coefficient after quantization (scalar or vector). It is extremely important to note that in order for the synthesis filter $1/\hat{A}(z)$ to be stable, the roots of $\hat{A}(z)$ have to lie inside the unit circle. Even if the LPC coefficients are such that $1/A(z)$ is stable, the quantized set of LPC coefficients (which will be the one that will be used by the voice decoder) may lead to an unstable synthesis filter. Furthermore, checking the stability (finding the roots of a 10th-order polynomial) is no trivial task. Hence alternative methods of representing LPC coefficients that guarantee stability of the LPC synthesis filter after quantization have been developed, the popular ones being reflection coefficients, log area ratios, arcsines of partial correlation coefficients, and line spectral frequencies.

Examining Eq. (11), it is easy to see that $|k_i| < 1.0$, and it is also observed that the LPC coefficients are derived from k_i. Hence k_i [called the reflection coefficients or *partial correlation* (PARCOR) coefficients] provide a suitable alternative for LPC coefficients, since k_i can be checked for stability by simply verifying whether $|k_i| > 1.0$. It can however be shown that the LPC spectral distortion introduced by quantizing k_i depends on $|k_i|$ and that values of $|k_i|$ closer to unity are more sensitive than those reflection coefficients that have smaller magnitudes. In order to normalize this, a nonlinear transformation is performed on k_i such as

$$l_i = \log \frac{1 - k_i}{1 + k_i}$$

where l_i are called *log-area ratio* (LAR) coefficients. Another nonlinear transformation that is quite often used is the arcsine transformation $s_i = \sin^{-1} k_i$, where s_i are called arcsine coefficients. Both LAR and arcsine functions tend to emphasize larger values of $|k_i|$ and deemphasize smaller values, thereby performing a transformation similar to the expansion function in PCM systems. Such a transformation allows more accurate quantization of larger values of $|k_i|$ than of smaller values.

Perhaps the most popular alternative to LPC coefficients is the *line spectral frequency* (LSF) representation. Here, odd and even polynomials are formed using $A(z)$ as follows:

$$P(z) = A(z) + z^{-(L+1)}A(z^{-1})$$
$$Q(z) = A(z) - z^{-(L+1)}A(z^{-1})$$

It can be shown that the roots of $P(z)$ and $Q(z)$ lie on the unit circle, and that the roots of $P(z)$ and $Q(z)$ alternate on the unit circle in the complex z plane. Furthermore, $z = -1$ and $z = +1$ are roots of $P(z)$ and $Q(z)$ respectively. Hence $P(z)$ and $Q(z)$ can be written as

$$P(z) = (1 + z^{-1})P_1(z)$$
$$Q(z) = (1 - z^{-1})Q_1(z)$$

where

$$P_1(z) = \sum_{i=0}^{L} p_i z^{-i}$$
$$Q_1(z) = \sum_{i=0}^{L} q_i z^{-i}$$

where it can be shown that

$$\begin{aligned} &p_0 = 1, \qquad q_0 = 1 \\ &p_j = \alpha_j + \alpha_{L-j+1} - p_{j-1}, \qquad 1 \le j \le L \\ &q_j = \alpha_j - \alpha_{L-j+1} + q_{j-1}, \qquad 1 \le j \le L \end{aligned}$$

The roots of $P_1(z)$ and $Q_1(z)$ form the line spectral frequencies for the set of LPC coefficients, and they are on the unit circle. Such an ordering is essential to ensure stability of the synthesis filter. It can be shown that the coefficients of $P_1(z)$ and $Q_1(z)$ are such that $p_j = p_{L-j}$ and $q_j = q_{L-j}$. Hence $P_1(z)$ and $Q_1(z)$ take the form

$$\begin{aligned} P_1(z) &= 1 + p_1 z^{-1} + \cdots + p_{L/2-1} z^{L/2-1} + p_{L/2} z^{L/2} \\ &\quad + p_{L/2-1} z^{L/2+1} + \cdots + p_1 z^{L-1} + z^L \\ &= z^{-L/2}[(z^{L/2} + z^{-L/2}) + p_1(z^{L/2-1} + z^{-(L/2-1)}) \\ &\quad + \cdots + p_j(z^{L/2-j} + z^{-(L/2-j)}) + \cdots + p_{L/2}] \\ Q_1(z) &= 1 + q_1 z^{-1} + \cdots + q_{L/2-1} z^{L/2-1} + q_{L/2} z^{L/2} \\ &\quad + q_{L/2-1} z^{L/2+1} + \cdots + q_1 z^{L-1} + z^L \\ &= z^{-L/2}[(z^{L/2} + z^{-L/2}) + q_1(z^{L/2-1} + z^{-(L/2-1)}) \\ &\quad + \cdots + q_j(z^{L/2-j} + z^{-(L/2-j)}) + \cdots + q_{L/2}] \end{aligned}$$

Since the roots of $P_1(z)$ and $Q_1(z)$ lie on the unit circle, $P_1(z)$ and $Q_1(z)$ need to be evaluated only at $z = e^{jw}$. On observing that $e^{jx} + e^{-jx} = 2 \cos x$, it is clear that it is necessary to find the roots of the following two equations:

$$\begin{aligned} &\cos\left(\frac{wL}{2}\right) + p_1 \cos\left[w\left(\frac{L}{2} - 1\right)\right] + \cdots + p_j \cos\left[w\left(\frac{L}{2} - j\right)\right] \\ &\quad + \cdots + p_{L/2} = 0 \\ &\cos\left(\frac{wL}{2}\right) + q_1 \cos\left[w\left(\frac{L}{2} - 1\right)\right] + \cdots + q_j \cos\left[w\left(\frac{L}{2} - j\right)\right] \\ &\quad + \cdots + q_{L/2} = 0 \end{aligned}$$

The above two equations are solved for $L/2$ values of w in the range of 0 to π. Let $w_{p1}, w_{p2}, \ldots, w_{pL/2}$ be the roots of $P_1(z)$ in the range $0 \le w_{pi} \le \pi$, and $w_{q1}, w_{q2}, \ldots, w_{qL/2}$ be the roots of $Q_1(z)$ in the range $0 \le w_{qi} \le \pi$. The w_{pi} and w_{qi}, which are expressed in radians, are converted to a vector of line spectral frequencies as [$2\pi w_{p1}$, $2\pi w_{q1}$, $2\pi w_{p2}$, $2\pi w_{q2}$, . . ., $2\pi w_{pL/2}$, $2\pi w_{qL/2}$].

It is noted from the equations above that it is only necessary to find roots of two L/2th-order polynomials rather than one Lth-order polynomial. Several methods of finding roots of the above equations have been reported in the literature, such as finding zero-crossing points, phase reversal tracking, and use of Chebychev polynomials (11). In all methods, the complexity of finding roots of the above two equations is significantly less than that of finding the roots of a Lth-order polynomial. Furthermore, after quantization of the line spectral frequencies, it is only necessary to check the ordering to verify whether the synthesis filter is stable or not.

In addition to the ordering property, another attractive feature of LSFs is their localized spectral sensitivity, that is, a small quantization error in a particular LSF element will result in deviation from unquantized LPC spectrum only in the frequencies surrounding the quantized LSF. Such a property has been very effectively exploited in split-vector quantization of line spectral frequencies (12) to achieve transparent quantization. Here the L-element LSF vector is split into subvectors and each subvector is independently vector-quantized. Such a scheme permits different-size codebooks to be used for different subvectors, depending on importance placed on the subband of frequency encompassed by the subvector. It is to be noted that the LSF vector that is formed after quantization has to possess the ordering property in order to retain stability of the LPC synthesis filter. Such a constraint forces a partial search on the individual codebooks, thereby resulting in increased quantization error. Attempts to circumvent this problem include training the codebook vectors in a constrained fashion, and populating the codebooks after training in a manner such that complete search is possible (13).

EXCITATION MODELING

The LPC analysis described serves to model the vocal tract parameters [namely, $A(z)$ of Eq. (6)] of the human speech production mechanism. In order to reproduce speech at the remote decoder, in addition to vocal tract parameters, it is necessary to model, digitize, and transmit the excitation parameters [namely, $E(z)$ of Eq. (6)] of the human speech production mechanism using as few bits as possible. From a speech encoder point of view, $E(z)$ can be treated as the output of a system $A(z)$ whose input is speech $S(z)$; in other words, $E(z)$ is the LPC residual. From a decoder point of view, $E(z)$ can be interpreted as the input to the LPC synthesis filter $1/A(z)$ whose output is the speech signal $S(z)$. Hence the terms "LPC residual" and "excitation sequence" are used interchangeably.

Perhaps the simplest excitation model that can be used to synthesize intelligible speech is the two-state excitation model, as was demonstrated by Dudley in his vocoder apparatus in 1939. The model is illustrated in Fig. 7. Here a periodic excitation (typically a set of equally spaced impulses) whose period is equal to the reciprocal of the fundamental frequency of the segment of speech under consideration is used for voiced speech, and for unvoiced speech a noiselike excitation is used that is independent of the talker or speech material under consideration.

The fundamental frequency or pitch is typically estimated by determining the peak of the autocorrelation of the LPC residual signal for a given range of autocorrelation lags. Typically the range of autocorrelation lags over which the pitch period is determined is between 20 and 120, corresponding to pitch frequencies between 400 Hz and 66 Hz for 8-kHz-sampled speech signals. In order to reduce the complexity associated with the computation of autocorrelation function for about a 100 lag values (between 20 and 120), the LPC residual signal is low-pass filtered to less than 100 Hz and downsampled (decimated) by 4. The *simplified inverse filtering technique* (SIFT) of pitch estimation is based on this approach (14). Other approaches have been reported in the literature, such as picking the peak of the cepstrum of the speech signal within a given range; computing the average of the magnitudes of the differences (known as the AMD function or AMDF) between speech samples with different offsets (belonging to a given range) and picking the offset with least AMD as the pitch value (14). A more comprehensive view of pitch estimation techniques is provided in Ref. 15. The decision as to whether a speech frame is voiced or unvoiced is usually made—using features such as energy measurements, zero-crossing rate, peak value of the autocorrelation function, and magnitude of the first reflection coefficient—by evaluating a weighted distortion measure and decision thresholds chosen to obtain a low misclassification error (16,17).

The two-state excitation model illustrated in Fig. 7 can produce intelligible-quality speech at very low bit rates of 2400 bits/s or less, but has serious limitations in producing high-quality speech that will display naturalness and the characteristics of the talker. The reasons are multifold:

1. The LPC residuals for voiced sounds are not impulses, and hence when the LPC synthesis filter is driven by a series of impulses, the resulting speech gets noticeably degraded.
2. The LPC residual for unvoiced sounds is not truly bandlimited white noise; it has a spectral tilt to it depending on the inadequacies of the LPC modeling. Hence when the LPC synthesis filter is driven by bandlimited white noise, the resulting speech gets noticeably degraded.
3. The LPC technique is often unable to model poles that are close to each other.
4. The all-pole model is not accurate for nasal sounds, which are characterized by zeros as well.
5. The LPC coefficients need to be quantized before transmitting them on a digital channel and hence suffers quantization errors, resulting in shifting of the resonant frequencies.
6. The model heavily relies upon voice–unvoiced classification of segments of speech and hence depends on the accuracy of that classification. For some plosives and voiced fricatives, an accurate classification is difficult, since they do not bear the characteristics of completely voiced or unvoiced sounds.

It is clear that in order to produce natural-sounding speech, the LPC residual signal has to be adequately represented and transmitted to the voice decoder, for use as the excitation sequence for the LPC synthesis filter at the remote voice decoder. (If the prediction residual obtained using quantized LPC parameters at the encoder is used at the remote decoder, then exact reconstruction of original speech is possible.) The bulk of the research in speech coding over the last decade has been devoted to arriving at good excitation models rather than to improving the deficiencies of the LPC analysis. Part of the motivation behind that approach is that it might lead to a single unified technique that will provide a solution to all of the inadequacies mentioned above.

The inadequacies of the two-state excitation model (in conjunction with LPC analysis) were initially addressed by representing the LPC residual (and hence the excitation sequence input to the voice decoder) as a combination of periodic and aperiodic signals. Such schemes not only led to elimination of excitation sequences that were purely based on pitch estimation and voiced–unvoiced decisions, but also consequently led to reduction of some of the buzzy artifacts in voiced segments of reconstructed speech due to excessive periodicity (18–22). Among these, two schemes, namely the *residual excited linear predictive* (RELP) coder and the *mixed excited linear predictive* (MELP) coders, gained prominence in this effort. They are described briefly below.

Open-Loop Excitation Modeling

The RELP and MELP coders mentioned above belong to a class of excitation modeling techniques called *open-loop modeling* because their objective is to best represent the LPC residual signal. No feedback is provided as to whether the model parameters yield good-quality speech in a perceptual sense.

Residual Excited Linear Predictive Vocoder. Here the LPC residual signal is low-pass filtered, and the decimated residual signal is then encoded using ADM techniques described above and transmitted to the remote voice decoder (18). A block diagram of the RELP encoder and decoder is shown in Fig. 8. The low-pass filter has a cutoff frequency that is at least high enough to accommodate the second harmonic of the highest possible fundamental frequency. Hence, for low-bit-rate applications, the RELP coder typically uses cutoff frequencies in the vicinity of 800 Hz to 1000 Hz, based on the assumption that the highest possible human fundamental frequency is about 400 Hz. At the receiver, the low-pass-filtered residual signal is recovered using an adaptive delta demodulator and then processed using a nonlinear device to generate higher harmonics of the residual signal. The spectral flattener shown in Fig. 8 actually contains a nonlinear device that generates higher harmonics of the residual signal with decreasing strengths; hence a double differencer is used to enhance the higher harmonics (effectively provide a flat spectrum at higher harmonics); this is followed by a high-pass filter so as to only use the spectrally flattened excitation spectrum for higher harmonics. This is illustrated in Fig. 8. The lower harmonics will be the same as obtained using an ADM. Finally, a controlled amount of white noise is mixed with the spectrally flattened excitation signal before using it to drive the LPC synthesis filter so as to reduce any buzziness in voiced segments of reconstructed speech.

It is extremely important to note that there is no explicit pitch extraction or transmission associated with RELP coding, thereby making the scheme pitch-independent and immune to pitch determination and tracking errors.

Mixed Excitation Linear Predictive Coder. One of the shortcomings of the RELP coder is the bit rate necessary to perform an ADM coding of the low-pass-filtered decimated residual signal. Typical bit rates in the range of 6 kbit/s to 7 kbit/s have been reported in the literature as being necessary to encode the residual signal in order to produce natural-sounding speech. For speech coders that are required to operate at bit rates of 4.8 kbit/s and below, RELP coders are unsuitable. The MELP coders (19,20) provides a low-bit-rate alternative to the RELP, and at the same time removes the inadequacies

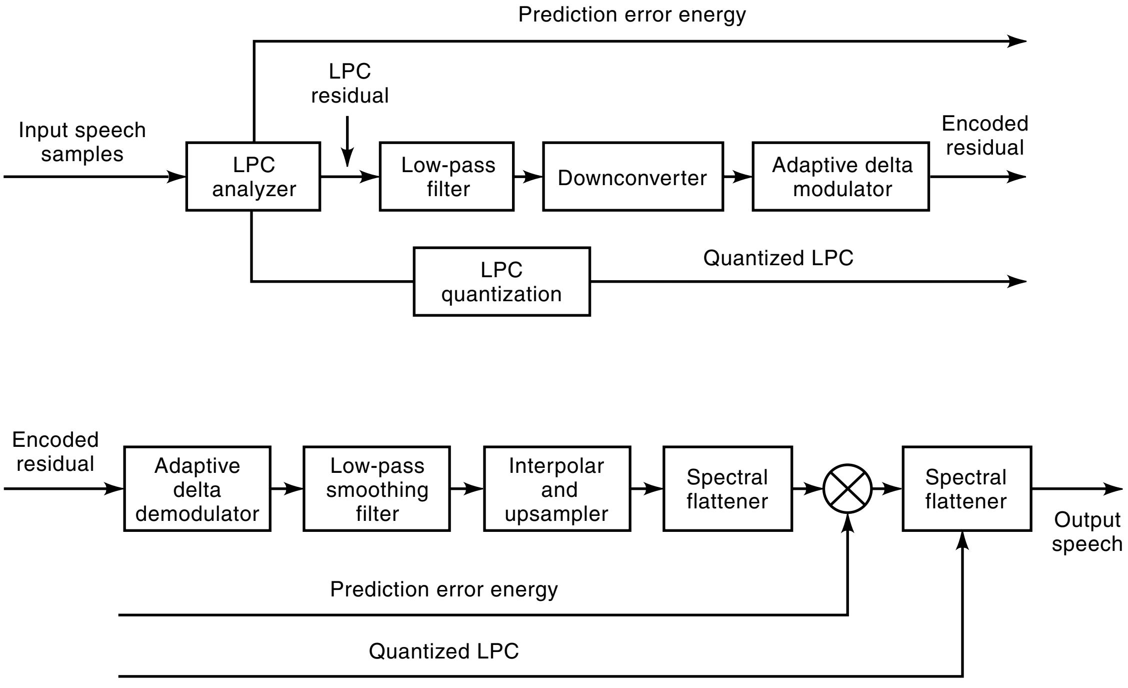


Figure 8. Residual excited linear prediction (RELP) coder.

of the two-state excitation modeling, both in removing buzziness in reconstructed speech and in providing robustness to voicing decision errors. Unlike RELP coders, where pitch estimation was absent, MELP coders perform pitch estimation and incorporate frequency-dependent voicing measures to model the LPC residual signal. Here the residual signal is modeled as a sum of aperiodic and periodic components. The periodic component is generated by an impulse train passed through a low-pass filter, and the aperiodic component is generated using white noise and a high-pass filter. The gains of the two components are chosen so that the overall excitation spectrum is flat. A more sophisticated mixed excitation scheme (see Fig. 9) includes a method of introducing a controlled amount of jitter into the impulse train to reflect the amount of periodicity in the LPC residual signal. Furthermore, an adaptive spectral enhancer is used to emphasize the energy in formant regions of speech. A MELP coder operating at 2.4 kbit/s was recently standardized by the U.S. Department of Defense for use in the military.

Closed-Loop Excitation Modeling—Analysis-by-Synthesis Coding

One of the disadvantages of the open-loop excitation modeling schemes described in the preceding subsection is that the excitation parameters are extracted in such a way that they best represent the LPC residual signal, and not necessarily something that will result in reconstructed speech that is close to original speech signal. Furthermore, the perception-based residual modeling in RELP and MELP is optimal in the LPC residual domain. A desirable alternative is to replicate the voice decoder operation in the encoder and extract encoder parameters that will minimize a perceptually weighted distortion measure between original speech and reconstructed speech. Such a technique, whereby the analysis parameters are optimized for synthesizing speech in the encoder, is referred to as *analysis-by-synthesis coding*. This is a powerful technique in that it not only optimizes parameters in the speech domain, but also permits perceptual weighting to be performed in the speech domain, which is highly desirable.

The perceptual weighting of the difference between the original speech signal and reconstructed speech signal is a key feature of most analysis-by-synthesis coders. Here the difference signal is typically passed through a time-varying pole–zero filter that will shape the spectral distortions (due to imprecise modeling and quantization) to be strong in formant regions and weak in valley regions, thereby striving to achieve a constant signal-to-noise ratio across the frequency spectrum of interest. Such perceptual weighting essentially exploits noise masking properties of the human auditory system, thereby making quantization noise in spectral valleys inaudible.

In general the perceptual weighting filter that is used in most speech coders is of the form

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)} \tag{12}$$

where $A(z)$ is the unquantized LPC spectrum. While the actual effect of $\gamma_1,\gamma_2 \in [0,\ 1]$ is to broaden the poles and zeros (peaks and valleys of the spectrum) in a controlled manner, the real purpose of using such a weighted filter is to perform optimization that will match the synthesized speech more closely at spectral valleys than at spectral peaks, thereby achieving the desired constancy in signal-to-noise ratio.

Perhaps the optimal choice of a weighted error criterion would be to perform an analysis-by-synthesis coding that optimizes the difference between original speech signal and synthesized speech signal after passing through the human speech perception mechanism. This is illustrated in Fig. 10. In the $f(X)$ figure is in general a nonlinear function of the vector X; it represents the transfer function of the human speech perception mechanism, including the cochlea of the human ear (23). $d(X_1, X_2)$ is a decision function, which ideally should represent the decision-making process in the human brain. However, in order to reduce the complexity of the search procedure and for the sake of analytical tractability, a simplified weighted distortion measure such as the weighted linear minimum mean squared error is used as optimization criterion. Here $f(X)$ is replaced by WX, where W is an $N \times N$ matrix whose entries represent the impulse response of the filter $W(z)$ of Eq. (12), which coarsely represents the human speech perception mechanism. The distortion measure $d(\cdot\,,\ \cdot)$ is typically chosen to be an L_2 norm, which again is a highly simplified version of the complex decision-making process of the human brain.

In practice, analysis-by-synthesis coders are implemented as a sequential optimization process (rather than a joint opti-

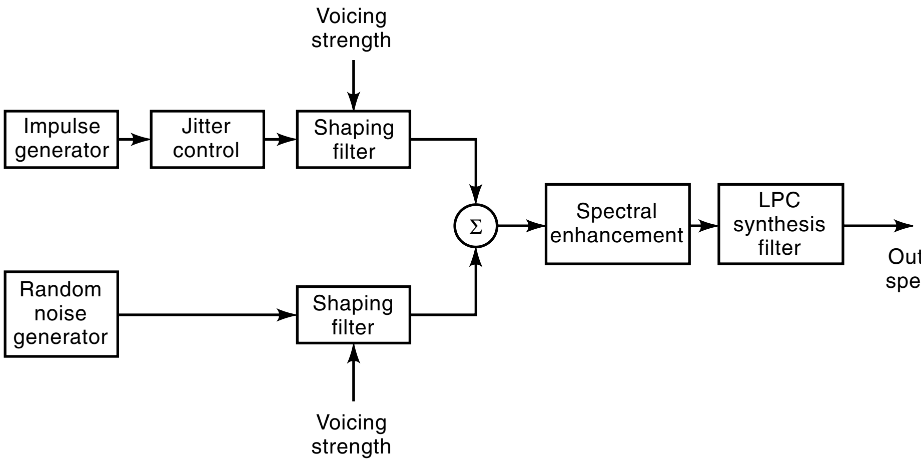


Figure 9. Mixed excitation linear prediction (MELP) coder.

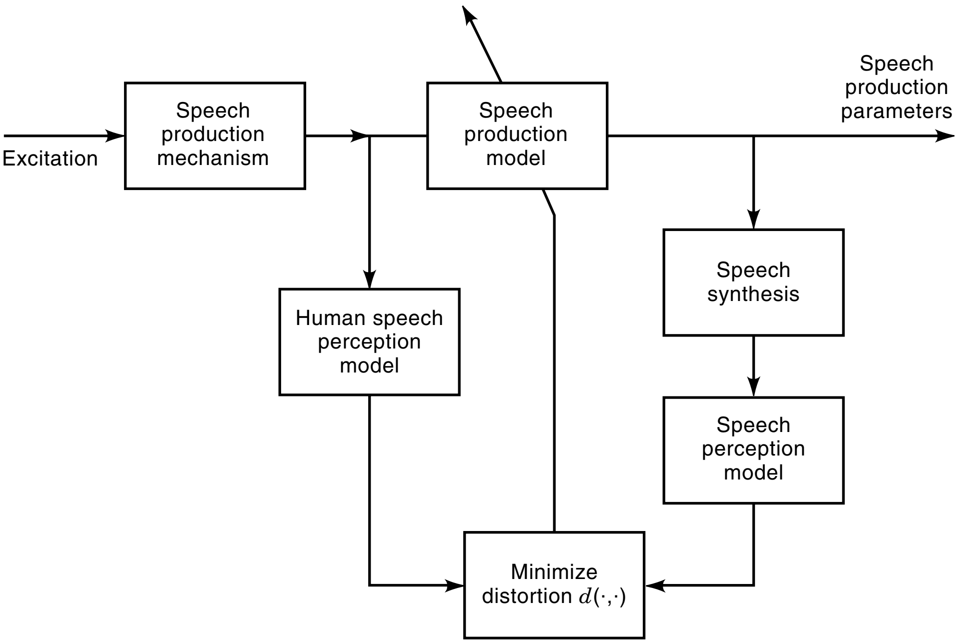

Excitation
Speech production mechanism
Speech production model
Speech production parameters
Human speech perception model
Speech synthesis
Speech perception model
Minimize distortion $d(\cdot,\cdot)$

Figure 10. The ideal analysis-by-synthesis coding method.

mization process) whereby closed-loop optimization is only performed to extract parameters of the excitation model. This is illustrated in Fig. 11, where the LPC analysis is performed in an open-loop fashion and then the excitation parameters are extracted using the closed-loop analysis-by-synthesis approach.

The earliest pioneering and practical work in analysis-by-synthesis coding that used the perceptual weighting criterion was reported by Atal and Remde (24), wherein a multipulse excitation scheme was proposed to represent the LPC residual signal. Here the LPC residual signal is represented by a sparse sequence of pulses separated by zeros. This is described in the following section.

Multipulse Linear Predictive Coding Method. The fundamental principle behind the multipulse LPC (MPLPC) technique is that only a fraction of the prediction residual samples that are perceptually important yield a high degree of naturalness in reconstructed speech, and hence it is not necessary to transmit all prediction residual samples. It is indeed true that if all prediction residual samples were transmitted to a remote voice decoder with a very high signal-to-quantization-noise ratio, then the reconstructed speech would be very close to the original speech. The objective here, however, is to reduce the bit rate and still achieve natural-sounding speech.

Here a subframe sequence of N prediction residual samples $\{r_0, r_1, \ldots, r_{N-1}\}$ that is obtained using quantized LPC

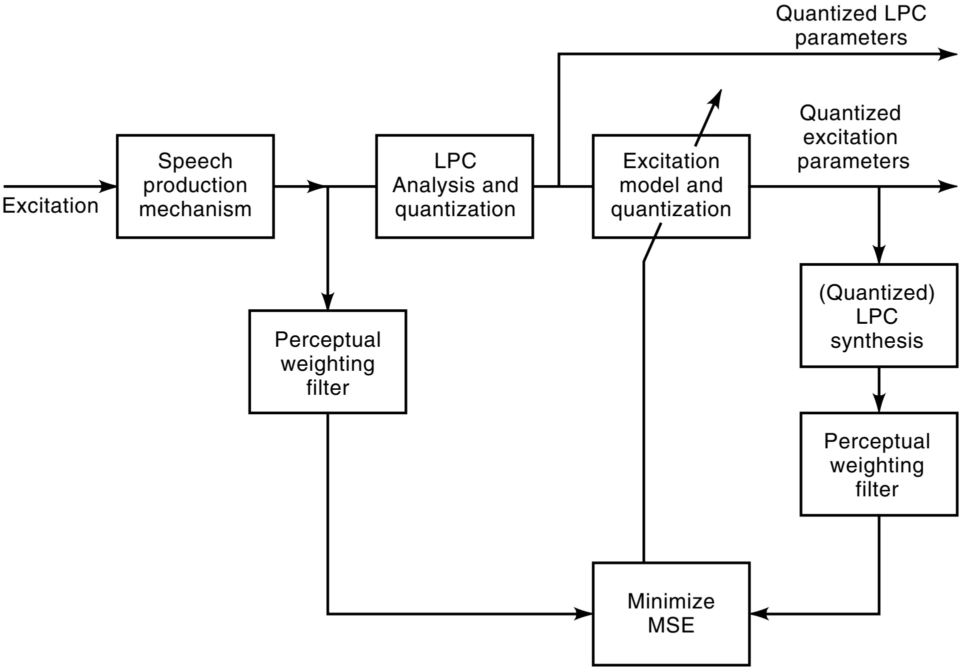

Quantized LPC parameters
Quantized excitation parameters
Excitation
Speech production mechanism
LPC Analysis and quantization
Excitation model and quantization
Perceptual weighting filter
(Quantized) LPC synthesis
Perceptual weighting filter
Minimize MSE

Figure 11. A practical analysis-by-synthesis coding method.

coefficients is represented by a set of P ($P \ll N$) pulses with positions $p_1, p_2, \ldots, p_p$, $p_i \in [0, N-1]$, and nonzero amplitudes $g_1, g_2, \ldots, g_p$. There are

$$\binom{N}{P}$$

possible combinations of P pulse positions from among the N positions. The pulse positions p_i and their amplitudes g_i for a given combination are determined in a closed-loop analysis-by-synthesis method using the weighted mean squared error described above. Specifically, the cost function

$$J(s_\mathrm{w}(n), \tilde{s}_k(n)) = \sum_{j=0}^{N-1} [s_\mathrm{w}(n) - \tilde{s}_k(n)]^2$$

is minimized with respect to pulse positions and amplitudes, where $s_\mathrm{w}(n)$ is the perceptually weighted input speech signal and $\tilde{s}_k(n)$ is the perceptually weighted synthesized speech signal for the kth

$$\left(k \in \left[0, \binom{N}{P} - 1\right]\right)$$

combination of pulse positions and amplitudes. First, the perceptually weighted speech signal $s_\mathrm{w}(n)$ is obtained by passing the original speech signal $s(n)$ through the weighting filter $A(z/\gamma_1)/A(z/\gamma_2)$ as follows:

$$s_\mathrm{w}(n) = s(n) + \sum_{i=1}^{L} \alpha_i \gamma_1^i s(n-1) - \sum_{i=1}^{L} \alpha_i \gamma_2^i s_\mathrm{w}(n-i)$$

and subtracting the zero-input response of the weighting filter. Then $\tilde{s}_k(n)$ is obtained by passing the kth combination of pulse positions and amplitudes through the LPC synthesis filter $1/\hat{A}(z)$ and then performing perceptual weighting on the synthesized speech using the weighting filter $A(z/\gamma_1)/A(z/\gamma_2)$ in a manner similar to that for the original speech. In practice, however, these two steps are combined, and hence the chosen amplitudes are passed through a linear FIR filter whose impulse response represents the combination

$$\frac{A(z/\gamma_1)}{A(z/r_2)} \frac{1}{\hat{A}(z)}$$

Typically the truncated impulse response $h(n)$, $n = 0, 1, \ldots, N-1$, of the combined synthesis and weighting filter is obtained by filtering a signal consisting of the coefficients of the filter $A(z/\gamma_1)$ extended by zeros through two filters $1/\hat{A}(z)$ and $1/A(z/r_2)$:

$$\tilde{s}_k(n) = \sum_{j=1}^{P} g_{jk} h(n - p_{jk})$$

where p_{jk} and g_{jk} are the pulse positions and amplitudes corresponding to the kth combination. Minimizing $J(s_\mathrm{w}(n), \tilde{s}_k(n))$ individually with respect to g_{ik}, $i = 1, \ldots, P$, according to

$$\frac{\partial}{\partial g_{ik}} \sum_{n=0}^{N-1} [s_\mathrm{w}(n) - \tilde{s}_k(n)]^2 = 0$$

yields a set of P simultaneous equations

$$\sum_{n=0}^{N-1} s_\mathrm{w}(n) h(n - p_{ik}) = \sum_{j=1}^{P} g_{jk} \sum_{n=0}^{N-1} h(n - p_{jk}) h(n - p_{ik})$$
$$1 \le i \le P$$

which in matrix form can be written as

$$\boldsymbol{Y}_k = A_k \boldsymbol{G}_k$$

where

$$\boldsymbol{Y}_k = \begin{bmatrix} \sum_{n=0}^{N-1} s_\mathrm{w}(n) h(n - p_{1k}) \\ \sum_{n=0}^{N-1} s_\mathrm{w}(n) h(n - p_{2k}) \\ \vdots \\ \sum_{n=0}^{N-1} s_\mathrm{w}(n) h(n - p_{Pk}) \end{bmatrix}$$

$$A_k = \begin{bmatrix} \sum_{n=0}^{N-1} h(n - p_{1k}) h(n - p_{1k}) & \sum_{n=0}^{N-1} h(n - p_{1k}) h(n - p_{2k}) & \cdots & \sum_{n=0}^{N-1} h(n - p_{1k}) h(n - p_{Pk}) \\ \sum_{n=0}^{N-1} h(n - p_{2k}) h(n - p_{1k}) & \sum_{n=0}^{N-1} h(n - p_{2k}) h(n - p_{2k}) & \cdots & \sum_{n=0}^{N-1} h(n - p_{2k}) h(n - p_{Pk}) \\ \vdots & \vdots & & \vdots \\ \sum_{n=0}^{N-1} h(n - p_{Pk}) h(n - p_{1k}) & \sum_{n=0}^{N-1} h(n - p_{1k}) h(n - p_{2k}) & \cdots & \sum_{n=0}^{N-1} h(n - p_{Pk}) h(n - p_{Pk}) \end{bmatrix}$$

and

$$\boldsymbol{G}_k = \begin{bmatrix} g_{1k} \\ g_{2k} \\ \vdots \\ g_{Pk} \end{bmatrix}$$

Hence the optimal set of amplitudes for the kth combination is given by

$$\boldsymbol{G}_k^* = A_k^{-1} \boldsymbol{Y}_k$$

and the resulting mean squared error is given by

$$E_k = \sum_{n=0}^{N-1} s_\mathrm{w}^2(n) - \boldsymbol{Y}_k^\mathrm{t} A_k^{-1} \boldsymbol{Y}_k \qquad (13)$$

Since A_k is a symmetric positive definite matrix, it is easy to see that the second term $\boldsymbol{Y}_k^\mathrm{t} A_k^{-1} \boldsymbol{Y}_k$ is positive. Hence minimizing $J(s_\mathrm{w}(n), \tilde{s}_k(n))$ is equivalent to maximizing the second term $\boldsymbol{Y}_k^\mathrm{t} A_k^{-1} \boldsymbol{Y}_k$ in the E_k expression in Eq. (13). Hence in practice, the second term is evaluated for all possible

$$\binom{N}{P}$$

combinations of P pulses, and the combination $k = k^*$ that yields the maximum value of $\boldsymbol{Y}_k^\mathrm{t} A_k^{-1} \boldsymbol{Y}_k$ is chosen as the optimal combination that yields the least mean squared error in a perceptually weighted sense.

The MPLPC technique has the distinct advantage in that it does not depend on the pitch estimation or voiced–unvoiced decision. Hence it provides a unified framework for representing the excitation of the decoder for all types of speech segments. It has two drawbacks, however: the computational complexity necessary to determine the optimal pulse positions and their amplitudes, and the bit rate necessary to transmit them. As an example, if it is determined that five nonzero samples ($P = 5$) have to be identified in a duration of 5 ms ($N = 40$ at 8 kHz sampling rate), then the number of possible combinations for which E_k has to be computed is

$$\binom{40}{5} = 658,008$$

Hence less complex suboptimal schemes have been proposed in the literature.

The earliest proposal was to perform sequential search, that is, determine one pulse location and amplitude at a time. Here the optimal first pulse position p_1^* (and its amplitude g_1^*) is determined from among N possible choices by computing the second term of the equation for $\tilde{s}_k(n)$ with $P = 1$, namely

$$W_k = \frac{[\sum_{n=0}^{N-1} s_{\mathrm{w}}(n)h(n-k)]^2}{\sum_{n=0}^{N-1} h^2(n-k)}$$

for $k = 0, 1, \ldots, N - 1$ and determining the value of k ($= p_1^*$) for which W_k is maximum. The corresponding amplitude g_1^* is determined using

$$g_1^* = \frac{\sum_{n=0}^{N-1} s_{\mathrm{w}}(n)h(n-p_1^*)}{\sum_{n=0}^{N-1} h^2(n-p_1^*)}$$

Subsequent pulse positions p_m^* and g_m^*, $2 \le m \le P$, are obtained one by one by minimizing the cost function

$$J(s_{\mathrm{w}}(n), \tilde{s}_m(n)) = \sum_{j=0}^{N-1} [s_{\mathrm{w}}(n) - \tilde{s}_m(n)]^2$$

where

$$\tilde{s}_m(n) = \sum_{j=1}^{m-1} g_j^* h(n-p_j^*) + g_m h(n-p_m)$$

with respect to g_m and p_m, or equivalently, computing

$$W_{km} = \frac{[\sum_{n=0}^{N-1} s_{\mathrm{w}m}(n)h(n-k)]^2}{\sum_{n-0}^{N-1} h^2(n-k)}, \qquad 0 \le k \le N-1$$
$$k \notin \{p_1^*, p_2^*, \ldots, p_{m-1}^*\}$$

where

$$s_{\mathrm{w}m}(n) = s_{\mathrm{w}}(n) - \sum_{j=1}^{m-1} g_j^* h(n-p_j^*)$$

and finding the value of k ($=p_m^*$) for which W_{km} is maximized. The corresponding amplitude g_m^* is obtained using

$$g_m^* = \frac{\sum_{n=0}^{N-1} s_{\mathrm{w}m}(n)h(n-p_m^*)}{\sum_{n=0}^{N-1} h^2(n-p_m^*)}$$

The total number of pulse positions searched in this sequential procedure is $NP + P(P - 1)/2$. For the above example of $N = 40$ and $P = 5$, this turns out to be 210, as compared to 658,008 for the optimal full-blown search. Hence a significant saving in complexity is achieved in sequential search. This saving is in addition to the significant savings achieved by not needing to invert matrices using Cholesky decomposition for every possible combination of pulse position as in optimal search.

Regular Pulse Excitation Coding. Another popular analysis-by-synthesis technique that has been reported in the literature and that reduces the computational complexity and the bit rate is the *regular pulse excitation* (RPE) coding technique, which is used as the basis for the design of the Global System for Mobile (GSM) Communications full-rate speech coder. Here the spacing between the nonzero pulses is held constant (25). This implies that the only position that needs to be transmitted to the decoder is the position of the first pulse relative to the start of a speech subframe, thereby achieving a significant reduction in bit rate or equivalently bandwidth. It is reasonable to expect then that in MPLPC the pulse positions that get chosen to be transmitted in voiced speech segments will include pitch pulses, in the absence of which a large mean squared error would result. In RPE coding, the constraint imposed by equal spacing regardless of pitch period causes a severely suboptimal grid of pulses to be selected. Hence in both MPLP and RPE coding, there is a strong incentive (in terms of optimal selection of pulses) to first perform long-term prediction that removes periodicity in the LPC residual signal (and hence eliminates the strong pitch pulses) and then perform the coding. This would require fewer pulses and fewer bits to transmit each pulse amplitude, thereby resulting in bit-rate reduction.

Modeling Periodic Component of Excitation—Long-Term Prediction. As described above, long term prediction (LTP) is performed essentially to remove the periodic component from the residual signal and then model the LTP residual using techniques such as MPLPC and RPE described above. The problem of LTP is typically formulated as follows.

Let $e(n)$ denote the LPC residual at time instant n, let $e(n)$ be predicted from $e(n - D - M)$, $e(n - D - M + 1), \ldots,$ $e(n - D - 1)$, $e(n - D)$, $e(n - D + 1), \ldots, e(n - D + M)$ and let the LTP residual be denoted by $w(n)$. Then

$$e(n) = \sum_{i=-M}^{M} \beta_i e(n-D-i) + w(n) \tag{14}$$

If the signal is truly periodic with period equal to ΔT, where T is the sampling interval (125 μs for 8 kHz sampling rate) and Δ a positive integer, then $w(n) = 0$ when $\beta_0 = 1.0$, $D = \Delta$, and $M = 0$. Since speech is a slowly varying nonstationary process, it is required to estimate D and β_i on short segments (typically every 5 ms, corresponding to 40 samples) of residual signal. While it is desirable to optimize the value of M, for reasons of complexity M is chosen to be less than or equal to 1. Estimation of D and β_i can be performed in an open-loop or closed-loop fashion. In the open-loop method, the objective is to estimate D and β_i that minimize the mean square of the LTP error $w(n)$. In the closed-loop method, the objective is to estimate D and β_i that when used in the voice decoder will minimize a perceptually weighted distortion between reconstructed speech and original speech at the input of the voice encoder. As will be described later, some speech coders (open-loop and closed-loop) try to estimate D with fractional resolution (as opposed to integer resolution) by either interpolating the residual signal itself or interpolating the autocorrelation function of the residual signal.

Open-Loop Long-Term Prediction. For the predictor formulation in Eq. (14) with $M = 1$, the equivalent of the Yule–Walker equations can be written as

$$\sum_n e_D(n)e(n) = \left(\sum_n e_D(n)e_D^T(n)\right)\begin{bmatrix}\beta_{-1}\\ \beta_0\\ \beta_1\end{bmatrix}$$

based on which the optimum value of $\boldsymbol{\beta}^* = [\beta_{-1}\ \beta_0\ \beta_1]^{\mathrm{t}}$ is obtained as

$$\boldsymbol{\beta}^* = \Phi^{-1}\sum_n \boldsymbol{e}_D(n)e(n) \tag{15}$$

for a given value of D, where

$$\boldsymbol{e}_D(n) = \begin{bmatrix} e(n-D-1)\\ e(n-D)\\ e(n-D+1)\end{bmatrix}$$

and

$$\Phi = \sum \boldsymbol{e}_D(n)\boldsymbol{e}_D^{\mathrm{t}}(n)$$

The resulting mean squared error is given by

$$E_D = \sum_n e^2(n) - \left(\sum_n \boldsymbol{e}_D(n)e(n)\right)^{\mathrm{t}}\Phi^{-1}\left(\sum_n \boldsymbol{e}_D(n)e(n)\right)$$

Since Φ (and hence Φ^{-1}) is a positive definite matrix, minimizing E_D is equivalent to maximizing $[\sum_n \boldsymbol{e}_D(n)e(n)]^{\mathrm{t}}\ \Phi^{-1}[\sum_n \boldsymbol{e}_D(n)e(n)]$. Hence in practice, $[\sum_n \boldsymbol{e}_D(n)e(n)]^{\mathrm{t}}\ \Phi^{-1}[\sum_n \boldsymbol{e}_D(n)e(n)]$ is computed for all possible values of D, and the value of D that yields the maximum value is chosen as the optimal delay value. For this optimal value of D, the LTP coefficient vector $\boldsymbol{\beta}^* = [\beta_{-1}\ \beta_0\ \beta_1]^{\mathrm{t}}$ is computed according to Eq. (15).

Closed-Loop Long-Term Prediction. LTP using closed-loop search is based an analysis-by-synthesis approach, whereby the LTP parameters (D, $\boldsymbol{\beta}$) are optimized for reconstructing speech over permissible values of D (and corresponding $\boldsymbol{\beta}$) by comparison with the original speech using a perceptual weighted filter

$$\frac{A(z/\gamma_1)}{A(z/\gamma_2)}$$

Essentially, the cost function $J(D, \boldsymbol{\beta}) = \|s_{\mathrm{w}}(n) - z(n) - \tilde{s}_D\boldsymbol{\beta}^{(n)}\|^2$ is minimized with respect to D and $\boldsymbol{\beta}$, where $s_{\mathrm{w}}(n)$ is the perceptually weighted input speech, typically derived as

$$s_{\mathrm{w}}(n) = s(n) + \sum_{i=1}^{L}\alpha_i\gamma_1^i s(n-i) - \sum_{i=1}^{L}\alpha_i\gamma_2 s_{\mathrm{w}}(n-i)$$

Here s_{w} is the input speech signal and α_i ($i = 1, 2, \ldots, L$) are the unquantized LPC coefficients. $z(n)$ is the zero-input response of the perceptually weighted synthesis filter

$$\frac{A(z/\gamma_1)}{A(z/\gamma_2)}\cdot\hat{A}(z)$$

which is subtracted from $s_{\mathrm{w}}(n)$ in the equation above to remove the contribution of the previous frame in the optimization process. $\hat{A}(z)$ is the quantized LPC filter that is actually used at the remote decoder to synthesize speech.

Let $\tilde{s}_{D,}\boldsymbol{\beta}(n)$ be the convolution of the truncated impulse response of the perceptually weighted synthesis filter and the past synthetic residual at signal delay D, computed as

$$\tilde{s}_{D,\boldsymbol{\beta}}(n) = \sum_{i=-1}^{1}\beta_i\sum_{j=0}^{N-1}\tilde{e}(n-D-i-j)h(j)\qquad n = 0, 1, 2, \ldots, N-1 \tag{16a}$$

where $h(0), h(1), \ldots, h(N-1)$ is the truncated impulse response of the weighted synthesis filter $[A(z/\gamma_1)/A(z/\gamma_2)]\hat{A}(z)$, which is computed in a manner similar to that described for multipulse coding above.

From an analysis similar to that for open-loop LTP, it can be shown that minimization of $J(D, \boldsymbol{\beta})$ above is equivalent to the maximization of

$$\left(\sum_{n=0}^{N-1}\boldsymbol{e}_{\mathrm{w}}^{(D)}(n)[\boldsymbol{s}_{\mathrm{w}}(n) - z(n)]\right)^{\mathrm{t}}\Phi_D^{-1}\left(\sum_{n=0}^{N-1}\boldsymbol{e}_{\mathrm{w}}^{(D)}(n)[\boldsymbol{s}_{\mathrm{w}}(n) - z(n)]\right) \tag{16b}$$

for all possible values of D (typically between $D = 20$ and $D = 128$), where

$$\boldsymbol{e}_{\mathrm{w}}^{(D)}(n) = \begin{bmatrix}\sum_{j=0}^{N-1}\tilde{e}(n-D+1-j)h(j)\\ \sum_{j=0}^{N-1}\tilde{e}(n-D-j)h(j)\\ \sum_{j=0}^{N-1}\tilde{e}(n-D-1-j)h(j)\end{bmatrix}$$

and

$$\Phi_D = \sum_{n=0}^{N-1}\boldsymbol{e}_{\mathrm{w}}^{(D)}(n)[\boldsymbol{e}_{\mathrm{w}}^{(D)}(n)]^{\mathrm{t}}$$

Once the value of $D = D^*$ that maximizes Eq. (16b) is obtained, the optimal vector $\boldsymbol{\beta}^* = [\beta_{-1}^*, \beta_0^*, \beta_1^*]$ is obtained as

$$\boldsymbol{\beta}^* = \Phi_{D^*}^{-1}\sum_{n=0}^{N-1}\boldsymbol{e}_{\mathrm{w}}^{(D^*)}(n)[s_{\mathrm{w}}(n) - z(n)]$$

In most speech coders, in order to reduce the complexity of determining the vector of long-term coefficients, $\boldsymbol{\beta}$, corresponding to delays of D, $D - 1$, and $D + 1$, a scalar coefficient corresponding to the delay D alone is computed. In this case, for open-loop LTP, the solution for β_0 is obtained by first maximizing

$$\frac{[\sum_{n=0}^{N-1} \tilde{e}(n-D)e(n)]^2}{\sum_{n=0+}^{N-1} \tilde{e}^2(n-D)} \tag{17}$$

for permissible values of D and then computing optimal $\beta_0 = \beta_0^*$ as

$$\beta_0^* = \frac{\sum_{n=0}^{N-1} \tilde{e}(n-D^*)e(n)}{\sum \tilde{e}^2(n-D^*)}$$

where D^* is the value of D that maximizes Eq. (17).

For closed-loop LTP, the solution for β_0 is obtained by first maximizing

$$\frac{\left(\sum_{n=0}^{N-1}\sum_{j=0}^{N-1} \tilde{e}(n-D-j)h(j)[\boldsymbol{s}_{\mathrm{w}}(n) - z(n)]\right)^2}{\sum_{n=0}^{N-1}\left(\sum_{j=0}^{N-1} \tilde{e}(n-D-j)h(j)\right)^2} \tag{18}$$

and computing the optimal $\beta_0 = \beta_0^*$ as

$$\beta_0^* = \frac{\sum_{n=0}^{N-1}\sum_{j=0}^{N-1} \tilde{e}(n-D^*-j)h(j)[S_{\mathrm{w}}(n) - z(n)]}{\sum_{n=0}^{N-1}\left(\sum_{j=0}^{N-1} \tilde{e}(n-D^*-j)h(j)\right)^2}$$

where D^* is the value of D that maximizes Eq. (18).

One advantage of having a vector of long-term coefficients (corresponding to delay values of D, $D - 1$, and $D + 1$) rather than a single coefficient β_0 is that it covers the case where the delay D is not an integer but fractional. It is important to note that the role of D is to represent the fundamental frequency f_0 of the talker for the speech segment under consideration.

However, as noted above, D is computed with the resolution of the sampling period. In reality, the LPC prediction residual is periodic with a duration $1/f_0$ that is not an integer multiple of sampling period ($1/f_s$, where f_s is the sampling frequency, typically 8 kHz). Hence, many speech coders attempt to estimate the correct fundamental frequency by computing D with a fractional resolution rather than an integer resolution. Computation of fractional values of D is typically performed in one of the following two ways. In the more rigorous method, the LPC residual signal (actually the previous excitation sequence) is upsampled by using interpolation filters and then maximizing the equations above for all fractional values within the permissible delay range (fractional resolution is typically one-fourth or one-eighth). In a less rigorous method, the terms in the above equation itself are interpolated around an integer value of D, and the fractional value of D for which the expression above is maximized is obtained. A typical interpolation filter is implemented using a finite impulse response (FIR) filter based on a truncated windowed sinc function such as a Hamming-windowed sinc function.

It is important to note in the above analysis that $\tilde{e}(n)$ is the excitation used to drive the LPC synthesis filter in the previous subframe. As evident from Eq. (16a), for values of $D < N + 1$, past synthetic excitation does not exist. In such cases the previous excitation is repeated with a periodicity of D and essentially extended in order to compute the optimal set of long-term coefficients. This idea is a key component to the concept of adaptive codebook (69), as will be mentioned in the next section.

Modeling the Aperiodic Component of Excitation. As described above, the periodic component modeling using LTP permits efficient modeling of the LTP residual using MP–LPC and RPE techniques. In fact, multipulse modeling of an aperiodic component of the excitation sequence is employed in the Inmarsat full-rate aeronautical speech-coding standard, operating at 9.6 kbit/s. The full-rate GSM speech-coding standard, operating at 13 kbit/s, uses the RPE technique after LTP to model the periodic component of the LPC residual signal.

However, by far the most popular analysis-by-synthesis technique, which has gained widespread importance and been employed by many speech coders, including many regional and international speech-coding standards, is the code-excited linear prediction (CELP) technique (26,68). Here, short segments of the LTP residual signal are approximated by an entry in the codebook of stored vectors. Essentially, a VQ of the LTP residual is performed, but the choice of a codebook entry is based upon a synthesizing speech signal using each entry of the codebook and comparing it with original speech in a perceptually weighted domain. Even the estimation of the periodic component of the LPC residual in an analysis-by-synthesis coder can be treated as selection from a codebook whose entries for any subframe of speech consists of previous excitation signals delayed by different amounts (integer and/or fractional). Obviously, the entries of such a codebook change from one subframe to another; hence it is called an *adaptive codebook* (69). It is noted that there is no need for an explicit long-term prediction filter when an adaptive codebook search is performed.

The power of the CELP technique is that it encompasses most of the excitation modeling techniques, including multipulse and RPE modeling, as special cases. The CELP technique has also been sometimes referred to as vector excitation coding (VXC) and stochastically excited linear prediction (SELP), depending on the nature of the codebooks.

A typical CELP encoder block diagram is shown in Fig. 12. The aperiodic component of the excitation model is selected from a codebook of stored vectors, whose kth vector is denoted by $\boldsymbol{C}_k = [C_{0k}, C_{1k}, \ldots, C_{N-1,k}]^{\mathrm{t}}$, where N is the dimension of vectors in the codebook (typically $N = 40$, equivalent to 5 ms). Each N-dimensional vector in the stored codebook represents the shape of an N-sample signal. Depending on the signal strength and the LTP prediction gain, the energy in the N-sample LTP prediction residual varies. Therefore, the appropriate gain value is computed after the shape is computed. Such a procedure is referred to as gain–shape representation of signals.

The objective here is to find an entry (vector) in the codebooks (and the associated gain) that when used in conjunction with the periodic component of an excitation sequence is input to the LPC synthesis filter, producing speech that is close to original speech in a perceptually weighted sense. The problem is formulated as the minimization of $J(\boldsymbol{C}_k, g_c)$ with respect to $\boldsymbol{C}_k$ and g_c:

$$J(\boldsymbol{C}_k, g_{\mathrm{c}}) = \|\boldsymbol{S}_{\mathrm{w}} - \boldsymbol{Z} - H(g_{\mathrm{c}}\boldsymbol{C}_k + \hat{\beta}_0^* \tilde{\boldsymbol{e}}^{(D^*)})\|^2 \tag{19}$$

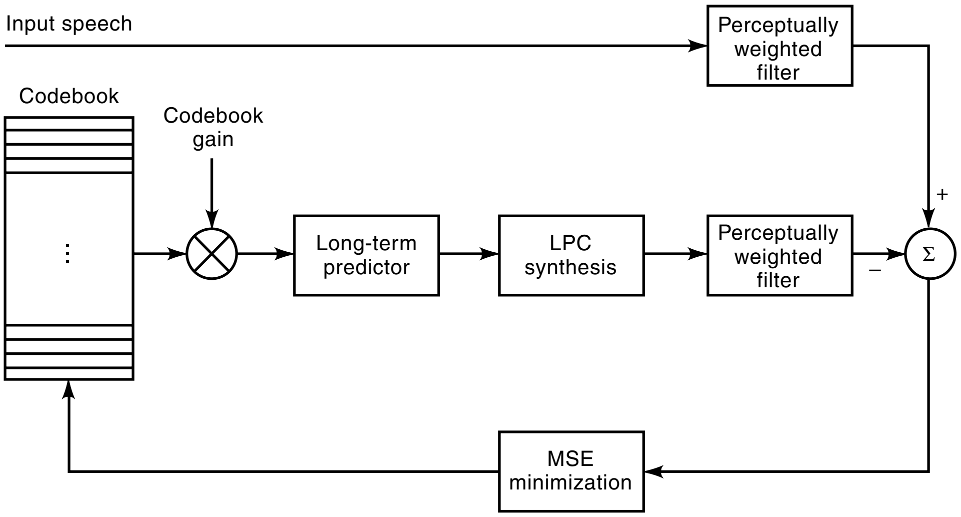


Figure 12. Code-excited linear predictive (CELP) coder.

where

$$\boldsymbol{S}_{\mathrm{w}} = [s_{\mathrm{w}}(0) \quad s_{\mathrm{w}}(1) \quad \cdots \quad s_{\mathrm{w}}(N-1)]^{\mathrm{t}}$$
$$\boldsymbol{Z} = [z(0) \quad z(1) \quad \cdots \quad z(N-1)]^{\mathrm{t}}$$

H is an $N \times N$ lower triangular matrix whose jth row contains the truncated impulse response of the weighted synthesis filter, that is,

$$H = \begin{bmatrix} h(0) & 0 & 0 & \dots & 0 \\ h(1) & h(0) & 0 & \dots & 0 \\ h(2) & h(1) & h(0) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ h(N-1) & h(N-2) & h(N-3) & \dots & h(0) \end{bmatrix}$$

β_0^* is the quantized value of β_0^* as computed using the above equation corresponding to a delay D^*, and $\tilde{\boldsymbol{e}}^{(D^*)} = [\tilde{\boldsymbol{e}}(-D^*), \tilde{\boldsymbol{e}}(1 - D^*), \ldots, \tilde{\boldsymbol{e}}(N - 1 - D^*)]$ is the vector of aperiodic components of excitation based on past excitation. When $\boldsymbol{D}^*$ is fractional (noninteger), $\tilde{\boldsymbol{e}}(j - D^*)$ is obtained using an interpolation filter on past excitation. $s_{\mathrm{w}}(n)$, $z(n)$, and $h(n)$ are the same as those obtained for the LTP described above. The minimization of $J(\boldsymbol{C}_k, g_{\mathrm{c}})$ can be shown to be the same as the maximization of

$$\frac{[\boldsymbol{C}_k^{\mathrm{t}} H^{\mathrm{t}} (\boldsymbol{S}_{\mathrm{w}} - \boldsymbol{Z} - H \hat{\beta}_0^* \tilde{\boldsymbol{e}}^{(D^*)})]^2}{\boldsymbol{C}_k^{\mathrm{t}} H^{\mathrm{t}} H \boldsymbol{C}_k} \tag{20}$$

Therefore, the above expression is computed for each vector $\boldsymbol{C}_k$ ($k = 0, 1, \ldots, 2^{B-1}$) in the codebook, and the entry $\boldsymbol{C}_k = \boldsymbol{C}^*$k that maximizes the above expression is chosen as the shape vector that best represents the aperiodic component of the excitation sequence. The corresponding gain g_{c}^* is computed as

$$g_{\mathrm{c}}^* = \frac{\boldsymbol{C}_{k^*}^* H^{\mathrm{t}} (\boldsymbol{S}_{\mathrm{w}} - \boldsymbol{Z} - H \hat{\beta}_0^* \tilde{\boldsymbol{e}}^{(D^*)})}{\boldsymbol{C}_{k^*} H^{\mathrm{t}} H \boldsymbol{C}_{k^*}} \tag{21}$$

It is noted that the gain term g_{c}^* above has to be quantized before transmission to the remote decoder. An alternative approach (especially when the number of bits allocated for gain quantization is few) is to select the gain value from a small table that minimizes Eq. 19 above and simply transmit its index to the remote decoder. Therefore, in the CELP coder, the optimal excitation to the LPC synthesis filter $1/\hat{A}(z)$ is given by

$$\tilde{e}(n) = \hat{\beta}_0^* \tilde{e}(n - D^*) + g_{\mathrm{c}}^* C_{nk^*}, \qquad n = 0, 1, 2, \ldots, N-1$$

The original proposal for CELP, which is attributed to Atal and Schroeder (26), suggested the use of unstructured codebooks whose entries were Gaussian random numbers. This resulted in codebook search complexities that were prohibitively high, but the promise that the technique held intrigued many researchers. As a result, numerous articles were published on CELP whose codebook search complexities were reduced, usually by having structured codebooks. Overlapped codebooks (70) have been proposed, whereby a given entry of the codebook is formed by a cyclic shift of the previous entry. Sparse excitation codebooks (71) have also been proposed, where many entries are zero and there are some constraints on the positions, magnitudes, and signs of the nonzero entries. A further simplification was achieved when the amplitude of nonzero pulses was constrained to have a magnitude of 1. Another type of codebook that has gained significant attention because of its low complexity and absence storage requirement, and that spans a significant portion of signal space, is the algebraic codebook (72), which is used in the ITU-T 8 kbit/s toll-quality speech coding standard.

Another approach to generating excitation codebooks uses centroids of vectors obtained from a large corpus of speech material, very similarly to the generation of VQ codebooks (73); this has led to sophisticated codebook generation principles similar to that used in VQ, such as the use of multistage VQ (or equivalently, multiple codebooks), which inherently has a reduced search complexity and reduced storage in comparison with full VQ with a single codebook. One such coder is the 8 kbit/s vector sum excited linear predictive (VSELP) coder, which was selected for the full-rate North American digital cellular time-division multiple-access (TDMA) standard. Here, the two stochastic codebooks are used to model the aperiodic component of the LPC residual signal. The excitation sequence in VSELP is formed by adding vectors from the two stochastic excitation codebooks.

TRANSFORMED-DOMAIN SPEECH CODING

The waveform coders and parametric coders based on LPC analysis, described above, process speech signals in the time domain. However, there are many speech coders that have gained widespread use that perform processing in a transformed domain such as the frequency domain, the quefrency (log-magnitude) domain, and other unitarily transformed domains. Here speech is first transformed into (or represented in) the desired domain using the appropriate transform [discrete Fourier transform (DFT) or fast Fourier transform (FFT) for the frequency domain, log magnitude of frequency-domain spectrum for the quefrency domain, and discrete cosine transform (DCT), Walsh–Hadamard transform (WHT), or Karhunen–Loeve Transform (KLT) for other unitarily transformed domains] and then analyzed accordingly. A primary motivation behind adopting transformed-domain coding is to exploit the human perception mechanism, which is better understood in transformed domains than in the time domain. Subband coding, multiband excited (MBE) coding, and sinusoidal transform coding (STC) are popular examples of frequency-domain speech coding techniques. The adaptive transform coder (ATC) with DCT is a popular unitary transform speech coding technique. The homomorphic vocoder is a good example of quefrency-domain coding.

Subband Coding

Here the speech signal is first transformed into the frequency domain and the spectrum is divided into frequency bands, which in general have unequal width (27). Division of the speech spectrum into bands is achieved using bandpass filter banks such as the lossless quadrature mirror filter (28) banks. Depending on the width of the bandpass filter, the output of the filter is downsampled or decimated after transforming each band into baseband. Depending on the energy in the passband of the filter bank, the downsampled speech samples are encoded as in PCM or ADPCM (as described above in the section "Waveform Coders") with different numbers of bits per sample. For example, the downsampled speech samples belonging to the lower frequency bands are usually allocated more bits per sample than higher frequency bands, since the lower bands usually carry more energy. Furthermore, for human perception, proper representation of the lower frequency bands is more critical than that of higher frequency bands. It is noted that the SBC can still be considered as a sample-by-sample processing technique because of the way in which encoding is performed. A simplified block diagram of a typical subband coder is given in Fig. 13. An excellent treatise on subband coding techniques is provided in Ref. 29. ITU-T has standardized a wideband speech coder, operating at 64 kbit/s and below, which uses subband coding as described in ITU-T G.722 (61).

Adaptive Transform Coding

Here a block of speech signals is transformed using DCT, WHT, or KLT, and the resulting block is quantized and transmitted (34). Each transformed element is quantized with a different number of bits, depending on its perceptual importance. The KLT yields the maximally decorrelated transformed sequence and hence the optimal one. However, the derivation of the basis vectors for KLT is computationally expensive and data-dependent. Hence less computationally expensive transforms, although suboptimal, that are data-independent, such as DCT, are used in practical speech coding implementations. A significant advantage of this approach is the uncorrelatedness of the transformed sequence, which allows quantization effects of each transformed element uncorrelated with each other. Furthermore, bit assignment to the transformed vector can be made adaptive, based on its perceptual importance. A simplified block diagram of an ATC coder is shown in Fig. 14.

Sinusoidal Transform Coding

The basic idea behind STC (33) is that speech is reconstructed as a sum of sinusoids whose amplitudes, frequencies, and phases are interpolated between sets of encoded parameters. These parameters are regularly updated by applying short-term Fourier analysis to representative speech segments at the encoder. The resulting spectra will normally exhibit magnitude peaks located, in principle, at harmonics of the pitch frequency for voiced speech and randomly for unvoiced speech. The required parameters are derived from the spectra at the frequencies identified by the peaks. Like the MBE coder, the STC coder is a parametric frequency-domain coder. Unlike the MBE coder, the original version of STC declared an entire frame of speech as voiced or unvoiced. Modern STC coders however have a low pass/high pass mixing of excitation. In the general form of STC, the frequencies are not necessarily harmonically related, and hence it has the capability to produce natural-sounding speech at moderate bit rates. A simplified block diagram of STC is shown in Fig. 15.

Multiband Excitation Coding

Here frames of speech (typically of 20 ms duration) are represented in the frequency domain as a set of parameters that describe the fundamental frequency, the magnitudes and phases of its harmonics, and a decision about whether each harmonic is voiced or unvoiced. The voice–unvoiced decision for a harmonic (or for a band encompassing several harmonics) is unique to the MBE coder, in contrast with other traditional vocoders, where the entire frame of speech is declared as voiced or unvoiced (30). The MELP coder described in the preceding section can be treated as a special case of MBE coding where the frequency band below a given cutoff frequency is treated as voiced and the frequency band above is treated as unvoiced. At the decoder of an MBE coder, speech is synthesized using the parameters received from the encoder; specifically, the voice decoder generates speech samples whose spectrum will comprise periodic and noisy contributions as indicated by the voiced–unvoiced decisions on a per-harmonic basis. A typical MBE voice encoder is shown in Fig. 16.

The fact that human speech has sounds consisting of periodic and aperiodic components at the same instant of time was recognized as early as 1939 (2). The MBE concept effectively utilizes this feature to produce good-quality speech at very low bit rates. The MBE coders have also typically exhibited good performance in presence of background noise. The

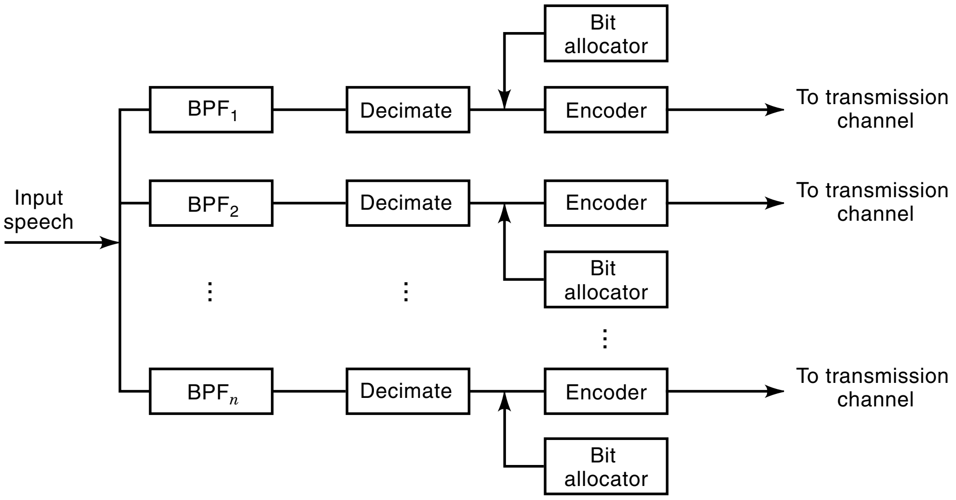


Figure 13. Block diagram of a subband coder.

MBE coders are being used in many mobile satellite communication systems, including the Inmarsat-M and Inmarsat-Mini-M systems (31,32).

Homomorphic Coding

The homomorphic vocoder is a transformed-domain vocoder where speech is processed in the quefrency domain, or equivalently, a cepstral domain. A signal $y(n)$ is said to be a cepstral-domain representation of $x(n)$ if $x(n)$ has undergone the following transformation:

$$y(n) = \Im^{-1}\{\log |\Im\{x(n)\}|\}$$

where $\mathfrak{F}$ represents the Fourier transform and $\mathfrak{F}^{-1}$ the inverse Fourier transform. The principle behind homomorphic vocoding is to separate the vocal tract spectrum from the excitation spectrum. It is noted that under the assumptions of linearity of the vocal tract system, the output of the human speech production system can be written as

$$S(\omega) = E(\omega)V(\omega)$$

where $S(\omega)$ is the speech spectrum, $E(\omega)$ is the excitation spectrum, and $V(\omega)$ is the vocal tract spectrum. The cepstral-domain transformation above permits separation of the spectrum because of the logarithmic transformation involved. At the decoder, an inverse transformation is applied to bring it back to the time domain.

SPEECH CODING STANDARDS

The commercialization of digital speech coding has accelerated in the last decade with the adoption of new speech coding standards and the introduction of major new technologies into commercial networks. This has been motivated by capacity limitations on major international and transcontinental transmission facilities, explosive growth of wireless communications, higher demand for integrated services such as voice, video, and data, and increased interest in communication privacy. In the previous sections, various speech coding techniques were discussed, during which it was mentioned that several of these techniques were part of regional and international speech coding standards. Some of these standards will be discussed briefly in this section.

Speech Coder Attributes: Quality and Bit Rate

Before embarking on a review of speech coding technology standards and their evolution, it is important to define attributes to be employed in determining the state of technology at a given time. As discussed in the introduction, two competing attributes are most important in this regard: the transmis-

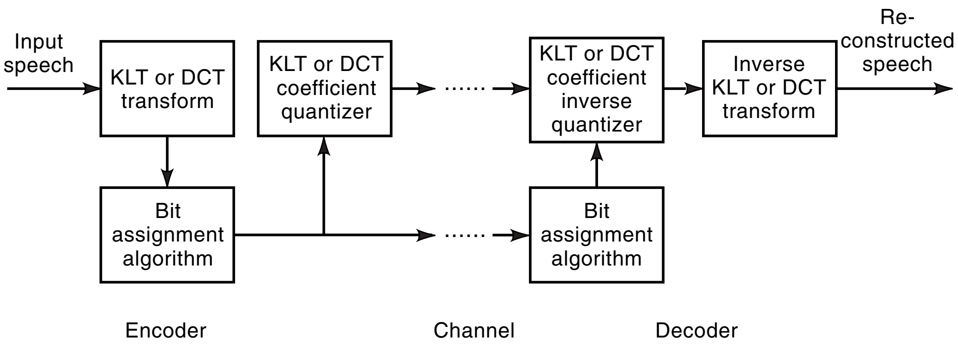


Figure 14. Block diagram of adaptive transform coder.

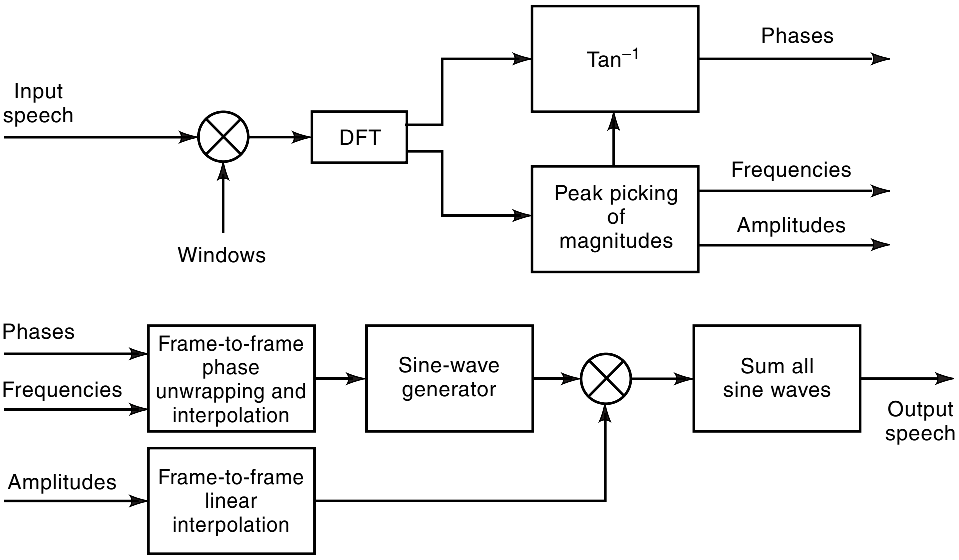


Figure 15. Block diagram of sinusoidal transform coder.

sion rate of the technique or class of techniques in question, and the end-to-end transmission quality.

Often, the term telecommunications quality, toll quality, or more recently wireline quality, is applied to speech coding technology that introduces little or no perceptible distortion. (This does not mean that no degradation is measurable.) The type of technology used in networks today, such as 64 kbit/s PCM and 32 kbit/s ADPCM as described above in the section "Waveform Coders," are good examples of wireline-quality coding.

In a similar fashion, the term cellular quality is used in this article to indicate transmission performance that is less than wireline, and specifically quality that is associated with a perceptible degradation to users. Cellular quality is not annoying in most instances, and typically, all speaker features such as identity and intonation are preserved. Simply, in this article, cellular quality is defined as being equivalent to that of full-rate digital standards, such as North American full-rate digital cellular speech coding standard, the 8 kbit/s VSELP and full-rate GSM speech coding standard, and the 13 kbit/s regular pulse excited with long-term prediction (RPE-LTP). These technologies were discussed in the section "Excitation Modeling." As will be discussed in the section entitled "Wireline or Toll Quality Speech Coding Standards," speech coding technology has improved significantly over time, whereby it is possible to obtain toll-quality speech at 8 kbit/s and below. These are currently being deployed in cellular systems and therefore the definitions for terms such as "cellular quality" is expected to change over time.

Communications quality means performance that is associated with perceptible degradation that can be annoying in some instances. With communications quality, speaker features are occasionally lost, but intelligibility is preserved in most instances. Unlike wireline and cellular quality, some perceptible loss in naturalness is also experienced. Here, communications quality is defined as being perceptually equivalent to that of Federal Standard 1016, the 4.8 kbit/s CELP coder mentioned above in the sub-subsection "Modeling the Aperiodic Component of Excitation." Typically, communications quality has been considered as the lower bound of commercial acceptability.

Intelligible quality provides another step down in service performance, manifested by a typical loss of speaker identity and a measurable, but not unacceptable, loss in intelligibility. Naturalness is also typically lost with intelligible-quality coders. It is useful to think of intelligible quality as being equivalent to that of Federal Standard 1015, the 2.4 kbit/s LPC-10e mentioned above in the section "Excitation Modeling."

Finally, for the sake of completeness, it is useful to define one more performance range: synthetic quality. In synthetic-quality coders, reproduction of input speaker naturalness is not possible, and these coding techniques are typically both speaker- and vocabulary-dependent. Coders exhibiting synthetic quality operate (today) by encoding speech with a few hundred bits per second.

A pictorial representation of the first four of the five quality descriptors against the mean opinion scores (MOS) and equivalent-Q scales is given in Fig. 17. The actual MOS scores or equivalent Q values for the different quality coders could be different from those shown in Fig. 17 for any given test (35), depending on factors such as input speech spectral shaping, input speech level, and type of listening instrument used

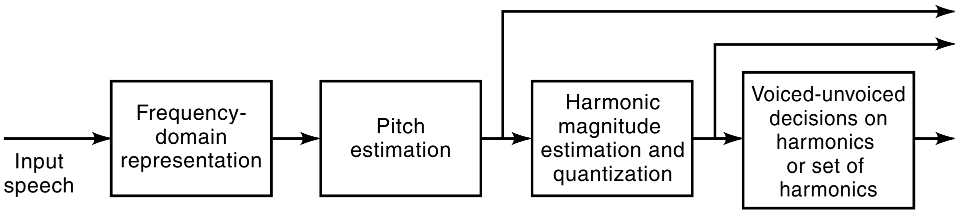

Input
speech
Frequency-domain representation
Pitch estimation
Harmonic magnitude estimation and quantization
Voiced-unvoiced decisions on harmonics or set of harmonics

Figure 16. Block diagram of a multiband excitation coder.

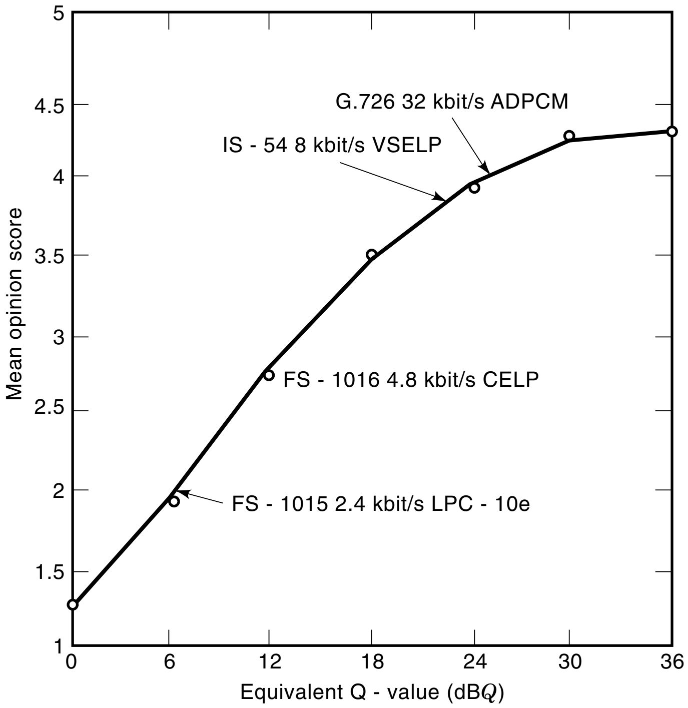


Figure 17. MOS versus Q value for four of the five quality attributes.

(36). However, the important thing to be observed from Fig. 17 is the *relative* performance of the different voice coding technologies (ordinal presentation) and their relative performance difference, both of which are less dependent on the factors mentioned above.

MOS represents averaged opinions of circuit quality by mapping expressed rating of excellent, good, fair, poor, and bad to 5, 4, 3, 2, and 1, respectively. The *Q value* is the ratio of the speech level to the multiplicative noise level (expressed in decibels) that is derived when random noise with an amplitude proportional to the instantaneous speech amplitude is added to the speech signal as specified in ITU-T Recommendation P.810. For a given speech coder, the equivalent Q values are obtained by means of subjective tests as described in the next section. It is worth noting that the technologies shown in Fig. 17 belong to diverse generations of voice coders.

Evolution of Speech Coding Technology and Standards

The first introduction of digital voice encoding technology into commercial service occurred in the early 1970s with the adoption of 64 kbit/s PCM as a standard for the transport of voice and voiceband services over the public switched telephone network (PSTN). Since then, the digitization of international and transcontinental transmission facilities and the associated rapid growth in voice and voiceband data traffic has highlighted the need of further efficiency improvements in transmitting voice signals. This need was fulfilled by the evolution of technology that made possible in the early 1980s the delivery of wireline-quality digital voice at one-half the PCM rates, using 32 kbit/s ADPCM.

Since the early 1980s, pressure to improve the transmission efficiency of voice signals has continued to rise, despite the rapid expansion of wireline network capacity. As a consequence, in the early 1990s, 16 kbit/s *low-delay code-excited linear prediction* (LD-CELP) was developed and adopted by the International Telecommunication Union (ITU) as a new wireline-quality voice compression standard. This was followed by the selection of the 8 kbit/s *conjugate-structured algebraic code-excited linear prediction* (CS-ACELP) coder as a new world standard by the ITU in 1995.

Wireline-quality ITU standards have stimulated and preceded (in terms of setting objectives that can only be met by a technology of the future), rather than followed, the evolution of voice coding technology. In other words, in the area of speech coding, ITU has had the tradition of outlining requirements and objectives for future applications that have stimulated speech coding researchers to conduct research to meet the objectives. Thus, by noting the year of adoption of different technologies as ITU standards, it is possible to quantitatively observe the evolution of wireline-quality voice coding technology over time. This is depicted in Fig. 18(a). From this

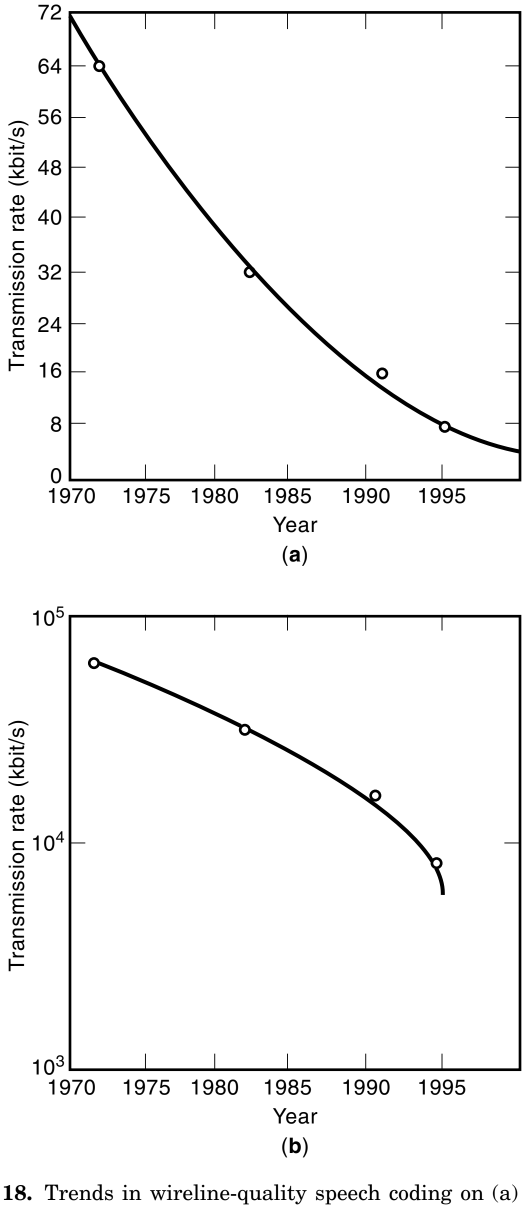


Figure 18. Trends in wireline-quality speech coding on (a) a linear scale, (b) a logarithmic scale.

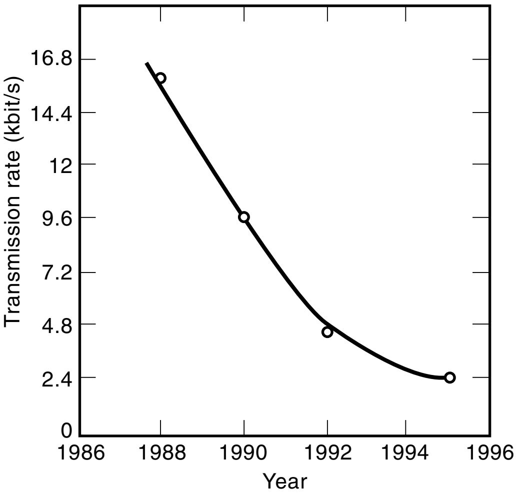


Figure 19. Trends in communications-quality speech coding on a linear scale.

figure it can be seen that in the early stages of voice coding (early 1970s to early 1980s), technology improvement resulted in the ability to reduce voice coding rates by approximately 3.2 (kbit/s)/yr. In the 1990s this rate slowed down to 1.8 (kbit/s)/yr, although the ability to halve the transmission rate actually accelerated. This is more clearly seen in Fig. 18(b), where the relationship shown in Fig. 18(a) is plotted on a logarithmic scale.

Somewhat similar behavior can also be observed when considering the use of communications-quality coding for providing commercial service (Fig. 19). In this case, in the mid-1980s, when communications-quality coders were first introduced into commercial service (principally for mobile–satellite applications by Inmarsat), it was possible to improve efficiency at approximately 3.2 (kbit/s)/yr, although this rate has recently slowed down to more like 0.8 (kbit/s)/yr.

From the above it can be seen that over the past decade it has been possible to reduce voice transmission rates while maintaining quality, a trend that is expected to continue in the near future. Nonetheless, even though these relationships appear to relate transmission rates monotonically with time, in reality, when examined in detail, they reveal a series of step functions whereby the ability of technology to deliver lower rates remains constant until some breakthrough causes the bit rate to suddenly drop. Consequently, at any one time, it is not readily obvious whether technology has reached the flat part of the curve, or is about to make a major breakthrough and permit a further steep reduction in bit rate to occur.

Simultaneously, for wireless applications such as cellular, mobile satellite, aeronautical, maritime, and military voice communications, where bandwidth is scarce and often expensive, lower-bit-rate speech coders (as low as 2.4 kbit/s) were explored. Such efforts have led to speech coding standards, which, among others, include the North American Digital Cellular Standard, which uses a 8 kbit/s VSELP speech coder; the full-rate European Digital Cellular Standard, which employs a 13 kbit/s RPE-LTP speech coder; the Japanese Digital Cellular Standard, which employs a modified 6.4 kbit/s VSELP speech coder, the International Maritime Satellite (Inmarsat) Aeronautical Standard, which employs a 9.6 kbit/s MPLPC; the Inmarsat-M Standard, which employs a 4.15 kbit/s improved multiband excited (IMBE) speech coder; the Inmarsat-Mini-M Standard, which employs a 3.6 kbit/s advanced multiband excited (AMBE) speech coder; the US Department of Defense (DoD) Federal Standard FS1016, which employs a 4.8 kbit/s CELP speech coder; DoD Federal Standard FS1015, which employs a 2.4 kbit/s LPC-based speech coder (LPC-10); and the newly standardized 2.4 kbit/s MELP coder as a replacement for FS1016.

As evident from Fig. 18(a), speech coding has advanced to a stage where it is possible to obtain wireline quality speech at bit rates used for wireless standards. This has led to the adoption of new second generation digital cellular standards such as Enhanced Full Rate GSM standard which employs a 12.2 kbit/s (13 kbit/s after CRC and repetition bits) Algebraic-CELP (ACELP) speech coder, the Enhanced Full Rate North American Digital Cellular TDMA standard which also employs an 8 kbit/s ACELP speech coder, and Enhanced Variable Rate Coder for North American Digital Cellular CDMA standard which employs a variable rate coder based on Relaxation-CELP (RCELP) technique. Another recent speech coding standard that is being used for multimedia internet and video-telephone applications is the dual-rate ITU-T G.723.1 standard that employs multipulse techniques and ACELP techniques of excitation modeling, depending on whether it operates at higher rate or lower rate.

In the sequel, some basic information about the speech-coding technologies involved in some ITU, GSM, Inmarsat, and DoD standards are discussed. In describing these technologies, emphasis is placed on the key features associated with each of the technologies, and no attempt is made to give all the details in their development, which are available in the references.

Wireline-Quality Speech Coding Standards

64 kbit/s ITU-T Pulse-Code-Modulated Speech Coder (Recommendation G.711). The PCM system as described in ITU-T Recommendation G.711 (6) consists of a prefilter, a sampler, and an analog-to-digital converter at the encoder, and a digital-to-analog converter and a low-pass filter at the decoder. The continuous-time speech is typically low-pass-filtered with a cutoff frequency slightly less than 4 kHz and then sampled at a rate of 8000 samples/s. Each sample is then quantized using 8 bits and transmitted to the decoder. The decoder then converts the digital stream to the corresponding amplitude, and the discrete-time signal is then passed through a low-pass filter to obtain a reconstructed continuous-time speech signal. As described above in the subsection "Pulse Code Modulation," ITU-T Recommendation G.711 provides two encoding laws, the A law and the μ law, to enhance the dynamic range of the signal without sacrificing the signal-to-quantization-noise ratio. Both encoding laws exploit the fact that the instantaneous amplitude of the speech signal is less that 25% of its maximum amplitude for more than 50% of the time and hence finer quantization can be performed on small-amplitude samples and coarser quantization on larger-amplitude samples. The mapping tables of these encoding algorithms are provided in Tables 1 and 2 of ITU-T Recommendation G.711.

32 kbit/s ITU-T Adaptive Differential Pulse-Code-Modulated Speech Coder (Recommendation G.726). During the period of 1982–1990, ITU-T (then called CCITT) adopted several ADPCM algorithms. First, the 32 kbit/s ADPCM algorithm described in G.721 was approved. Later G.723 was standardized which basically was an adaptation of the 32 kbit/s algorithm in G.721 to 40 kbit/s to handle voice-band data and 24 kbit/s to handle network congestion. In 1990 CCITT combined G.721 and G.723 and added another ADPCM rate at 16 kbit/s to handle overload situations, resulting in a new recommendation ITU-T G.726 (7), which defines an ADPCM voice coding algorithm operating at 40, 32, 24, and 16 kb/s.

The basic components of the G.726 ADPCM coder are an adaptive sample-by-sample predictor, an adaptive quantizer, and an adaptive inverse quantizer. The difference signal obtained by subtracting the predicted and inverse quantized signals from the original signal is then adaptively quantized and forms the ADPCM output bitstream. The G.726 ADPCM encoder is similar in principle to that in Fig. 5. As described above in the subsection "Adaptive Differential Pulse Code Modulation," in order to prevent the effect of accumulation of quantization errors, a replica of the remote voice decoder is included in the encoder structure. The adaptive predictor is a pole–zero filter as described in Eq. (4), with $N_1 = 2$ and $N_2 = 6$. Such a pole–zero predictor is called an autoregressive moving average (ARMA) predictor and denoted in particular by ARMA(2,6), showing the numbers of coefficients in the autoregressive and moving-average portions of the predictor. It is noted that $\beta_0 = 1$ in Eq. (4) and in G.726. The ARMA coefficients are updated on a sample-by-sample basis (at both encoder and decoder), thereby making the predictor adaptive. Since ADPCM employs backward prediction and sample-by-sample processing, the algorithmic delay is equal to 0.125 ms. ADPCM is used in standalone coders, in T1 and E1 multichannel transcoders, and in digital circuit multiplication equipment (DCME) systems such as ITU-T Recommendation G.763 (37).

In addition, an embedded version [ITU-T Recommendation G.727 (8)] of 32 kbit/s ADPCM encoding with voice quality indistinguishable from that of 32 kbit/s G.726 is used in packet circuit multiplication equipment (PCME). Embedding permits certain enhancement bits to be dropped in the network during congestion without informing the encoder and without any exchange of control information.

The ADPCM techniques defined in ITU-T Recommendations G.726 and G.727 provide the ability to vary the transmission rate among four different bit rates. The highest rate (40 kbit/s) is employed when high-speed voice-band data are being transmitted; the lowest two (24 kbit/s and 16 kbit/s) rates are employed dynamically as part of an overload traffic control strategy. Consequently, 24 kbit/s and 16 kbit/s ADPCM are not steady-state encoding rates under normal operating conditions.

16 kbit/s Toll-Quality ITU-T Standard (Recommendation G.728): Low-Delay CELP Coding. CELP coders have been demonstrated to produce very high-quality speech at 16 kbit/s. However, like many other parametric speech coders, they contribute a delay typically well above 10 ms. In many practical situations such as PSTNs and more complicated networks where tandem encoding are necessary, such long delays contribute to a significant impairment of the performance of the network and in many cases are unacceptable. Indiscriminate deployment of long-delay parametric speech coders in PSTN trunks could require substantial revision of echo control procedures in both networks and terminal equipment. The LD-CELP (38) coder was introduced in an effort to meet the performance requirement specified by CCITT, which was to achieve toll-quality speech at 16 kbit/s with a total delay no greater than 5 msec.

A block diagram of an LD-CELP coder is shown in Fig. 20. The essence of CELP, which was described in the sub-section "Modeling the Aperiodic Component of Excitation," is retained in LD-CELP. The main difference is that CELP uses forward adaptation for computing the coefficients of the short-term prediction filter, whereas LD-CELP uses a backward adaptive short-term predictor. In a backward adaptive configuration, the coefficients of the short-term filter are not derived from the original speech, but instead from the past reconstructed speech. Since both encoder and decoder have access to the past reconstructed speech, information about the short-term filter coefficients no longer need be transmitted to the decoder. Thus, in contrast to CELP, where the prediction coefficients, the gain, and the excitation sequence have to be transmitted, LD-CELP requires transmission of the excitation sequence only (see Fig. 20). The predictor coefficients are obtained by performing LPC analysis on previously quantized speech, and the gain is obtained by using the gain information embedded in previously quantized excitation.

For the 16 kbit/s LD-CELP coder, the excitation vector in the excitation codebook has a dimension (or block size) of five samples. The long-term predictor (pitch predictor) present in the conventional CELP coder is eliminated, and a 50th-order LPC analysis is used. The LPC predictor coefficients are updated once every four speech vectors (2.5 ms) by performing LPC analysis on previously synthesized speech. The excitation gain is updated once every vector by using a 10th-order adaptive linear predictor in the logarithmic domain. The coefficients of this log-gain predictor are updated once every four vectors by performing LPC analysis on the logarithmic gains of previously quantized and scaled excitation vectors. The 10th-order perceptual weighting filter is also updated once every four vectors by using a 10th-order LPC analysis of input speech. To reduce the complexity (in terms of codebook search time) and algorithmic delay, the 10 bits that are available to represent blocks of five samples (at 16 kbit/s) are used to encode a product code with a 3 bit gain codebook and a 7 bit shape codebook.

The three gain bits consist of a sign bit and two magnitude bits. The sign bit has the effect of doubling the shape codebook size while retaining the same search complexity. In LD-CELP, the shape codebook is closed-loop optimized by a codebook design algorithm based on the perceptually weighted criterion used by the LD-CELP encoder. This in contrast to conventional CELP coders, which use Gaussian random numbers to populate the codebook. The shape codebook design algorithm is similar to the LBG algorithm for vector quantizer design (39). After the shape codebook is designed, pseudo-gray coding (40) is used to assign codebook indices. With Gray-coded codebook indices, a single bit error will result in a decoded codevector close to the transmitted one. Such a technique significantly improves the performance of the coder under noisy channel conditions.

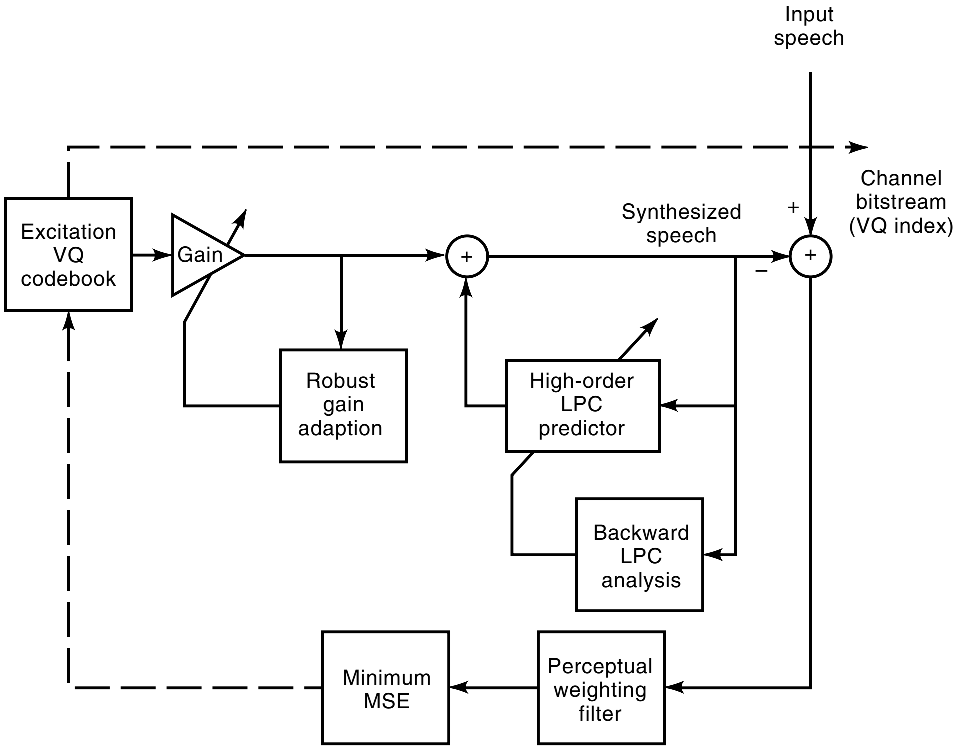


Figure 20. Block diagram of ITU-T 16 kbit/s LD-CELP coder.

Finally, an adaptive postfilter is used at the decoder to increase the perceptual quality of the synthesized output. The postfilter essentially consists of a short-term and a long-term postfilter; the short-term postfilter parameters are derived from the LPC analysis performed in the decoder for synthesis, and the long-term postfilter parameter is obtained by performing pitch extraction based on previously reconstructed speech.

The 16 kbit/s LD-CELP was adopted as an ITU-T standard for toll-quality speech coding at 16 kbit/s under Recommendation G.728 in 1992.

8 kbit/s Toll-Quality ITU-T Standard (Recommendation G.729): Conjugate-Structured Algebraic CELP. The 8 kbit/s CSACELP (41) coder was standardized by the ITU as a new world standard for toll-quality speech coding in 1995. The CSACELP, as its name indicates, also belongs to the CELP family of coders. Here the coder operates on speech frames of 10 ms and looks ahead 5 ms for LPC analysis. Hence the algorithmic delay of the coder is 15 ms. Every speech frame is divided into two equal subframes of 5 ms each. Linear prediction is performed using a Levinson–Durbin algorithm that uses bandwidth-expanded autocorrelation coefficients. LPC-to-LSF conversion is performed using Chebychev polynomials. The 10th-order LSF vector is then quantized using a predictive two-stage VQ with 18 bits.

In comparison with the traditional CELP approach, the excitation sequence to the decoder is determined using two codebooks: a fixed codebook and an adaptive codebook (see Fig. 21). The fixed codebook has an algebraic structure that helps determine four nonzero pulses per subframe of speech, and their positions, using 17 bits. As illustrated in the table below, every 5 ms (or every 40 samples), three pulses are chosen from three mutually exclusive sets each of which contains eight possible positions, thereby requiring 3 bits each to convey the chosen pulse position to the remote decoder. The fourth pulse is allowed to occur in any of the remaining 16 pulse positions, thereby requiring 4 bits to convey its pulse position to the remote decoder. Associated with each pulse position is sign information that also has to be conveyed to the remote decoder:

Pulse ID	Positions
1	0, 5, 10, 15, 20, 25, 30, 35
2	1, 6, 11, 16, 21, 26, 31, 36
3	2, 7, 12, 17, 22, 27, 32, 37
4	3, 4, 8, 9, 13, 14, 18, 19, 23, 24, 28, 29, 33, 34, 38, 39

The gains of adaptive and fixed codebooks are vector-quantized using 7 bits per subframe using a conjugate-structured codebook. While the algebraic structured codebook significantly reduces the complexity of the algorithm, the conjugate-structured codebook increases the robustness of the coder against channel errors.

The adaptive codebook index (or equivalently the optimal delay) for the first subframe, T_1, is transmitted using 8 bits. The 8 bits represent a fractional delay with sample resolution $\frac{1}{3}$ in the range $[19\frac{1}{3}, 84\frac{2}{3}]$ and integer delay in the range [85, 143]. For the second subframe, the adaptive codebook index always represents fractional delay with sample resolution $\frac{1}{3}$ in the range $[\text{int}(T_1) - 5\frac{2}{3}, \text{int}(T_1) + 4\frac{2}{3}]$, which is transmitted using 5 bits. As described above under "Closed-Loop Long-Term Prediction," fractional delays are obtained by interpolating the autocorrelation function of the residual using a Hamming windowed sinc function. With an additional parity bit for adaptive codebook indices, a total of 80 bits is transmitted every 10 ms, yielding a bit rate of 8 kbit/s.

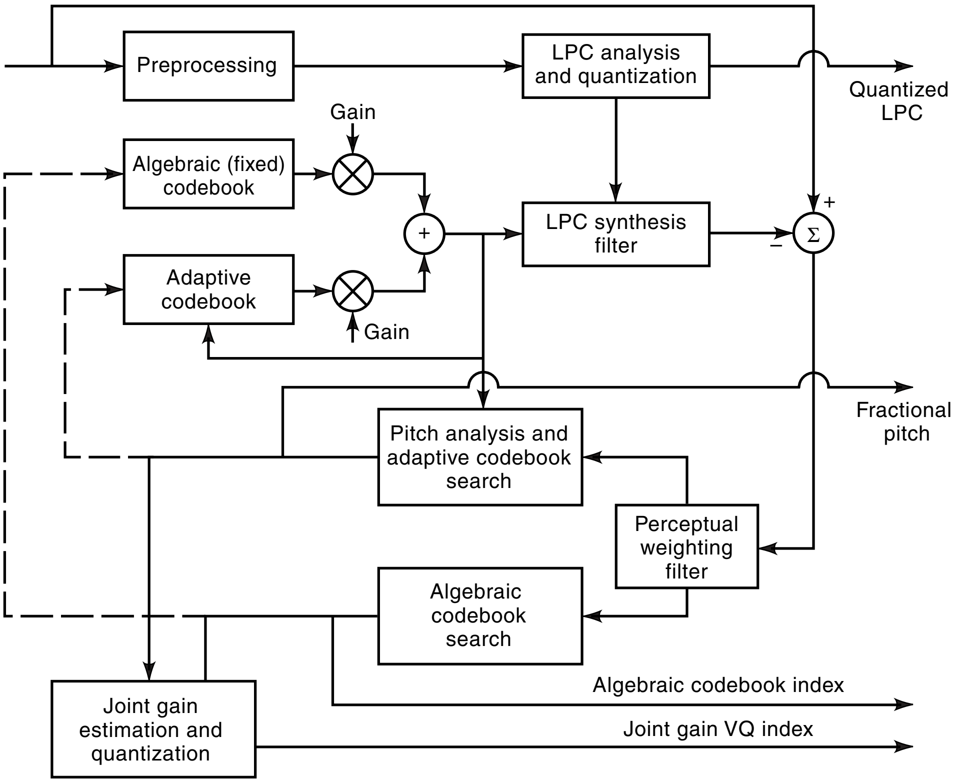

Preprocessing
LPC analysis and quantization
Quantized LPC
Gain
Algebraic (fixed) codebook
+
LPC synthesis filter
+
−
Σ
Adaptive codebook
Gain
Fractional pitch
Pitch analysis and adaptive codebook search
Perceptual weighting filter
Algebraic codebook search
Algebraic codebook index
Joint gain estimation and quantization
Joint gain VQ index

Figure 21. Block diagram of ITU-T 8 kbit/s CSACELP coder.

Since 1995, the ITU has been active in the process of standardizing a 4 kbit/s toll-quality speech coding standard with the objective of standardizing the algorithm in the year 2000. Once again, the requirements and objectives for such a standard were proposed early enough (42) to provide a clear target for speech coding researchers.

Cellular-Quality Speech Coding Standards

North American Full-Rate Digital Cellular TDMA Standard (IS 54): Vector-Sum-Excited Linear Predictive Coder. The VSELP (43) coder operating at 8 kbit/s has been adopted as a standard for North American TDMA digital cellular communications. The VSELP coder, like the CELP coder, falls into the class of analysis-by-synthesis coders. The VSELP coder was designed to accomplish the highest possible speech quality with robustness to channel errors while maintaining a reasonable computational complexity at 8 kbit/s. The VSELP speech coder achieves these goals through efficient utilization of structured excitation codebooks. The structured codebooks contribute to maintaining reasonable computational complexity while increasing robustness to channel errors.

The excitation sequence for the decoder in VSELP is derived from three codebooks, namely, one adaptive codebook that is associated with the fundamental frequency of the speech signal, and two stochastic codebooks. As the name implies, the excitation sequence is derived as a weighted sum of the three vectors in the three codebooks. The codewords in the stochastic (fixed) codebooks are formed so that a single bit error in a VSELP codeword on the channel does not affect the output of the vector sum. The frame size for the VSELP coder is 20 ms, and the subframe size is 5 ms.

A 10th-order LPC analysis is performed, and, as described above in the subsection “Autocorrelation Method of Linear Predictive Coding Analysis,” the LPC coefficients are represented as reflection coefficients. The 10 reflection coefficients are scalar quantized using 38 bits, the bit allocation being such that the first reflection coefficient represented uses 6 bits whereas the last uses only 2 bits. Excitation parameters are updated and transmitted every subframe of 5 ms. The adaptive codebook is searched for 128 possible lags using the closed-loop search as described above under “Closed-Loop Long-Term Prediction.” Hence the adaptive codebook index is transmitted every 5 ms using 7 bits.

The two stochastic codebooks contain 128 entries each (hence the need to transmit 14 bits every 5 ms), and each entry is 40 samples wide. The codebook entries are formed by linearly combining seven basis vectors so that when the codebooks are Gray-coded, a bit error in the transmitted codebook index will only lead to a selection of a codebook entry in the decoder that differed in only one basis vector. Thus robustness to channel errors is achieved.

The three codebook gains are (jointly) transmitted every 5 ms using 8 bits, and an overall energy of the speech frame using 5 bits per 20 ms is also transmitted. With an additional spare bit, the IS-54 VSELP coder transmits 160 bits every 20 ms, thereby achieving a bit rate of 8 kbit/s.

Modified versions of VSELP have also been used in the full-rate Japanese digital cellular standard and the GSM half-rate cellular standard.

While the 8 kbit/s VSELP technology is historically interesting, because of advancements in speech coding technology that yielded better quality speech at same bit rate, an enhanced version of full rate coder based on ACELP technology has been recently standardized for the North American Digital Cellular System which provides toll-quality speech.

GSM Full-Rate Standard: Regular Pulse Excitation Coding with Long Term Prediction. The RPE-LTP coder operating at 13 kb/

s (44) was adopted as the full-rate standard for GSM TDMA digital cellular communications. The RPE-LTP coder, like the CELP coder, falls into the class of analysis-by-synthesis coders. It processes speech in frames of 20 ms duration and subframes of 5 ms duration.

An eighth-order LPC analysis is performed every 20 ms (but interpolated every 5 ms), and the LPC coefficients are represented in the LAR domain as described above in the subsection “Quantization and Alternative Representation of Linear Predictive Coding Parameters.”. The eight LAR coefficients are scalar quantized using a total of 36 bits, and, as in VSELP, the bit allocation is different for different LAR coefficients. The first LAR coefficient is quantized using 6 bits, whereas the last coefficient is quantized using only 3 bits.

For every subframe of 5 ms (40 samples), the LTP lag and LTP gain (D^*, β_0^*) described above under “Closed-Loop Long-Term Prediction” are quantized using 7 and 2 bits respectively. A sequence of thirteen equally spaced pulses is chosen from four possible candidates in a closed-loop manner as described above in the sub-subsection “Modeling the Aperiodic Component of Excitation,” and the chosen candidate is indicated using 2 bits. As described above in the sub-section “Regular Pulse Excitation Coding,” because of the uniform and known spacing between pulses, the RPE coder has the advantage over the traditional multipulse coders that the individual pulse positions need not be transmitted. The normalized pulse amplitudes are quantized with the adaptive PCM technique, using 3 bits each. The normalizing factor, which is the maximum of all amplitudes, is transmitted using 6 bits. Therefore, a total of $7 + 2 + 2 + (13 \times 3) + 6 = 56$ bits are used to represent the excitation sequence every 5 ms.

Hence a total of 260 bits (36 bits for LAR coefficients and $224 = 56 \times 4$ bits for excitation) are transmitted every 20 ms, thereby achieving a bit rate of 13 kbit/s. A detailed description of the RPE-LTP algorithm can be found in ETSI Recommendation GSM TS 06.10.

For reasons similar to that of North American Digital Cellular system, the European Telecommunications Standards Institute (ETSI) has recently standardized a 12.2 kbit/s Enhanced Full Rate (EFR) coder based on ACELP technology that provides toll quality speech.

Communications-Quality Speech Coding Standards

Inmarsat Full-Rate Aeronautical Standard: The Multipulse Excited Linear Predictive Coder. The Inmarsat aeronautical system employs the MPLPC operating at 9.6 kbit/s (35). The Inmarsat full-rate aeronautical standard processes speech in frames of 20 ms duration and models excitation in subframes of 4 ms duration. A 10th-order LPC analysis is performed every 20 ms, and the LPC coefficients are represented as reflection coefficients for quantization as described above in the subsection “Quantization and Alternative Representations of Linear Predictive Coding Parameters.” The ten reflection coefficients are scalar quantized using 40 bits. Unlike VSELP or GSM, long-term prediction is performed here only every 20 ms rather than every subframe. The LTP lag and gain (D^*, β_0^*) described above under “Closed-Loop Long-Term Prediction” are quantized using 6 and 2 bits, respectively. Multipulse excitation analysis as described in the sub-subsection “Multipulse Linear Predictive Coding Analysis” is performed on the residual signal after long-term prediction. In order to reduce the complexity, the sequential search approach described in the sub-subsection “Multipulse Linear Predictive Coding Method” is used. Here, for every 4 ms (32 samples) duration of the residual signal, three pulses that are subjectively more important are chosen (using analysis by synthesis). The positions of the three pulses (p_1^*, p_2^*, p_3^*) are quantized using 5 bits each. The amplitude of first two pulses (g_1^*, g_2^*) are quantized using 4 bits each, and the amplitude of the third pulse (g_3^*) is quantized using 3 bits. Overall, the LTP residual is quantized using 26 bits every 4 ms.

Hence a total of 192 bits (40 bits for reflection coefficients, 6 bits for LTP lag, 2 bits for LTP gain, $130 = 26 \times 5$ bits for excitation, and 14 bits for error control) are transmitted every 20 ms, thereby achieving a bit rate of 9.6 kbit/s.

The United States Department of Defense 4.8 kbit/s CELP FS1016 Coder. The FS1016 coder (45), which is primarily used in military applications such as the US Department of Defense, operates at 4.8 kbit/s and is based on the CELP structure. It uses a 30 ms frame size with four 7.5 ms subframes. CELP analysis consists of three basic functions: (1) short-term linear prediction, (2) long-term adaptive codebook search, and (3) innovation stochastic codebook search. CELP synthesis consists of the corresponding three synthesis functions performed in reverse order with the addition of a post-filter to enhance reconstructed speech.

A tenth-order LPC analysis is used to model the speech signal’s short-term spectrum, or formant structure. The corresponding LSF parameters are scalar quantized using 34 bits per frame. Every 7.5 ms subframe of long-term signal periodicity (pitch) is modeled by an adaptive codebook. The optimal adaptive codebook index D^* for the first and third subframes are represented using 8 bits each, whereas the second and fourth subframe are represented using 6 bits each. The adaptive codebook index represents the optimal pitch in fractional resolution as described above under “Closed-Loop Long-Term Prediction.” The adaptive codebook gain β_0^* represents using 5 bits per subframe. The residual from the short-term LPC parameters and pitch VQ is vector quantized using a fixed stochastic codebook of size 512, thereby requiring 9 bits per subframe to transmit the optimal stochastic codebook index. The optimal scaled excitation vectors from the adaptive and stochastic codebooks are selected by minimizing a time-varying perceptually weighted distortion measure that improves subjective quality by exploiting masking properties of the human ear. The optimal stochastic codebook gain (g_c^* in the sub-subsection “Modeling the Aperiodic Component of Excitation”) is quantized using 5 bits per subframe. Hence a total of 104 (28 + 20 + 36 + 20) are used every 30 ms to represent the excitation sequence to the LPC synthesis filter at the remote decoder.

This, together with 34 bits for the LSF quantizer, 1 bit for frame synchronization, 4 bits for error control, and 1 bit for future expansion, leads to 144 bits every 30 ms, for a bit rate of 4.8 kbit/s.

Inmarsat-M and Inmarsat-Mini-M Speech Coding Standards: Multiband Excitation Coding. Inmarsat standardized the IMBE coder operating at 4.15 kbit/s for Inmarsat-M service (which uses a briefcase-size terminal) and later for the AMBE coder operating at 3.6 kbit/s for Inmarsat-Mini-M service (31,32) (which uses a notebook-size terminal), both of which are based on the basic MBE (30) speech model. The principles of the two coders are essentially the same as described above in the subsection "Multiband Excitation Coding"; however, they differ in the way the parameters are extracted and quantized. The encoder extracts pitch information every 20 ms and performs voice–unvoiced decision on groups of harmonics. The magnitudes of the harmonics of the pitch frequency are either scalar or vector quantized, depending on the number of harmonics and their location. The IMBE coder provides communications-quality speech, and the AMBE vocoder achieves close to cellular quality for certain types of filtered speech (31).

The United States Department of Defense 2.4 kbit/s Mixed Excitation Linear Predictive Coder. With increased evidence of rapid advances in speech coding technology, the US Department of Defense sought, during the period 1994–1996, a 2.4 kbit/s coder whose performance would be subjectively equivalent that of the 4.8 kbit/s FS1016 coder. This resulted in the very recent selection of a 2.4 kbit/s MELP coder (46) among other competing technologies. The structure of the MELP coder is similar to that described in the sub-subsection on that topic (see Fig. 9).

Here a frame size of 22.5 ms is used, and the LPC coefficients are represented in the LSF domain. A 25 bit multistage VQ is used to quantize the LSFs. The MSVQ uses joint optimization for both codebook design and search, using an *M*-best algorithm. The 25 bit codebook consists of four stages of 7, 6, 6, and 6 bits, respectively. The gain is transmitted twice per frame of 22.5 ms, the gain for the first subframe is coded with 3 bits covering a small dynamic range based on neighboring subframe values. The gain for the second subframe is coded using 5 bits for the full dynamic range of speech. Pitch and overall voicing is quantized using 7 bits per frame. This is true for both voiced and unvoiced speech.

For voiced speech, a Fourier analysis is performed on the LPC residual signal, and the magnitudes of the first 10 harmonics are quantized using 8 bits. A bandpass voicing measure to reflect the frequency band over which the speech signal is estimated to be periodic is conveyed to the decoder using 4 bits. Finally, a bit indicating the degree of periodicity is also transmitted to the remote decoder, which then controls the amount of jitter in the synthesized speech signal.

For unvoiced speech, Fourier magnitudes, bandpass voicing measure, and periodicity flag are not transmitted. Instead, the 13 bits are used to perform error control using Hamming codes.

Overall, 53 bits are used to quantize the LPC and excitation parameters, and with one additional bit for synchronization, 54 bits every 22.5 ms yields a rate of 2.4 kbit/s.

Intelligible-Quality Speech Coding Standard

The United States Department of Defense 2.4 kbit/s LPC-10e FS1015 Coder. The DoD LPC-10 (47) FS1015 vocoder uses a 22.5 ms frame length for analysis and performs a modified covariance analysis to obtain the LPC parameters. It uses the two-state excitation model described in the section "Excitation Modeling" (see also Fig. 7). Pitch and voicing decisions are made using the average-magnitude-difference function algorithm and a voicing detector. Pitch and voicing decisions are smoothed using dynamic programming techniques that employ two frames of delay.

For voiced speech, the ten LPC coefficients are scalar quantized using a total of 41 bits. Pitch and voicing decisions are quantized using 7 bits, and gain information is quantized using 5 bits.

For unvoiced speech, only four LPC coefficients are transmitted, using 20 bits (5 bits per coefficient). Pitch and gain are quantized using 7 and 5 bits, respectively. Similarly to the US DoD MELP coder described above, the unused 21 bits are used for error control during unvoiced speech.

With the addition of a synchronization bit and a total of 54 bits per 22.5 ms, the LPC-10 FS1015 coder operates at 2400 bits/s.

SPEECH CODER PERFORMANCE ASSESSMENT

In the previous section it was mentioned that two attributes, namely, *speech quality* and *bit rate,* are predominantly used to characterize speech coder performance. Furthermore it was mentioned that the speech quality produced by a speech coder could be broadly categorized as, wireline (toll) quality, cellular quality, communications quality, intelligible quality, and synthetic quality. However, the nonlinearity of low-rate parametric coders has rendered analytical or objective methods questionable for applying that classification under the variety of source and channel conditions over which the speech coder is to be assessed. For this reason, subjective tests as described in the ITU-T P.800 (62) series recommendations have remained the only reliable way to conduct speech coder performance assessment. While ITU has also recently standardized an objective measurement tool (ITU-T Recommendation P.861 (63)), that tool has still not gained widespread usage to the extent of replacing the subjective test, the primary reason being that its accuracy has been recognized to be technology-dependent.

While the subjective assessment ought to be conducted with the intent of capturing all types of impairments anticipated in the system, the primary intent is to capture communication impairments, if any, because of speech coding. In general, communications impairment factors can be divided into three types, depending on the affected direction of the communication link (48). The first type comprises impairments that cause an increase in listening difficulty when the communications link is unidirectional and no assistance is given to the listener by the talker. The second type comprises impairments that cause difficulty while talking only. The third type comprises impairments that cause difficulty while conversing, or factors associated with the alternation of the talking and listening roles of the participants.

Digital speech coding systems typically give rise to impairments of the first type, in view of the modeling distortions and quantization noise introduced by the encoding and decoding processes (49). Consequently, listening tests are often used to evaluate the transmission performance of such systems. This will be discussed in further detail in the subsection below. Telephone handsets (particularly with respect to the effect of sidetones), and loading coils with unbalanced 4-to-2 wire terminations without echo control can give rise to the second type of impairments, since the presence of echo and sidetone may increase the talking difficulty in a telephone conversation (50). Echo suppressors (51), on the other hand, are an example of devices that introduce impairments of the third type, which cause difficulty in conversing, since these devices operate by disallowing fully bidirectional communication. Similarly, circuits with long propagation delay introduce impairments of the third type, because they alter the perceived dynamics of conversational communication.

Listener Opinion Tests

For listener opinion tests, the method recommended by the ITU-T in Recommendation P.830 is frequently employed. Typically naive (untrained) listeners, or *subjects,* are invited to assess the quality of speech material (typically in the form of sentence pairs) passed through the speech coder under consideration. In generating suitable speech material, a set of phonetically balanced sentences uttered by a variety of talkers, both male and female, is normally required (52). It is common to employ one set of recordings obtained using a microphone appropriate to the various systems under evaluation, and then use the same recordings for several experiments in which the same type of microphone would normally be employed.

According to P.800, listening-only methods can be classified into three groups: *absolute category rating* (ACR), *degradation category rating* (DCR), and *comparison category rating* (ccr). The first is an absolute rating method, while the other two are relative rating methods. For subjective tests that require better discrimination accuracy, *paired comparison rating* (PCR) is sometimes used. In that case, ACR and DCR tests are the most commonly employed.

The ACR test is characterized by a single-stimulus presentation: a sentence pair is played to the subject through headphones or telephone handsets, and he is requested to express his opinion on the quality of the speech material on an absolute five-point scale: {excellent, good, fair, poor, bad}. Typically a set of phonetically balanced speech material from a number of talkers of both gender is passed through all speech coders under consideration. The performance of each speech coder is typically evaluated by first mapping the five-point scale to {5, 4, 3, 2, 1} and averaging the scores provided by various subjects across all talkers and sentence pairs to yield a MOS for the speech coder under consideration.

The DCR test is characterized by a dual-stimulus presentation to the subject. Here the same sentence pair is presented twice—first as an unprocessed or reference sample, and second as a processed or test sample—to the subject, whereupon he is requested to express his opinion on the degradation of processed speech compared to unprocessed speech on a relative five-point scale: {degradation is inaudible, degradation is audible but not annoying, degradation is slightly annoying, degradation is annoying, degradation is very annoying}. The unprocessed speech sample is essentially the input to the speech coder under consideration, and the processed speech is its output. Similarly to the MOS of an ACR test, a degradation MOS (DMOS) score is computed for the DCR test.

CCR tests are similar to DCR tests in that they are dual-stimulus tests. However, the CCR method uses a bipolar seven-point scale where the subjects are requested to quantify their preference towards *test,* or *reference,* stimuli. While the reference speech sample in DCR tests is always the unprocessed source signal, in CCR tests the reference may be either processed or unprocessed speech. In CCR the dual stimuli can be presented in any random order.

Finally, paired-comparison tests are a simpler case of the CCR method, where a binary scale is used and the subject is asked which sample is preferred (reference or test).

The selection of a suitable experimental approach, particularly the choice between absolute and relative or rank order designs, is very important and is influenced by the type of systems being evaluated, the overall test objective, and the number of conditions to be assessed (49). Generally, in the absence of background noise, if the speech coders and test conditions to be assessed result in outputs that are degraded in an entirely different manner from one another, then ACR tests are preferable, since they are absolute or single-stimulus by design (i.e., each sample is listened to and rated without a direct comparison with other reference samples). The DCR test is generally a more sensitive test than the ACR test, since minor degradations introduced by the speech coder can be penalized heavily by the subject that, in the absence of a reference signal (as in the real world), would have gone unnoticed. Hence rating speech on an absolute scale is preferred.

However, for the evaluation of coded speech quality in the presence of acoustic (e.g., vehicular) noise, a DCR test is usually chosen. This selection is in accordance with currently revised CCITT procedures, as the distortion of high levels of background noise is believed to be more effectively measured with a dual-stimulus assessments. The primary reason here is that, if ACR were used under high levels of background noise, then even the reference unprocessed noisy sample would not be rated as excellent or good, and hence the dynamic range provided by the five-point scale is not utilized. Furthermore, most low-bit-rate speech coders are optimized to work well for speech types of signals, and hence the presence of background noises that are not produced by human speech can result in distorted output that is more annoying to the human ear than where low-level background noise is present. Hence a reference speech sample that reflects the actual background noise at the input of the speech coder, as provided by the DCR test, is desirable.

CCR tests are very useful in comparing two systems that are close to each other. Another important application of CCR is to the case where a speech coder performs noise cancellation, due to which the processed speech sample may actually be perceived to sound better than an unprocessed noisy speech sample.

CONCATENATED SPEECH CODING

Future long-distance, and especially international, telephone calls will involve an increasing number of multilink circuits of cellular, mobile satellite, and private and PSTN connections. Calls will thus be established over multilink circuits employing different types of speech coding technologies operating at different bit rates. Since the very early 1980s one of the most unappreciated implications of integrating a variety of voice coding technologies into the network has been the associated reduction in end-to-end quality. The Integrated Services Digital Network (ISDN) infrastructure was not designed to provide switching capability at transmission rates below 64 kbit/s, and thus has been outpaced by today's modern coding technology. This means that interconnection of different voice coding technologies is typically possible only after conversion to 64 kbit/s PCM. The implication of this is, that unlike data transmission, where throughput is limited by the capacity of the least transparent link, voice quality is reduced disproportionately and below the quality of the worst-performing link.

The characteristics of interconnected voice links were recently a subject of investigation (53–55), resulting in considerable attention being given to the interconnection characteristics of new voice coding technology. These concerns have, of course, been heightened by the wider mix of voice coding technologies arising from a proliferation of proprietary, regional, and international standards; by increasing wireless network access, promising to reach 50% by the end of the century; by the increasing use of wireless local loops in developing countries as a means to accelerate network deployment; and by the accelerating telecommunications network deregulation and privatization, resulting in a larger number of network links with disparate technologies being encountered in end-to-end connections.

The proliferation of voice standards and their effect on transmission planning can be better visualized by considering a few interconnectivity scenarios and making some assumptions regarding the different types of voice technology that might be used in the network. Several foreseen network scenarios are presented in Fig. 22 which involve an international link as part of a multilink connection. Network configuration A represents a call initiated from a wireline user in a foreign country that is destined to a North American or European digital cellular user. The international link uses a DCME that employs 32 kbit/s ADPCM speech coding as described in Recommendation G.726. DCME equipment has been slowly migrating to the use of 16 kbit/s LD-CELP, which is shown in configuration B. Configuration C is the case of a call from a North American cellular user to a European cellular user. The results of subjective tests (55) using such interconnections indicated that voice quality in the link of configuration C was degraded by more than 1.0 in MOS with respect to the weakest portion of the link, namely, 13 kbit/s RPE-LTP. Since the results shown were obtained with an accuracy of 0.1 in MOS at a 95% level of confidence, and since an MOS drop of more than 0.3 is typically associated with a new class of service quality, it can be seen that the effect of PSTN interconnections is to produce an end-to-end connection whose quality differs markedly from that where no such interconnections exist.

An even more interesting scenario is the case of a mobile satellite user in country A calling a cellular user local to the mobile satellite user, where the gateway for the mobile satellite system is situated in country B, and the international PSTN link between countries A and B is similar to that in configuration C. For this local call the speech undergoes a concatenation of three voice coding technologies.

The implication of these observations is that some reduction of the number of different voice coding technologies is likely to occur or, as a minimum, future introduction of voice coding into the network is likely to place some constraints on the use of technologies whose quality and transmission rates are dissimilar. In the past decade, interconnectivity of voice encoding technologies has received increasing attention. Consequently, the development of low-rate coders that remain robust to such interconnected configurations will challenge researchers in the future, as it is doing at present.

FUTURE TRENDS IN SPEECH CODING

As evident from previous sections, over the last decade, a significant amount of research effort has been directed towards

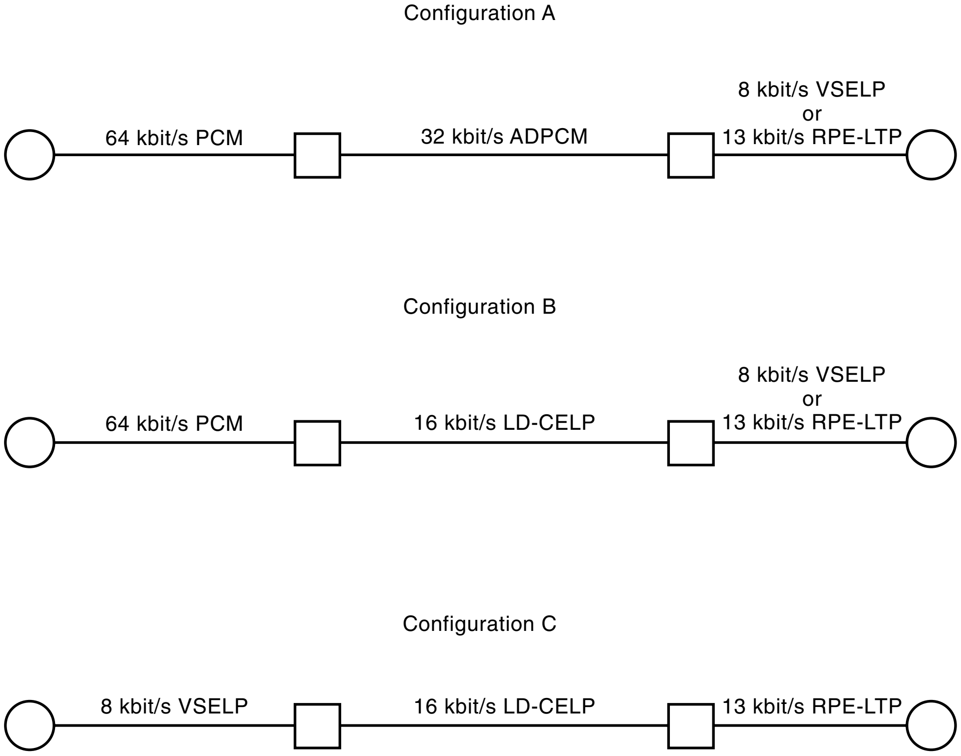


Figure 22. Some interconnection scenarios to demonstrate concatenated speech coding.

better modeling of the human speech production system, better representations of parameters of such model, efficient quantizations of these representations, and, most importantly, better representation of the LPC residual signal. It is very evident that CELP analysis-by-synthesis coders have enjoyed tremendous success in achieving better than communications-quality speech at bit rates as low as 4.8 kbit/s. A good number of candidates for the 4 kbit/s ITU toll-quality standardization effort are also CELP-based. More recently, however, non-CELP-based coders have been playing an important role in achieving communications-quality (or better) speech below 4 kbit/s. The 3.6 kbit/s AMBE coder (described in Section 6.5.3) and 2.4 kbit/s MELP coder (described herein) are two good examples of the trend. In addition, one emerging technology that is receiving considerable attention among many speech coding researchers and that has shown significant promise in achieving high quality speech at low bit rates is the *prototype waveform interpolation* (PWI) technique (56). PWI uses a powerful model that transforms and decomposes a segment of speech signal into slowly and rapidly evolving waveforms and encodes them separately.

Vector excitation modeling similar to that used in CELP is expected to continue its dominance in producing high-quality speech at low bit rates. However, rather than performing a VQ on the shape of the excitation waveform, future very low-bit-rate (4 kbit/s and less) speech coding technologies are expected to perform VQ on a set of parameters that are representative (in a perceptually significant sense) of the excitation waveform (46,57).

In recent years many researchers have been focusing their work on understanding and utilizing human speech perception models (including the decision-making process in the human brain) and integrating these models towards development of better speech coders. While most of these efforts were initially directed towards obtaining high-quality audio coders, the resulting techniques show significant promise of being applicable to speech coders as well. It is the combination of source and auditory coding that perhaps holds the greatest promise for permitting high-quality, very low-bit-rate speech coding to be realized (such as toll quality at 2 kbit/s or below).

In summary, advancement in several areas will hold the key to the success of speech coding technology in the future: (1) a perceptually weighted filter in the analysis-by-synthesis loop of the encoder that better reflects the human speech perception mechanism; (2) quantization and coding of only those parameters that are important to the human ear, based on masking properties of the human ear; (3) postfiltering of reconstructed speech, taking into account the loudness properties of human ear; and (4) providing robustness in the performance of speech coders in the presence of strong background noise for mobile applications. In addition, for wireless applications such as cellular and mobile–satellite systems where the total channel bandwidth is limited, robustness against channel errors becomes a key performance parameter, and the speech coder (also referred to as the source coder) no longer enjoys the luxury of being blind to the transmission channel characteristics. As a result, source-dependent channel coding and combined source and channel coding schemes (58,59) are gaining increased importance.

A more recent application of speech coding that is expected to grow faster than any other is the Internet, where compressed speech is transmitted as packets on an existing packet-switched network infrastructure. An issue of importance here (as in any packet network) is judicious concealment of the effects of missing packets. Research in this direction is expected to gain momentum as well (60).

CONCLUSIONS

The concept of speech coding and the technical realization of a real-time speech coding apparatus dates back to 1939. Digital speech coding technology operating at 64 kbit/s was introduced into commercial service in the early 1970's. Lower-rate digital speech coding technology (16 kbit/s or less) has evolved significantly over the last decade while maintaining voice quality. The ability to transmit voice at 8 kbit/s with toll quality was unthinkable in 1992, when 16 kbit/s voice technology delivering this quality was considered a breakthrough. The introduction of 4.8 kbit/s voice coding as a commercial service, the project of standardizing a 4 kbit/s toll-quality voice coder by the year 2000, and the potential of introducing 2.4 kbit/s voice coding for commercial satellite-based mobile services in the future are highly remarkable.

Although these achievements did not come without engineering costs, such as increase in complexity, these costs have, in general, been compensated by advancements in DSP technology that can provide the necessary horse-power to execute sophisticated speech coding algorithms in real-time digital signal processors (DSPs).

BIBLIOGRAPHY

1. J. Flanagan, *Speech Analysis, Synthesis and Perception,* New York: Springer-Verlag, 1972.
2. H. Dudley, Remaking speech, *J. Acoust. Soc. Amer.,* **11**: 169–177, 1939.
3. G. Fant, *Acoustic Theory of Speech Production,* Gravenhage, The Netherlands: Mouton, 1960.
4. W. Koemig, H. K. Dunn, and L. Y. Lacy, The sound spectrograph, *J. Acoust. Soc. Amer.,* **17**: 19–49, 1946.
5. P. E. Papamichalis, *Practical Approaches to Speech Coding,* Englewood Cliffs, NJ: Prentice-Hall, 1987.
6. ITU-T Recommendation G.711, *Pulse Code Modulation (PCM) for Voice Frequencies,* (Red Book), Malaga-Torremolinos, 1984.
7. ITU-T Recommendation G.726, *40-, 32-, 24-, and 16-kb / s Adaptive Differential Pulse Code Modulation* (Blue Book), Geneva, 1990.
8. ITU-T Recommendation G.727. *5-, 4-, 3-, and 2-bits per sample Embedded Adaptive Differential Pulse Code Modulation* (Blue Book), Geneva, 1991.
9. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals,* Englewood Cliffs, NJ: Prentice-Hall,1975.
10. W. P. LeBlanc et al., Efficient search and design procedures for robust multi-stage VQ of LPC parameters for 4 kbps speech coding, *IEEE Trans. Speech Audio Process.,* **1**: 373–385, 1993.
11. P. Kabel and R. Ramachandran, The computation of line spectral frequencies using Chebychev polynomials, *IEEE Trans. Acoust., Speech Signal Process.,* **ASSP-34**: 1419–1426, 1986.
12. K. Paliwal and B. Atal, Efficient vector quantization of LPC parameters at 24 bits/frame, *Proc. Int. Conf. Acoust., Speech Signal Process.,* 1991, pp. 661–663.
13. C. S. Ravishankar, B. R. U. Bhaskar, and S. Dimolitsas, A 1200 bps voice coder based upon split VQ of line spectral frequencies, *Proc. 1993 IEEE Speech Coding Workshop,* St. Adele, 1993, pp. 37–38.

14. R. Schafer and J. Markel, *Speech Analysis,* New York: IEEE Press, 1979.

15. W. Hess, *Pitch Determination of Speech Signal,* New York: Springer-Verlag, 1983.

16. B. S. Atal and L. R. Rabiner, A pattern recognition approach to voiced–unvoiced–silence classification with applications to speech recognition, *IEEE Trans. Acoust., Speech Signal Process.,* **ASSP-24**: 201–212, 1976.

17. J. P. Campbell and T. E. Tremain, Voiced/unvoice classification of speech with applications to US Government LPC-10E algorithm, *Proc. ICASSP,* Tokyo, 1986. pp. 472–476.

18. C. K. Un and D. T. Magill, The residual-excited linear prediction vocoder with transmission below 9.6 Kb/s, *IEEE Trans. Commun.,* **COM-23**: 1466–1473, 1995.

19. J. Makhoul et al., A mixed source model for speech compression and synthesis, *J. Acoust. Soc. Amer.,* **64**: 1577–1581, 1978.

20. A. McCree and T. Barnwell, III, A new mixed excitation LPC vocoder, *Proc. ICASSP,* 1991, pp. 593–596.

21. V. Viswanathan et al., A harmonic deviations linear predictive vocoder for improved narrowband speech transmission, *Proc. ICASSP,* 1982, pp. 610–613.

22. C. S. Ravishankar, B. R. U. Bhaskar, and S. Dimolitsas, A 1200 bps voice coder based upon alternate transmission of LPC and residual information, *Proc. 1995 IEEE Speech Coding Workshop,* Annapolis, MD, 1995, pp. 111–112.

23. J. B. Allen and S. T. Neely, Micromechanical models of the cochlea, *Phys. Today,* **45** (7): 40–47, 1992.

24. B. Atal and J. Remde, A new model for LPC excitation for producing natural sounding speech at low bit rates, *Proc. Int. Conf. Acoust., Speech Signal Process.,* 1982, pp. 614–617.

25. P. Kroon, E. F. Deprettere, and R. J. Sluyter, Regular-pulse excitation—A novel approach to effective and efficient multipulse coding of speech, *IEEE Trans. Acoust. Speech Signal Process.,* **ASSP-34**: 1054–1063, 1986.

26. M. R. Schroeder and B. S. Atal, Code excited linear prediction (CELP): High quality speech at very low bit rates, *Proc. Int. Conf. Acoust., Speech Signal Process.,* 1985, pp. 937–940.

27. R. Crochiere and L. Rabiner, *Multirate Digital Signal Processing,* Englewood Cliffs, NJ: Prentice-Hall, 1983.

28. P. P. Vaidyanathan, Quadrature mirror filter banks for M-band extensions and perfect reconstruction techniques, *Acoust. Speech Signal Process Mag.,* **4** (3): 4–20, 1987.

29. R. V. Cox et al., New directions in sub-band coding, *IEEE Trans. Sel. Areas Commun.,* **6**: 391–409, 1988.

30. D. W. Griffin and J. S. Lim, A new model based speech analysis/synthesis system, *Proc. Int. Conf. Acoust., Speech Signal Process.,* 1985, pp. 513–516.

31. S. Dimolitsas et al., Evaluation of voice codec performance for the Inmarsat mini-M system, *Proc., 10th Int. Digital Satellite Conf.,* Brighton, England, 1995.

32. S. Dimolitsas, F. L. Corcoran, and C. Ravishankar, Voice transmission quality of mobile satellite communications systems, *Int. J. Satellite Commu.,* **12**: 361–368, 1994.

33. R. McAulay and T. Quatieri, Speech analysis/synthesis based on a sinusoidal representation, *IEEE Trans. Acous. Speech Signal Process.,* **ASSP-34**: 744, 1986.

34. R. Zelinski and P. Noll, Adaptive transform coding speech signals, *IEEE Trans. Acoust. Speech Signal Process.,* **ASSP-25**: 299–309, 1977.

35. C. S. Ravishankar and S. Dimolitsas, Voice coding technology for digital aeronautical communications, *Air Traffic Control Q.,* **4** (3): 197–221, 1997.

36. S. Dimolitsas, F. L. Corcoran, and C. Ravishankar, Correlation between headphone and telephone-handset listener opinion scores for single stimulus voice coder assessments, *IEEE Lett. Signal Process.,* **2** (3): 41–43, 1995.

37. ITU-T Recommendation G.763, *Digital Circuit Multiplication Equipment Using 32 kb/s ADPCM and Digital Speech Interpolation,* Geneva, 1991.

38. J. H. Chen and R. Cox, The creation and evolution of 16 kbps LD-CELP: From concept to standard, in *Speech Communication,* Amsterdam: Elsevier/North-Holland, 1993, pp. 103–111.

39. Y. Linde, A. Buzo, and A. Gray, An algorithm for vector quantizer design, *IEEE Trans. Commun.,* **COM-28**: 84–95, 1980.

40. J. R. B. De Marca and N. S. Jayant, An algorithm for assigning binary indices to the codevectors of a multi-dimensional quantizer, *Proc. Int. Conf. Commun.,* Seattle, WA, pp. 1128–1132, 1987.

41. R. Salami et al., Description of the proposed ITU-T 8 kb/s speech coding standard, *Proc. 1995 IEEE Speech Coding Workshop,* Annapolis, MD, 1995, pp. 3–5.

42. S. Dimolitsas, C. S. Ravishankar, and G. Schröder, Current objectives for 4 kbit/s wireline-quality speech coding standardization, *IEEE Lett. Signal Process.,* **1** (11): 157–159, 1994.

43. I. Gerson and M. Jasiuk, Vector sum excited linear prediction (VSELP) speech coding at 8 Kb/s, *Proc. Int. Conf. Acoust., Speech Signal Process.,* Albuquerque, NM, 1990, pp. 461–464.

44. K. Hellwig et al., Speech coder for the European mobile radio system, *Proc. GLOBECOM,* Dallas, TX, 1989, pp. 1065–1069.

45. J. P. Campbell, Jr., T. E. Tremain, and V. C. Welch, The DoD 4.8 kb/s standard (the proposed federal standard FS1016), in B. S. Atal, V. Cuperman, and A. Gersho (eds), *Advances in Speech Coding,* Norwell, MA: Kluwer, 1991, pp. 121–133.

46. A. McCree et al., A 2.4 kbps MELP coder candidate for the new US federal standard, *Proc. Int. Conf. Acoust., Speech Signal Process.,* 1996, pp. 200–203.

47. T. E. Tremain, The government standard linear predictive coding algorithm: LPC-10, *Speech Technol.,* **1** (2): 40–49, 1982.

48. D. L. Richards, *Telecommunications by Speech,* New York: Wiley, 1973.

49. S. Dimolitsas, Subjective assessment methods for the measurement of digital speech coder quality, in B. S. Atal, V. Cuperman, and A. Gersho (eds.), *Speech and Audio Coding for Wireless Applications,* Norwell, MA: Kluwer, 1992.

50. ITU-T Recommendation G.131, *Stability and Echo* (Red Book), Malaga Torremonilos, 1984, Vol. III.1, pp. 183–194.

51. ITU-T Recommendation G.164, *Echo Suppressors* (Red Book), Malaga Torremonilos, 1984, Vol. III.1, pp. 225–258.

52. IEEE Recommended Practice for Speech Quality Measurements, *IEEE Trans. Audio Electroacoust.,* **AU-17**: 225–246, 1969.

53. S. Dimolitsas, F. L. Corcoran, and M. Baraniecki, Transmission quality of North American cellular, personal communications, and public switched telephone networks, *IEEE Trans Veh. Techol.,* **32**: 245–251, 1994.

54. S. Dimolitsas, F. L. Corcoran, and C. Ravishankar, Voice quality of interconnected PCS, Japanese cellular, and public switched telephone networks, *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.,* Detroit, MI, 1995, pp. 273–276.

55. S. Dimolitsas, F. L. Corcoran, and C. Ravishankar, Voice quality of interconnected North American cellular, European cellular, and public switched telephone networks, *Proc. IEEE Veh. Technol. Conf., VTC'95,* Chicago, 1995, pp. 719–722.

56. W. B. Kleijn, Encoding speech using prototype waveforms, *IEEE Trans. Speech Audio Process.,* **1**: 386–399, 1993.

57. P. Lupini and V. Cuperman, Vector quantization of harmonic magnitudes for low-rate speech coders, *Proc. GLOBECOM,* 1994, pp. 858–862.

58. V. Vaishampayan and N. Farvardin, Joint design of block source codes and modulation signal sets, *IEEE Trans. Inf. Theory,* **38**: 1230–1248, 1992.

59. S. Hong, P. K. M. Ho, and V. Cuperman, Combined speech and channel coding for mobile radio communications, *IEEE Trans. Veh. Technol.,* **43**: 1078–1087, 1994.

60. A. Husain and V. Cuperman, Reconstruction of missing packets for CELP based speech coders, *Proc. Int. Conf. Acoust., Speech Signal Process.,* 1995, pp. 245–248.

61. ITU-T Recommendation G.722, *7 kHz Audio Coding within 64 kbps,* Melbourne (Blue Book), 1988.

62. ITU-T Recommendation P.800, *Methods of Subjective Determination of Transmission Quality,* 1996.

63. ITU-T Recommendation P.861, *Objective Quality Measurement of Telephone Band (300–3400 Hz) Speech Coders,* 1996.

64. N. S. Jayant and P. Noll, *Digital Coding of Waveforms,* Englewood Cliffs, NJ: Prentice Hall, 1984.

65. J. Makhoul, Linear prediction: A tutorial review, *Proc. IEEE,* **63** (4): 561–580, 1975.

66. B. S. Atal and S. L. Hanauer, Speech analysis and synthesis by linear prediction of the speech wave, *J. Acoustical Society of America,* **50**: 637–655, 1971.

67. F. I. Itakura and S. Saito, Analysis-synthesis telephony based on the maximum likelihood method, *Proc. 6th Int. Cong. Acous.,* Tokyo, Japan, 1968, pp. C17–20.

68. B. S. Atal and M. R. Schroeder, Stochastic coding of speech signals at very low bit rates, *Proc. Int. Conf. Commun.,* 1984, pp. 1610–1613.

69. W. B. Kleijn, D. J. Krasinski, and R. H. Ketchum, Improved speech quality and efficient vector quantization in SELP, *Proc. Int. Conf. Acoust., Speech Signal Process.,* 1988, pp. 155–158.

70. D. Lin, New approaches to stochastic coding of speech sources at very low bit rates, in *Signal Processing III: Theories and Applications,* Elsevier-North Holland, 1986.

71. G. Davidson and A. Gersho, Complexity reduction methods for vector excitation coding, *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.,* 1986, pp. 3055–3058.

72. J. P. Adoul et al., Fast CELP Coding Based on Algebraic Codes, *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.,* 1987, pp. 1957–1960.

73. A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression,* Norwell, MA: Kluwer, 1992.

CHANNASANDRA RAVISHANKAR
Hughes Network Systems

SPIROS DIMOLITSAS
Lawrence Livermore National Laboratory

SPEECH CODING. See COMPANDORS; DATA CODING AND COMPRESSION FOR NETWORKING.