2 Digital speech coding

The human vocal and auditory organs form one of the most useful and complex communication systems in the animal kingdom. All speech (voice) sounds are formed by blowing air from the lungs through the vocal cords (also called the vocal fold), which act like a valve between the lung and vocal tract. After leaving the vocal cords, the blown air continues to be expelled through the vocal tract towards the oral cavity and eventually radiates out from the lips (see Figure 2.1). The vocal tract changes its shape with a relatively slow period (10 ms to 100 ms) in order to produce different sounds [1] [2].

In relation to the opening and closing vibrations of the vocal cords as air blows over them, speech signals can be roughly categorized into two types of signals: voiced speech and unvoiced speech. On the one hand, voiced speech, such as vowels, exhibit some kind of semi-periodic signal (with time-varying periods related to the pitch); this semi-periodic behavior is caused by the up–down valve movement of the vocal fold (see Figure 2.2(a)). As a voiced speech wave travels past, the vocal tract acts as a resonant cavity, whose resonance produces large peaks in the resulting speech spectrum. These peaks are known as formants (see Figure 2.2(b)).

On the other hand, the hiss-like fricative or explosive unvoiced speech, e.g., the sounds, such as s, f, and sh, are generated by constricting the vocal tract close to the lips (see Figure 2.3(a)). Unvoiced speech tends to have a nearly flat or high-pass spectrum (see Figure 2.3(b)). The energy in the signal is also much lower than that in voiced speech.

The speech sounds can be converted into electrical signals by a transducer, such as a microphone, which transforms the acoustic waves into an electrical current. Since most human speech contains signals below 4 kHz then, according to the sampling theorem [4] [5], the electrical current can be sampled (analog-to-digital converted) at 8 kHz as discrete data, with each sample typically represented by eight bits. This 8-bit representation, in fact, provides 14-bit resolution by the use of quantization step sizes which decrease logarithmically with signal level (the so-called A-law or μ-law [2]). Since human ears are less sensitive to changes in loud sounds than to quiet sounds, low-amplitude samples can be represented with greater accuracy than high-amplitude samples. This corresponds to an uncompressed rate of 64 kilobits per second (kbps).

In the past two to three decades, there have been great efforts towards further reductions in the bitrate of digital speech for communication and for computer storage [6] [7]. There are many practical applications of speech compression, for example, in digital cellular technology, where many users share the same frequency bandwidth and good compression allows more users to share the system than otherwise possible. Another example is in digital voice storage (e.g., answering machines). For a given memory size, compression [3] allows longer messages to be stored. Speech coding techniques can have the following attributes [2]:

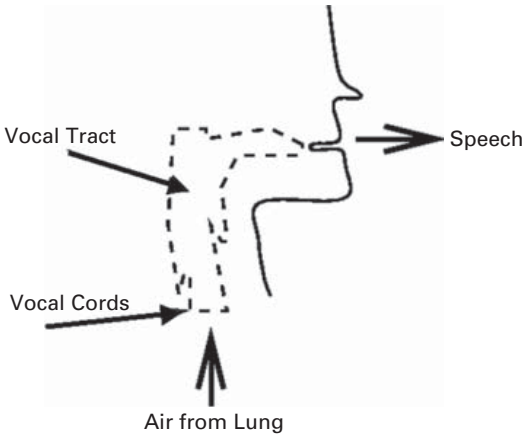

Figure 2.1 The human speech-production mechanism [3].

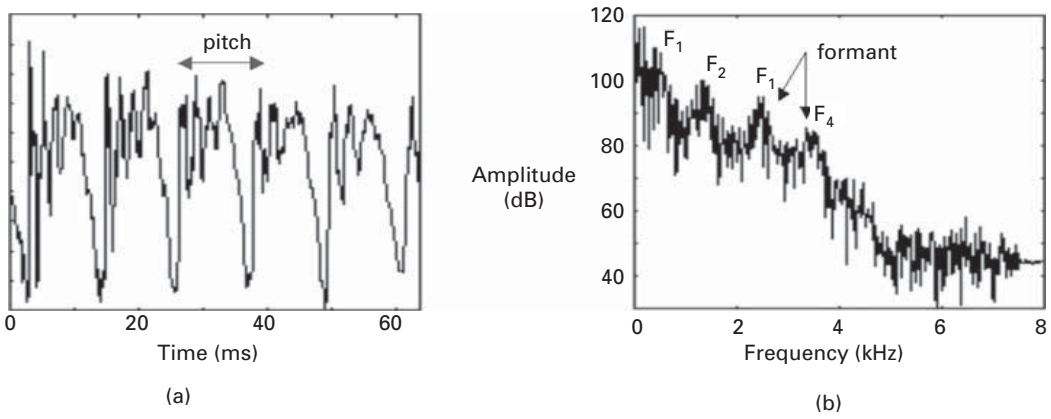

Figure 2.2 Voiced speech can be considered as a kind of semi-periodic signal.

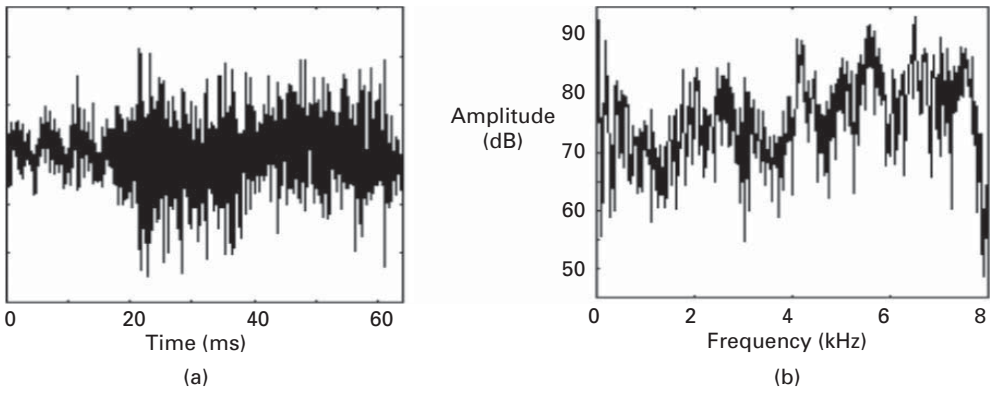

Figure 2.3 Hiss-like fricative or explosive unvoiced speech is generated by constricting the vocal tract close to the lips.

(1) *Bitrate* This is 800 bps – 16 kbps, most 4.8 kbps or higher, normally the sample-based waveform coding (e.g., ADPCM-based G.726 [8]) has a relatively higher bitrate, while block-based parametric coding has a lower bitrate.
(2) *Delay* The lower-bitrate parametric coding has a longer delay than waveform coding; the delay is about 3–4 times the block (frame) size.
(3) *Quality* The conventional objective mean square error (MSE) is only applicable to waveform coding and cannot be used to measure block-based parametric coding, since the reconstructed (synthesized) speech waveform after decoding is quite different from the original waveform. The subjective mean opinion score (MOS) test [9], which uses 20–60 untrained listeners to rate what is heard on a scale from 1 (unacceptable) to 5 (excellent), is widely used for rating parametric coding techniques.
(4) *Complexity* This used to be an important consideration for real-time processing but is less so now owing to the availability of much more powerful CPU capabilities.

2.1 LPC modeling and vocoder

With current speech compression techniques (all of which are lossy), it is possible to reduce the rate to around 8 kbps with almost no perceptible loss in quality. Further compression is possible at the cost of reduced quality. All current low-rate speech coders are based on the principle of *linear predictive coding (LPC)* [10] [11], which assumes that a speech signal $s(n)$ can be approximated as an auto-regressive (AR) formulation

$$\hat{s}(n) = e(n) + \sum_{k=1}^{p} a_k s(n-k) \tag{2.1}$$

or as an all-pole vocal tract filter, $H(z)$:

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}, \tag{2.2}$$

where the residue signal $e(n)$ is assumed to be white noise and the linear regression coefficients $\{a_k\}$ are called LPC coefficients. The LPC-based speech coding system is illustrated in Figure 2.4 [12]; note that a speech "codec" consists of an encoder and a decoder. This LPC modeling, which captures the formant structure of the short-term speech spectrum, is also called short-term prediction (STP).

Commonly the LPC analysis on synthesis filter has order p equal to 8 or 10 and the coefficients $\{a_k\}$ are derived on the basis of a 20–30 ms block of data (frame). More specifically, the LPC coefficients can be derived by solving a least squares solution assuming that $\{e(n)\}$ are estimation errors, i.e., solving the following normal (Yule–Walker) linear equation:

$$\begin{bmatrix} r_s(1) \\ r_s(2) \\ r_s(3) \\ \vdots \\ r_s(p) \end{bmatrix} = \begin{bmatrix} r_s(0) & r_s(1) & r_s(2) & \cdots & r_s(p-1) \\ r_s(1) & r_s(0) & r_s(1) & \cdots & r_s(p-2) \\ r_s(2) & r_s(1) & r_s(0) & \cdots & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_s(p-1) & r_s(p-2) & r_s(p-3) & \cdots & r_s(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix} \tag{2.3}$$

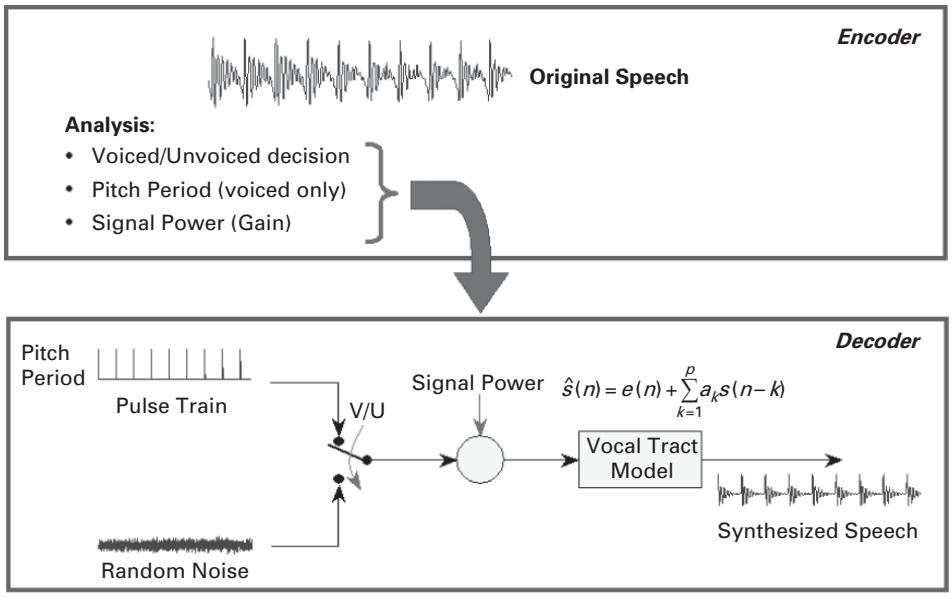

Figure 2.4 A typical example of a LPC-based speech codec.

where the autocorrelation $r_s(k)$ is defined as

$$r_s(k) = \sum_{n=0}^{N-k-1} s(n)s(n+k). \tag{2.4}$$

Owing to the special Toeplitz matrix structure of the Yule–Walker linear equation, the LPC coefficients $\{a_k\}$ can be solved using the efficient Levinson–Durbin recursion algorithm [12].

A complete LPC-based voice coder (vocoder) consists of an analysis performed in the encoder (see the upper part of Figure 2.4), which determines the LPC coefficients $\{a_k\}$ and the gain parameter G (a side product of the Levinson–Durbin recursion in solving for the $\{a_k\}$ coefficients) and, for each frame, a voice/unvoiced decision with pitch period estimation. As shown in Figure 2.5, this is achieved through a simplified (ternary valued) autocorrelation (AC) calculation method. The pitch-period search is confined to $F_s/350$ and $F_s/80$ samples (i.e., 23–100 samples) or, equivalently, the pitch frequency is confined to between 80 to 350 Hz. Pitch period estimation is sometimes called long-term prediction (LTP), since it captures the long-term correlation, i.e., periodicity, of the speech signals. The autocorrelation function $R(k)$ is given by

$$R(k) = \sum_{m=0}^{N-k-1} x^c(m)x^c(m+k), \tag{2.5}$$

where

$$x^c(n) = \begin{cases} +1 & \text{if} \quad x(n) > C_L, \\ -1 & \text{if} \quad x(n) < -C_L, \\ 0 & \text{otherwise}, \end{cases}$$

where C_L denotes the threshold, which is equal to 30% of the maximum of the absolute value of $\{x(n)\}$ within this frame. The 10 LPC coefficients $\{a_k\}$, together with the pitch period and gain parameters, are derived on the basis of 180 samples (22.5 ms) per frame and are encoded at 2.4

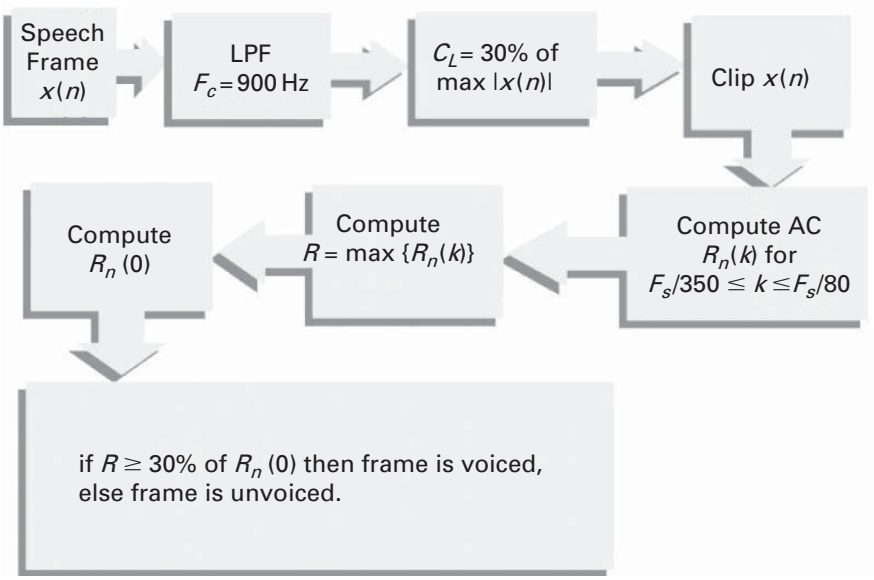

Figure 2.5 The voiced or unvoiced decision with pitch period estimation is achieved through a simplified autocorrelation calculation method (http://www.ee.ucla.edu/~ingrid/ee213a/speech/speech.html).

kbps for transmission or storage (according to the LPC-10 or FS-1015 standards) [13] [14]. The decoder is responsible for synthesizing the speech using the coefficients and parameters in the flow chart shown in the lower part of Figure 2.4. The 2.4 kbps FS-1015 was used in various low-bitrate and secure applications, such as in defense or underwater communications, until 1996, when the 2.4 kbps LPC-based standard was replaced with the new mixed-excitation linear prediction (MELP) coder [15][16] by the United States Department of Defense Voice Processing Consortium (DDVPC). The MELP coder is based on the LPC model with additional features that include mixed excitation, aperiodic pulses, adaptive spectral enhancement, pulse dispersion filtering, and Fourier magnitude modeling.

Even though the speech synthesized from the LPC vocoder is quite intelligible it does sound somewhat unnatural, with MOS values [9] ranging from 2.7 to 3.3. This unnatural speech quality results from the over-simplified representation (i.e., one impulse per pitch period) of the residue signal $e(n)$, which can be calculated from Eq. (2.5) after the LPC coefficients have been derived (see Figure 2.6). To improve speech quality, many other (hybrid) speech coding standards have been finalized, all having more sophisticated representations of the residue signal $e(n)$, as shown in Figure 2.6:

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k). \tag{2.6}$$

To further improve the representation of the residue signal $e(n)$, long-term prediction (LTP) can be applied by first removing the periodic redundancy caused by the semi-periodic pitch movement. More specifically, each frame of speech (20 or 30 ms) is divided into four uniform subframes, each with N_{sf} samples, taking each of the subframe samples backwards to find the best-correlated counterpart (which has a time lag of p samples) having the necessary gain factor β. The LTP-filtered signal is called the excitation $u(n)$ and has an even smaller dynamic range; it can thus be encoded more effectively (see Figure 2.7). Different

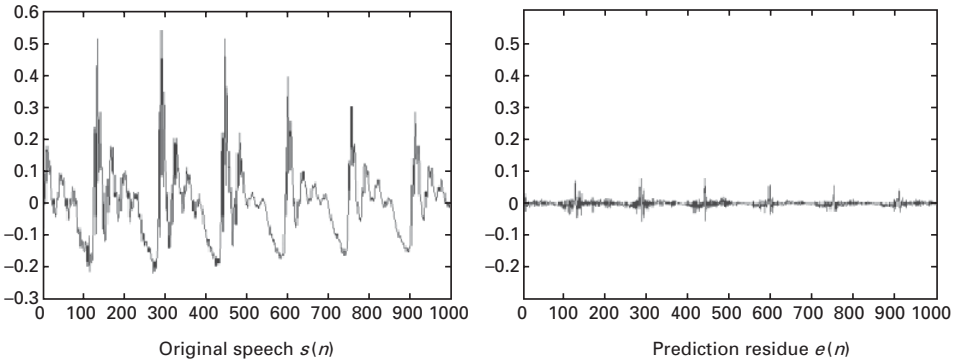

Figure 2.6 The prediction residue signal $e(n)$ of LPC can be calculated from Eq. (2.5) after the LPC coefficients have been derived.

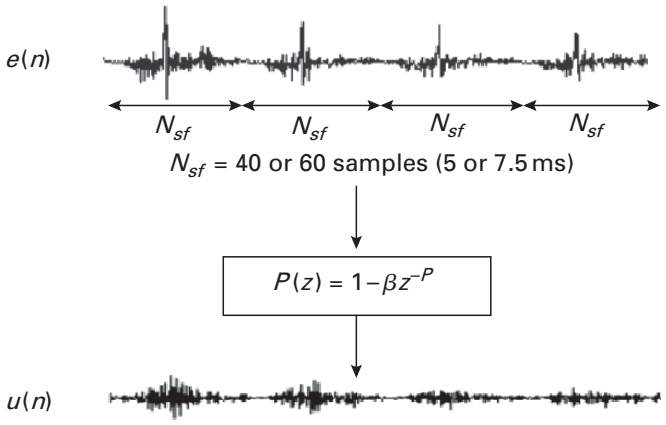

Figure 2.7 The LTP-filtered signal, the excitation $u(n)$, has an even smaller dynamic range than the unfiltered signal and can thus be encoded more effectively.

encoding of the excitation signals (with also some slight variations in STP analysis) leads to different speech coding standards (see Table 2.1), e.g.,

(1) *Regular pulse excitation (RPE)* This is used mainly to encode the magnitude of selected (uniformly decimated) samples; e.g., GSM [17] [18] [19].
(2) *Code-excited linear prediction (CELP)* This is used mainly to encode excitations based on pre-clustered codebook entries, i.e., magnitude and locations are both important; e.g., CELP [20], G.728 [21] [22], and VSELP [23].
(3) *Multiple pulse coding (MPC)* This is used mainly to encode the locations of selected samples (pulses with sufficiently large magnitude); e.g., G.723.1 [24] and G.729 [25].

2.2 Regular pulse excitation with long-term prediction

The global system for mobile communications (GSM) [17] [18] [19] standard, the digital cellular phone protocol defined by the European Telecommunication Standards Institute (ETSI, http://www.etsi.org/), derives eight-order LPC coefficients from 20 ms frames and

Table 2.1 Various encodings of excitation signals (with also some slight variations in STP analysis) and the corresponding speech coding standards

Standards	Year	Bitrate (kbps)	MOS
PCM (PSTN)	1972	64	4.4
LPC-10 (FS-1015)	1976 (1996)	2.4	2.7 (3.3)
G.726 (ADPCM, G.721)	1990	16, 24, 32, 40	4.1 (32 kbps)
GSM (RPE-LTP)	1987	13	3.7
CELP (FS-1016)	1991	4.8	3.2
G.728 (LD-CELP)	1992	16	4
VSELP (IS-54)	1992	8	3.5
G.723.1 (MPC-MLQ)	1995	6.3/5.3	3.98/3.7
G.729 (CS-ACELP)	1995	8	4.2

Table 2.2 There are 260 bits allocated for each GSM frame (20 ms), resulting in a total bitrate of 13 kbps

Parameters	Bits per subframe	Bits per frame
LPC coefficients	–	36
LTP lag	7	28
LTP gain	2	8
ORPE subsequence scaling factor	6	24
ORPE subsequence index	2	8
ORPE subsequence values	39	156
Total	56	260

uses a regular pulse excitation (RPE) encoder over the excitation signal $u(n)$ after redundancy removal with long-term prediction (LTP). More specifically, GSM sorts each subframe (5 ms, 40 samples) after LTP into four sequences:

(1) sequence 1: 0 3 6 9 . . . 36
(2) sequence 2: 1 4 7 10 . . . 37
(3) sequence 3: 2 5 8 11 . . . 38
(4) sequence 4: 3 6 9 12 . . . 39

Only one sequence, the highest-energy sequence among the four, per subframe is selected for encoding. Each sample of the selected sequence is quantized at three bits (instead of the original sampling of 13 bits). This is called an optimized RPE (ORPE) sequence. The bit allocation of each GSM frame (20 ms) is shown in Table 2.2 for the case when 260 bits are used per frame, resulting in a total bitrate of 13 kbps. The overall operations of a GSM encoder are shown in Figure 2.8 and those of a GSM decoder in Figure 2.9.

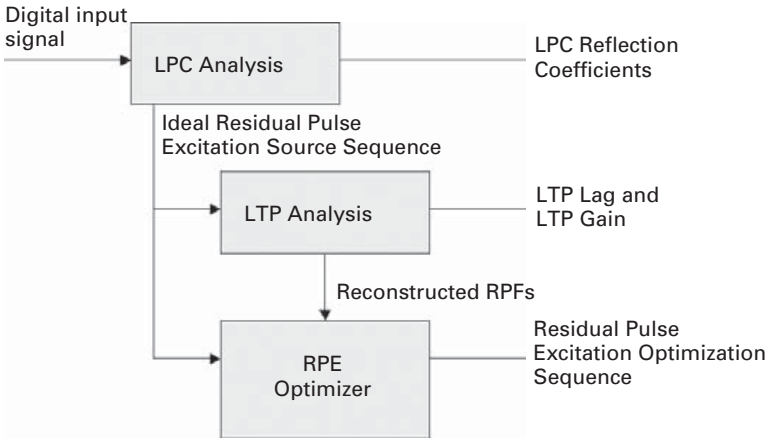

Figure 2.8 A GSM encoder.

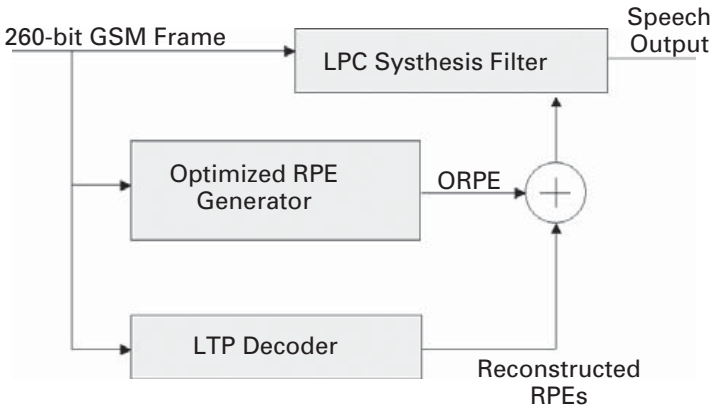

Figure 2.9 A GSM decoder.

2.3 Code-excited linear prediction (CELP)

The RPE uses a downsampled version of excitation signals to represent the complete excitation, while a code-excited linear prediction (CELP) coder uses a codebook entry from a vector quantized (VQ) codebook to represent the excitation; see Figure 2.10. In this figure, $P(z)$ is the LTP filter and $1/P(z)$ is used to compensate for the difference operation performed in the LTP filtering (i.e., recovering $u(n)$ back to $e(n)$); the $1/A(z)$ filter synthesizes the speech $\hat{s}(n)$ to be compared with the original speech $s(n)$. The objective of encoding the excitations is to choose the codebook entry (codeword) that minimizes the weighted error between the synthesized and original speech signals. This technique, referred to as *analysis by synthesis*, is widely used in CELP-based speech coding standards. The analysis by synthesis technique simulates the decoder in the encoder so that the encoder can choose the optimal configuration, or tune itself for the best parameters, to minimize the weighted error calculated from the *original* speech and the *reconstructed* speech (see Figure 2.11).

The perceptual weighting filter $A(Z)/A(Z/\gamma)$, $\gamma \approx 0.7$, is used to provide different weighting on the error signals by allowing for more error around the resonant formant

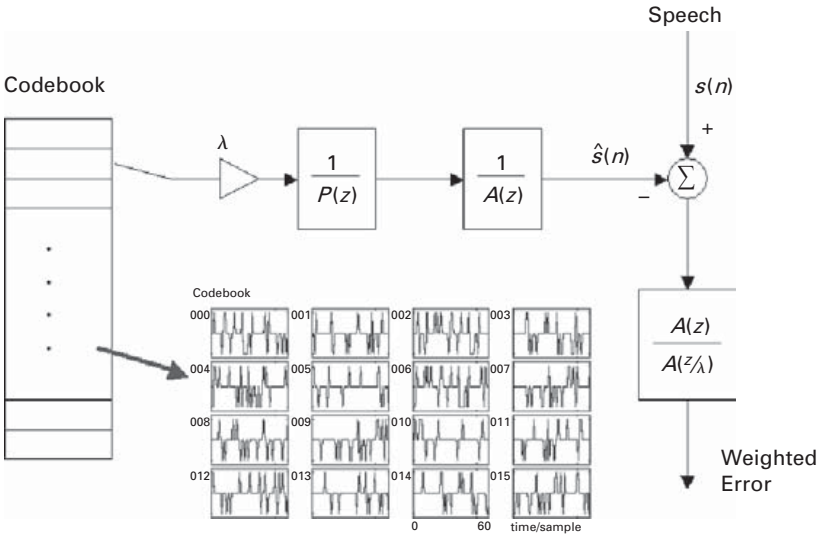

Figure 2.10 A CELP coder uses a codebook entry from a vector-quantized codebook to represent the excitation.

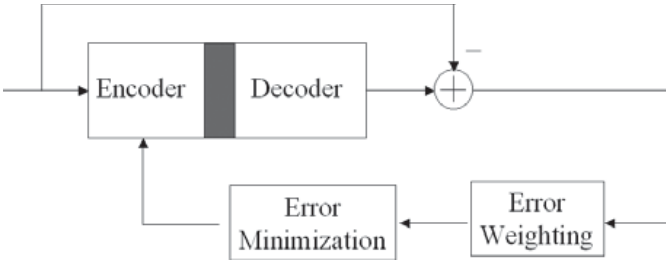

Figure 2.11 The analysis by synthesis technique simulates the decoder in the encoder so that the encoder can choose the optimal configuration to minimize the weighted error calculated from the original speech and the reconstructed speech.

frequencies (by widening the bandwidth of spectral resonance), since human ears are less sensitive to the error around those frequencies. A typical perceptual weighting filter frequency response, given the original LPC filter frequency response, is presented in Figure 2.12 for various values of γ.

The US Federal Standard FS-1016 [20] is based on CELP techniques with 4.8 kbps data rate. The index of the chosen codeword is encoded for transmission or storage. An effective algorithm for the codeword search was developed so that a CELP coder can be implemented in real time using digital signal processors. In FS-1016, 10 LPC coefficients were derived from each 30 ms frame and there are 512 codewords in the excitation codebook, each codeword having 7.5 ms (60 samples) of ternary valued (+1, 0, −1) excitation data. The FS-1016 is currently not widely used since its successor, MELP [15], provides better performance in all applications, even though CELP was used in some secure applications as well as adopted in MPEG-4 for encoding natural speech for 3G cellular phones.

Another CELP-based speech coder is the low-delay CELP G.728 (LD-CELP) [21] [22], which provides 16 kbps speech with a quality similar to that of the 32 kbps speech provided by the ADPCM waveform-based G.726 speech coding. The G.728 speech coder is widely used in

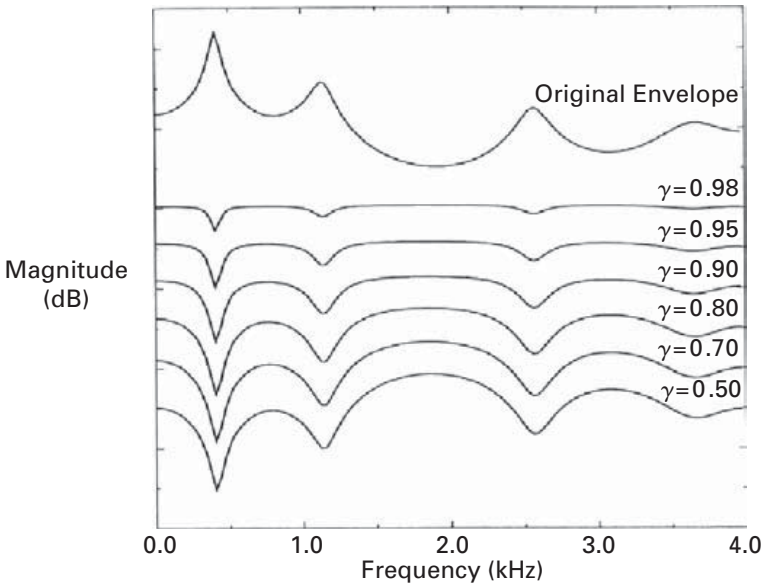

Figure 2.12 A typical perceptual weighting filter frequency response, given the original LPC filter frequency response, for various values of γ.

voice over cable or voice over IP (VoIP) teleconferencing applications through packet networks. For G.728, 50th-order LPC coefficients were derived recursively, on the basis of its immediate 16 past outputs (there was no need to transmit the LPC coefficients since they can be computed in the receiver side in a recursive backward-adaptive fashion) and there are 1024 codewords contained in the excitation codebook, each codeword having only five samples of excitation data.

The major drawback of CELP-based speech coding is the very large computational requirements. To overcome this requirement, the vector sum excited linear prediction (VSELP) speech coder [23], which also falls into the CELP class, utilizes a codebook with a structure that allows for a very efficient search procedure. This VSELP coder (see Figure 2.13), at 8 kbps with 20 ms/frame, was selected by the Telecommunication Industry Association (TIA, http://www.tiaonline.org/) as the standard for use in North American TDMA IS-54 digital cellular telephone systems.

As shown in Figure 2.13, the VSELP codec contains two separate codebooks ($k = 1$ or 2), each of which can contribute $2^M = 2^7 = 128$ codevectors (in fact there are only 64 distinct patterns since $u(n)$ and $-u(n)$ are regarded as the same signal pattern) if constructed as a linear combination of $M = 7$ basis vectors $\{v_{k,m}(n), m = 1, 2, \ldots, M\}$,

$$u_{k,i}(n) = \sum_{m=1}^{M} \theta_{i,k} v_{k,m}(n) \qquad \text{and} \qquad u(n) = v_i u_{1,i}(n) + v_2 u_{2,j}(n), \tag{2.7}$$

where $u_{k,i}(n)$ denotes the ith codevector in the kth codebook; $v_{k,m}(n)$ denotes the mth basis vector of the kth codebook and $\theta_{i,m} = \pm 1$; $u(n)$ denotes the resulting combined excitation signal, which should be further compensated by inverse LTP filtering,

$$\frac{1}{p(z)} = \frac{1}{1 - \beta z^{-T}},$$

to recover $e(n)$ from $u(n)$. The pitch pre-filter and spectral post-filter are also used to further finetune the estimated parameters for a better synthesis.

The bit allocation of each VSELP frame (20 ms long) is shown in Table 2.3; 160 bits are used per frame, resulting in a total bitrate of 8 kbps.

Table 2.3 There are 160 bits allocated for each VSELP frame (20 ms), resulting in a total bitrate of 8 kbps

Parameters	Bits per subframe	Bits per frame
LPC coefficients	–	38
Energy – $R_q(0)$	–	5
Excitation codes (I, H)	7+7	56
Lag (L)	7	28
GS-P0-P1 code	8	32
<unused>	–	1
Total	29	160

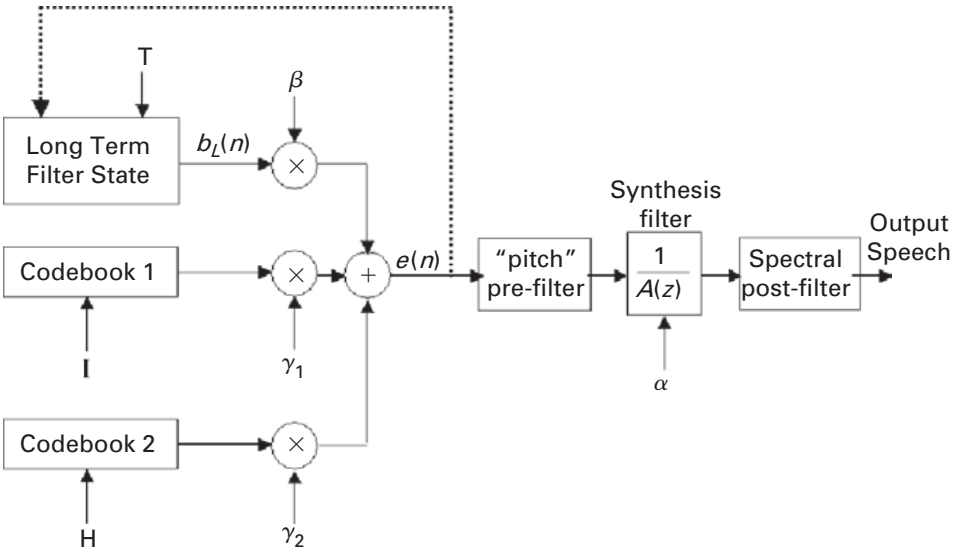

Figure 2.13 The VSELP codec contains two separate codebooks, each of which can contribute $2^7 = 128$ distinct code vectors.

2.4 Multiple-pulse-excitation coding

Another main category of speech coding is based on encoding only the locations of large enough pulses (i.e., the excitations $u(n)$ after LTP). The speech coder G.723.1 [24] [26] [27] provides near toll-phone quality transmitted speech signals. This type of speech coder has been standardized for internet speech transmission. The G.723.1 processes the speech using 30 ms frames, each 7.5 ms subframe containing 60 samples.

The 10th-order LPC analysis is based on a 60 sample subframe ($i = 0, 1, 2, 3$) but is only performed on the last subframe ($i = 3$) of each frame. The LPC coefficients are then converted into line spectral pairs (LSPs), which are defined to be the roots of $P(z)$ and $Q(z)$, where

$$\begin{aligned} P(z) &= A(z) + z^{-p}A(z^{-1}), \\ Q(z) &= A(z) - z^{-p}A(z^{-1}). \end{aligned} \tag{2.8}$$

This ensures better stability during the quantization process.

The LSP vectors for the other three subframes of each frame can be derived using linear interpolation between the current frame's LSP vector, e.g., the $\{P_n\}$, and the previous frame's LSP vector, $\{P_{n-1}\}$, i.e.,

$$p_{ni} = \begin{cases} 0.75p_{n-1} + 0.25p_n, & i = 0, \\ 0.50p_{n-1} + 0.50p_n, & i = 1, \\ 0.25p_{n-1} + 0.75p_n, & i = 2, \\ p_n, & i = 3. \end{cases} \tag{2.9}$$

The speech coder G.723.1 also introduces a more efficient LTP technique to improve the accuracy of (pitch) redundancy removal based on open and closed loop analyses. This gives a lower encoding bitrate with better speech synthesis quality. A formant perceptual weighting filter, $W_i(z)$ (similar to the one used in the CELP analysis by synthesis technique), where

$$W_i(z)\frac{A(z/r_1)}{A(z/r_2)} = \frac{1 - \sum_{j=1}^{10} a_{ij}z^{-j}(0.9)^j}{1 - \sum_{j=1}^{10} a_{ij}z^{-j}(0.5)^j}, \qquad 0 \le i \le 3,\ r_1 = 0.9,\ r_2 = 0.5 \tag{2.10}$$

is constructed for every subframe, using the unquantized LPC coefficients $\{a_{ij}\}$ derived from the interpolated LSPs $\{P_n\}$ of each subframe, i.e., every subframe of input speech signal is first filtered to obtain perceptually weighted speech, $f(n)$;

$$CR_{OL}(j) = \frac{\left(\sum_{n=0}^{119} f(n)f(n-j)\right)^2}{\sum_{n=0}^{119} f(n-j)f(n-j)}, \qquad 18 \le j \le 142. \tag{2.11}$$

The open-loop LTP estimates the pitch period for every two subframes (120 samples) and a cross-correlation criterion is used on perceptually weighted speech, $f(n)$ (see Eq. 2.11). The index j which maximizes CR_{OL} is selected and named $\hat{L}$. The open-loop LTP analysis is followed by a closed-loop LTP analysis, which estimates the pitch lag around the open-loop pitch lag $\hat{L}$ calculated earlier. More specifically, for subframes 0 and 2 the closed-loop pitch lag is selected in the range ± 1 of the open-loop pitch lag (coded with seven bits) and, for subframes 1 and 3, the lag differs from the subframe's open-loop pitch lag by -1, 0, $+1$ or $+2$ (coded with two bits).

Instead of directly determining the pulses from the excitation signal $u(n)$ resulting from the speech $s(n)$ passing through the STP and LTP operations, G.723.1 tries to determine a pure multiple-pulse signal $v(n)$, which can be filtered by a 20th-order FIR weighted synthesis filter $h(n)$ to produce $v'(n)$ so as to approximate $u(n)$, another effective use of the analysis by synthesis technique:

$$v'(n) = \sum_{j=0}^{19} h(j)v(n-j), \qquad 0 \le n \le 59, \tag{2.12}$$

where $v(n)$ consists only of multiple pulses, i.e., six pulses for subframes 0 and 2, with magnitudes of either $+G$ or $-G$, the gain factor analyzed for this frame; and five pulses for subframes 1 and 3, also with magnitudes of either $+G$ or $-G$. This results in a 6.3 kbps multiple pulse coding (MPC) data rate (see Table 2.4), since

Table 2.4 The bit allocation for a 30 ms G.723.1 frame, which results in a 6.3 kbps multiple-pulse coding (MPC) data rate

Parameters	Subframe 0	Subframe 1	Subframe 2	Subframe 3	Total
LPC indices					24
Adaptive codebook lags	7	2	7	2	18
Combined gains	12	12	12	12	48
Pulse positions	20	18	20	18	73
Pulse signs	6	5	6	5	22
Grid index	1	1	1	1	4
Total					189

Table 2.5 The 5.3 kbps G.723.1 version locates at most four pulses from each subframe, and the four pulses have to be limited to one of four predefined groups

Sign	Positions
±1	0, 8, 16, 24, 32, 40, 48, 56
±1	2, 10, 18, 26, 34, 42, 50, 58
±1	4, 12, 20, 28, 36, 44, 52
±1	6, 14, 22, 30, 38, 46, 54

$$189\frac{\text{bits}}{\text{frame}} \times 33\frac{\text{frames}}{\text{seconds}} = 6.3 \text{ kbps}.$$

To reduce the bitrate further, G.723.1 also offers a 5.3 kbps version by locating at most four pulses from each subframe; the four pulses have to be limited to one of four predefined groups, as shown in Table 2.5.

Another multiple-pulse-coding-based speech coder is called the conjugate structure algebraic code-excited linear prediction (CS-ACELP) G.729 [25] [26] [27] and can achieve 32 kbps G.726 ADPCM toll-phone quality with only 8 kbps. It has been adopted in several Internet-based VoIP or session initiation protocol (SIP) phones. The G.729 uses 10 ms frames, with 5 ms (40 sample) subframes for excitation signal representation. Similarly to G.723.1, this speech codec allows four nonzero pulses (unit magnitude) and each pulse has to be chosen from a predefined group. This is called interleaved single-pulse permutation (ISPP), and the groups are as follows:

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

References

[1] J. D. Gibson, T. Berger, T. Lookabaugh, D. Lindbergh, and R. L. Baker, *Digital Compression for Multimedia: Principles and Standards*, Morgan Kauffman, 1998.

[2] W. B. Kleijn and K. K. Paliwal, *Speech Coding and Synthesis*, Elsevier Science, 1995.

[3] "Speech compression, by Data-Compression.com," http://www.data-compression.com/speech.html.

[4] J. Jerri, "The Shannon sampling theorem – its various extensions and applications: a tutorial review," *Proc. IEEE*, 65(11): 1565–1596, November 1977.

[5] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, Second edition, Prentice Hall, 1999.

[6] A. Gersho, "Advances in speech and audio compression," *Proc. IEEE*, 82(6): 900–918, June 1994.

[7] A. S. Spanias, "Speech coding: a tutorial review," *Proc. IEEE*, 82(10): 1542–1582, October 1994.

[8] "40, 32, 24, 16 kbit/s adaptive differential pulse code modulation (ADPCM)," ITU G.726: December 1990, http://www.itu.int/rec/T-REC-G.726/e.

[9] "Methods for subjective determination of transmission quality," ITU Recommendation P.800; August 1996, http://www.itu.int/rec/T-REC-P.800-199608-I/en.

[10] S. Saito and F. Itakura, "The theoretical consideration of statistically optimum methods for speech spectral density," Report No. 3107, Electrical Communication Laboratory, NTT, Tokyo, December 1966.

[11] B. S. Atal, "The history of linear prediction," *IEEE Signal Process. Mag.*, 23(2): 154–161, March 2006.

[12] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, 1978.

[13] T. E. Tremain, "The Government Standard linear predictive coding algorithm: LPC-10," *Speech Technol.*, 40–49, April 1982.

[14] J. P. Campbell Jr. and T.E. Tremain, "Voiced/unvoiced classification of speech with applications to the US government LPC-10E algorithm," in *Proc. Int Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. II, pp. 473–476, 1986.

[15] L. M. Supplee, R. P. Cohn, J. S. Collura, and A. V. McCree, "MELP: the new Federal Standard at 2400 bps," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. II: pp. 1591–1594, April 1997.

[16] M. A. Kohler, "A comparison of the new 2400 bps MELP Federal Standard with other standard coders," in *Proc. Int Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. II, pp. 1587–1590, April 1997.

[17] "Global system for mobile (GSM) world," http://www.gsmworld.com/index.shtml.

[18] P. Vary, K. Hellwig, R. Hofmann, R. J. Sluyter, C. Galand, and M. Rosso, "Speech codec for the European mobile radio system," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. I, pp. 227–230, April 1988.

[19] K. Hellwig, P. Vary, D. Massaloux, J.P. Petit, C. Galand, and M. Rosso, "Speech codec for the European mobile radio system," in *Proc. IEEE Global Communications Conf. (GLOBECOM)*, Vol. 2, pp. 1065–1069, November 1989.

[20] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): high-quality speech at very low bitrates," in *Proc. ICASSP'85*, pp. 937–940, March 1985.

[21] A. Kumar and A. Gersho, "LD-CELP speech coding with nonlinear prediction," *IEEE Signal Processing Letters*, 4(4): 89–91, April 1997.

[22] "Coding of speech at 16 kbit/s using low-delay code excited linear prediction," ITU-T Recommendation G.728, http://www.itu.int/rec/T-REC-G.728/e.

[23] I. A. Gerson, M. A. Jasiuk, "Vector sum excited linear prediction (VSELP) speech coding at 8 kbps," in *Proc. Int. Conf. on Acoustics, Speech, Signal Processing (ICASSP)*, pp. 461–464, April 1990.

[24] ITU-T Recommendation G.723.1, "Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbps," http://www.itu.int/rec/T-REC-G.723.1/e.

[25] "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP), ITU-T Recommendation G.729, http://www.itu.int/rec/T-REC-G.729/e.

[26] R. V. Cox, "Three new speech coders from the ITU cover a range of applications," *IEEE Commun. Mag*., 35(9): 40–47, September 1997.

[27] R. V. Cox and P. Kroon, "Low bit-rate speech coders for multimedia communication," *IEEE Commun. Mag*., 34(12): 34–42, December 1996.