Chan, W.; Gerson, I. & Miki, T. “Half-Rate Standards”
Mobile Communications Handbook
Ed. Suthan S. Suthersan
Boca Raton: CRC Press LLC, 1999

Half-Rate Standards

Wai-Yip Chan
Illinois Institute of Technology

Ira Gerson
Motorola Corporate Systems Research Laboratories

Toshio Miki
NTT Mobile Communication Network, Inc.

30.1 Introduction

A half-rate speech coding standard specifies a procedure for digital transmission of speech signals in a digital cellular radio system. The speech processing functions that are specified by a half-rate standard are depicted in Fig. 30.1. An input speech signal is processed by a *speech encoder* to generate

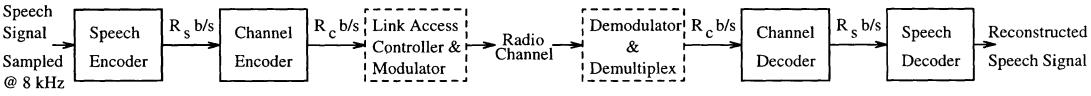

FIGURE 30.1: Digital speech transmission for digital cellular radio. Boxes with solid outlines represent processing modules that are specified by the half-rate standards.

a digital representation at a *net bit rate* of R_s bits per second. The encoded bit stream representing the input speech signal is processed by a *channel encoder* to generate another bit stream at a *gross bit rate* of R_c bits per second, where $R_c > R_s$. The channel encoded bit stream is organized into data frames, and each frame is transmitted as payload data by a radio-link access controller and modulator. The net bit rate R_s counts the number of bits used to describe the speech signal, and the difference between the gross and net bit rates ($R_c - R_s$) counts the number of error protection bits needed by the *channel decoder* to correct and detect transmission errors. The output of the channel decoder is given

to the *speech decoder* to generate a *quantized* version of the speech encoder's input signal. In current digital cellular radio systems that use time-division multiple access (TDMA), a voice connection is allocated a fixed transmission rate (i.e., R_c is a constant). The operations performed by the speech and channel encoders and decoders and their input and output data formats are governed by the half-rate standards.

Globally, three major TDMA cellular radio systems have been developed and deployed. The initial digital speech services offered by these cellular systems were governed by *full-rate standards.* Because of the rapid growth in demand for cellular services, the available transmission capacity in some areas is frequently saturated, eroding customer satisfaction. By providing essentially the same voice quality but at half the gross bit rates of the full-rate standards, half-rate standards can readily double the number of callers that can be serviced by the cellular systems. The gross bit rates of the full-rate and half-rate standards for the European Groupe Speciale Mobile (GSM), Japanese Personal Digital Cellular[1] (PDC), and North American cellular (IS-54) systems are listed in Table 30.1. The three systems were developed and deployed under different time tables. Their disparate full- and half-bit rates partly reflect this difference. At the time of writing (January, 1995), the European and the Japanese systems have each selected an algorithm for their respective half-rate **codec.** Standardization of the North American half-rate codec has not reached a conclusion as none of the candidate algorithms has fully satisfied the standard's requirements. Thus, we focus here on the Japanese and European half-rate standards and will only touch upon the requirements of the North American standard.

TABLE 30.1 Gross Bit Rates Used for Digital Speech Transmission in Three TDMA Cellular Radio Systems

	Gross Bit Rate, b/s	
Standard Organization and Digital Cellular System	Full Rate	Half Rate
European Telecommunications Standards Institute (ETSI), GSM	22,800	11,400
Research & Development Center for Radio Systems (RCR), PDC	11,200	5,600
Telecommunication Industries Association (TIA), IS-54	13,000	6,500

30.2 Speech Coding for Cellular Mobile Radio Communications

Unlike the relatively benign transmission media commonly used in the public-switched telephone network (PSTN) for analog and digital transmission of speech signals, mobile radio channels are impaired by various forms of fading and interference effects. Whereas proper engineering of the radio link elements (modulation, power control, diversity, equalization, frequency allocation, etc.) ameliorates fading effects, burst and isolated bit errors still occur frequently. The net effect is such that speech communication may be required to be operational even for bit-error rates greater than 1%. In order to furnish reliable voice communication, typically half of the transmitted payload bits are devoted to error correction and detection.

It is common for low-bit-rate speech codecs to process samples of the input speech signal one frame

[1] Personal Digital Cellular was formerly Japanese Digital Cellular (JDC).

at a time, e.g., 160 samples processed once every 20 ms. Thus, a certain amount of time is required to gather a block of speech samples, encode them, perform channel encoding, transport the encoded data over the radio channel, and perform channel decoding and speech synthesis. These processing steps of the speech codec add to the overall end-to-end transmission delay. Long transmission delay hampers conversational interaction. Moreover, if the cellular system is interconnected with the PSTN and a four-wire to two-wire (analog) circuit conversion is performed in the network, feedbacks called *echoes* may be generated across the conversion circuit. The echoes can be heard by the originating talker as a delayed and distorted version of his/her speech and can be quite annoying. The annoyance level increases with the transmission delay and may necessitate (at additional costs) the deployment of **echo cancellers**.

A consequence of user mobility is that the level and other characteristics of the acoustic background noise can be highly variable. Though acoustic noise can be minimized through suitable acoustic transduction design and the use of adaptive filtering/cancellation techniques [9, 13, 15], the speech encoding algorithm still needs to be robust against background noise of various levels and kinds (e.g., babble, music, noise bursts, and colored noise).

Processing complexity directly impacts the viability of achieving a circuit realization that is compact and has low-power consumption, two key enabling factors of equipment portability for the end user. Factors that tend to result in low complexity are fixed-point instead of floating-point computation, lack of complicated arithmetic operations (division, square roots, transcendental functions), regular algorithm structure, small data memory, and small program memory. Since, in general, better speech quality can be achieved with increasing speech and channel coding delay and complexity, the digital cellular mobile-radio environment imposes conflicting and challenging requirements on the speech codec.

30.3 Codec Selection and Performance Requirements

The half-rate speech coding standards are drawn up through competitive testing and selection. From a set of candidate codec algorithms submitted by contending organizations, the one algorithm that meets basic selection criteria and offers the best performance is selected to form the standard. The codec performance measures and codec testing and selection procedures are set out in a test plan under the auspices of the organization (Table 30.1) responsible for the standardization process (see, e.g., [16]). Major codec characteristics evaluated are speech quality, delay, and complexity. The full-rate codec is also evaluated as a *reference codec*, and its evaluation scores form part of the selection criteria for the codec candidates.

The speech quality of each candidate codec is evaluated through listening tests. To conduct the tests, each candidate codec is required to process speech signals and/or encoded bit streams that have been preprocessed to simulate a range of operating conditions: variations in speaker voice and level, acoustic background noise type and level, channel error rate, and stages of **tandem coding**. During the tests, subjects listen to processed speech signals and judge their quality levels or annoyance levels on a five-point opinion scale. The opinion scores collected from the tests are suitably averaged over all trials and subjects for each test condition (see [11], for mean opinion score (MOS) and degradation mean opinion score). The categorical opinion scales of the subjects are also calibrated using *modulated noise reference units* (*MNRUs*) [3]. Modulated noise better resembles the distortions created by speech codecs than noise that is uncorrelated with the speech signal. Modulated noise is generated by multiplying the speech signal with a noise signal. The resultant modulated noise is scaled to a desired power level and then added to the uncoded (clean) speech signal. The ratio between the power level of the speech signal and that of the modulated noise is expressed in decibels

and given the notation *dBQ*. Under each test condition, subjects are presented with speech signals processed by the codecs as well as speech signals corrupted by modulated noise. Through presenting a range of modulated-noise levels, the subjects' opinions are calibrated on the dBQ scale. Thereafter, the mean opinion scores obtained for the codecs can also be expressed on that scale.

For each codec candidate, a profile of scores is compiled, consisting of speech quality scores, delay measurements, and complexity estimates. Each candidate's score profile is compared with that of the reference codec, ensuring that basic requirements are satisfied (see, e.g., [12]). An overall figure of merit for each candidate is also computed from the profile. The candidates, if any, that meet the basic requirements then compete on the basis of maximizing the figure of merit.

Basic performance requirements for each of the three half-rate standards are summarized in Table 30.2. In terms of speech quality, the GSM and PDC half-rate codecs

are permitted to underperform their respective full-rate codecs by no more than 1 dBQ averaging over all test conditions and no more than 3 dBQ within each test condition. More stringently, the North American half-rate codec is required to furnish a speech-quality profile that is statistically equivalent to that of the North American full-rate codec as determined by a specific statistical procedure for multiple comparisons [16]. Since various requirements on the half-rate standards are set relative to their full-rate counterparts, an indication of the *relative* speech quality between the three half-rate standards can be deduced from the test results of De Martino [2] comparing the three full-rate codecs. The maximum delays in Table 30.2 apply to the total of the delays through the speech and channel encoders and decoders (Fig. 30.1). Codec complexity is computed using a formula that counts the computational operations and memory usage of the codec algorithm. The complexity of the half-rate codecs is limited to 3 or 4 times that of their full-rate counterparts.

TABLE 30.2 Basic Performance Requirements for the Three Half-Rate Standards

	Basic performance requirements		
Digital Cellular Systems	Min. Speech Quality, dBQ Rel. to Full Rate	Max. Delay, ms	Max. Complexity Rel. to Full Rate
Japanese (PDC)	−1 average, −3 maximum	94.8	3×
European (GSM)	−1 average, −3 maximum	90	4×
North American (IS-54)	Statistically equivalent	100	4×

30.4 Speech Coding Techniques in the Half-Rate Standards

Existing half-rate and full-rate standard coders can be characterized as *linear-prediction based analysis-by-synthesis* (LPAS) speech coders [4]. LPAS coding entails using a time-varying all-pole filter in the decoder to synthesize the quantized speech signal. A short segment of the signal is synthesized by driving the filter with an *excitation* signal that is either *quasiperiodic* (for *voiced* speech) or *random* (for *unvoiced* speech). In either case, the excitation signal has a *spectral envelope* that is relatively flat. The synthesis filter serves to shape the spectrum of the excitation input so that the spectral envelope of the synthesized output resembles the filter's magnitude frequency response. The magnitude response often has prominent peaks; they render the *formants* that give a speech signal its phonetic character. The synthesis filter has to be adapted to the current frame of input speech signal. This is accomplished with the encoder performing a linear prediction (LP) analysis of the frame: the inverse of the all-pole synthesis filter is applied as an LP *error filter* to the frame, and the values of the filter parameters are

computed to minimize the energy of the filter's output error signal. The resultant filter parameters are quantized and conveyed to the decoder for it to update the synthesis filter.

Having executed an LP analysis and quantized the synthesis filter parameters, the LPAS encoder performs analysis-by-synthesis (ABS) on the input signal to find a suitable excitation signal. An ABS encoder maintains a *copy* of the decoder. The encoder examines the possible outputs that can be produced by the decoder copy in order to determine how best to instruct (using transmitted information) the actual decoder so that it would output (synthesize) a good approximation of the input speech signal. The decoder copy tracks the state of the actual decoder, since the latter evolves (under ideal channel conditions) according to information received from the encoder. The details of the ABS procedure vary with the particular excitation model employed in a specific coding scheme. One of the earliest seminal LPAS schemes is *code excited linear prediction* (*CELP*) [4]. In CELP, the excitation signal is obtained from a **codebook** of *code vectors*, each of which is a candidate for the excitation signal. The encoder searches the codebook to find the one code vector that would result in a best match between the resultant synthesis output signal and the encoder's input speech signal. The matching is considered best when the energy of the difference between the two signals being matched is minimized. A *perceptual weighting filter* is usually applied to the difference signal (prior to energy integration) to make the minimization more relevant to human perception of speech fidelity. Regions in the frequency spectrum where human listeners are more sensitive to distortions are given relatively stronger weighting by the filter and vice versa. For instance, the concentration of spectral energy around the formant frequencies gives rise to stronger *masking* of coder noise (i.e., rendering the noise less audible) and, therefore, weaker weighting can be applied to the formant frequency regions. For masking to be effective, the weighting filter has to be adapted to the time-varying speech spectrum. Adaptation is achieved usually by basing the weighting filter parameters on the synthesis filter parameters.

The CELP framework has evolved to form the basis of a great variety of speech coding algorithms, including all existing full- and half-rate standard algorithms for digital cellular systems. We outline next the basic CELP encoder-processing steps, in a form suited to our subsequent detailed descriptions of the PDC and GSM half-rate coders. These steps have accounted for various computational efficiency considerations and may, therefore, deviate from a conceptual functional description of the encoder constituents.

1. LP analysis on the current frame of input speech to determine the coefficients of the all-pole synthesis filter;
2. quantization of the LP filter parameters;
3. determination of the open-loop **pitch period** or lag;
4. adapting the perceptual weighting filter to the current LP information (and also pitch information when appropriate) and applying the adapted filter to the input speech signal;
5. formation of a filter cascade (which we shall refer to as *perceptually weighted synthesis filter*) consisting of the LP synthesis filter, as specified by the quantized parameters in step 2, followed by the perceptual weighting filter;
6. subtraction of the *zero-input response* of the perceptually weighted synthesis filter (the filter's decaying response due to past input) from the perceptually weighted input speech signal obtained in step 4;
7. an *adaptive codebook* is searched to find the most suitable periodic excitation, i.e., when the perceptually weighted synthesis filter is driven by the best code vector from the adaptive codebook, the output of the filter cascade should best match the difference signal obtained in step 6;

8. one or more nonadaptive excitation codebooks are searched to find the most suitable random excitation vectors that, when added to the best periodic excitation as determined in step 7 and with the resultant sum signal driving the filter cascade, would result in an output signal best matching the difference signal obtained in step 6.

Steps 1–6 are executed once per frame. Steps 7 and 8 are executed once for each of the *subframes* that together constitute a frame. Step 7 may be skipped depending on the pitch information from step 3, or if step 7 were always executed, a *nonperiodic excitation* decision would be one of the possible outcomes of the search process in step 7. Integral to steps 7 and 8 is the determination of gain (scaling) parameters for the excitation vectors. For each frame of input speech, the filter and excitation and gain parameters determined as outlined are conveyed as encoded bits to the speech decoder.

In a properly designed system, the data conveyed by the channel decoder to the speech decoder should be free of errors most of the time, and the speech signal synthesized by the speech decoder would be identical to that as determined in the speech encoder's ABS operation. It is common to enhance the quality of the synthesized speech by using an adaptive *postfilter* to attenuate coder noise in the perceptually sensitive regions of the spectrum. The postfilter of the decoder and the perceptual weighting filter of the encoder may seem to be functionally identical. The weighting filter, however, influences the selection of the best excitation among available choices, whereas the postfilter actually shapes the spectrum of the synthesized signal. Since postfiltering introduces its own distortion, its advantage may be diminished if tandem coding occurs along the end-to-end communication path. Nevertheless, proper design can ensure that the net effect of postfiltering is a reduction in the amount of audible codec noise [1]. Excepting postfiltering, all other speech synthesis operations of an LPAS decoder are (effectively) duplicated in the encoder (though the converse is not true). Using this fact, we shall illustrate each coder in the sequel by exhibiting only a block diagram of its encoder or decoder but not both.

30.5 Channel Coding Techniques in the Half-Rate Standards

Crucial to the maintenance of quality speech communication is the ability to transport coded speech data across the radio channel with minimal errors. Low-bit-rate LPAS coders are particularly sensitive to channel errors; errors in the bits representing the LP parameters in one frame, for instance, could result in the synthesis of nonsensical sounds for longer than a frame duration. The error rate of a digital cellular radio channel with no channel coding can be catastrophically high for LPAS coders. The amount of tolerable transmission delay is limited by the requirement of interactive communication and, consequently, *forward error control* is used to remedy transmission errors. "Forward" means that channel errors are remedied in the receiver, with no additional information from the transmitter and, hence, no additional transmission delay. To enable the channel decoder to correct channel errors, the channel encoder conveys more bits than the amount generated by the speech encoder. The additional bits are for error *protection*, as errors may or may not occur in any particular transmission epoch. The ratio of the number of encoder input (information) bits to the number of encoder output (code) bits is called the (channel) *coding rate*. This is a number no more than one and generally decreases as the error protection power increases. Though a lower channel coding rate gives more error protection, fewer bits will be available for speech coding. When the channel is in good condition and, hence, less error protection is needed, the received speech quality could be better if bits devoted to channel coding were used for speech coding. On the other hand, if a high channel coding rate were used, there would be uncorrected errors under poor channel conditions and speech quality would suffer.

Thus, when nonadaptive forward error protection is used over channels with nonstationary statistics, there is an inevitable tradeoff between quality degradation due to uncorrected errors and that due to expending bits on error protection (instead of on speech encoding).

Both the GSM and PDC half-rate coders use *convolutional coding* [14] for error correction. Convolutional codes are sliding or sequential codes. The encoder of a rate m/n, $m < n$ convolutional code can be realized using m shift registers. For every m information bits input to the encoder (one bit to each of the m shift registers), n code bits are output to the channel. Each code bit is computed as a modulo-2 sum of a subset of the bits in the shift registers. Error protection overhead can be reduced by exploiting the unequal sensitivity of speech quality to errors in different positions of the encoded bit stream. A family of *rate-compatible punctured convolutional codes* (RCPCCs) [10] is a collection of related convolutional codes; all of the codes in the collection except the one with the lowest rate are derived by *puncturing* (dropping) code bits from the convolutional code with the lowest rate. With an RCPCC, the channel coding rate can be varied on the fly (i.e., variable-rate coding) while a sequence of information bits is being encoded through the shift registers, thereby imparting on different segments in the sequence different degrees of error protection.

For decoding a convolutional coded bit stream, the *Viterbi algorithm* [14] is a computationally efficient procedure. Given the output of the demodulator, the algorithm determines the most likely sequence of data bits sent by the channel encoder. To fully utilize the error correction power of the convolutional code, the amplitude of the demodulated *channel symbol* can be quantized to more bits than the minimum number required, i.e., for subsequent *soft decision decoding*. The minimum number of bits is given by the number of channel-coded bits mapped by the modulator onto each channel symbol; decoding based on the minimum-rate bit stream is called *hard decision* decoding. Although soft decoding gives better error protection, decoding complexity is also increased.

Whereas convolutional codes are most effective against randomly scattered bit errors, errors on cellular radio channels often occur in bursts of bits. These bursts can be broken up if the bits put into the channel are rearranged after demodulation. Thus, in *block interleaving*, encoded bits are read into a matrix by row and then read out of the matrix by column (or vice versa) and then passed on to the modulator; the reverse operation is performed by a *deinterleaver* following demodulation. Interleaving increases the transmission delay to the extent that enough bits need to be collected in order to fill up the matrix.

Owing to the severe nature of the cellular radio channel and limited available transmission capacity, uncorrected errors often remain in the decoded data. A common countermeasure is to append an error detection code to the speech data stream prior to channel coding. When residual channel errors are detected, the speech decoder can take various remedial measures to minimize the negative impact on speech quality. Common measures are repetition of speech parameters from the most recent good frames and gradual muting of the possibly corrupted synthesized speech.

The PDC and GSM half-rate standard algorithms together embody some of the latest advances in speech coding techniques, including: *multimodal coding* where the coder configuration and bit allocation change with the type of speech input; *vector quantization* (*VQ*) [5] of the LP filter parameters; higher precision and improved coding efficiency for pitch-periodic excitation; and postfiltering with improved tandeming performance. We next explore the more distinctive features of the PDC and GSM speech coders.

30.6 The Japanese Half-Rate Standard

An algorithm was selected for the Japanese half-rate standard in April 1993, following the evaluation of 12 submissions in a first round, and four final candidates in a second round [12]. The selected

algorithm, called pitch synchronous innovation CELP[2] (PSI-CELP), met all of the basic selection criteria and scored the highest among all candidates evaluated. A block diagram of the PSI-CELP encoder is shown in Fig. 30.2, and bit allocations are summarized in Table 30.3. The complexity of the coder is estimated to be approximately 2.4 times that of the PDC full-rate coder. The frame size of the coder is 40 ms, and its subframe size is 10 ms. These sizes are longer than those used in most existing CELP-type standard coders. However, LP analysis is performed twice per frame in the PSI-CELP coder.

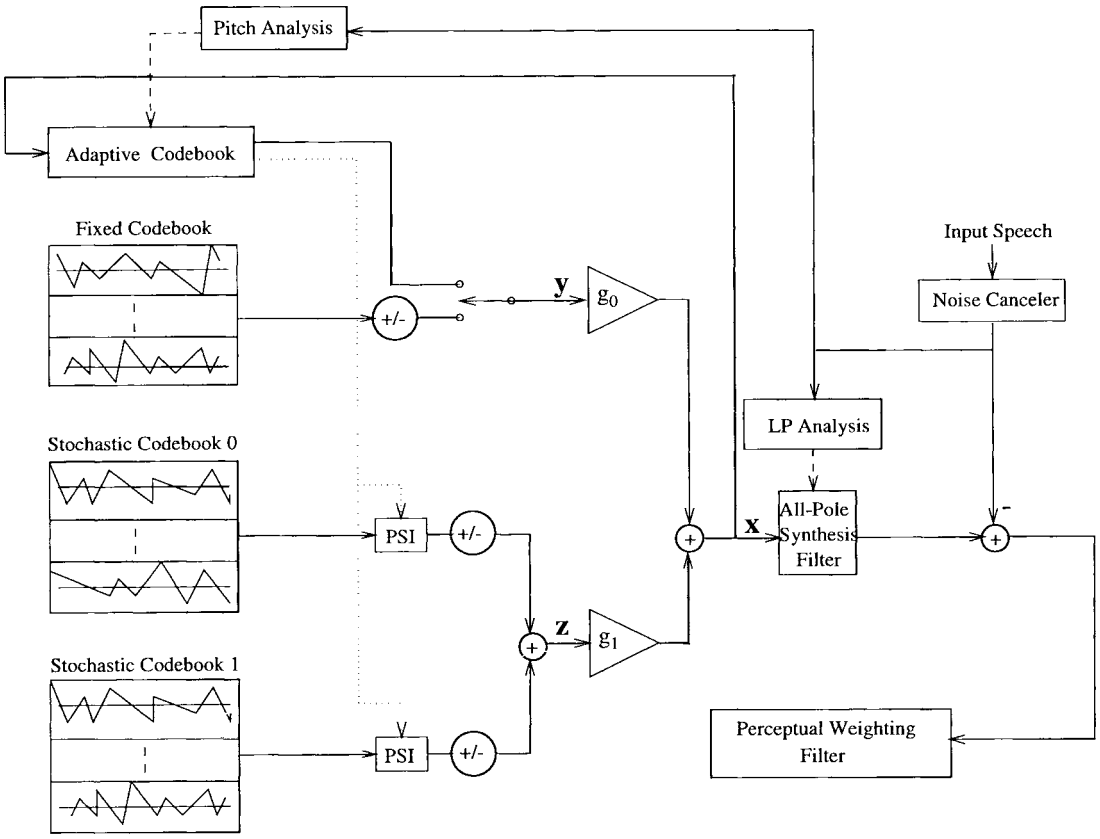

FIGURE 30.2: Basic structure of the PSI-CELP encoder.

A distinctive feature of the PSI-CELP coder is the use of an adaptive noise canceller [13, 15] to suppress noise in the input signal prior to coding. The input signal is classified into various modes, depending on the presence or absence of background noise and speech and their relative power levels. The current active mode determines whether *Kalman filtering* [9] is applied to the input signal

[2]There were two candidate algorithms named PSI-CELP in the PDC half-rate competition. The algorithm described here was contributed by NTT Mobile Communications Network, Inc. (NTT DoCoMo).

TABLE 30.3 Bit Allocations for the PSI-CELP Half- Rate PDC Speech Coder

Parameter	Bits	Error Protected Bits
LP synthesis filter	31	15
Frame energy	7	7
Periodic excitation	8×4	8×4
Stochastic excitation	10×4	0
Gain	7×4	3×4
Total	138	66

and whether the parameters of the Kalman filter are adapted. Kalman filtering is applied when a significant amount of background noise is present or when both background noise and speech are strongly present. The filter parameters are adapted to the statistics of the speech and noise signals in accordance with whether they are both present or only noise is present.

The LP filter parameters in the PSI-CELP coder are encoded using VQ. A tenth-order LP analysis is performed every 20 ms. The resultant filter parameters are converted to 10 *line spectral frequencies* (LSFs).[3] The LSF parameters have a naturally increasing order, and together are treated as the ordered components of a vector. Since the speech spectral envelope tends to evolve slowly with time, there is intervector dependency between adjacent LSF vectors that can be exploited. Thus, the two LSF vectors for each 40-ms frame are paired together and jointly encoded. Each LSF vector in the pair is split into three subvectors. The pair of subvectors that cover the same vector component indexes are combined into one composite vector and vector quantized. Altogether, 31 b are used to encode a pair of LSF vectors. This three-way *split VQ*[4] scheme embodies a compromise between the prohibitively high complexity of using a large vector dimension and the performance gain from exploiting intra- and intervector dependency.

The PSI-CELP encoder uses a perceptual weighting filter consisting of a cascade of two filter sections. The sections exploit the pitch-harmonic structure and the LP spectral-envelope structure of the speech signal, respectively. The pitch-harmonic section has four parameters, a pitch lag and three coefficients, whose values are determined from an analysis of the periodic structure of the input speech signal. Pitch-harmonic weighting reduces the amount of noise in between the pitch harmonics by aggregating coder noise to be closer to the harmonic frequencies of the speech signal. In high-pitched voice, the harmonics are spaced relatively farther apart, and pitch-harmonic weighting becomes correspondingly more important.

The excitation vector $\boldsymbol{x}$ (Fig. 30.2) is updated once every subframe interval (10 ms) and is constructed as a *linear combination* of two vectors

$$\boldsymbol{x} = g_0 \boldsymbol{y} + g_1 \boldsymbol{z} \tag{30.1}$$

where g_0 and g_1 are scalar gains, $\boldsymbol{y}$ is labeled as the *periodic* component of the excitation and $\boldsymbol{z}$ as the *stochastic* or *random* component. When the input speech is voiced, the ABS operation would find a value for $\boldsymbol{y}$ from the *adaptive codebook* (Fig. 30.2). The codebook is constructed out of past samples of the excitation signal $\boldsymbol{x}$; hence, there is a feedback path into the adaptive codebook in Fig. 30.2. Each code vector in the adaptive codebook corresponds to one of the 192 possible pitch lag L values available for encoding; the code vector is populated with samples of $\boldsymbol{x}$ beginning with the Lth sample backward in time. L is not restricted to be an integer, i.e., *fractional pitch period* is

[3] Also known as line spectrum pairs (LSPs).

[4] Matrix quantization is another possible description.

permitted. Successive values of L are more closely spaced for smaller values of L; short, medium, and long lags are quantized to one-quarter, one-half, and one sampling-period resolution, respectively. As a result, the *relative* quantization error in the encoded pitch frequency (which is the reciprocal of the encoded pitch lag) remains roughly constant with increasing pitch frequency. When the input speech is unvoiced, $\boldsymbol{y}$ would be obtained from the fixed codebook (Fig. 30.2). To find the best value for $\boldsymbol{y}$, the encoder searches through the aggregate of 256 code vectors from both the adaptive and fixed codebooks. The code vector that results in a synthesis output most resembling the input speech is selected. The best code vector thus chosen also implicitly determines the voicing condition (voiced/unvoiced) and the pitch lag value L^* most appropriate to the current subframe of input speech. These parameters are said to be determined in a *closed-loop* search.

The stochastic excitation $\boldsymbol{z}$ is formed as a sum of two code vectors, each selected from a *conjugate codebook* (Fig. 30.2) [13]. Using a pair of conjugate codebooks each of size 16 code vectors (4 b) has been found to improve robustness against channel errors, in comparison with using one single codebook of size 256 code vectors (8 b). The synthesis output due to $\boldsymbol{z}$ can be decomposed into a sum of two orthogonal components, one of which points in the same direction as the synthesis output due to the periodic excitation $\boldsymbol{y}$ and the other component points in a direction orthogonal to the synthesis output due to $\boldsymbol{y}$. The latter synthesis output component of $\boldsymbol{z}$ is kept, whereas the former component is discarded. Such decomposition enables the two gain factors g_0 and g_1 to be separately quantized. For voiced speech, the conjugate code vectors are preprocessed to produce a set of *pitch synchronous innovation* (PSI) vectors. The first L^* samples of each code vector are treated as a fundamental period of samples. The fundamental period is replicated until there are enough samples to populate a subframe. If L^* is not an integer, interpolated samples of the code vectors are used (upsampled versions of the code vectors can be precomputed). PSI has been found to reinforce the periodicity and substantially improve the quality of synthesized voiced speech.

The postfilter in the PSI-CELP decoder has three sections, for enhancing the formants, the pitch harmonics, and the high frequencies of the synthesized speech, respectively. Pitch-harmonic enhancement is applied only when the adaptive codebook has been used. Formant enhancement makes use of the decoded LP synthesis filter parameters, whereas a refined pitch analysis is performed on the synthesized speech to obtain the values for the parameters of the pitch-harmonic section of the postfilter. A first-order high-pass filter section compensates for the low-pass spectral tilt [1] of the formant enhancement section.

Of the 138 speech data bits generated by the speech encoder every 40-ms frame, 66 b (Table 30.3) receive error protection and the remaining 72 speech data bits of the frame are not error protected. An error detection code of 9 *cyclic redundancy check* (*CRC*) bits is appended to the 66 b and then submitted to a rate 1/2, punctured convolutional encoder to generate a sequence of 152 channel coded bits. Of the unprotected 72 b, the 40 b that index the excitation codebooks (Table 30.3) are remapped or *pseudo-Gray coded* [17] so as to equalize their channel error sensitivity. As a result, a bit error occurring in an index word is likely to cause about the same amount of degradation regardless of the bit error position in the index word. For each speech frame, the channel encoder emits 224 b of payload data. The payload data from two adjacent frames are interleaved before transmission over the radio link.

Uncorrected errors in the most critical 66 b are detected with high probability as a CRC error. A finite state machine keeps track of the recent history of CRC errors. When a sequence of CRC errors is encountered, the power level of the synthesized speech is progressively suppressed, so that muting is reached after four consecutive CRC errors. Conversely, following the cessation of a sequence of CRC errors, the power level of the synthesized speech is ramped up gradually.

30.7 The European GSM Half-Rate Standard

A *vector sum excited linear prediction* (VSELP) coder, contributed by Motorola, Inc., was selected in January 1994 by the main GSM technical committee as a basis for the GSM half-rate standard. The standard was finally approved in January 1995. VSELP is a generic name for a family of algorithms from Motorola; the North American full-rate and the Japanese full-rate standards are also based on VSELP. All VSELP coders make use of the basic idea of representing the excitation signal by a linear combination of *basis vectors* [6]. This representation renders the excitation codebook search procedure very computationally efficient. A block diagram of the GSM half-rate decoder is depicted in Fig. 30.3 and bit allocations are tabulated in Table 30.4. The coder's frame size is 20 ms, and each frame comprises four subframes of 5 ms each. The coder has been optimized for execution on a processor with 16-b word length and 32-b accumulator. The GSM standard is a *bit exact* specification: in addition to specifying the codec's processing steps, the numerical formats and precisions of the codec's variables are also specified.

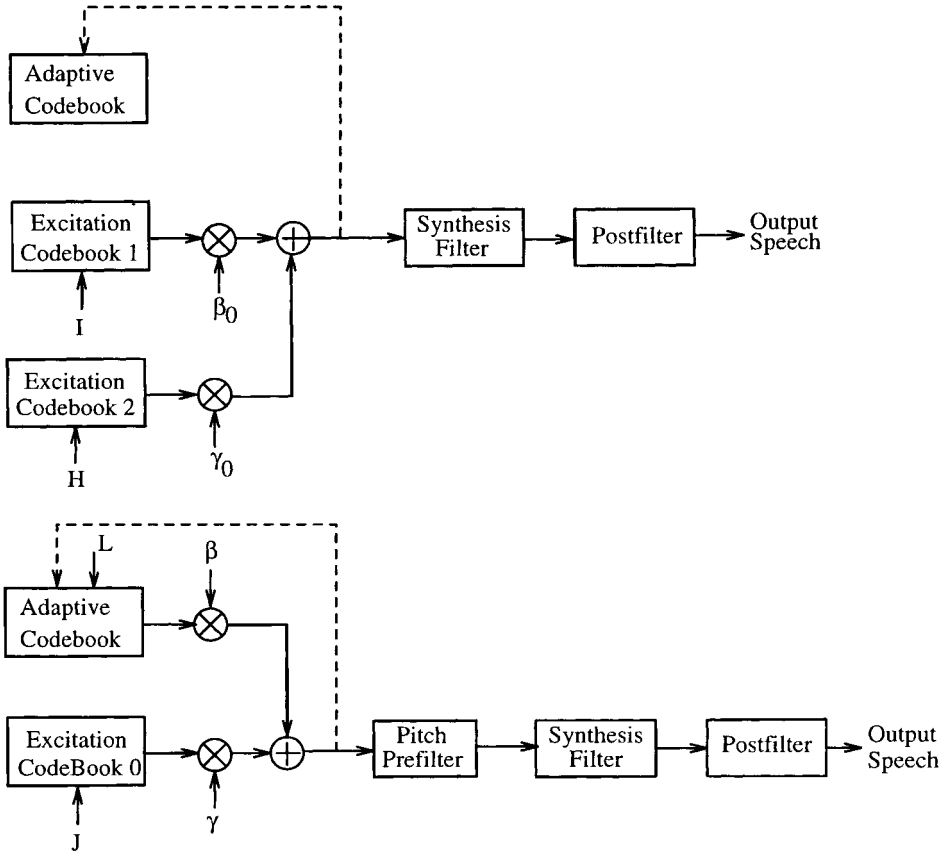

FIGURE 30.3: Basic structure of the GSM VSELP decoder. Top is for mode 0 and bottom is for modes 1, 2, and 3.

TABLE 30.4 Bit Allocations for the VSELP Half-Rate GSM Coder

Parameter	Bits/subframe	Bits/frame
LP synthesis filter		28
Soft interpolation		1
Frame energy		5
Mode selection		2
Mode 0		
Excitation code I	7	28
Excitation code H	7	28
Gain code G_s, P_0	5	20
Mode 1, 2, and 3		
Pitch lag L (first subframe)		8
Difference lag (subframes 2, 3, 4)	4	12
Excitation code J	9	36
Gain code G_s, P_0	5	20
Total		112

The synthesis filter coefficients in GSM VSELP are encoded using the *fixed point lattice technique* (FLAT) [8] and vector quantization. FLAT is based on the *lattice filter* representation of the linear prediction error filter. The tenth-order lattice filter has 10 stages, with the ith stage, $i \in \{1, \ldots, 10\}$, containing a *reflection coefficient* parameter r_i. The lattice filter has an *order-recursion* property such that the best prediction error filters of all orders less than ten are all embedded in the best tenth-order lattice filter. This means that once the values of the lower order reflection coefficients have been optimized, they do not have to be reoptimized when a higher order predictor is desired; in other words, the coefficients can be optimized sequentially from low to high orders. On the other hand, if the lower order coefficients were suboptimal (as in the case when the coefficients are quantized), the higher order coefficients could still be selected to minimize the prediction *residual* (or error) energy at the output of the higher order stages; in effect, the higher order stages can compensate for the suboptimality of lower order stages.

In the GSM VSELP coder, the ten reflection coefficients $\{r_1, \ldots, r_{10}\}$ that have to be encoded for each frame are grouped into three coefficient vectors $\boldsymbol{v}_1 = [r_1 r_2 r_3]$, $\boldsymbol{v}_2 = [r_4 r_5 r_6]$, $\boldsymbol{v}_3 = [r_7 r_8 r_9 r_{10}]$. The vectors are quantized sequentially, from $\boldsymbol{v}_1$ to $\boldsymbol{v}_3$, using a b_i-bit VQ codebook C_i for $\boldsymbol{v}_i$, where b_i, $i = 1, 2, 3$ are 11, 9, and 8 b, respectively. The vector $\boldsymbol{v}_i$ is quantized to minimize the prediction error at the energy output of the jth stage of the lattice filter where r_j is the highest order coefficient in the vector $\boldsymbol{v}_i$. The computational complexity associated with quantizing $\boldsymbol{v}_i$ is reduced by searching only a small subset of the code vectors in C_i. The subset is determined by first searching a *prequantizer* codebook of size c_i bits, where c_i, $i = 1, \ldots, 3$ are 6, 5, and 4 b, respectively. Each code vector in the prequantizer codebook is associated with $2^{b_i - c_i}$ code vectors in the target codebook. The subset is obtained by pooling together all of the code vectors in C_i that are associated with the top few best matching prequantizer code vectors. In this way, a factor of reduction in computational complexity of nearly $2^{b_i - c_i}$ is obtained for the quantization of $\boldsymbol{v}_i$.

The half-rate GSM coder changes its configuration of excitation generation (Fig. 30.3) in accordance with a *voicing mode* [7]. For each frame, the coder selects one of four possible voicing modes depending on the values of the *open-loop* pitch-prediction gains computed for the frame and its four subframes. Open loop refers to determining the pitch lag and the pitch-predictor coefficient(s) via a direct analysis of the input speech signal or, in the case of the half-rate GSM coder, the perceptually weighted (LP-weighting only) input signal. Open-loop analysis can be regarded as the opposite of closed-loop analysis, which in our context is synonymous with ABS. When the pitch-prediction gain for the frame is weak, the input speech signal is deemed to be unvoiced and mode 0 is used. In this mode, two 7-b *trained* codebooks (excitation codebooks 1 and 2 in Fig. 30.3) are used, and the excitation signal for each subframe is formed as a linear combination of two code vectors, one from

each of the codebooks. A trained codebook is one designed by applying the coder to a representative set of speech signals while optimizing the codebook to suit the set. Mode 1, 2, or 3 is chosen depending on the strength of the pitch-prediction gains for the frame and its subframes. In these modes, the excitation signal is formed as a linear combination of a code vector from an 8-b adaptive codebook and a code vector from a 9-b trained codebook (Fig. 30.3). The code vectors that are summed together to form the excitation signal for a subframe are each scaled by a gain factor (β and γ in Fig. 30.3). Each mode uses a gain VQ codebook specific to that mode.

As depicted in Fig. 30.3, the decoder contains an adaptive pitch prefilter for the voiced modes and an adaptive postfilter for all modes. The filters enhance the perceptual quality of the decoded speech and are not present in the encoder. It is more conventional to locate the pitch prefilter as a section of the postfilter; the distinctive placement of the pitch prefilter in VSELP was chosen to reduce artifacts caused by the time-varying nature of the filter. In mode 0, the encoder uses an LP spectral weighting filter in its ABS search of the two excitation codebooks. In the other modes, the encoder uses a pitch-harmonic weighting filter in cascade with an LP spectral weighting filter for searching excitation codebook 0, whereas only LP spectral weighting is used for searching the adaptive codebook. The pitch-harmonic weighting filter has two parameters, a pitch lag and a coefficient, whose values are determined in the aforementioned open-loop pitch analysis.

A code vector in the 8-b adaptive codebook has a dimension of 40 (the duration of a subframe) and is populated with past samples of the excitation signal beginning with the Lth sample back from the present time. L can take on one of 256 different integer and fractional values. The best adaptive code vector for each subframe can be selected via a complete ABS; the required exhaustive search of the adaptive codebook is, however, computationally expensive. To reduce computation, the GSM VSELP coder makes use of the aforementioned open-loop pitch analysis to produce a list of *candidate lag values*. The open-loop pitch-prediction gains are ranked in decreasing order, and only the lags corresponding to top-ranked gains are kept as candidates. The final decisions for the four L values of the four subframes in a frame are made jointly. By assuming that the four L values can not vary over the entire range of all possible 256 values in the short duration of a frame, the L of the first subframe is coded using 8 b, and the L of each of the other three subframes is coded *differentially* using 4 b. The 4 b represent 16 possible values of deviation relative to the lag of the previous subframe. The four lags in a frame trace out a *trajectory* where the change from one time point to the next is restricted; consequently, only 20 b are needed instead of 32 b for encoding the four lags. Candidate trajectories are constructed by linking top ranked lags that are commensurate with differential encoding. The best trajectory among the candidates is then selected via ABS.

The trained excitation codebooks of VSELP have a special vector sum structure that facilitates fast searching [6]. Each of the 2^b code vectors in a b-bit trained codebook is formed as a linear combination of b *basis vectors*. Each of the b scalar weights in the linear combination is restricted to have a binary value of either 1 or -1. The 2^b code vectors in the codebook are obtained by taking all 2^b possible combinations of values of the weights. A substantial storage saving is incurred by storing only b basis vectors instead of 2^b code vectors. Computational saving is another advantage of the vector-sum structure. Since filtering is a linear operation, the synthesis output due to each code vector is a linear combination of the synthesis outputs due to the individual basis vectors, where the same weight values are used in the output linear combination as in forming the code vector. A vector sum codebook can be searched by first performing synthesis filtering on its b basis vectors. If, for the present subframe, another trained codebook (mode 0) or an adaptive codebook (mode 1, 2, 3) had been searched, the filtered basis vectors are further orthogonalized with respect to the signal synthesized from that codebook, i.e., each filtered basis vector is replaced by its own component that is orthogonal to the synthesized signal. Further complexity reduction is obtained by examining the code vectors in a sequence such that two successive code vectors differ in only one of the b scalar

weight values; that is, the entire set of 2^b code vectors is searched in a *Gray coded* sequence. With successive code vectors differing in only one term in the linear combination, it is only necessary in the codebook search computation to progressively track the difference [6].

The total energy of a speech frame is encoded with 5 b (Table 30.4). The two gain factors (β and γ in Fig. 30.3) for each subframe are computed after the excitation codebooks have been searched and are then transformed to parameters G_s and P_0 to be vector quantized. Each mode has its own 5-b gain VQ codebook. G_s represents the energy of the subframe relative to the total frame energy, and P_0 represents the fraction of the subframe energy due to the first excitation source (excitation codebook 1 in mode 0, or the adaptive codebook in the other modes).

An *interpolation bit* (Table 30.4) transmitted for each frame specifies to the decoder whether the LP synthesis filter parameters for each subframe should be obtained from interpolating between the decoded filter parameters for the current and the previous frames. The encoder determines the value of this bit according to whether interpolation or no interpolation results in a lower prediction residual energy for the frame. The postfilter in the decoder operates in concordance with the actual LP parameters used for synthesis.

The speech encoder generates 112 b of encoded data (Table 30.4) for every 20-ms frame of the speech signal. These bits are processed by the channel encoder to improve, after channel decoding at the receiver, the uncoded bit-error rate and the detectability of uncorrected errors. Error detection coding in the form of 3 CRC bits is applied to the most critical 22 data bits. The combined 25 b plus an additional 73 speech data bits and 6 *tail bits* are input to an RCPCC encoder (the tail bits serve to bring the channel encoder and decoder to a fixed terminal state at the end of the payload data stream). The 3 CRC bits are encoded at rate 1/3 and the other 101 b are encoded at rate 1/2, generating a total of 211 channel coded bits. These are finally combined with the remaining 17 (uncoded) speech data bits to form a total of 228 b for the payload data of a speech frame. The payload data from two speech frames are interleaved for transmission over four timeslots of the GSM TDMA channel.

With the Viterbi algorithm, the channel decoder performs soft decision decoding on the demodulated and deinterleaved channel data. Uncorrected channel errors may still be present in the decoded speech data after Viterbi decoding. Thus, the channel decoder classifies each frame into three integrity categories: bad, unreliable, and reliable, in order to assist the speech decoder in undertaking error concealment measures. A frame is considered bad if the CRC check fails or if the received channel data is close to more than one candidate sequence. The latter evaluation is based on applying an adaptive threshold to the metric values produced by the Viterbi algorithm over the course of decoding the most critical 22 speech data bits and their 3 CRC bits. Frames that are not bad may be classified as unreliable, depending on the metric values produced by the Viterbi algorithm and on channel reliability information supplied by the demodulator.

Depending on the recent history of decoded data integrity, the speech decoder can take various error concealment measures. The onset of bad frames is concealed by repetition of parameters from previous reliable frames, whereas the persistence of bad frames results in power attenuation and ultimately muting of the synthesized speech. Unreliable frames are decoded with normality constraints applied to the energy of the synthesized speech.

30.8 Conclusions

The half-rate standards employ some of the latest techniques in speech and channel coding to meet the challenges posed by the severe transmission environment of digital cellular radio systems. By halving the bit rate, the voice transmission capacity of existing full-rate digital cellular systems can be doubled. Although advances are still being made that can address the needs of quarter-rate speech transmission,

much effort is currently devoted to enhancing the speech quality and robustness of full-rate (GSM and IS-54) systems, aiming to be closer to *toll quality*. On the other hand, the imminent introduction of competing wireless systems that use different modulation schemes [e.g., coded division multiple access (CDMA)] and/or different radio frequencies [e.g., personal communications systems (PCS)] is poised to alleviate congestion in high-user-density areas.

Defining Terms

Codebook: An ordered collection of all possible values that can be assigned to a scalar or vector variable. Each unique scalar or vector value in a codebook is called a *codeword*, or *code vector* where appropriate.

Codec: A contraction of *(en)coder–decoder*, used synonymously with the word *coder*. The encoder and decoder are often designed and deployed as a pair. A half-rate standard codec performs speech as well as channel coding.

Echo canceller: A signal processing device that, given the source signal causing the echo signal, generates an estimate of the echo signal and subtracts the estimate from the signal being interfered with by the echo signal. The device is usually based on a discrete-time adaptive filter.

Pitch period: The fundamental period of a voiced speech waveform that can be regarded as periodic over a short-time interval (quasiperiodic). The reciprocal of pitch period is *pitch frequency* or simply, *pitch*.

Tandem coding: Having more than one encoder–decoder pair in an end-to-end transmission path. In cellular radio communications, having a radio link at each end of the communication path could subject the speech signal to two passes of speech encoding–decoding. In general, repeated encoding and decoding increases the distortion.

Acknowledgment

The authors would like to thank Erdal Paksoy and Mark A. Jasiuk for their valuable comments.

References

[1] Chen, J.-H. and Gersho, A., Adaptive postfiltering for quality enhancement of coded speech. *IEEE Trans. Speech & Audio Proc.*, 3(1), 59–71, 1995.

[2] De Martino, E., Speech quality evaluation of the European, North-American and Japanese speech codec standards for digital cellular systems. In *Speech and Audio Coding for Wireless and Network Applications*, Atal, B.S., Cuperman, V., and Gersho, A., Eds., 55–58, Kluwer Academic Publishers, Norwell, MA, 1993.

[3] Dimolitsas, S., Corcoran, F.L., and Baraniecki, M.R., Transmission quality of North American cellular, personal communications, and public switched telephone networks. *IEEE Trans. Veh. Tech.*, 43(2), 245–251, 1994.

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

[5] Gersho, A. and Gray, R.M., *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, Norwell, MA, 1991.

[6] Gerson, I.A. and Jasiuk, M.A., Vector sum excited linear prediction (VSELP) speech coding at 8 kbps. In *Proceedings, IEEE Intl. Conf. Acoustics, Speech, & Sig. Proc.*, 461–464, April, 1990.

[7] Gerson, I.A. and Jasiuk, M.A., Techniques for improving the performance of CELP—type speech coders. *IEEE J. Sel. Areas Comm.*, 10(5), 858–865, 1992.
[8] Gerson, I.A., Jasiuk, M.A., Nowack, J.M., Winter, E.H., and Müller, J.-M., Speech and channel coding for the half-rate GSM channel. In *Proceedings, ITG-Report 130 on Source and Channel Coding*, 225–232. Munich, Germany, Oct., 1994.
[9] Gibson, J.D., Koo, B., and Gray, S.D., Filtering of colored noise for speech enhancement and coding. *IEEE Trans. Sig. Proc.*, 39(8), 1732–1742, 1991.
[10] Hagenauer, J., Rate-compatible punctured convolutional codes (RCPC codes) and their applications. *IEEE Trans. Comm.*, 36(4), 389–400, 1988.
[11] Jayant, N.S. and Noll, P., *Digital Coding of Waveforms*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
[12] Masui, F. and Oguchi, M.,. Activity of the half rate speech codec algorithm selection for the personal digital cellular system. *Tech. Rept. of IEICE*, RCS93-77(11), 55–62 (in Japanese), 1993.
[13] Ohya, T., Suda, H., and Miki, T., 5.6 kbits/s PSI-CELP of the half-rate PDC speech coding standard. In *Proceedings, IEEE Veh. Tech. Conf.*, 1680–1684, June, 1994.
[14] Proakis, J.G., *Digital Communications*, 3rd ed., McGraw-Hill, New York, 1995.
[15] Suda, H., Ikeda, K., and Ikedo, J., Error protection and speech enhancement schemes of PSI-CELP, *NTT R & D.* (Special issue on PSI-CELP speech coding system for mobile communications), 43(4), 373–380, (in Japanese), 1994.
[16] Telecommunication Industries Association (TIA). Half-rate speech codec test plan V6.0. TR45.3.5/93.05.19.01, 1993.
[17] Zeger, K. and Gersho, A., Pseudo-Gray coding. *IEEE Trans. Comm.*, 38(12), 2147–2158, 1990.

Further Information

Additional technical information on speech coding can be found in the books, periodicals, and conference proceedings that appear in the list of references. Other relevant publications not represented in the list are *Speech Communication*, Elsevier Science Publishers; *Advances in Speech Coding*, B. S. Atal, V. Cuperman, and A, Gersho, eds., Kluwer Academic Publishers; and *Proceedings of the IEEE Workshop on Speech Coding.*