SPEECH CODING: FUNDAMENTALS AND APPLICATIONS

Mark Hasegawa-Johnson
University of Illinois at
Urbana-Champaign
Urbana, IL

Abeer Alwan
University of California at Los
Angeles
Los Angles, CA

1. INTRODUCTION

Speech coding is the process of obtaining a compact representation of voice signals for efficient transmission over band-limited wired and wireless channels and/or storage. Today, speech coders have become essential components in telecommunications and in the multimedia infrastructure. Commercial systems that rely on efficient speech coding include cellular communication, voiceover internet protocol (VOIP), videoconferencing, electronic toys, archiving, and digital simultaneous voice and data (DSVD), as well as numerous PC-based games and multimedia applications.

Speech coding is the art of creating a minimally redundant representation of the speech signal that can be efficiently transmitted or stored in digital media, and decoding the signal with the best possible perceptual quality. Like any other continuous-time signal, speech may be represented digitally through the processes of sampling and quantization; speech is typically quantized using either 16-bit uniform or 8-bit companded quantization. Like many other signals, however, a sampled speech signal contains a great deal of information that is either redundant (nonzero mutual information between successive samples in the signal) or perceptually irrelevant (information that is not perceived by human listeners). Most telecommunications coders are *lossy*, meaning that the synthesized speech is perceptually similar to the original but may be physically dissimilar.

A speech coder converts a digitized speech signal into a coded representation, which is usually transmitted in frames. A speech decoder receives coded frames and synthesizes reconstructed speech. Standards typically dictate the input–output relationships of both coder and decoder. The input–output relationship is specified using a reference implementation, but novel implementations are allowed, provided that input–output equivalence is maintained. Speech coders differ primarily in bit rate (measured in bits per sample or bits per second), complexity (measured in operations per second), delay (measured in milliseconds between recording and playback), and perceptual quality of the synthesized speech. *Narrowband* (NB) coding refers to coding of speech signals whose bandwidth is less than 4 kHz (8 kHz sampling rate), while *wideband* (WB) coding refers to coding of 7-kHz-bandwidth signals (14–16 kHz sampling rate). NB coding is more common than WB coding mainly because of the narrowband nature of the wireline telephone channel (300–3600 Hz). More recently, however, there has been an increased effort in wideband speech coding because of several applications such as videoconferencing.

There are different types of speech coders. Table 1 summarizes the bit rates, algorithmic complexity, and standardized applications of the four general classes of coders described in this article; Table 2 lists a selection of specific speech coding standards. Waveform coders attempt to code the exact shape of the speech signal waveform, without considering the nature of human speech production and speech perception. These coders are high-bit-rate coders (typically above 16 kbps). Linear prediction coders (LPCs) assume that the speech signal is the output of a linear time-invariant (LTI) model of speech production. The transfer function of that model is assumed to be all-pole (autoregressive model). The excitation function is a quasiperiodic signal constructed from discrete pulses (1–8 per pitch period), pseudorandom noise, or some combination of the two. If the excitation is generated only at the receiver, based on a transmitted pitch period and voicing information, then the system is designated as an LPC vocoder. LPC vocoders that provide extra information about the spectral shape of the excitation have been adopted as coder standards between 2.0 and 4.8 kbps. LPC-based analysis-by-synthesis coders (LPC-AS), on the other hand, choose an excitation function by explicitly testing a large set of candidate excitations and choosing the best. LPC-AS coders are used in most standards between 4.8 and 16 kbps. Subband coders are frequency-domain coders that attempt to parameterize the speech signal in terms of spectral properties in different frequency bands. These coders are less widely used than LPC-based coders but have the advantage of being scalable and do not model the incoming signal as speech. Subband coders are widely used for high-quality audio coding.

This article is organized as follows. Sections 2, 3, 4 and 5 present the basic principles behind waveform coders, subband coders, LPC-based analysis-by-synthesis coders, and LPC-based vocoders, respectively. Section 6 describes the different quality metrics that are used to evaluate speech coders, while Section 7 discusses a variety of issues that arise when a coder is implemented in a communications network, including voiceover IP, multirate coding, and channel coding. Section 8 presents an overview of standardization activities involving speech coding, and we conclude in Section 9 with some final remarks.

2. WAVEFORM CODING

Waveform coders attempt to code the exact shape of the speech signal waveform, without considering in detail the nature of human speech production and speech perception. Waveform coders are most useful in applications that require the successful coding of both speech and nonspeech signals. In the public switched telephone network (PSTN), for example, successful transmission of modem and fax signaling tones, and switching signals is nearly as important as the successful transmission of speech. The most commonly used waveform coding algorithms are uniform 16-bit PCM, companded 8-bit PCM [48], and ADPCM [46].

Table 1. Characteristics of Standardized Narrowband Speech Coding Algorithms in Each of Four Broad Categories

Speech Coder Class	Rates (kbps)	Complexity	Standardized Applications	Section
Waveform coders	16–64	Low	Landline telephone	2
Subband coders	12–256	Medium	Teleconferencing, audio	3
LPC-AS	4.8–16	High	Digital cellular	4
LPC vocoder	2.0–4.8	High	Satellite telephony, military	5

2.1. Pulse Code Modulation (PCM)

Pulse code modulation (PCM) is the name given to memoryless coding algorithms that quantize each sample of $s(n)$ using the same reconstruction levels $\hat{s}_k$, $k = 0, \ldots, m, \ldots, K$, regardless of the values of previous samples. The reconstructed signal $\hat{s}(n)$ is given by

$$\hat{s}(n) = \hat{s}_m \quad \text{for} \quad s(n) \text{ s.t. } (s(n) - \hat{s}_m)^2 = \min_{k=0,\ldots,K} (s(n) - \hat{s}_k)^2 \tag{1}$$

Many speech and audio applications use an odd number of reconstruction levels, so that background noise signals with a very low level can be quantized exactly to $\hat{s}_{K/2} = 0$. One important exception is the A-law companded PCM standard [48], which uses an even number of reconstruction levels.

2.1.1. Uniform PCM.

Uniform PCM is the name given to quantization algorithms in which the reconstruction levels are uniformly distributed between $S_{\max}$ and $S_{\min}$. The advantage of uniform PCM is that the quantization error power is independent of signal power; high-power signals are quantized with the same resolution as low-power signals. Invariant error power is considered desirable in many digital audio applications, so 16-bit uniform PCM is a standard coding scheme in digital audio.

The error power and SNR of a uniform PCM coder vary with bit rate in a simple fashion. Suppose that a signal is quantized using B bits per sample. If zero is a reconstruction level, then the quantization step size Δ is

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

Assuming that quantization errors are uniformly distributed between $\Delta/2$ and $-\Delta/2$, the quantization error power is

$$\begin{aligned} 10 \log_{10} E[e^2(n)] = 10 \log_{10} \frac{\Delta^2}{12} &\approx \text{constant} \\ &+ 20 \log_{10}(S_{\max} - S_{\min}) - 6B \end{aligned} \tag{3}$$

2.1.2. Companded PCM.

Companded PCM is the name given to coders in which the reconstruction levels $\hat{s}_k$ are not uniformly distributed. Such coders may be modeled using a compressive nonlinearity, followed by uniform PCM, followed by an expansive nonlinearity:

$$\begin{aligned} s(n) \rightarrow \boxed{\text{compress}} \rightarrow t(n) \rightarrow \boxed{\text{uniform PCM}} \\ \rightarrow \hat{t}(n) \rightarrow \boxed{\text{expand}} \rightarrow \hat{s}(n) \end{aligned} \tag{4}$$

Q1

Table 2. •A Representative Sample of Speech Coding Standards

Application	Rate (kbps)	BW (kHz)	Standards Organization	Standard Number	Algorithm	Year
Landline	64	3.4	ITU	G.711	μ-law or A-law PCM	1988
telephone	32	3.4	ITU	G.726	ADPCM	1990
	16–40	3.4	ITU	G.727	ADPCM	1990
Tele conferencing	48–64	7	ITU	G.722	Split-band ADPCM	1988
	16	3.4	ITU	G.728	Low-delay CELP	1992
Digital	13	3.4	ETSI	Full-rate	RPE-LTP	1992
cellular	12.2	3.4	ETSI	EFR	ACELP	1997
	7.9	3.4	TIA	IS-54	VSELP	1990
	6.5	3.4	ETSI	Half-rate	VSELP	1995
	8.0	3.4	ITU	G.729	ACELP	1996
	4.75–12.2	3.4	ETSI	AMR	ACELP	1998
	1–8	3.4	CDMA-TIA	IS-96	QCELP	1993
Multimedia	5.3–6.3	3.4	ITU	G.723.1	MPLPC, CELP	1996
	2.0–18.2	3.4–7.5	ISO	MPEG-4	HVXC, CELP	1998
Satellite	4.15	3.4	INMARSAT	M	IMBE	1991
telephony	3.6	3.4	INMARSAT	Mini-M	AMBE	1995
Secure	2.4	3.4	DDVPC	FS1015	LPC-10e	1984
communications	2.4	3.4	DDVPC	MELP	MELP	1996
	4.8	3.4	DDVPC	FS1016	CELP	1989
	16–32	3.4	DDVPC	CVSD	CVSD	

It can be shown that, if small values of $s(n)$ are more likely than large values, expected error power is minimized by a companding function that results in a higher density of reconstruction levels $\hat{x}_k$ at low signal levels than at high signal levels [78]. A typical example is the μ-law companding function [48] (Fig. 1), which is given by

$$t(n) = S_{\max}\frac{\log(1+\mu|s(n)/S_{\max}|)}{\log(1+\mu)}\operatorname{sign}(s(n)) \tag{5}$$

where μ is typically between 0 and 256 and determines the amount of nonlinear compression applied.

2.2. Differential PCM (DPCM)

Successive speech samples are highly correlated. The long-term average spectrum of voiced speech is reasonably well approximated by the function $S(f) = 1/f$ above about 500 Hz; the first-order intersample correlation coefficient is approximately 0.9. In differential PCM, each sample $s(n)$ is compared to a prediction $s_p(n)$, and the difference is called the prediction residual $d(n)$ (Fig. 2). $d(n)$ has a smaller dynamic range than $s(n)$, so for a given error power, fewer bits are required to quantize $d(n)$.

Accurate quantization of $d(n)$ is useless unless it leads to accurate quantization of $s(n)$. In order to avoid amplifying the error, DPCM coders use a technique copied by many later speech coders; the encoder includes an embedded decoder, so that the reconstructed signal $\hat{s}(n)$ is known at the encoder. By using $\hat{s}(n)$ to create $s_p(n)$, DPCM coders avoid amplifying the quantization error:

$$d(n) = s(n) - s_p(n) \tag{6}$$

$$\hat{s}(n) = \hat{d}(n) + s_p(n) \tag{7}$$

$$e(n) = s(n) - \hat{s}(n) = d(n) - \hat{d}(n) \tag{8}$$

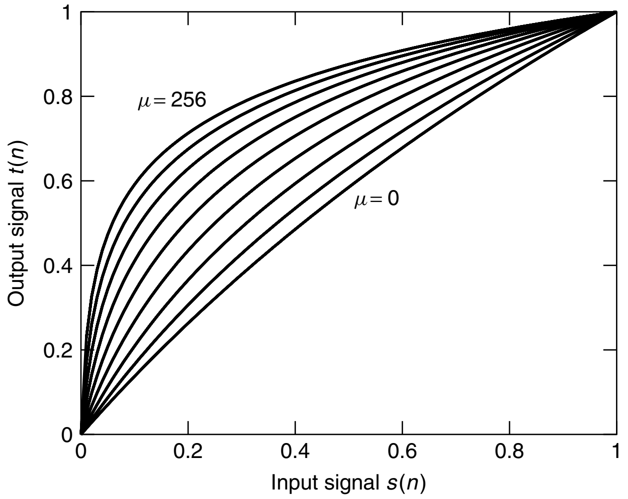

Figure 1. μ-law companding function, $\mu = 0, 1, 2, 4, 8, \ldots, 256$.

Two existing standards are based on DPCM. In the first type of coder, continuously varying slope delta modulation (CVSD), the input speech signal is upsampled to either 16 or 32 kHz. Values of the upsampled signal are predicted using a one-tap predictor, and the difference signal is quantized at one bit per sample, with an adaptively varying Δ. CVSD performs badly in quiet environments, but in extremely noisy environments (e.g., helicopter cockpit), CVSD performs better than any LPC-based algorithm, and for this reason it remains the U.S. Department of Defense recommendation for extremely noisy environments [64,96].

DPCM systems with adaptive prediction and quantization are referred to as adaptive differential PCM systems (ADPCM). A commonly used ADPCM standard is G.726, which can operate at 16, 24, 32, or 40 kbps (2–5 bits/sample) [45]. G.726 ADPCM is frequently used at 32 kbps in landline telephony. The predictor in G.726 consists of an adaptive second-order IIR predictor in series with an adaptive sixth-order FIR predictor. Filter coefficients are adapted using a computationally simplified gradient descent algorithm. The prediction residual is quantized using a semilogarithmic companded PCM quantizer at a rate of 2–5 bits per sample. The quantization step size adapts to the amplitude of previous samples of the quantized prediction error signal; the speed of adaptation is controlled by an estimate of the type of signal, with adaptation to speech signals being faster than adaptation to signaling tones.

3. SUBBAND CODING

In subband coding, an analysis filterbank is first used to filter the signal into a number of frequency bands and then bits are allocated to each band by a certain criterion. Because of the difficulty in obtaining high-quality speech at low bit rates using subband coding schemes, these techniques have been used mostly for wideband medium to high bit rate speech coders and for audio coding.

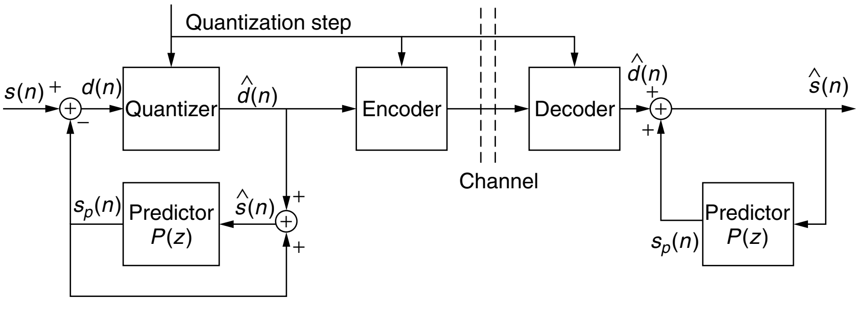

Figure 2. Schematic of a DPCM coder.

For example, G.722 is a standard in which ADPCM speech coding occurs within two subbands, and bit allocation is set to achieve 7-kHz audio coding at rates of 64 kbps or less.

In Refs. 12,13, and 30 subband coding is proposed as a flexible scheme for robust speech coding. A speech production model is not used, ensuring robustness to speech in the presence of background noise, and to nonspeech sources. High-quality compression can be achieved by incorporating masking properties of the human auditory system [54,93]. In particular, Tang et al. [93] present a scheme for robust, high-quality, scalable, and embedded speech coding. Figure 3 illustrates the basic structure of the coder. Dynamic bit allocation and prioritization and embedded quantization are used to optimize the perceptual quality of the embedded bitstream, resulting in little performance degradation relative to a nonembedded implementation. A subband spectral analysis technique was developed that substantially reduces the complexity of computing the perceptual model.

The encoded bitstream is embedded, allowing the coder output to be scalable from high quality at higher bit rates, to lower quality at lower rates, supporting a wide range of service and resource utilization. The lower bit rate representation is obtained simply through truncation of the higher bit rate representation. Since source rate adaptation is performed through truncation of the encoded stream, interaction with the source coder is not required, making the coder ideally suited for rate adaptive communication systems.

Even though subband coding is not widely used for speech coding today, it is expected that new standards for wideband coding and rate-adaptive schemes will be based on subband coding or a hybrid technique that includes subband coding. This is because subband coders are more easily scalable in bit rate than standard CELP techniques, an issue which will become more critical for high-quality speech and audio transmission over wireless communication channels and the Internet, allowing the system to seamlessly adapt to changes in both the transmission environment and network congestion.

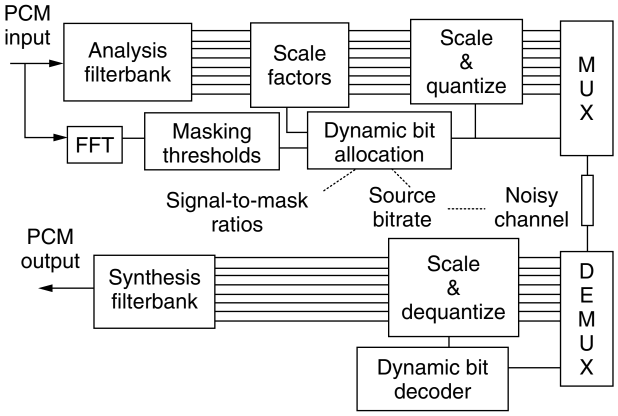

Figure 3. Structure of a perceptual subband speech coder [93].

4. LPC-BASED ANALYSIS BY SYNTHESIS

An analysis-by-synthesis speech coder consists of the following components:

- A model of speech production that depends on certain parameters θ:
$$\hat{s}(n) = f(\theta) \tag{9}$$
- A list of K possible parameter sets for the model
$$\theta_1, \ldots, \theta_k, \ldots, \theta_K \tag{10}$$
- An error metric $|E_k|^2$ that compares the original speech signal $s(n)$ and the coded speech signal $\hat{s}(n)$. In LPC-AS coders, $|E_k|^2$ is typically a perceptually weighted mean-squared error measure.

A general analysis-by-synthesis coder finds the optimum set of parameters by synthesizing all of the K different speech waveforms $\hat{s}_k(n)$ corresponding to the K possible parameter sets θ_k, computing $|E_k|^2$ for each synthesized waveform, and then transmitting the index of the parameter set which minimizes $|E_k|^2$. Choosing a set of transmitted parameters by explicitly computing $\hat{s}_k(n)$ is called "closed loop" optimization, and may be contrasted with "open-loop" optimization, in which coder parameters are chosen on the basis of an analytical formula without explicit computation of $\hat{s}_k(n)$. Closed-loop optimization of all parameters is prohibitively expensive, so LPC-based analysis-by-synthesis coders typically adopt the following compromise. The gross spectral shape is modeled using an all-pole filter $1/A(z)$ whose parameters are estimated in open-loop fashion, while spectral fine structure is modeled using an excitation function $U(z)$ whose parameters are optimized in closed-loop fashion (Fig. 4).

4.1. The Basic LPC Model

In LPC-based coders, the speech signal $S(z)$ is viewed as the output of a linear time-invariant (LTI) system whose

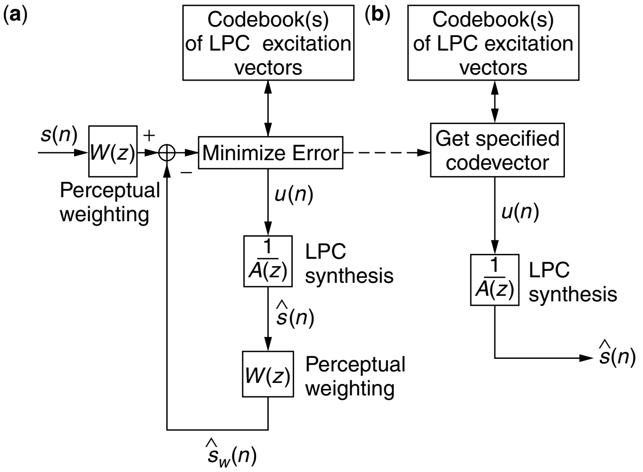

Figure 4. General structure of an LPC-AS coder (**a**) and decoder (**b**). LPC filter $A(z)$ and perceptual weighting filter $W(z)$ are chosen open-loop, then the excitation vector $u(n)$ is chosen in a closed-loop fashion in order to minimize the error metric $|E|^2$.

input is the excitation signal $U(z)$, and whose transfer function is represented by the following:

$$S(z) = \frac{U(z)}{A(z)} = \frac{U(z)}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{11}$$

Most of the zeros of $A(z)$ correspond to resonant frequencies of the vocal tract or *formant frequencies*. Formant frequencies depend on the geometry of the vocal tract; this is why men and women, who have different vocal-tract shapes and lengths, have different formant frequencies for the same sounds.

The number of LPC coefficients (p) depends on the signal bandwidth. Since each pair of complex-conjugate poles represents one formant frequency and since there is, on average, one formant frequency per 1 kHz, p is typically equal to $2BW$ (in kHz) $+2 - 4$. Thus, for a 4 kHz speech signal, a 10th–12th-order LPC model would be used.

This system is excited by a signal $u(n)$ that is uncorrelated with itself over lags of less than $p + 1$. If the underlying speech sound is unvoiced (the vocal folds do not vibrate), then $u(n)$ is uncorrelated with itself even at larger time lags, and may be modeled using a pseudo-random-noise signal. If the underlying speech is voiced (the vocal folds vibrate), then $u(n)$ is quasiperiodic with a fundamental period called the "pitch period."

4.2. Pitch Prediction Filtering

In an LPC-AS coder, the LPC excitation is allowed to vary smoothly between fully voiced conditions (as in a vowel) and fully unvoiced conditions (as in /s/). Intermediate levels of voicing are often useful to model partially voiced phonemes such as /z/.

The partially voiced excitation in an LPC-AS coder is constructed by passing an uncorrelated noise signal $c(n)$ through a pitch prediction filter [2,79]. A typical pitch prediction filter is

$$u(n) = gc(n) + bu(n - T_0) \tag{12}$$

where T_0 is the pitch period. If $c(n)$ is unit variance white noise, then according to Eq. (12) the spectrum of $u(n)$ is

$$|U(e^{j\omega})|^2 = \frac{g^2}{1 + b^2 - 2b\cos\omega T_0} \tag{13}$$

Figure 5 shows the normalized magnitude spectrum $(1 - b)|U(e^{j\omega})|$ for several values of b between 0.25 and 1. As shown, the spectrum varies smoothly from a uniform spectrum, which is heard as unvoiced, to a harmonic spectrum that is heard as voiced, without the need for a binary voiced/unvoiced decision.

In LPC-AS coders, the noise signal $c(n)$ is chosen from a "stochastic codebook" of candidate noise signals. The stochastic codebook index, the pitch period, and the gains b and g are chosen in a closed-loop fashion in order to minimize a perceptually weighted error metric. The search for an optimum T_0 typically uses the same algorithm as the search for an optimum $c(n)$. For this reason, the list of excitation samples delayed by different candidate values of T_0 is typically called an "adaptive codebook" [87].

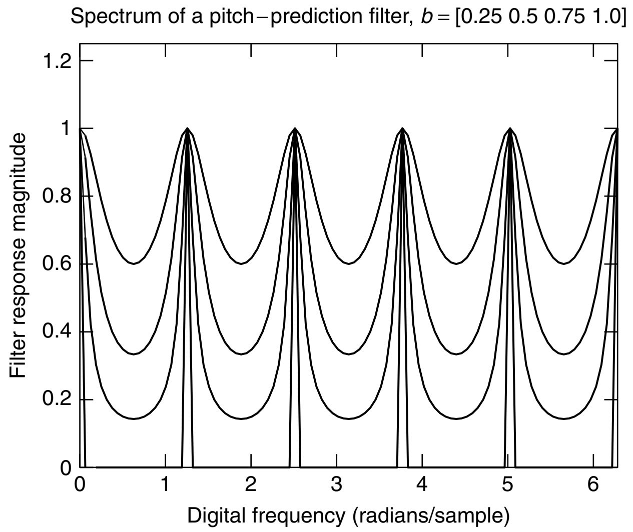

Figure 5. Normalized magnitude spectrum of the pitch prediction filter for several values of the prediction coefficient.

4.3. Perceptual Error Weighting

Not all types of distortion are equally audible. Many types of speech coders, including LPC-AS coders, use simple models of human perception in order to minimize the audibility of different types of distortion. In LPC-AS coding, two types of perceptual weighting are commonly used. The first type, perceptual weighting of the residual quantization error, is used during the LPC excitation search in order to choose the excitation vector with the least audible quantization error. The second type, adaptive postfiltering, is used to reduce the perceptual importance of any remaining quantization error.

4.3.1. Perceptual Weighting of the Residual Quantization Error. The excitation in an LPC-AS coder is chosen to minimize a perceptually weighted error metric. Usually, the error metric is a function of the time domain waveform error signal

$$e(n) = s(n) - \hat{s}(n) \tag{14}$$

Early LPC-AS coders minimized the mean-squared error

$$\sum_n e^2(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} |E(e^{j\omega})|^2\, d\omega \tag{15}$$

It turns out that the MSE is minimized if the error spectrum, $E(e^{j\omega})$, is white — that is, if the error signal $e(n)$ is an uncorrelated random noise signal, as shown in Fig. 6.

Not all noises are equally audible. In particular, noise components near peaks of the speech spectrum are hidden by a "masking spectrum" $M(e^{j\omega})$, so that a shaped noise spectrum at lower SNR may be less audible than a white-noise spectrum at higher SNR (Fig. 7). The audibility of noise may be estimated using a noise-to-masker ratio $|E_w|^2$:

$$|E_w|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|E(e^{j\omega})|^2}{|M(e^{j\omega})|^2} d\omega \tag{16}$$

The masking spectrum $M(e^{j\omega})$ has peaks and valleys at the same frequencies as the speech spectrum, but the difference in amplitude between peaks and valleys is somewhat smaller than that of the speech spectrum. A variety of algorithms exist for estimating the masking spectrum, ranging from extremely simple to extremely complex [51]. One of the simplest model masking spectra that has the properties just described is as follows [2]:

$$M(z) = \frac{|A(z/\gamma_2)|}{|A(z/\gamma_1)|}, \quad 0 < \gamma_2 < \gamma_1 \leq 1 \tag{17}$$

where $1/A(z)$ is an LPC model of the speech spectrum. The poles and zeros of $M(z)$ are at the same frequencies as the poles of $1/A(z)$, but have broader bandwidths. Since the zeros of $M(z)$ have broader bandwidth than its poles, $M(z)$ has peaks where $1/A(z)$ has peaks, but the difference between peak and valley amplitudes is somewhat reduced.

The noise-to-masker ratio may be efficiently computed by filtering the speech signal using a perceptual weighting filter $W(z) = 1/M(z)$. The perceptually weighted input speech signal is

$$S_w(z) = W(z)S(z) \tag{18}$$

Likewise, for any particular candidate excitation signal, the perceptually weighted output speech signal is

$$\hat{S}_w(z) = W(z)\hat{S}(z) \tag{19}$$

Given $s_w(n)$ and $\hat{s}_w(n)$, the noise-to-masker ratio may be computed as follows:

$$|E_w|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |S_w(e^{j\omega}) - \hat{S}_w(e^{j\omega})|^2 d\omega = \sum_n (s_w^2(n) - \hat{s}_w^2(n)) \tag{20}$$

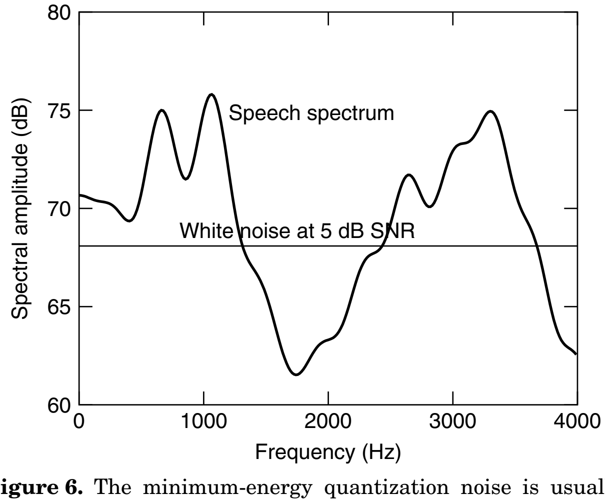

Figure 6. The minimum-energy quantization noise is usually white noise.

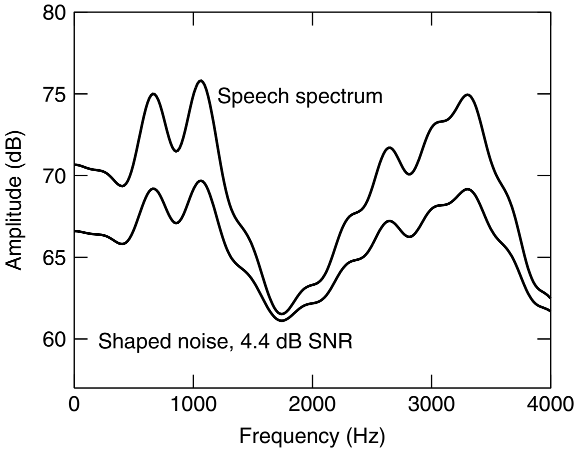

Figure 7. Shaped quantization noise may be less audible than white quantization noise, even at slightly lower SNR.

4.3.2. Adaptive Postfiltering. Despite the use of perceptually weighted error minimization, the synthesized speech coming from an LPC-AS coder may contain audible quantization noise. In order to minimize the perceptual effects of this noise, the last step in the decoding process is often a set of adaptive postfilters [11,80]. Adaptive postfiltering improves the perceptual quality of noisy speech by giving a small extra emphasis to features of the spectrum that are important for human-to-human communication, including the pitch periodicity (if any) and the peaks in the spectral envelope.

A pitch postfilter (or long-term predictive postfilter) enhances the periodicity of voiced speech by applying either an FIR or IIR comb filter to the output. The time delay and gain of the comb filter may be set equal to the transmitted pitch lag and gain, or they may be recalculated at the decoder using the reconstructed signal $\hat{s}(n)$. The pitch postfilter is applied only if the proposed comb filter gain is above a threshold; if the comb filter gain is below threshold, the speech is considered unvoiced, and no pitch postfilter is used. For improved perceptual quality, the LPC excitation signal may be interpolated to a higher sampling rate in order to allow the use of fractional pitch periods; for example, the postfilter in the ITU G.729 coder uses pitch periods quantized to $\frac{1}{8}$ sample.

A short-term predictive postfilter enhances peaks in the spectral envelope. The form of the short-term postfilter is similar to that of the masking function $M(z)$ introduced in the previous section; the filter has peaks at the same frequencies as $1/A(z)$, but the peak-to-valley ratio is less than that of $A(z)$.

Postfiltering may change the gain and the average spectral tilt of $\hat{s}(n)$. In order to correct these problems, systems that employ postfiltering may pass the final signal through a one-tap FIR preemphasis filter, and then modify its gain, prior to sending the reconstructed signal to an A/D converter.

4.4. Frame-Based Analysis

The characteristics of the LPC excitation signal $u(n)$ change quite rapidly. The energy of the signal may change from zero to nearly full amplitude within one millisecond Q4 at the release of a plosive• sound, and a mistake of more than about 5 ms in the placement of such a sound is clearly audible. The LPC coefficients, on the other hand, change relatively slowly. In order to take advantage of the slow rate of change of LPC coefficients without sacrificing the quality of the coded residual, most LPC-AS coders encode speech using a frame–subframe structure, as depicted in Fig. 8. A frame of speech is approximately 20 ms in length, and is composed of typically three to four subframes. The LPC excitation is transmitted only once per subframe, while the LPC coefficients are transmitted only once per frame. The LPC coefficients are computed by analyzing a window of speech that is usually longer than the speech frame (typically 30–60 ms). In order to minimize the number of future samples required to compute LPC coefficients, many LPC-AS coders use an asymmetric window that may include several hundred milliseconds of past context, but that emphasizes the samples of the current frame [21,84].

The perceptually weighted original signal $s_w(n)$ and weighted reconstructed signal $\hat{s}_w(n)$ in a given subframe are often written as L-dimensional row vectors S and $\hat{S}$, where the dimension L is the length of a subframe:

$$S_w = [s_w(0), \ldots, s_w(L-1)], \quad \hat{S}_w = [\hat{s}_w(0), \ldots, \hat{s}_w(L-1)] \tag{21}$$

The core of an LPC-AS coder is the closed-loop search for an optimum coded excitation vector U, where U is typically composed of an "adaptive codebook" component representing the periodicity, and a "stochastic codebook" component representing the noiselike part of the excitation. In general, U may be represented as the weighted sum of several "shape vectors" X_m, $m = 1, \ldots, M$, which may be drawn from several codebooks, including possibly multiple adaptive codebooks and multiple stochastic codebooks:

$$U = GX, \quad G = [g_1, g_2, \ldots], \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \end{bmatrix} \tag{22}$$

The choice of shape vectors and the values of the gains g_m are jointly optimized in a closed-loop search, in order to minimize the perceptually weighted error metric $|S_w - \hat{S}_w|^2$.

The value of S_w may be computed prior to any codebook search by perceptually weighting the input speech vector. The value of $\hat{S}_w$ must be computed separately for each candidate excitation, by synthesizing the speech signal $\hat{s}(n)$, and then perceptually weighting to obtain $\hat{s}_w(n)$. These operations may be efficiently computed, as described below.

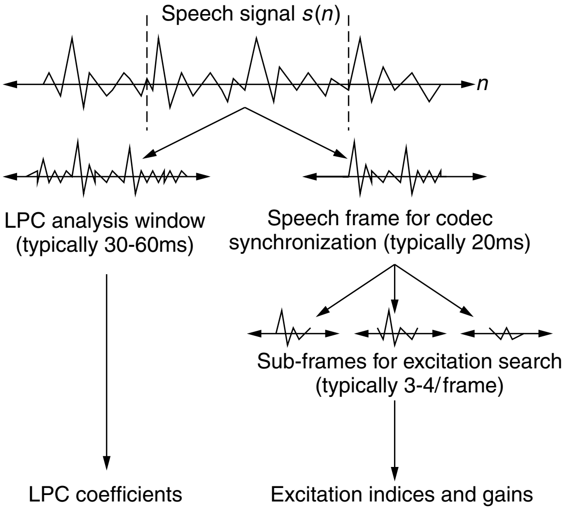
Speech signal $s(n)$
n
LPC analysis window (typically 30-60ms)
Speech frame for codec synchronization (typically 20ms)
Sub-frames for excitation search (typically 3-4/frame)
LPC coefficients
Excitation indices and gains

Figure 8. The frame/subframe structure of most LPC analysis by synthesis coders.

4.4.1. Zero State Response and Zero Input Response.

Let the filter $H(z)$ be defined as the composition of the LPC synthesis filter and the perceptual weighting filter, thus $H(z) = W(z)/A(z)$. The computational complexity of the excitation parameter search may be greatly simplified if $\hat{S}_w$ is decomposed into the zero input response (ZIR) and zero state response (ZSR) of $H(z)$ [97]. Note that the weighted reconstructed speech signal is

$$\hat{S}_w = [\hat{s}_w(0), \ldots, \hat{s}_w(L-1)], \quad \hat{s}_w(n) = \sum_{i=0}^{\infty} h(i)u(n-i) \tag{23}$$

where $h(n)$ is the infinite-length impulse response of $H(z)$. Suppose that $\hat{s}_w(n)$ has already been computed for $n < 0$, and the coder is now in the process of choosing the optimal $u(n)$ for the subframe $0 \le n \le L-1$. The sum above can be divided into two parts: a part that depends on the current subframe input, and a part that does not:

$$\hat{S}_w = \hat{S}_{\text{ZIR}} + UH \tag{24}$$

where $\hat{S}_{\text{ZIR}}$ contains samples of the zero input response of $H(z)$, and the vector UH contains the zero state response. The zero input response is usually computed by implementing the recursive filter $H(z) = W(z)/A(z)$ as the sequence of two IIR filters, and allowing the two filters to run for L samples with zero input. The zero state response is usually computed as the matrix product UH, where

$$H = \begin{bmatrix} h(0) & h(1) & \ldots & h(L-1) \\ 0 & h(0) & \ldots & h(L-2) \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & h(0) \end{bmatrix},$$
$$U = [u(0), \ldots, u(L-1)] \tag{25}$$

Given a candidate excitation vector U, the perceptually weighted error vector E may be defined as

$$E_w = S_w - \hat{S}_w = \tilde{S} - UH \tag{26}$$

where the target vector $\tilde{S}$ is

$$\tilde{S} = S_w - \hat{S}_{\mathrm{ZIR}} \tag{27}$$

The target vector needs to be computed only once per subframe, prior to the codebook search. The objective of the codebook search, therefore, is to find an excitation vector U that minimizes $|\tilde{S} - UH|^2$.

4.4.2. Optimum Gain and Optimum Excitation.

Recall that the excitation vector U is modeled as the weighted sum of a number of codevectors X_m, $m = 1, \ldots, M$. The perceptually weighted error is therefore:

$$|E|^2 = |\tilde{S} - GXH|^2 = \tilde{S}\tilde{S}' - 2GXH\tilde{S}' + GXH(GXH)' \tag{28}$$

where prime denotes transpose. Minimizing $|E|^2$ requires optimum choice of the shape vectors X and of the gains G. It turns out that the optimum gain for each excitation vector can be computed in closed form. Since the optimum gain can be computed in closed form, it need not be computed during the closed-loop search; instead, one can simply assume that each candidate excitation, if selected, would be scaled by its optimum gain. Assuming an optimum gain results in an extremely efficient criterion for choosing the optimum excitation vector [3].

Suppose we define the following additional bits of notation:

$$R_X = XH\tilde{S}', \quad \Sigma = XH(XH)' \tag{29}$$

Then the mean-squared error is

$$|E|^2 = \tilde{S}\tilde{S}' - 2GR_X + G\Sigma G' \tag{30}$$

For any given set of shape vectors X, G is chosen so that $|E|^2$ is minimized, which yields

$$G = R_X'\Sigma^{-1} \tag{31}$$

If we substitute the minimum MSE value of G into Eq. (30), we get

$$|E|^2 = \tilde{S}\tilde{S}' - R_X'\Sigma^{-1}R_X \tag{32}$$

Hence, in order to minimize the perceptually weighted MSE, we choose the shape vectors X in order to maximize the covariance-weighted sum of correlations:

$$X_{\mathrm{opt}} = \arg\max(R_X'\Sigma^{-1}R_X) \tag{33}$$

When the shape matrix X contains more than one row, the matrix inversion in Eq. (33) is often computed using approximate algorithms [4]. In the VSELP coder [25], X is transformed using a modified Gram–Schmidt orthogonalization so that Σ has a diagonal structure, thus simplifying the computation of Eq. (33).

4.5. Types of LPC-AS Coder

4.5.1. Multipulse LPC (MPLPC).

In the multipulse LPC algorithm [4,50], the shape vectors are impulses. U is typically formed as the weighted sum of 4–8 impulses per subframe.

The number of possible combinations of impulses grows exponentially in the number of impulses, so joint optimization of the positions of all impulses is usually impossible. Instead, most MPLPC coders optimize the pulse positions one at a time, using something like the following strategy. First, the weighted zero state response of $H(z)$ corresponding to each impulse location is computed. If C_k is an impulse located at $n = k$, the corresponding weighted zero state response is

$$C_kH = [0, \ldots, 0, h(0), h(1), \ldots, h(L - k - 1)] \tag{34}$$

The location of the first impulse is chosen in order to optimally approximate the target vector $\tilde{S}_1 = \tilde{S}$, using the methods described in the previous section. After selecting the first impulse location k_1, the target vector is updated according to

$$\tilde{S}_m = \tilde{S}_{m-1} - C_{k_{m-1}}H \tag{35}$$

Additional impulses are chosen until the desired number of impulses is reached. The gains of all pulses may be reoptimized after the selection of each new pulse [87].

Variations are possible. The multipulse coder described in ITU standard G.723.1 transmits a single gain for all the impulses, plus sign bits for each individual impulse. The G.723.1 coder restricts all impulse locations to be either odd or even; the choice of odd or even locations is coded using one bit per subframe [50]. The regular pulse excited LPC algorithm, which was the first GSM full-rate speech coder, synthesized speech using a train of impulses spaced one per 4 samples, all scaled by a single gain term [65]. The alignment of the pulse train was restricted to one of four possible locations, chosen in a closed-loop fashion together with a gain, an adaptive codebook delay, and an adaptive codebook gain.

Singhal and Atal demonstrated that the quality of MPLPC may be improved at low bit rates by modeling the periodic component of an LPC excitation vector using a pitch prediction filter [87]. Using a pitch prediction filter, the LPC excitation signal becomes

$$u(n) = bu(n - D) + \sum_{m=1}^{M} c_{k_m}(n) \tag{36}$$

where the signal $c_k(n)$ is an impulse located at $n = k$ and b is the pitch prediction filter gain. Singhal and Atal proposed choosing D before the locations of any impulses are known, by minimizing the following perceptually weighted error:

$$|E_D|^2 = |\tilde{S} - bX_DH|^2, X_D = [u(-D), \ldots, u((L-1) - D)] \tag{37}$$

The G.723.1 multipulse LPC coder and the GSM (Global System for Mobile Communication) full-rate RPE-LTP (regular-pulse excitation with long-term prediction) coder both use a closed-loop pitch predictor, as do all standardized variations of the CELP coder (see Sections 4.5.2 and 4.5.3). Typically, the pitch delay and gain are optimized first, and then the gains of any additional excitation vectors (e.g., impulses in an MPLPC algorithm) are selected to minimize the remaining error.

4.5.2. Code-Excited LPC (CELP). LPC analysis finds a filter $1/A(z)$ whose excitation is uncorrelated for correlation distances smaller than the order of the filter. Pitch prediction, especially closed-loop pitch prediction, removes much of the remaining intersample correlation. The spectrum of the pitch prediction residual looks like the spectrum of uncorrelated Gaussian noise, but replacing the residual with real noise (noise that is independent of the original signal) yields poor speech quality. Apparently, some of the temporal details of the pitch prediction residual are perceptually important. Schroeder and Atal proposed modeling the pitch prediction residual using a stochastic excitation vector $c_k(n)$ chosen from a list of stochastic excitation vectors, $k = 1, \ldots, K$, known to both the transmitter and receiver [85]:

$$u(n) = bu(n - D) + gc_k(n) \tag{38}$$

The list of stochastic excitation vectors is called a *stochastic codebook*, and the index of the stochastic codevector is chosen in order to minimize the perceptually weighted error metric $|E_k|^2$. Rose and Barnwell discussed the similarity between the search for an optimum stochastic codevector index k and the search for an optimum predictor delay D [82], and Kleijn et al. coined the term "adaptive codebook" to refer to the list of delayed excitation signals $u(n - D)$ which the coder considers during closed-loop pitch delay optimization (Fig. 9).

The CELP algorithm was originally not considered efficient enough to be used in real-time speech coding, but a number of computational simplifications were proposed that resulted in real-time CELP-like algorithms. Trancoso and Atal proposed efficient search methods based on the truncated impulse response of the filter $W(z)/A(z)$, as discussed in Section 4.4 [3,97]. Davidson and Lin separately proposed center clipping the stochastic codevectors, so that most of the samples in each codevector are zero [15,67]. Lin also proposed structuring the stochastic codebook so that each codevector is a slightly-shifted version of the previous codevector; such a codebook is called an *overlapped codebook* [67]. Overlapped stochastic codebooks are rarely used in practice today, but overlapped-codebook search methods are often used to reduce the computational complexity of an adaptive codebook search. In the search of an overlapped codebook, the correlation R_X and autocorrelation Σ introduced in Section 4.4 may be recursively computed, thus greatly reducing the complexity of the codebook search [63].

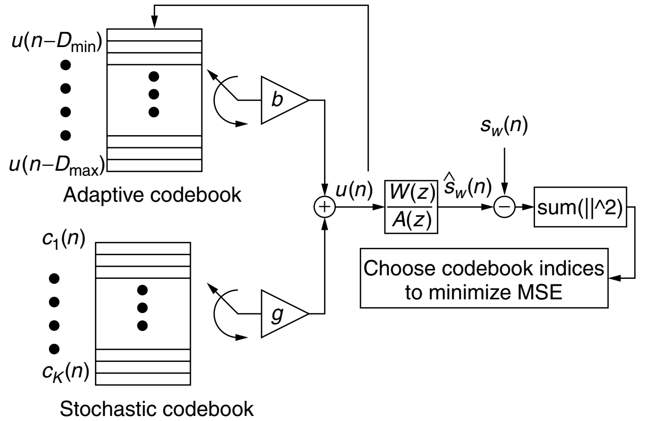

Figure 9. The code-excited LPC algorithm (CELP) constructs an LPC excitation signal by optimally choosing input vectors from two codebooks: an "adaptive" codebook, which represents the pitch periodicity; and a "stochastic" codebook, which represents the unpredictable innovations in each speech frame.

Most CELP coders optimize the adaptive codebook index and gain first, and then choose a stochastic codevector and gain in order to minimize the remaining perceptually weighted error. If all the possible pitch periods are longer than one subframe, then the entire content of the adaptive codebook is known before the beginning of the codebook search, and the efficient overlapped codebook search methods proposed by Lin may be applied [67]. In practice, the pitch period of a female speaker is often shorter than one subframe. In order to guarantee that the entire adaptive codebook is known before beginning a codebook search, two methods are commonly used: (1) the adaptive codebook search may simply be constrained to only consider pitch periods longer than L samples — in this case, the adaptive codebook will lock onto values of D that are an integer multiple of the actual pitch period (if the same integer multiple is not chosen for each subframe, the reconstructed speech quality is usually good); and (2) adaptive codevectors with delays of $D < L$ may be constructed by simply repeating the most recent D samples as necessary to fill the subframe.

4.5.3. SELP, VSELP, ACELP, and LD-CELP. Rose and Barnwell demonstrated that reasonable speech quality is achieved if the LPC excitation vector is computed completely recursively, using two closed-loop pitch predictors in series, with no additional information [82]. In their "self-excited LPC" algorithm (SELP), the LPC excitation is initialized during the first subframe using a vector of samples known at both the transmitter and receiver. For all frames after the first, the excitation is the sum of an arbitrary number of adaptive codevectors:

$$u(n) = \sum_{m=1}^{M} b_m u(n - D_m) \tag{39}$$

Kleijn et al. developed efficient recursive algorithms for searching the adaptive codebook in SELP coder and other LPC-AS coders [63].

Just as there may be more than one adaptive codebook, it is also possible to use more than one stochastic codebook. The vector-sum excited LPC algorithm (VSELP) models the LPC excitation vector as the sum of one adaptive and two stochastic codevectors [25]:

$$u(n) = bu(n - D) + \sum_{m=1}^{2} g_m c_{k_m}(n) \tag{40}$$

The two stochastic codebooks are each relatively small (typically 32 vectors), so that each of the codebooks may be searched efficiently. The adaptive codevector and the two stochastic codevectors are chosen sequentially. After selection of the adaptive codevector, the stochastic codebooks are transformed using a modified Gram–Schmidt orthogonalization, so that the perceptually weighted speech vectors generated during the first stochastic codebook search are all orthogonal to the perceptually weighted adaptive codevector. Because of this orthogonalization, the stochastic codebook search results in the choice of a stochastic codevector that is jointly optimal with the adaptive codevector, rather than merely sequentially optimal. VSELP is the basis of the Telecommunications Industry Associations digital cellular standard IS-54.

The algebraic CELP (ACELP) algorithm creates an LPC excitation by choosing just one vector from an adaptive codebook and one vector from a fixed codebook. In the ACELP algorithm, however, the fixed codebook is composed of binary-valued or trinary-valued algebraic codes, rather than the usual samples of a Gaussian noise process [1]. Because of the simplicity of the codevectors, it is possible to search a very large fixed codebook very quickly using methods that are a hybrid of standard CELP and MPLPC search algorithms. ACELP is the basis of the ITU standard G.729 coder at 8 kbps. ACELP codebooks may be somewhat larger than the codebooks in a standard CELP coder; the codebook in G.729, for example, contains 8096 codevectors per subframe.

Most LPC-AS coders operate at very low bit rates, but require relatively large buffering delays. The low-delay CELP coder (LD-CELP) operates at 16 kbps [10,47] and is designed to obtain the best possible speech quality, with the constraint that the total algorithmic delay of a tandem coder and decoder must be no more than 2 ms. LPC analysis and codevector search are computed once per 2 ms (16 samples). Transmission of LPC coefficients once per two milliseconds would require too many bits, so LPC coefficients are computed in a recursive backward-adaptive fashion. Before coding or decoding each frame, samples of $\hat{s}(n)$ from the previous frame are windowed, and used to update a recursive estimate of the autocorrelation function. The resulting autocorrelation coefficients are similar to those that would be obtained using a relatively long asymmetric analysis window. LPC coefficients are then computed from the autocorrelation function using the Levinson–Durbin algorithm.

4.6. Line Spectral Frequencies (LSFs) or Line Spectral Pairs (LSPs)

Linear prediction can be viewed as an inverse filtering procedure in which the speech signal is passed through an all-zero filter $A(z)$. The filter coefficients of $A(z)$ are chosen such that the energy in the output, that is, the residual or error signal, is minimized. Alternatively, the inverse filter $A(z)$ can be transformed into two other filters $P(z)$ and $Q(z)$. These new filters turn out to have some interesting properties, and the representation based on them, called the *line spectrum pairs* [89,91], has been used in speech coding and synthesis applications.

Let $A(z)$ be the frequency response of an LPC inverse filter of order p:

$$A(z) = -\sum_{i=0}^{p} a_i z^{-i}$$

with $a_0 = -1$. The a_i values are real, and all the zeros of $A(z)$ are inside the unit circle.

If we use the lattice formulation of LPC, we arrive at a recursive relation between the mth stage $[A_m(z)]$ and the one before it $[A_{m-1}(z)]$ For the pth-order inverse filter, we have

$$A_p(z) = A_{p-1}(z) - k_p z^{-p} A_{p-1}(z^{-1})$$

By allowing the recursion to go one more iteration, we obtain

$$A_{p+1}(z) = A_p(z) - k_{p+1} z^{-(p+1)} A_p(z^{-1}) \tag{41}$$

If we choose $k_{p+1} = \pm 1$ in Eq. (41), we can define two new polynomials as follows:

$$P(z) = A(z) - z^{-(p+1)} A(z^{-1}) \tag{42}$$

$$Q(z) = A(z) + z^{-(p+1)} A(z^{-1}) \tag{43}$$

Physically, $P(z)$ and $Q(z)$ can be interpreted as the inverse transfer function of the vocal tract for the *open-glottis* and *closed-glottis* boundary conditions, respectively [22], and $P(z)/Q(z)$ is the driving-point impedance of the vocal tract as seen from the glottis [36].

If p is odd, the formulae for p_n and q_n are as follows:

$$\begin{aligned} P(z) &= A(z) + z^{-(p+1)} A(z^{-1}) \\ &= \prod_{n=1}^{(p+1)/2} (1 - e^{jp_n} z^{-1})(1 - e^{-jp_n} z^{-1}) \end{aligned} \tag{44}$$

$$\begin{aligned} Q(z) &= A(z) - z^{-(p+1)} A(z^{-1}) \\ &= (1 - z^{-2}) \prod_{n=1}^{(p-1)/2} (1 - e^{jq_n} z^{-1})(1 - e^{-jq_n} z^{-1}) \end{aligned} \tag{45}$$

The LSFs have some interesting characteristics: the frequencies $\{p_n\}$ and $\{q_n\}$ are related to the formant frequencies; the dynamic range of $\{p_n\}$ and $\{q_n\}$ is limited and the two alternate around the unit circle $(0 \le p_1 \le q_1 \le p_2 \ldots)$; $\{p_n\}$ and $\{q_n\}$ are correlated so that intraframe prediction is possible; and they change slowly from one frame to another, hence, interframe prediction is also possible. The interleaving nature of the $\{p_n\}$ and $\{q_n\}$ allow for efficient iterative solutions [58].

Almost all LPC-based coders today use the LSFs to represent the LP parameters. Considerable recent research has been devoted to methods for efficiently quantizing the LSFs, especially using vector quantization (VQ) techniques. Typical algorithms include predictive VQ, split VQ [76], and multistage VQ [66,74]. All of these methods are used in the ITU standard ACELP coder G.729: the moving-average vector prediction residual is quantized using a 7-bit first-stage codebook, followed by second-stage quantization of two subvectors using independent 5-bit codebooks, for a total of 17 bits per frame [49,84].

5. LPC VOCODERS

5.1. The LPC-10e Vocoder

The 2.4-kbps LPC-10e vocoder (Fig. 10) is one of the earliest and one of the longest-lasting standards for low-bit-rate digital speech coding [8,16]. This standard was originally proposed in the 1970s, and was not officially replaced until the selection of the MELP 2.4-kbps coding standard in 1996 [64]. Speech coded using LPC-10e sounds metallic and synthetic, but it is intelligible.

In the LPC-10e algorithm, speech is first windowed using a Hamming window of length 22.5ms. The gain (G) and coefficients (a_i) of a linear prediction filter are calculated for the entire frame using the Levinson–Durbin recursion. Once G and a_i have been computed, the LPC residual signal $d(n)$ is computed:

$$d(n) = \frac{1}{G}\left(s(n) - \sum_{i=1}^{p} a_i s(n-i)\right) \quad (46)$$

The residual signal $d(n)$ is modeled using either a periodic train of impulses (if the speech frame is voiced) or an uncorrelated Gaussian random noise signal (if the frame is unvoiced). The voiced/unvoiced decision is based on the average magnitude difference function (AMDF),

$$\Phi_d(m) = \frac{1}{N-|m|}\sum_{n=|m|}^{N-1} |d(n) - d(n-|m|)| \quad (47)$$

The frame is labeled as voiced if there is a trough in $\Phi_d(m)$ that is large enough to be caused by voiced excitation. Only values of m between 20 and 160 are examined, corresponding to pitch frequencies between 50 and 400 Hz. If the minimum value of $\Phi_d(m)$ in this range is less than a threshold, the frame is declared voiced, and otherwise it is declared unvoiced [8].

If the frame is voiced, then the LPC residual is represented using an impulse train of period T_0, where

$$T_0 = \arg\min_{m=20}^{160} \Phi_d(m) \quad (48)$$

If the frame is unvoiced, a pitch period of $T_0 = 0$ is transmitted, indicating that an uncorrelated Gaussian random noise signal should be used as the excitation of the LPC synthesis filter.

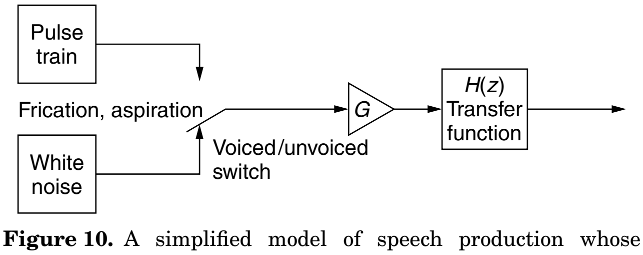

Figure 10. A simplified model of speech production whose parameters can be transmitted efficiently across a digital channel.

5.2. Mixed-Excitation Linear Prediction (MELP)

The mixed-excitation linear prediction (MELP) coder [69] was selected in 1996 by the United States Department of Defense Voice Processing Consortium (DDVPC) to be the U.S. Federal Standard at 2.4 kbps, replacing LPC-10e. 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 [70]. The synthesis model for the MELP coder is illustrated in Fig. 11. LP coefficients are converted to LSFs and a multistage vector quantizer (MSVQ) is used to quantize the LSF vectors. For voiced segments a total of 54 bits that represent: LSF parameters (25), Fourier magnitudes of the prediction residual signal (8), gain (8), pitch (7), bandpass voicing (4), aperiodic flag (1), and a sync bit are sent. The Fourier magnitudes are coded with an 8-bit VQ and the associated codebook is searched with a perceptually-weighted Euclidean distance. For unvoiced segments, the Fourier magnitudes, bandpass voicing, and the aperiodic flag bit are not sent. Instead, 13 bits that implement forward error correction (FEC) are sent. The performance of MELP at 2.4 kbps is similar to or better than that of the federal standard at 4.8 kbps (FS 1016) [92]. Versions of MELP coders operating at 1.7 kbps [68] and 4.0 kbps [90] have been reported.

5.3. Multiband Excitation (MBE)

In multiband excitation (MBE) coding the voiced/unvoiced decision is not a binary one; instead, a series of voicing decisions are made for independent harmonic intervals [31]. Since voicing decisions can be made in different frequency bands individually, synthesized speech may be partially voiced and partially unvoiced. An improved version of the MBE was introduced in the late 1980s [7,35] and referred to as the IMBE coder. The IMBE at 2.4 kbps produces better sound quality than does the LPC-10e. The IMBE was adopted as the Inmarsat-M coding standard for satellite voice communication at a total rate of 6.4 kbps, including 4.15 kbps of source coding and 2.25 kbps of channel coding [104]. The Advanced MBE (AMBE) coder was adopted as the Inmarsat Mini-M standard at a 4.8 kbps total data rate, including 3.6 kbps of speech and 1.2 kbps of channel coding [18,27]. In [14] an enhanced multiband excitation (EMBE) coder was presented. The distinguishing features of the EMBE coder include signal-adaptive multimode spectral modeling and parameter quantization, a two-band signal-adaptive

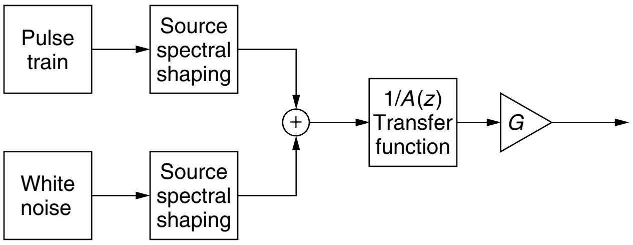

Figure 11. The MELP speech synthesis model.

frequency-domain voicing decision, a novel VQ scheme for the efficient encoding of the variable-dimension spectral magnitude vectors at low rates, and multiclass selective protection of spectral parameters from channel errors. The 4-kbps EMBE coder accounts for both source (2.9 kbps) and channel (1.1 kbps) coding and was designed for satellite-based communication systems.

5.4. Prototype Waveform Interpolative (PWI) Coding

A different kind of coding technique that has properties of both waveform and LPC-based coders has been proposed [59,60] and is called *prototype waveform interpolation* (PWI). PWI uses both interpolation in the frequency domain and forward–backward prediction in the time domain. The technique is based on the assumption that, for voiced speech, a perceptually accurate speech signal can be reconstructed from a description of the waveform of a single, representative pitch cycle per interval of 20–30 ms. The assumption exploits the fact that voiced speech can be interpreted as a concentration of slowly evolving pitch cycle waveforms. The prototype waveform is described by a set of linear prediction (LP) filter coefficients describing the formant structure and a prototype excitation waveform, quantized with analysis-by-synthesis procedures. The speech signal is reconstructed by filtering an excitation signal consisting of the concatenation of (infinitesimal) sections of the instantaneous excitation waveforms. By coding the voiced and unvoiced components separately, a 2.4-kbps version of the coder performed similarly to the 4.8-kbps FS1016 standard [61].

Recent work has aimed at reducing the computational complexity of the coder for rates between 1.2 and 2.4 kbps by including a time-varying waveform sampling rate and a cubic B-spline waveform representation [62,86].

6. MEASURES OF SPEECH QUALITY

Deciding on an appropriate measurement of quality is one of the most difficult aspects of speech coder design, and is an area of current research and standardization. Early military speech coders were judged according to only one criterion: intelligibility. With the advent of consumer-grade speech coders, intelligibility is no longer a sufficient condition for speech coder acceptability. Consumers want speech that sounds "natural." A large number of subjective and objective measures have been developed to quantify "naturalness," but it must be stressed that any scalar measurement of "naturalness" is an oversimplification. "Naturalness" is a multivariate quantity, including such factors as the metallic versus breathy quality of speech, the presence of noise, the color of the noise (narrowband noise tends to be more annoying than wideband noise, but the parameters that predict "annoyance" are not well understood), the presence of unnatural spectral envelope modulations (e.g., flutter noise), and the absence of natural spectral envelope modulations.

6.1. Psychophysical Measures of Speech Quality (Subjective Tests)

The final judgment of speech coder quality is the judgment made by human listeners. If consumers (and reviewers) like the way the product sounds, then the speech coder is a success. The reaction of consumers can often be predicted to a certain extent by evaluating the reactions of experimental listeners in a controlled psychophysical testing paradigm. Psychophysical tests (often called "subjective tests") vary depending on the quantity being evaluated, and the structure of the test.

6.1.1. Intelligibility. Speech coder intelligibility is evaluated by coding a number of prepared words, asking listeners to write down the words they hear, and calculating the percentage of correct transcriptions (an adjustment for guessing may be subtracted from the score). The diagnostic rhyme test (DRT) and diagnostic alliteration test (DALT) are intelligibility tests which use a controlled vocabulary to test for specific types of intelligibility loss [101,102]. Each test consists of 96 pairs of confusable words spoken in isolation. The words in a pair differ in only one distinctive feature, where the distinctive feature dimensions proposed by Voiers are voicing, nasality, sustention●, sibilation, graveness, and compactness. In the DRT, the words in a pair differ in only one distinctive feature of the initial consonant; for instance, "jest" and "guest" differ in the sibilation of the initial consonant. In the DALT, words differ in the final consonant; for instance, "oaf" and "oath" differ in the graveness of the final consonant. Listeners hear one of the words in each pair, and are asked to select the word from two written alternatives. Professional testing firms employ trained listeners who are familiar with the speakers and speech tokens in the database, in order to minimize test-retest variability. Q5

Intelligibility scores quoted in the speech coding literature often refer to the composite results of a DRT. In a comparison of two federal standard coders, the LPC 10e algorithm resulted in 90% intelligibility, while the FS-1016 CELP algorithm had 91% intelligibility [64]. An evaluation of waveform interpolative (WI) coding published DRT scores of 87.2% for the WI algorithm, and 87.7% for FS-1016 [61].

6.1.2. Numerical Measures of Perceptual Quality. Perhaps the most commonly used speech quality measure is the mean opinion score (MOS). A mean opinion score is computed by coding a set of spoken phrases using a variety of coders, presenting all of the coded speech together with undegraded speech in random order, asking listeners to rate the quality of each phrase on a numerical scale, and then averaging the numerical ratings of all phrases coded by a particular coder. The five-point numerical scale is associated with a standard set of descriptive terms: 5 = excellent, 4 = good, 3 = fair, 2 = poor, and 1 = bad. A rating of 4 is supposed to correspond to standard toll-quality speech, quantized at 64 kbps using ITU standard G.711 [48].

Mean opinion scores vary considerably depending on background noise conditions; for example, CVSD performs significantly worse than LPC-based methods in quiet recording conditions, but significantly better under extreme noise conditions [96]. Gender of the speaker may also affect the relative ranking of coders [96]. Expert listeners tend to give higher rankings to speech coders

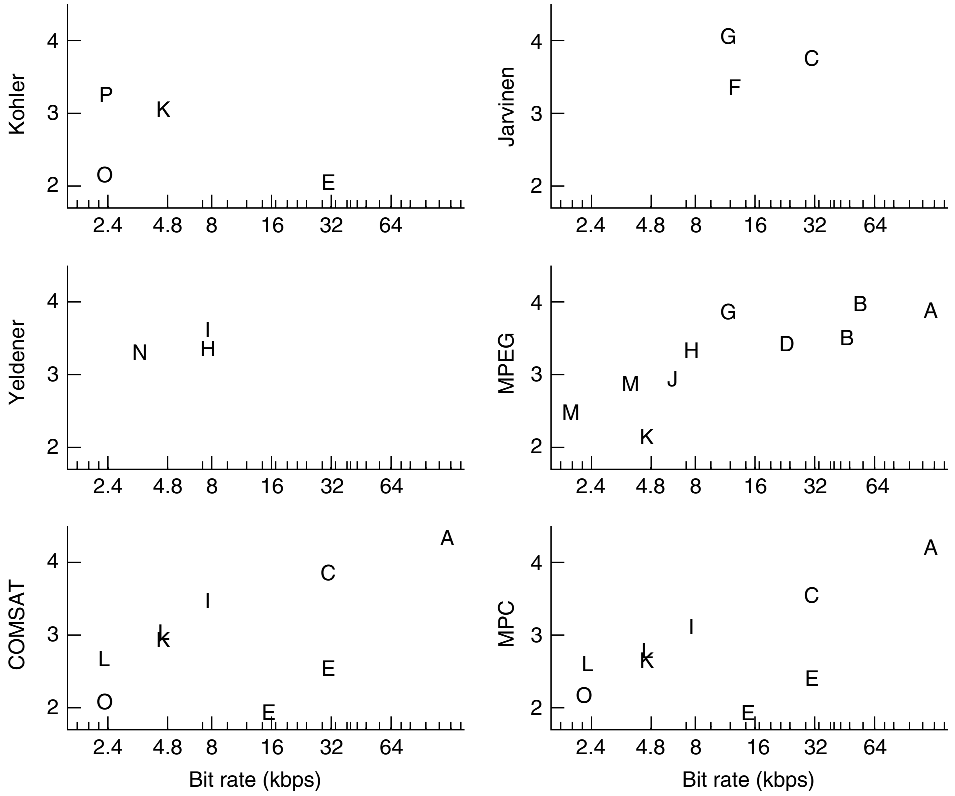

Figure 12. Mean opinion scores from five published studies in quiet recording conditions — Jarvinen [53], Kohler [64], MPEG [39], Yeldener [107], and the COMSAT and MPC sites from Tardelli et al. [96]: (A) unmodified speech, (B) ITU G.722 subband ADPCM, (C) ITU G.726 ADPCM, (D) ISO MPEG-II layer 3 subband audio coder, (E) DDVPC CVSD, (F) GSM full-rate RPE-LTP, (G) GSM EFR ACELP, (H) ITU G.729 ACELP, (I) TIA IS54 VSELP, (J) ITU G.723.1 MPLPC, (K) DDVPC FS-1016 CELP, (L) sinusoidal transform coding, (M) ISO MPEG-IV HVXC, (N) Inmarsat mini-M AMBE, (O) DDVPC FS-1015 LPC-10e, (P) DDVPC MELP.

with which they are familiar, even when they are not consciously aware of the order in which coders are presented [96]. Factors such as language and location of the testing laboratory may shift the scores of all coders up or down, but tend not to change the rank order of individual coders [39]. For all of these reasons, a serious MOS test must evaluate several reference coders in parallel with the coder of interest, and under identical test conditions. If an MOS test is performed carefully, intercoder differences of approximately 0.15 opinion points may be considered significant. Figure 12 is a plot of MOS as a function of bit rate for coders evaluated under quiet listening conditions in five published studies (one study included separately tabulated data from two different testing sites [96]).

The diagnostic acceptability measure (DAM) is an attempt to control some of the factors that lead to variability in published MOS scores [100]. The DAM employs trained listeners, who rate the quality of standardized test phrases on 10 independent perceptual scales, including six scales that rate the speech itself (fluttering, thin, rasping, muffled, interrupted, nasal), and four scales that rate the background noise (hissing, buzzing, babbling, rumbling). Each of these is a 100-point scale, with a range of approximately 30 points between the LPC-10e algorithm (50 points) and clean speech (80 points) [96]. Scores on the various perceptual scales are combined into a composite quality rating. DAM scores are useful for pointing out specific defects in a speech coding algorithm. If the only desired test outcome is a relative quality ranking of multiple coders, a carefully controlled MOS test in which all coders of interest are tested under the same conditions may be as reliable as DAM testing [96].

6.1.3. Comparative Measures of Perceptual Quality. It is sometimes difficult to evaluate the statistical significance of a reported MOS difference between two coders. A more powerful statistical test can be applied if coders are evaluated in explicit A/B comparisons. In a comparative test, a listener hears the same phrase coded by two different coders, and chooses the one that sounds better. The result of a comparative test is an apparent preference score, and an estimate of the significance of the observed preference; for example, in a 1999 study, WI coding at 4.0 kbps was preferred to 4 kbps HVXC 63.7% of the time, to 5.3 kbps G.723.1 57.5% of the time (statistically significant differences), and to 6.3 kbps G.723.1 53.9% of the time (not statistically significant) [29]. It should be noted that "statistical significance" in such a test refers only to the probability that the same listeners listening to the same waveforms will show the same preference in a future test.

6.2. Algorithmic Measures of Speech Quality (Objective Measures)

Psychophysical testing is often inconvenient; it is not possible to run psychophysical tests to evaluate every proposed adjustment to a speech coder. For this reason, a number of algorithms have been proposed that approximate, to a greater or lesser extent, the results of psychophysical testing.

The signal-to-noise ratio of a frame of N speech samples starting at sample number n may be defined as

$$\mathrm{SNR}(n) = \frac{\sum_{m=n}^{n+N-1} s^2(m)}{\sum_{m=n}^{n+N-1} e^2(m)} \tag{49}$$

High-energy signal components can mask quantization error, which is synchronous with the signal component, or separated by at most a few tens of milliseconds. Over longer periods of time, listeners accumulate a general perception of quantization noise, which can be modeled as the average log segmental SNR:

$$\mathrm{SEGSNR} = \frac{1}{K}\sum_{k=0}^{K-1} 10\log_{10}\mathrm{SNR}(kN) \tag{50}$$

High-amplitude signal components tend to mask quantization error components at nearby frequencies and times. A high-amplitude spectral peak in the speech signal is able to mask quantization error components at the same frequency, at higher frequencies, and to a much lesser extent, at lower frequencies. Given a short-time speech spectrum $S(e^{j\omega})$, it is possible to compute a short-time "masking spectrum" $M(e^{j\omega})$ which describes the threshold energy at frequency ω below which noise components are inaudible. The perceptual salience of a noise signal $e(n)$ may be estimated by filtering the noise signal into K different subband signals $e_k(n)$, and computing the ratio between the noise energy and the masking threshold in each subband:

$$\mathrm{NMR}(n,k) = \frac{\sum_{m=n}^{n+N-1} e_k^2(m)}{\int_{\omega_k}^{\omega_{k+1}} |M(e^{j\omega})|^2\,d\omega} \tag{51}$$

where ω_k is the lower edge of band k, and ω_{k+1} is the upper band edge. The band edges must be close enough together that all of the signal components in band k are effective in masking the signal $e_k(n)$. The requirement of effective masking is met if each band is exactly one Bark in width, where the Bark frequency scale is described in many references [71,77].

Fletcher has shown that the perceived loudness of a signal may be approximated by adding the cube roots of the signal power in each one-bark subband, after properly accounting for masking effects [20]. The total loudness of a quantization noise signal may therefore be approximated as

$$\mathrm{NMR}(n) = \sum_{k=0}^{K-1}\left(\frac{\sum_{m=n}^{n+N-1} e_k^2[m]}{\int_{\omega_k}^{\omega_{k+1}} |M(e^{j\omega})|^2\,d\omega}\right)^{1/3} \tag{52}$$

The ITU perceptual speech quality measure (PSQM) computes the perceptual quality of a speech signal by filtering the input and quantized signals using a Bark-scale filterbank, nonlinearly compressing the amplitudes in each band, and then computing an average subband signal to noise ratio [51]. The development of algorithms that accurately predict the results of MOS or comparative testing is an area of active current research, and a number of improvements, alternatives, and/or extensions to the PSQM measure have been proposed. An algorithm that has been the focus of considerable research activity is the Bark spectral distortion measure [73,103,105,106]. The ITU has also proposed an extension of the PSQM standard called perceptual evaluation of speech quality (PESQ) [81], which will be released as ITU standard P.862.

7. NETWORK ISSUES

7.1. Voiceover IP

Speech coding for the voiceover Internet Protocol (VOIP) application is becoming important with the increasing dependency on the Internet. The first VoIP standard was published in 1998 as recommendation H.323 [52] by the International Telecommunications Union (ITU-T). It is a protocol for multimedia communications over local area networks using packet switching, and the voice-only subset of it provides a platform for IP-based telephony. At high bit rates, H.323 recommends the coders G.711 (3.4 kHz at 48, 56, and 64 kbps) and G.722 (wideband speech and music at 7 kHz operating at 48, 56, and 64 kbps) while at the lower bit rates G.728 (3.4 kHz at 16 kbps), G.723 (5.3 and 6.5 kbps), and G.729 (8 kbps) are recommended [52].

In 1999, a competing and simpler protocol named the Session Initiation Protocol (SIP) was developed by the Internet Engineering Task Force (IETF) Multiparty Multimedia Session Control working group and published as RFC 2543 [19]. SIP is a signaling protocol for Internet conferencing and telephony, is independent of the packet layer, and runs over UDP or TCP although it supports more protocols and handles the associations between Internet end systems. For now, both systems will coexist but it is predicted that the H.323 and SIP architectures will evolve such that two systems will become more similar.

Speech transmission over the Internet relies on sending "packets" of the speech signal. Because of network congestion, packet loss can occur, resulting in audible artifacts. High-quality VOIP, hence, would benefit from variable-rate source and channel coding, packet loss concealment, and jitter buffer/delay management. These are challenging issues and research efforts continue to generate high-quality speech for VOIP applications [38].

7.2. Embedded and Multimode Coding

When channel quality varies, it is often desirable to adjust the bit rate of a speech coder in order to match the channel capacity. Varying bit rates are achieved in one of two ways. In multimode speech coding, the transmitter and the receiver must agree on a bit rate prior to transmission of the coded bits. In embedded source coding, on the other hand, the bitstream of the coder operating at low bit rates is embedded in the bitstream of the coder operating at higher rates. Each increment in bit rate provides marginal improvement in speech quality. Lower bit rate coding is obtained by puncturing bits from the higher rate coder and typically exhibits graceful degradation in quality with decreasing bit rates.

ITU Standard G.727 describes an embedded ADPCM coder, which may be run at rates of 40, 32, 24, or 16 kbps (5, 4, 3, or 2 bits per sample) [46]. Embedded ADPCM algorithms are a family of variable bit rate coding algorithms operating on a sample per sample basis (as opposed to, e.g., a subband coder that operates on a frame-by-frame basis) that allows for bit dropping after encoding. The decision levels of the lower-rate quantizers are subsets of those of the quantizers at higher rates. This allows for bit reduction at any point in the network without the need of coordination between the transmitter and the receiver.

The prediction in the encoder is computed using a more coarse quantization of $\hat{d}(n)$ than the quantization actually transmitted. For example, 5 bits per sample may be transmitted, but as few as 2 bits may be used to reconstruct $\hat{d}(n)$ in the prediction loop. Any bits not used in the prediction loop are marked as "optional" by the signaling channel mode flag. If network congestion disrupts traffic at a router between sender and receiver, the router is allowed to drop optional bits from the coded speech packets.

Embedded ADPCM algorithms produce codewords that contain enhancement and core bits. The feedforward (FF) path of the codec utilizes both enhancement bits and core bits, while the feedback (FB) path uses core bits only. With this structure, enhancement bits can be discarded or dropped during network congestion.

An important example of a multimode coder is QCELP, the speech coder standard that was adopted by the TIA North American digital cellular standard based on code-division multiple access (CDMA) technology [9]. The coder selects one of four data rates every 20 ms depending on the speech activity; for example, background noise is coded at a lower rate than speech. The four rates are approximately 1 kbps (eighth rate), 2 kbps (quarter rate), 4 kbps (half rate), and 8 kbps (full rate). QCELP is based on the CELP structure but integrates implementation of the different rates, thus reducing the average bit rate. For example, at the higher rates, the LSP parameters are more finely quantized and the pitch and codebook parameters are updated more frequently [23]. The coder provides good quality speech at average rates of 4 kbps.

Another example of a multimode coder is ITU standard G.723.1, which is an LPC-AS coder that can operate at 2 rates: 5.3 or 6.3 kbps [50]. At 6.3 kbps, the coder is a multipulse LPC (MPLPC) coder while the 5.3-kbps coder is an algebraic CELP (ACELP) coder. The frame size is 30 ms with an additional lookahead of 7.5 ms, resulting in a total algorithmic delay of 67.5 ms. The ACELP and MPLPC coders share the same LPC analysis algorithm and frame/subframe structure, so that most of the program code is used by both coders. As mentioned earlier, in ACELP, an algebraic transformation of the transmitted index produces the excitation signal for the synthesizer. In MPLPC, on the other hand, minimizing the perceptual-error weighting is achieved by choosing the amplitude and position of a number of pulses in the excitation signal. Voice activity detection (VAD) is used to reduce the bit rate during silent periods, and switching from one bit rate to another is done on a frame-by-frame basis.

Multimode coders have been proposed over a wide variety of bandwidths. Taniguchi et al. proposed a multimode ADPCM coder at bit rates between 10 and 35 kbps [94]. Johnson and Taniguchi proposed a multimode CELP algorithm at data rates of 4.0–5.3 kbps in which additional stochastic codevectors are added to the LPC excitation vector when channel conditions are sufficiently good to allow high-quality transmission [55]. The European Telecommunications Standards Institute (ETSI) has recently proposed a standard for adaptive multirate coding at rates between 4.75 and 12.2 kbps.

7.3. Joint Source-Channel Coding

In speech communication systems, a major challenge is to design a system that provides the best possible speech quality throughout a wide range of channel conditions. One solution consists of allowing the transceivers to monitor the state of the communication channel and to dynamically allocate the bitstream between source and channel coding accordingly. For low-SNR channels, the source coder operates at low bit rates, thus allowing powerful forward error control. For high-SNR channels, the source coder uses its highest rate, resulting in high speech quality, but with little error control. An adaptive algorithm selects a source coder and channel coder based on estimates of channel quality in order to maintain a constant total data rate [95]. This technique is called *adaptive multirate* (AMR) coding, and requires the simultaneous implementation of an AMR source coder [24], an AMR channel coder [26,28], and a channel quality estimation algorithm capable of acquiring information about channel conditions with a relatively small tracking delay.

The notion of determining the relative importance of bits for further unequal error protection (UEP) was pioneered by Rydbeck and Sundberg [83]. Rate-compatible channel codes, such as Hagenauer's rate compatible punctured convolutional codes (RCPC) [34], are a collection of codes providing a family of channel coding rates. By puncturing bits in the bitstream, the channel coding rate of RCPC codes can be varied instantaneously, providing UEP by imparting on different segments different degrees of protection. Cox et al. [13] address the issue of channel coding and illustrate how RCPC codes can be used to build a speech transmission scheme for mobile radio channels. Their approach is based on a subband coder with dynamic bit allocation proportional to the average energy of the bands. RCPC codes are then used to provide UEP.

Relatively few AMR systems describing source and channel coding have been presented. The AMR systems [99,98,75,44] combine different types of variable rate CELP coders for source coding with RCPC and cyclic redundancy check (CRC) codes for channel coding and were presented as candidates for the European Telecommunications Standards Institute (ETSI) GSM AMR codec standard. In [88], UEP is applied to perceptually based audio coders (PAC). The bitstream of the PAC is divided into two classes and punctured convolutional codes are used to provide different levels of protection, assuming a BPSK constellation.

In [5,6], a novel UEP channel encoding scheme is introduced by analyzing how symbol-wise puncturing of symbols in a trellis code and the rate-compatibility constraint (progressive puncturing pattern) can be used to derive rate-compatible punctured trellis codes (RCPT). While conceptually similar to RCPC codes, RCPT codes are specifically designed to operate efficiently on large constellations (for which Euclidean and Hamming distances are no longer equivalent) by maximizing the residual Euclidean distance after symbol puncturing. Large constellation sizes, in turn, lead to higher throughput and spectral efficiency on high SNR channels. An AMR system is then designed based on a perceptually-based embedded subband encoder. Since perceptually based dynamic bit allocations lead to a wide range of bit error sensitivities (the perceptually least important bits being almost insensitive to channel transmission errors), the channel protection requirements are determined accordingly. The AMR systems utilize the new rate-compatible channel coding technique (RCPT) for UEP and operate on an 8-PSK constellation. The AMR-UEP system is bandwidth efficient, operates over a wide range of channel conditions and degrades gracefully with decreasing channel quality.

Systems using AMR source and channel coding are likely to be integrated in future communication systems since they have the capability for providing graceful speech degradation over a wide range of channel conditions.

8. STANDARDS

Standards for landline public switched telephone service (PSTN) networks are established by the International Telecommunication Union (ITU) (*http://www.itu.int*). The ITU has promulgated a number of important speech and waveform coding standards at high bit rates and with very low delay, including G.711 (PCM), G.727 and G.726 (ADPCM), and G.728 (LDCELP). The ITU is also involved in the development of internetworking standards, including the voiceover IP standard H.323. The ITU has developed one widely used low-bit-rate coding standard (G.729), and a number of embedded and multimode speech coding standards operating at rates between 5.3 kbps (G.723.1) and 40 kbps (G.727). Standard G.729 is a speech coder operating at 8 kbps, based on algebraic code-excited LPC (ACELP) [49,84]. G.723.1 is a multimode coder, capable of operating at either 5.3 or 6.3 kbps [50]. G.722 is a standard for wideband speech coding, and the ITU will announce an additional wideband standard within the near future. The ITU has also published standards for the objective estimation of perceptual speech quality (P.861 and P.862).

The ITU is a branch of the International Standards Organization (ISO) (*http://www.iso.ch*). In addition to ITU activities, the ISO develops standards for the Moving Picture Experts Group (MPEG). The MPEG-2 standard included digital audiocoding at three levels of complexity, including the layer 3 codec commonly known as MP3 [72]. The MPEG-4 motion picture standard includes a structured audio standard [40], in which speech and audio “objects” are encoded with header information specifying the coding algorithm. Low-bit-rate speech coding is performed using harmonic vector excited coding (HVXC) [43] or code-excited LPC (CELP) [41], and audiocoding is performed using time–frequency coding [42]. The MPEG homepage is at *drogo.cselt.stet.it/mpeg*.

Standards for cellular telephony in Europe are established by the European Telecommunications Standards Institute (ETSI) (*http://www.etsi.org*). ETSI speech coding standards are published by the Global System for Mobile Telecommunications (GSM) subcommittee. All speech coding standards for digital cellular telephone use are based on LPC-AS algorithms. The first GSM standard coder was based on a precursor of CELP called *regular-pulse excitation with long-term prediction* (RPE-LTP) [37,65]. Current GSM standards include the enhanced full-rate codec GSM 06.60 [32,53] and the adaptive multirate codec [33]; both standards use algebraic code-excited LPC (ACELP). At the time of writing, both ITU and ETSI are expected to announce new standards for wideband speech coding in the near future. ETSI’s standard will be based on GSM AMR.

The Telecommunications Industry Association (*http://www.tiaonline.org*) published some of the first U.S. digital cellular standards, including the vector-sum-excited LPC (VSELP) standard IS54 [25]. In fact, both the initial U.S. and Japanese digital cellular standards were based on the VSELP algorithm. The TIA has been active in the development of standard TR41 for voiceover IP.

The U.S. Department of Defense Voice Processing Consortium (DDVPC) publishes speech coding standards for U.S. government applications. As mentioned earlier, the original FS-1015 LPC-10e standard at 2.4 kbps [8,16], originally developed in the 1970s, was replaced in 1996 by the newer MELP standard at 2.4 kbps [92]. Transmission at slightly higher bit rates uses the FS-1016 CELP (CELP) standard at 4.8 kbps [17,56,57]. Waveform applications use the continuously variable slope delta modulator (CVSD) at 16 kbps. Descriptions of all DDVPC standards and code for most are available at *http://www.plh.af.mil/ddvpc/index.html*.

9. FINAL REMARKS

In this article, we presented an overview of coders that compress speech by attempting to match the time waveform as closely as possible (waveform coders), and coders that attempt to preserve perceptually relevant spectral properties of the speech signal (LPC-based and subband coders). LPC-based coders use a speech production model to parameterize the speech signal, while subband coders filter the signal into frequency bands and assign bits by either an energy or perceptual criterion. Issues pertaining to networking, such as voiceover IP and joint source–channel coding, were also touched on. There are several other coding techniques that we have not discussed in this article because of space limitations. We hope to have provided the reader with an overview of the fundamental techniques of speech compression.

Acknowledgments

This research was supported in part by the NSF and HRL. We thank Alexis Bernard and Tomohiko Taniguchi for their suggestions on earlier drafts of the article.

BIBLIOGRAPHY

1. J.-P. Adoul, P. Mabilleau, M. Delprat, and S. Morisette, Fast CELP coding based on algebraic codes, *Proc. ICASSP*, 1987, pp. 1957–1960.
2. B. S. Atal, Predictive coding of speech at low bit rates, *IEEE Trans. Commun.* **30**: 600–614 (1982).
3. B. S. Atal, High-quality speech at low bit rates: Multi-pulse and stochastically excited linear predictive coders, *Proc. ICASSP*, 1986, pp. 1681–1684.
4. B. S. Atal and J. R. Remde, A new model of LPC excitation for producing natural-sounding speech at low bit rates, *Proc. ICASSP*, 1982, pp. 614–617.
5. A. Bernard, X. Liu, R. Wesel, and A. Alwan, Channel adaptive joint-source channel coding of speech, *Proc. 32nd Asilomar Conf. Signals, Systems, and Computers*, 1998, Vol. 1, pp. 357–361.
6. A. Bernard, X. Liu, R. Wesel, and A. Alwan, Embedded joint-source channel coding of speech using symbol puncturing of trellis codes, *Proc. IEEE ICASSP*, 1999, Vol. 5, pp. 2427–2430.
7. M. S. Brandstein, P. A. Monta, J. C. Hardwick, and J. S. Lim, A real-time implementation of the improved MBE speech coder, *Proc. ICASSP*, 1990, Vol. 1: pp. 5–8.
8. J. P. Campbell and T. E. Tremain, Voiced/unvoiced classification of speech with applications to the U.S. government LPC-10E algorithm, *Proc. ICASSP*, 1986, pp. 473–476.
9. *CDMA, Wideband Spread Spectrum Digital Cellular System Dual-Mode Mobile Station-Base Station Compatibility Standard*, Technical Report Proposed EIA/TIA Interim Standard, Telecommunications Industry Association TR45.5 Subcommittee, 1992.
10. J.-H. Chen et al., A low delay CELP coder for the CCITT 16 kb/s speech coding standard, *IEEE J. Select. Areas Commun.* **10**: 830–849 (1992).
11. J.-H. Chen and A. Gersho, Adaptive postfiltering for quality enhancement of coded speech, *IEEE Trans. Speech Audio Process.* **3**(1): 59–71 (1995).
12. R. Cox et al., New directions in subband coding, *IEEE JSAC* **6**(2): 391–409 (Feb. 1988).
13. R. Cox, J. Hagenauer, N. Seshadri, and C. Sundberg, Subband speech coding and matched convolutional coding for mobile radio channels, *IEEE Trans. Signal Process.* **39**(8): 1717–1731 (Aug. 1991).
14. A. Das and A. Gersho, Low-rate multimode multiband spectral coding of speech, *Int. J. Speech Tech.* **2**(4): 317–327 (1999).
15. G. Davidson and A. Gersho, Complexity reduction methods for vector excitation coding, *Proc. ICASSP*, 1986, pp. 2055–2058.
16. DDVPC, *LPC-10e Speech Coding Standard*, Technical Report FS-1015, U.S. Dept. of Defense Voice Processing Consortium, Nov. 1984.
17. DDVPC, *CELP Speech Coding Standard*, Technical Report FS-1016, U.S. Dept. of Defense Voice Processing Consortium, 1989.
18. S. Dimolitsas, Evaluation of voice coded performance for the Inmarsat Mini-M system, *Proc. 10th Int. Conf. Digital Satellite Communications*, 1995.
19. M. Handley et al., *SIP: Session Initiation Protocol*, IETF RFC, March 1999, *http://www.cs.columbia.edu/hgs/sip/sip.html*.
20. H. Fletcher, *Speech and Hearing in Communication*, Van Nostrand, Princeton, NJ, 1953.
21. D. Florencio, Investigating the use of asymmetric windows in CELP vocoders, *Proc. ICASSP*, 1993, Vol. II, pp. 427–430.
22. S. Furui, *Digital Speech Processing, Synthesis, and Recognition*, Marcel Dekker, New York, 1989.
23. W. Gardner, P. Jacobs, and C. Lee, QCELP: A variable rate speech coder for CDMA digital cellular, in B. Atal, V. Cuperman, and A. Gersho, eds., *Speech and Audio Coding for Wireless and Network Applications*, Kluwer, Dordrecht, The Netherlands, 1993, pp. 85–93.
24. A. Gersho and E. Paksoy, An overview of variable rate speech coding for cellular networks, *IEEE Int. Conf. Selected Topics in Wireless Communications Proc.*, June 1999, pp. 172–175.
25. I. Gerson and M. Jasiuk, Vector sum excited linear prediction (VSELP), in B. S. Atal, V. S. Cuperman, and A. Gersho, eds., *Advances in Speech Coding*, Kluwer, Dordrecht, The Netherlands, 1991, pp. 69–80.
26. D. Goeckel, Adaptive coding for time-varying channels using outdated fading estimates, *IEEE Trans. Commun.* **47**(6): 844–855 (1999).
27. R. Goldberg and L. Riek, *A Practical Handbook of Speech Coders*, CRC Press, Boca Raton, FL, 2000.
28. A. Goldsmith and S. G. Chua, Variable-rate variable power MQAM for fading channels, *IEEE Trans. Commun.* 1218–1230 ●(1997). Q2
29. O. Gottesman and A. Gersho, Enhanced waveform interpolative coding at 4 kbps, *Proc. ICASSP*, Phoenix, AZ, ●1999. Q3
30. K. Gould, R. Cox, N. Jayant, and M. Melchner, Robust speech coding for the indoor wireless channel, *ATT Tech. J.* 64–73 ●(1993). Q2
31. D. W. Griffn and J. S. Lim, Multi-band excitation vocoder, *IEEE Trans. Acoust. Speech, Signal Process.* **36**(8): 1223–1235 (1988).
32. Special Mobile Group (GSM), *Digital Cellular Telecommunications System: Enhanced Full Rate (EFR) Speech Transcoding*, Technical Report GSM 06.60, European Telecommunications Standards Institute (ETSI), 1997.
33. Special Mobile Group (GSM), *Digital Cellular Telecommunications System (Phase 2+): Adaptive Multi-rate (AMR) Speech Transcoding*, Technical Report GSM 06.90, European Telecommunications Standards Institute (ETSI), 1998.
34. J. Hagenauer, Rate-compatible punctured convolutional codes and their applications, *IEEE Trans. Commun.* **36**(4): 389–400 (1988).
35. J. C. Hardwick and J. S. Lim, A 4.8 kbps multi-band excitation speech coder, *Proc. ICASSP*, 1988, Vol. 1, pp. 374–377.

36. M. Hasegawa-Johnson, Line spectral frequencies are the poles and zeros of a discrete matched-impedance vocal tract model, *J. Acoust. Soc. Am.* **108**(1): 457–460 (2000).

37. K. Hellwig et al., Speech codec for the european mobile radio system, *Proc. IEEE Global Telecomm. Conf.*, 1989.

38. O. Hersent, D. Gurle, and J.-P. Petit, *IP Telephony*, Addison-Wesley, Reading, MA, 2000.

39. ISO, *Report on the MPEG-4 Speech Codec Verification Tests*, Technical Report JTC1/SC29/WG11, ISO/IEC, Oct. 1998.

40. ISO/IEC, *Information Technology — Coding of Audiovisual Objects, Part 3: Audio, Subpart 1: Overview*, Technical Report ISO/JTC 1/SC 29/N2203, ISO/IEC, 1998.

41. ISO/IEC, *Information Technology — Coding of Audiovisual Objects, Part 3: Audio, Subpart 3: CELP*, Technical Report ISO/JTC 1/SC 29/N2203CELP, ISO/IEC, 1998.

42. ISO/IEC, *Information Technology — Coding of Audiovisual Objects, Part 3: Audio, Subpart 4: Time/Frequency Coding*, Technical Report ISO/JTC 1/SC 29/N2203TF, ISO/IEC, 1998.

43. ISO/IEC, *Information Technology — Very Low Bitrate*• Q6 Q2 *Audio-Visual Coding, Part 3: Audio, Subpart 2: Parametric Coding*, Technical Report ISO/JTC 1/SC 29/N2203PAR, ISO/IEC, 1998.

44. H. Ito, M. Serizawa, K. Ozawa, and T. Nomura, An adaptive multi-rate speech codec based on mp-celp coding algorithm for etsi amr standard, *Proc. ICASSP*, 1998, Vol. 1, pp. 137–140.

45. ITU-T, *40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)*, Technical Report G.726, International Telecommunications Union, Geneva, 1990.

46. ITU-T, *5-, 4-, 3- and 2-bits per Sample Embedded Adaptive Differential Pulse Code Modulation (ADPCM)*, Technical Report G.727, International Telecommunications Union, Geneva, 1990.

47. ITU-T, *Coding of Speech at 16 kbit/s Using Low-Delay Code Excited Linear Prediction*, Technical Report G.728, International Telecommunications Union, Geneva, 1992.

48. ITU-T, *Pulse Code Modulation (PCM) of Voice Frequencies*, Technical Report G.711, International Telecommunications Union, Geneva, 1993.

49. ITU-T, *Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)*, Technical Report G.729, International Telecommunications Union, Geneva, 1996.

50. ITU-T, *Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s*, Technical Report G.723.1, International Telecommunications Union, Geneva, 1996.

51. ITU-T, *Objective Quality Measurement of Telephone-Band (300-3400 Hz) speech codecs*, Technical Report P.861, International Telecommunications Union, Geneva, 1998.

52. ITU-T, *Packet Based Multimedia Communications Systems*, Technical Report H.323, International Telecommunications Union, Geneva, 1998.

53. K. Jarvinen et al., GSM enhanced full rate speech codec, *Proc. ICASSP*, 1997, pp. 771–774.

54. N. Jayant, J. Johnston, and R. Safranek, Signal compression based on models of human perception, *Proc. IEEE* **81**(10): 1385–1421 (1993).

55. M. Johnson and T. Taniguchi, Low-complexity multi-mode VXC using multi-stage optimization and mode selection, *Proc. ICASSP*, 1991, pp. 221–224.

56. J. P. Campbell Jr., T. E. Tremain, and V. C. Welch, The DOD 4.8 KBPS standard (proposed federal standard 1016), in B. S. Atal, V. C. Cuperman, and A. Gersho, ed., *Advances in Speech Coding*, Kluwer, Dordrecht, The Netherlands, 1991, pp. 121–133.

57. J. P. Campbell, Jr., V. C. Welch, and T. E. Tremain, An expandable error-protected 4800 BPS CELP coder (U.S. federal standard 4800 BPS voice coder), *Proc. ICASSP*, 1989, 735–738.

58. P. Kabal and R. Ramachandran, The computation of line spectral frequencies using chebyshev polynomials, *IEEE Trans. Acoust. Speech, Signal Process.* **ASSP-34**: 1419–1426 (1986).

59. W. Kleijn, Speech coding below 4 kb/s using waveform interpolation, *Proc. GLOBECOM* 1991, Vol. 3, pp. 1879–1883.

60. W. Kleijn and W. Granzow, Methods for waveform interpolation in speech coding, *Digital Signal Process.* 215–230 •(1991).

61. W. Kleijn and J. Haagen, A speech coder based on decomposition of characteristic waveforms, *Proc. ICASSP*, 1995, pp. 508–511.

62. W. Kleijn, Y. Shoham, D. Sen, and R. Hagen, A low-complexity waveform interpolation coder, *Proc. ICASSP*, 1996, pp. 212–215.

63. W. B. Kleijn, D. J. Krasinski, and R. H. Ketchum, Improved speech quality and efficient vector quantization in SELP, *Proc. ICASSP*, 1988, pp. 155–158.

64. M. Kohler, A comparison of the new 2400bps MELP federal standard with other standard coders, *Proc. ICASSP*, 1997, pp. 1587–1590.

65. 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. ASSP* **34**: 1054–1063 (1986).

66. W. LeBlanc, B. Bhattacharya, S. Mahmoud, and V. Cuperman, Efficient search and design procedures for robust multi-stage VQ of LPC parameters for 4kb/s speech coding, *IEEE Trans. Speech Audio Process.* **1**: 373–385 (1993).

67. D. Lin, New approaches to stochastic coding of speech sources at very low bit rates, in I. T. Young et al., ed., *Signal Processing III: Theories and Applications*, Elsevier, Amsterdam, 1986, pp. 445–447.

68. A. McCree and J. C. De Martin, A 1.7 kb/s MELP coder with improved analysis and quantization, *Proc. ICASSP*, 1998, Vol. 2, pp. 593–596.

69. A. McCree et al., A 2.4 kbps MELP coder candidate for the new U.S. Federal standard, *Proc. ICASSP*, 1996, Vol. 1, pp. 200–203.

70. A. V. McCree and T. P. Barnwell, III, A mixed excitation LPC vocoder model for low bit rate speech coding, *IEEE Trans. Speech Audio Processing* **3**(4): 242–250 (1995).

71. B. C. J. Moore, *An Introduction to the Psychology of Hearing*, Academic Press, San Diego, (1997).

72. P. Noll, MPEG digital audio coding, *IEEE Signal Process. Mag.* 59–81 •(1997). Q2

73. B. Novorita, Incorporation of temporal masking effects into bark spectral distortion measure, *Proc. ICASSP*, Phoenix, AZ, 1999, pp. 665–668.

74. E. Paksoy, W-Y. Chan, and A. Gersho, Vector quantization of speech LSF parameters with generalized product codes, *Proc. ICASSP*, 1992, pp. 33–36.

75. E. Paksoy et al., An adaptive multi-rate speech coder for digital cellular telephony, *Proc. of ICASSP*, 1999, Vol. 1, pp. 193–196.

76. K. K. Paliwal and B. S. Atal, Efficient vector quantization of LPC parameters at 24 bits/frame, *IEEE Trans. Speech Audio Process.* **1**: 3–14 (1993).

77. L. Rabiner and B.-H Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffis, NJ, 1993.

78. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, 1978.

79. R. P. Ramachandran and P. Kabal, Stability and performance analysis of pitch filters in speech coders, *IEEE Trans. ASSP* **35**(7): 937–946 (1987).

80. V. Ramamoorthy and N. S. Jayant, Enhancement of ADPCM speech by adaptive post-filtering, *AT&T Bell Labs. Tech. J.* 1465–1475 •(1984).

81. A. Rix, J. Beerends, M. Hollier, and A. Hekstra, Pesq — the new ITU standard for end-to-end speech quality assessment, *AES 109th Convention*, Los Angeles, CA, Sept. 2000.

82. R. C. Rose and T. P. Barnwell, III, The self-excited vocoder — an alternate approach to toll quality at 4800 bps, *Proc. ICASSP*, •1986.

83. N. Rydbeck and C. E. Sundberg, Analysis of digital errors in non-linear PCM systems, *IEEE Trans. Commun.* **COM-24**: 59–65 (1976).

84. R. Salami et al., Design and description of CS-ACELP: A toll quality 8 kb/s speech coder, *IEEE Trans. Speech Audio Process.* **6**(2): 116–130 (1998).

85. M. R. Schroeder and B. S. Atal, Code-excited linear prediction (CELP): High-quality speech at very low bit rates, *Proc. ICASSP*, 1985, pp. 937–940.

86. Y. Shoham, Very low complexity interpolative speech coding at 1.2 to 2.4 kbp, *Proc. ICASSP*, 1997, pp. 1599–1602.

87. S. Singhal and B. S. Atal, Improving performance of multi-pulse LPC coders at low bit rates, *Proc. ICASSP*, 1984, pp. 1.3.1–1.3.4.

88. D. Sinha and C.-E. Sundberg, Unequal error protection methods for perceptual audio coders, *Proc. ICASSP*, 1999, Vol. 5, pp. 2423–2426.

89. F. Soong and B.-H. Juang, Line spectral pair (LSP) and speech data compression, *Proc. ICASSP*, 1984, pp. 1.10.1–1.10.4.

90. J. Stachurski, A. McCree, and V. Viswanathan, High quality MELP coding at bit rates around 4 kb/s, *Proc. ICASSP*, 1999, Vol. 1, pp. 485–488.

91. N. Sugamura and F. Itakura, Speech data compression by LSP speech analysis-synthesis technique, *Trans. IECE*, **J64-A**(8): 599–606 (1981) (in Japanese).

92. L. Supplee, R. Cohn, and J. Collura, MELP: The new federal standard at 2400 bps, *Proc. ICASSP*, 1997, pp. 1591–1594.

93. B. Tang, A. Shen, A. Alwan, and G. Pottie, A perceptually-based embedded subband speech coder, *IEEE Trans. Speech Audio Process.* **5**(2): 131–140 (March 1997).

94. T. Taniguchi, ADPCM with a multiquantizer for speech coding, *IEEE J. Select. Areas Commun.* **6**(2): 410–424 (1988).

95. T. Taniguchi, F. Amano, and S. Unagami, Combined source and channel coding based on multimode coding, *Proc. ICASSP*, 1990, pp. 477–480.

96. J. Tardelli and E. Kreamer, Vocoder intelligibility and quality test methods, *Proc. ICASSP*, 1996, pp. 1145–1148.

97. I. M. Trancoso and B. S. Atal, Efficient procedures for finding the optimum innovation in stochastic coders, *Proc. ICASSP*, 1986, pp. 2379–2382.

Q2 98. A. Uvliden, S. Bruhn, and R. Hagen, Adaptive multi-rate. A speech service adapted to cellular radio network quality, *Proc. 32nd Asilomar Conf.*, 1998, Vol. 1, pp. 343–347.

99. J. Vainio, H. Mikkola, K. Jarvinen, and P. Haavisto, GSM EFR based multi-rate codec family, *Proc. ICASSP*, 1998, Vol. 1, pp. 141–144.

Q3 100. W. D. Voiers, Diagnostic acceptability measure for speech communication systems, *Proc. ICASSP*, 1977, pp. 204–207.

101. W. D. Voiers, Evaluating processed speech using the diagnostic rhyme test, *Speech Technol.* **1**(4): 30–39 (1983).

102. W. D. Voiers, Effects of noise on the discriminability of distinctive features in normal and whispered speech, *J. Acoust. Soc. Am.* **90**: 2327 (1991).

103. S. Wang, A. Sekey, and A. Gersho, An objective measure for predicting subjective quality of speech coders, *IEEE J. Select. Areas Commun.* 819–829 •(1992). Q2

104. S. W. Wong, An evaluation of 6.4 kbps speech codecs for Inmarsat-M system, *Proc. ICASSP*, •1991. Q3

105. W. Yang, M. Benbouchta, and R. Yantorno, Performance of the modified bark spectral distortion measure as an objective speech quality measure, *Proc. ICASSP*, 1998, pp. 541–544.

106. W. Yang and R. Yantorno, Improvement of MBSD by scaling noise masking threshold and correlation analysis with MOS difference instead of MOS, *Proc. ICASSP*, Phoenix, AZ, 1999, pp. 673–676.

107. S. Yeldener, A 4 kbps toll quality harmonic excitation linear predictive speech coder, *Proc. ICASSP*, 1999, pp. 481–484.

Please clarify the following queries:

Author Query 1: There are two tables represented for table no. 2. We have chosen the table on page 27. So, Please clarify us to set either on page 27 or the attached one.
Author Query 2: vol. no. ?
Author Query 3: pages ?
Author Query 4: plosive or explosive? pls. clarify.
Author Query 5: sustension or sibilation? Pls. clarify.
Author Query 6: Bitrate. Is it okay. Pls clarify