Speech Coding in Mobile Radio Communications

MADHUKAR BUDAGAVI AND JERRY D. GIBSON, FELLOW, IEEE

Speech coding, the efficient representation of speech in digital form, is one of the key technologies in current and evolving digital cellular and wireless voice communications offerings. The speech coders in existing standards exhibit a level of sophistication and performance unimaginable just 15 years ago. We outline the characteristics of the mobile communications problem with respect to speech coders and point out the principal issues in speech-coder design for these applications. Speech-coding methods in existing mobile communications standards are described and contrasted. The limitations imposed by the wireless channel and by background impairments are discussed, and approaches to addressing their resulting effects are presented. Suggestions for future research in speech coding for the mobile communications problem are outlined.

Keywords*—*Digital cellular, speech-coding standards.

I. INTRODUCTION

Advances in modulation methods, error-control coding, multiple access techniques, equalization algorithms, and digital signal processing (DSP) have led to extraordinary new wireless communications systems and services. Separately, speech-coding methods have evolved over the last 30 years such that toll-quality speech can be achieved at rates down to 8 kbits/s with single DSP-based implementations. Coupled together, these advances in digital communications and speech coding have facilitated the offering of wireless mobile voice communications throughout the world.

This paper addresses speech-coding techniques and algorithms for mobile voice communications, including an overview of currently important (and developing) speech-coding standards. Although wireless speech-coding standards are described, this paper is not offered as an exhaustive survey of speech-coding algorithms and standards. Several exceptional surveys and tutorials on speech-coding standards and methods have been published in recent years [1]–[4], along with several books [5]–[7]. Rather, we present a general view of how these speech coders approach the speech-coding problem and thus illuminate how and why they work. Building on this background, we then outline the relevant criteria and issues in the design and performance evaluation of speech coders for mobile communications, and present our view of how speech coders need to evolve to offer maximum flexibility and performance for mobile applications.

The outline of this paper is as follows. We begin with a high-level description of the mobile communications problem, with the goal of explaining the important parameters and issues as they pertain to speech-coder design and performance. We then present a view of the speech-coding problem that allows the roles of the component functions and parameters of the subsequently described speech coders to be better understood.

We then present a brief discussion of wireline speech-coding standards, a few details on low-tier personal communications systems (PCS's), and a development of full- and half-rate standards for digital cellular systems throughout the world. Sufficient details are presented that the reader can contrast the approach of each coder and identify important tradeoffs. The innovations that have allowed digital cellular and mobile voice communications to achieve the current state of the art are described next, followed by an examination of the limitations on the performance of present-day systems. We conclude with a view of possible new directions for research and development in speech coding for mobile applications.

II. THE MOBILE COMMUNICATIONS PROBLEM

Broadly, speech coding for mobile communications has some of the same goals as speech coding for wireline transmission; namely, the objective is to achieve highly intelligible, high-quality speech at the given bit rate with acceptable complexity. Delay, which is sometimes important in wired telephony, receives less emphasis in the mobile environment since there is already substantial delay due to time division multiplexing, interleaving, and error-control coding/decoding. Complexity is an issue in mobile communications because of the need to conserve battery power, and hence maintain moderate-weight handsets, while still having acceptable talk time between battery recharging. Complexity of the speech coder may become less of an issue, however, since as much as two-thirds of

Manuscript received November 20, 1997; revised February 7, 1998. This research was supported in part by the National Science Foundation under Grant NCR-9 796 255.

M. Budagavi is with the Department of Electrical Engineering, Texas A&M University, College Station, TX 77840 USA.

J. D. Gibson is with the Department of Electrical Engineering, Southern Methodist University, Dallas, TX 75275-0338 USA.

Publisher Item Identifier S 0018-9219(98)04240-6.

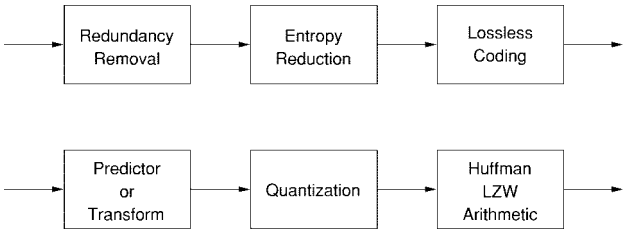

Fig. 1. Major steps in data compression.

the battery power in today's handsets is consumed by radio-frequency (RF) transmission and signal processing, and the expectation is that this may increase to as much as 90% by the third-generation systems.

Speech coders must exhibit acceptable performance in the presence of expected bit error rates or other undesirable channel behaviors no matter what the application. However, the mobile radio channel, with its propensity to experience deep (and perhaps rapidly varying) fades, presents quite a challenge to the communications system designer. The most direct effects of the rather poor quality channel on the speech coder are, first, to impose additional robustness requirements, and second, to demand a substantial fraction of the available channel bit rate for error-control coding.

As the data rate available for speech transmission, as opposed to error-control coding, decreases and the speech coders become more speech-model based, the effects of background noise of various kinds become significant. Specifically, if a speech coder incorporates an explicit, parsimonious speech model and expects only speech to be present at its input, natural background sounds can be turned into unnatural-sounding artifacts. Such performance can make the speech coder completely unacceptable to users, regardless of other positive traits.

Tandem performance with speech coders in other wireless systems, with speech coders in wireline systems, with conference bridges, and with voice-mail systems are all issues. This is because several different speech coders may be used in these series connections, and each of these may contribute a different type of distortion. The subsequent decoding/reencoding stages required by the tandem connection may result in seriously degraded performance as compared to performance over any one of the stages alone. In many applications, this tandem performance is not carefully evaluated.

III. The Speech-Coding Problem

A general source compression problem is often described as the three-step process illustrated in Fig. 1 [7]. Here, we see that a source first undergoes redundancy removal (such as linear prediction or a unitary transform), then entropy reduction (some form of quantization), and last lossless coding (such as Huffman, Lempel–Ziv, arithmetic, or simple fixed-length coding). All speech coders do not have all three of these operations. For example, G.711 log pulse code modulation (PCM) consists only of quantization (entropy reduction) followed by fixed-length lossless coding.

Adaptive differential (AD)PCM, as shown in Fig. 2, includes a redundancy removal step as implemented by linear prediction around the quantizer. The linear prediction or redundancy removal decreases the dynamic range of the signal into the quantizer, thus leading to lower distortion for the same bit rate compared to quantization and lossless coding alone—or, alternatively, a lower transmitted bit rate for a fixed distortion. ADPCM is an example of a time-domain waveform-following coder, and at the bit rates it is usually applied, it has the property that background sounds at the coder input are usually reproduced at the decoder output without unnatural-sounding artifacts. Furthermore, ADPCM has low complexity and low delay, the latter of which is given some importance in wireline telephony.

To develop the role of linear prediction in speech coding, it is instructive to view the decoder in the ADPCM system depicted in Fig. 2 as a speech synthesizer, with the quantized prediction error signal serving as the excitation and the feedback linear prediction structure shown functioning as a model of the vocal tract. It is this view that motivated the early work in linear predictive coding (LPC). To see the roles of the excitation and the linear prediction vocal tract model, we consider the voiced segment of speech shown in Fig. 3(a). In Fig. 3(b), we plot the speech spectrum obtained by taking an FFT of the windowed speech in Fig. 3(a), which is the very spikey signal in the figure. The multiple sharp peaks in Fig. 3(b) are harmonics of the speaker fundamental excitation or pitch harmonics. The smoother plot shown in Fig. 3(b) is the frequency response of an eighth-order linear prediction (LP) synthesizer, with the same structure as the decoder in Fig. 2. Here, the LP coefficients have been determined using a standard autocorrelation linear prediction calculation as in LPC and code-excited linear prediction (CELP) coders.

The peaks in the smooth plot are called formants. Since the smooth plot is the frequency response of the decoder and the decoder is a vocal tract model, we surmise by the process of elimination that the excitation must provide the pitch harmonics if we are to reproduce the speech spectrum well.

Using these analysis ideas, we can investigate the performance of G.726 ADPCM at 32 kbits/s. Shown in Fig. 3(c) is the spectrum of the reconstructed speech at the 32 kbits/s G.726 decoder output. Here, we see that the ADPCM coder provides a rather nice reproduction of the input speech spectral content. Thus, the quantized prediction error signal, or residual, does a good job of introducing the pitch harmonics, often called the spectral fine structure. Notice, however, the difference in the spectral envelope between Fig. 3(b) and (c). The eighth-order linear predictor frequency response in Fig. 3(c) has deemphasized spectral peaks relative to Fig. 3(b). Thus, while the ADPCM encoding has preserved the spectral fine structure fairly well, it has introduced distortions in the reconstructed speech that affect the spectral envelope.

The fundamental concept that is the basis for most linear predictive wireline and wireless speech coders today is analysis-by-synthesis. That is, for a given segment of

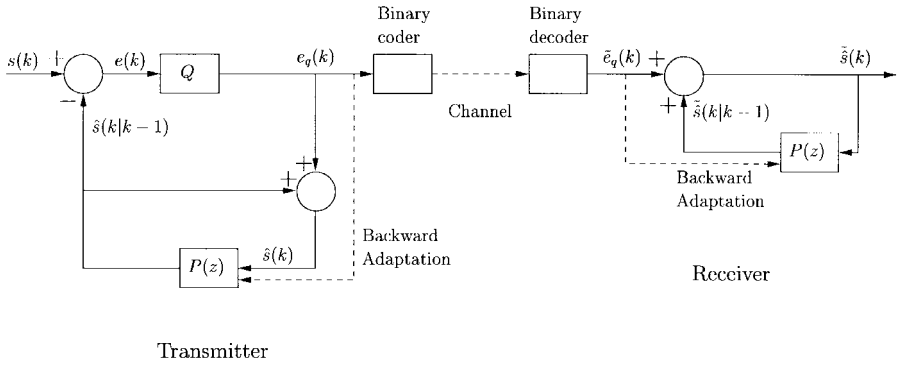

Fig. 2. Classic differential PCM.

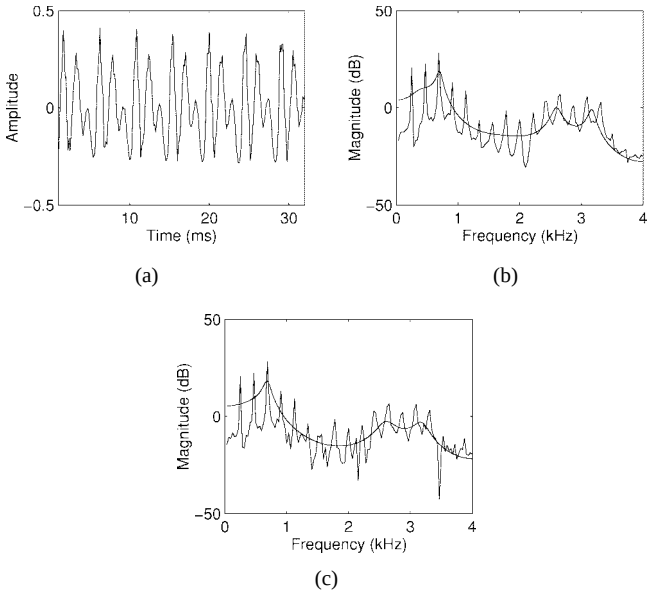

Fig. 3. LP analysis of clean and ADPCM encoded speech.

speech to be encoded, the encoding is accomplished by comparing this input segment to a representative set of segments stored in, or generated from, what is called a codebook. Once the best match is found, it is only necessary to transmit the codebook index and any associated model parameters. This is a very efficient coded representation of the segment. Complexity is an issue, however, since the analysis-by-synthesis approach implies that for each segment to be encoded, it is necessary to synthesize many versions of this segment in order to find the best match. Notice how distinctly different this approach is from PCM or ADPCM, where each sample is encoded as it is received and there is no explicit calculation of distortion induced during the encoding stage.

The analysis-by-synthesis paradigm is inherent in source-coding techniques motivated by information theory and rate distortion theory [12], and researchers in this field were among the first to invoke this method for speech coding. However, essential to the success of analysis-by-synthesis is the specification of an appropriate fidelity criterion or distortion measure, and the promise of the method for speech coding was firmly established by Atal and Schroeder [25] with their inclusion of a perceptually weighted distortion method within this paradigm. Since this early work, researchers have addressed the major issues of codebook design, search complexity, and perceptual distortion measures to the extent that today, we have an array of speech-coding standards based upon linear prediction analysis-by-synthesis (LPAS), most of which carry the designation CELP or a closely related variation of this nomenclature.

IV. Relevant Wireline Speech-Coding Standards

Early speech-coding standards were developed to meet the requirements of voice and multimedia terminals in wired telecommunications networks. Thus, much of the research in speech coding focused on meeting the requirements of wired networks. This research has had a significant impact on the speech-coding standards for wireless networks. Almost all of the current wireless speech-coding standards are based on ideas and algorithms that have their origins in speech-coding research for wired networks. We discuss these speech-coding standards briefly before discussing those for wireless networks [4], [7], [17].

One of the earliest speech coders standardized was the ADPCM-based G.721 coder operating at 32 kbits/s and now part of the G.726 standard. It was developed to double the capacity of standard telephone networks, which then used the 64 kbits/s A-law or μ-law quantized PCM (ITU G.711). The G.726 speech codec at 32 kbits/s achieves high quality, typically referred to as toll quality, with relatively low complexity. The G.726 codec is also used in low-tier PCS/cordless applications and is discussed in more detail in Section V.

The toll-quality speech-coding standard at 16 kbits/s is the low-delay CELP-based G.728 coder. G.728 uses backward-adaptive linear prediction and a short-block-length excitation signal (five-sample) to achieve low delay. Low delay is often specified in bidirectional networks to avoid echo problems.

More recently, the International Telecommunications Union has standardized two speech coders, G.729 and G.723.1, to operate at low bit rates. G.729 operates at 8 kbits/s and provides network-quality speech. It is based on the conjugate structure algebraic CELP (CS-ACELP) algorithm. The enhanced speech coders for digital cellular,

also recently standardized, use ACELP as well. Since G.729 is the first standard to be based on ACELP, it is discussed in more detail later. The G.723.1 standard specifies a dual-rate speech codec with rates of 6.3 and 5.3 kbits/s. It was designed for low-bit-rate video-telephony applications. It is also based on CELP, with the main difference in coding for the two rates coming from the type of fixed codebook used. The high-rate coder uses a multipulse maximum likelihood quantizer (MP-MLQ)-based fixed codebook excitation, whereas the low-rate coder uses an ACELP-based codebook.

The G.729 8-kbits/s coder is based on LPAS [4]. Processing of speech is done using 10-ms frames. LPC analysis is performed every frame to generate ten line spectral frequency parameters, which are quantized using a two-stage vector quantizer (VQ) and transmitted. Each speech frame is also split into two subframes, and the excitation processing is done every subframe (40 samples). The long-term predictor or pitch filter is implemented using the adaptive codebook approach. Fractional pitch delays of one-third sample time resolution are used. The adaptive codebook is searched using a hybrid open-loop/closed-loop approach. The search range for the best adaptive codebook delay is restricted to be around the candidate delay obtained using open-loop analysis. This reduces the complexity of the adaptive codebook search.

The fixed codebook used is an ACELP codebook, which has a special algebraic structure and provides advantages in terms of reduced search complexity and storage. The algebraic codebook is designed using an interleaved single-pulse permutation (ISPP) technique, which basically populates the codebook sparsely in a structured manner. The pulse amplitudes are restricted to be either $+1$ or -1. Each 40-sample code vector in the ACELP codebook is partitioned into four subsets, and a single pulse is selected from each subset. Thus, each code vector has four pulses. The code word corresponding to the selected code vector consists of the pulse locations and signs, thus obviating the need for storage of the codebook. The sparsity of the code vector reduces the number of computations by requiring fewer terms to be calculated in finding the correlation and energy terms involved in the search process. By performing the search in four nested loops and by changing only one pulse position at a time, this computational load is further reduced. A focused search is further employed to do a reduced codebook search (2–5% of full codebook search) but at the same time achieving speech quality close to the that obtained by a full search. The gains for the excitation vectors are vector quantized.

The G.729 coder was designed originally for PCS and the future public land mobile telecommunications system but is important in other applications as well.

V. Speech-Coding Standards for Low-Tier PCS/Cordless

Cordless and low-tier PCS systems such as CT-2 cordless telephony, digital European cordless telecommunication (DECT), U.S. enhanced personal wireless telecommunications (PWT-E), U.S. personal access communications services (PACS), and Japanese personal handyphone system (PHS) all make use of the G.726 32-kbits/s ADPCM coder [18], [19]. The predictor used is a backward adaptive pole-zero predictor with six zeroes and two poles. The prediction error is quantized using 4 bits/sample, giving a rate of 32 kbits/s for the input speech sampling rate of 8 kHz. The quantizer used is also backward adaptive and switches between two speeds of adaptation depending on the nature of the input (speech or voice-band data) signal.

The 32-kbits/s G.726 coder is the popular choice for low-tier PCS because of the requirements of low power (low complexity) and high quality in cordless/low-tier PCS applications and the fact that, as a waveform coder, background impairments are not changed into annoying artifacts.

VI. Speech-Coding Standards for Digital Cellular

In this section, we give an overview of some of the speech-coding standards used in digital cellular systems. Almost all of the digital cellular speech-coding standards are based on LPAS coding. The various coders significantly differ in the way they process the excitation to the LP synthesis filter. Efficient ways of representing the excitation and for searching the optimal excitation have played a major part in improving the efficiency of the speech coders. Hence, the emphasis in this section is on describing the various techniques used in the coders to process the excitation. Efficient techniques for coding the LP parameters, such as the use of line spectral pairs and vector quantization, have also played an important part in improving the coding efficiency of these coders.

In all of these coders, processing of speech is accomplished on a frame-by-frame basis, with LP analysis being done once every frame (typically 20 ms). The frame of speech is split into subframes, and excitation processing is performed every subframe (typically 5 ms).

A. Full-Rate Speech Coders

The full-rate speech coders standardized for the European, North American, and Japanese markets are described in this section. The bit rates of these coders are in the range of 6.7–13 kbits/s [20]–[22].

1) Global System for Mobile (GSM) RPE-LTP: The GSM full-rate speech coder, which has been standardized for the pan-European digital cellular system, operates at 13 kbits/s. It uses a single-tap long-term predictor (LTP) to exploit the pitch periodicity in the residual and the regular pulse excitation (RPE) technique to model the excitation signal [8]–[10]. RPE consists of modeling the excitation sequence using uniformly spaced pulses.

In every subframe, the LP residual is first calculated, and the lag and gain parameter of the LTP are computed and coded using seven and two bits, respectively. The LTP residual is obtained by filtering the LP residual using the LTP filter. The LTP residual is coded using a simplified RPE

technique, which consists of downsampling the residual by a factor of three to get four subsequences with varying phases. The subsequence with the largest energy is chosen as the excitation. The selected subsequence is coded with block adaptive PCM using three bits to quantize each normalized pulse amplitude and six bits for the normalizing factor.

Note that in RPE-LTP, excitation processing is done in an open-loop fashion. It was found in [8]–[10] that under fixed perceptual weighting filter assumptions, maximizing the residual energy provided performance that was very near to that provided by the optimal regular excitation based on closed-loop analysis-by-synthesis. Use of open-loop processing drastically reduces the computational requirements and thus the power requirements for the mobile stations. Voice activity detection is also used in the GSM cellular system to conserve mobile power [6], [7].

2) IS-54 VSELP: The original IS-54 full-rate speech coder for North American time division multiple access (TDMA) digital cellular is based on the vector sum excited linear prediction (VSELP) algorithm [6], [7]. The data rate of the coder is 7.95 kbits/s. The distinguishing characteristic of VSELP coding is the way in which the codebook vectors are generated. A code vector in VSELP is generated by taking a linear combination of a set of basis vectors. The linear combinations are constrained to be either additions or subtractions of the basis vectors. The codebook index for generating the code vector at the decoder is thus a binary vector specifying the polarity of the linear combination of the basis vectors. The advantage of using such a structured codebook is that computational complexity is reduced in the codebook search, as only the basis vectors (which are relatively very few) have to be passed through the synthesis filter. Error minimization can then be carried out using linear combinations of the filtered basis vectors (this is in contrast to having to filter every code vector when using an unstructured codebook). Another advantage is the improved error robustness, since a single bit error in the VSELP codebook index alters only one term in the linear combination used in generating the code vector.

The excitation in the IS-54 VSELP coder is generated as a weighted sum of three components from the following codebooks: the *adaptive* codebook, which implements the long-term or pitch synthesis filter, and two 7-bit VSELP codebooks. The adaptive codebook contribution is first found using a closed-loop search with integer lags. The contribution of the first VSELP codebook is then found, also by a closed-loop search, but with the codebook orthogonalized with respect to the selected adaptive codebook contribution. The second VSELP codebook is orthogonalized to both the selected adaptive codebook contribution and the selected first VSELP codevector, and a closed-loop search is used to find the optimal code vector from the second VSELP codebook. Orthogonalization facilitates the joint optimization of the selection of gains and the excitation components. Orthogonalization is computationally feasible in VSELP since only the basis vectors of the codebooks have to be orthogonalized. In general, orthogonalization is not practical if it has to be done for every code vector of the codebook.

The gains of the selected code vectors are quantized using a two-stage process, with the average speech energy being quantized once per frame in the first stage and the gain parameters being transformed and vector quantized once every subframe in the second stage.

3) Japanese Digital Cellular (JDC) VSELP: The JDC full-rate coder for use in the Japanese TDMA digital cellular system is similar to the IS-54 VSELP coder, the main difference being that only one VSELP codebook is used instead of the two used in IS-54 VSELP. The JDC VSELP codebook consists of nine basis vectors, thus requiring nine bits to transmit the codebook index.

4) IS-641 Enhanced Full-Rate Speech Codec: The IS-641 enhanced full-rate speech codec is for use in North American IS-136 TDMA cellular systems. The quality of the full-rate IS-54 VSELP coder is less than that of analog FM in strong signal coverage areas. Hence, a replacement codec with toll quality was envisaged. IS-641 is the resulting enhanced codec. The IS-641 coder structure is very similar to that of G.729, with a primary difference being an increase in frame size. The frame size in IS-641 is 20 ms, and short-term parameters, transmitted once per frame, are thus updated at a lower rate. This leads to a bit savings of 0.6 kbits/s. The data rate of the IS-641 speech coder is 7.4 kbits/s.

B. Half-Rate Speech Coders

Many TDMA cellular and PCS systems have planned from the beginning for the incorporation of half-rate speech coders, when they become available, and the need for half-rate speech coders has been necessitated by the rapid growth in demand for cellular services in some regions [23]. Due to the limited spectrum allocated for the cellular systems, this demand could be satisfied only by decreasing the bit rate used for speech coding. Half-rate speech coders operate at half the *gross* bit rates of those of the full-rate standards, thereby allowing double the number of users. They have roughly the same quality, in terms of mean opinion score (MOS), as that provided by the full-rate standards. Half-rate standards for GSM and personal digital cellular (PDC) have been standardized and are discussed in this subsection. In both of these coders, the reduction in the bit rate, with the goal of roughly the same quality, has come at the price of greatly increased complexity.

1) GSM VSELP: The half-rate GSM speech coder has a bit rate of 5.6 kbits/s. It is also based on the VSELP algorithm. Thus, only some important improvements in the half-rate coder are highlighted below.

The half-rate coder uses multimode coding, where the coder components and the bit allocation are adaptively changed to match the local characteristics of the speech. Four modes are defined based on the voicing activity present in the frame. Open-loop pitch prediction gains are used to select the modes. When the pitch prediction gain is low, indicating unvoiced speech, Mode 0 is used. In Mode 0, the excitation signal is formed as a linear combination of

two vectors, one each from two 7-bit VSELP codebooks. The adaptive codebook contribution is not used. In the other modes—Modes 1, 2, 3—the excitation signal is formed as a linear combination of code vectors from the adaptive codebooks and a 9-bit VSELP codebook. The gains for the code vectors are vector quantized using codebooks specific to the mode selected.

Another improvement incorporated in the half-rate coder is in the long-term predictor. The LTP uses subsample lags in the adaptive codebook to improve the quality of processed speech, as well as a hybrid open-loop/closed-loop adaptive codebook search to reduce computational requirements. The hybrid open-loop/closed-loop search uses a *frame lag trajectory* approach, where the sequence of subframe lags within a frame (a frame lag trajectory) is first globally optimized in an open-loop fashion over all the subframes in the frame and then a closed-loop search is employed to refine the lag estimates in each subframe. The deviations in the frame lag trajectories are restricted so that a delta lag encoding scheme, wherein the first subframe's lags are encoded independently and the subsequent frame's lags are differentially encoded, can be used. The use of differential encoding to improve the coding efficiency is based on the assumption that during voiced speech, there is a high degree of correlation in lags from subframe to subframe.

2) PDC Pitch Synchronous Innovation (PSI)-CELP: The PDC half-rate speech coder is based on the pitch synchronous innovation CELP (PSI-CELP) algorithm and has a bit rate of 3.45 kbits/s. The PSI-CELP coder achieves the same quality as that of the full-rate coder. The frame size used is 40 ms, which is double that used in most of the other digital cellular speech-coding standards; the corresponding subframe (excitation vector) size used is 10 ms. The increase in the excitation vector size is required so as to achieve the quantization efficiency required at the low bit rate. With the increase in the excitation vector size, the possibility of several pitch components in the excitation vector arises for voiced speech. It was found that the adaptive codebook is unable to remove pitch periodicity in the excitation vector completely, leading to residual periodicity. Traditional random codebooks do not model this residual periodicity effectively. Thus, to overcome this problem, pitch periodicity is introduced in the random codebooks, i.e., the innovation (excitation) is made pitch synchronous. Another difference between PSI-CELP and regular CELP is that a fixed codebook is used in conjunction with the adaptive codebook.

The excitation in the PSI-CELP coder is generated as a weighted sum of two components. The first component is obtained either from the adaptive codebook (ACB) or from the fixed codebook (FCB). The use of the FCB allows the coder to model unvoiced and transient parts of speech more effectively. The second component is obtained from a random codebook with a two-channel conjugate structure, where codebook vectors are generated as a sum of two vectors—one each from each of the two conjugate codebooks. The two-channel conjugate codebook is more robust to transmission errors, as a single bit error in the codeword affects the contribution of only one of the two codebooks. This is in contrast to when a single codebook is used, where a single bit error in the code word leads to a completely different code vector. Pitch synchronization is introduced in the random codebook only if the ACB is used in the first component of the excitation. Introduction of the pitch periodicity improves the quality of speech when it is predominantly voiced. Both the fixed and random codebooks are rotation codebooks, i.e., the code vectors are generated by rotating a reduced number of stored code vectors.

Open-loop preselection is used in the search process to reduce the number of codevectors searched closed-loop in the adaptive, fixed, and random codebooks. Orthogonalization is also adopted in the PSI-CELP coder. The random component of the excitation is orthogonalized with respect to the ACB/FCB component before the search. A delayed decision scheme is applied to excitation and gain code searches to find better code combinations. While searching for the best ACB/FCB vector, the second best choice is also retained. For each of the choices, the best random code vector and the gain codes are found. The combination that results in lower error is finally selected.

C. Variable-Rate Speech Coders

In variable-rate speech coding, the rate of the speech coder is matched to the time-varying local character of the speech signal, and/or changed based on the control inputs from the system that is used to transmit the encoded speech. In general, variable-bit-rate coding (based on source characteristics) can reduce, on the average, the bit rate required to transmit speech when compared to fixed-rate speech coders at roughly the same quality. This reduction in bit rate can be translated to an increase in capacity in code division multiple access (CDMA)-based digital cellular systems. Thus, in the North American digital cellular system based on CDMA technology, variable-bit-rate coding is exclusively used. In this section, we describe the two variable-bit-rate speech-coding standards, IS-96 QCELP and IS-127 enhanced variable rate (EVR) codec, for use in CDMA digital cellular systems.

1) IS-96B QCELP Coder: The IS-96B coder supports four bit rates—8, 4, 2, and 0.8 kbits/s—referred to as Rate 1, 1/2, 1/4, and 1/8 coders. The QCELP coder integrates all four rates with a scalable CELP architecture. Scalability in the bits used for transmitting the LP filter parameters (LSP's) is achieved by using a different number of bits to quantize the LSP's based on the rate of coding. Bit scalability of excitation parameters is achieved by using larger subframes (lower update rates) for lower rates. For Rate 1/8 coding, the adaptive codebook is not used, and the fixed codebook is replaced by a noise generator. The excitation search is done in a closed-loop fashion over both the codebook vectors and gains. The fixed codebook used is a circular codebook, which reduces the search complexity to a great extent.

Automatic rate selection in the QCELP coder is done by comparing the energy level of the frame with a set of three adaptive thresholds based on background noise energy estimates. If the input energy is greater than all the thresholds, the input signal is encoded at full rate; if the energy is less than all three thresholds, the signal is encoded at Rate 1/8; and if the energy is between the thresholds, the intermediate rates are chosen. Rate 1 corresponds to the case where there is active speech in the frame, and Rate 1/8 corresponds to the case where there is no speech and only background noise exists. In practice, these two rates are the ones that are used most frequently.

2) IS-127 EVR Codec: The reconstructed speech quality of the IS-96 QCELP coder is poorer than that of the IS-54 VSELP coder. Hence, the QCELP coder was targeted for replacement. The selected replacement coder is called the EVR coder and is based on the relaxed CELP (RCELP) coding algorithm. The EVR coder supports three rates, 8.5, 4, and 0.8 kbits/s, and these are called Rates 1, 1/2, and 1/8, respectively.

RCELP is based on the generalized analysis-by-synthesis principle [11], where the waveform matching constraints (weighted mean square error minimization) of the conventional analysis-by-synthesis are relaxed to allow for matching of modified speech. The aim of using the modifications is that they should facilitate improvements in coding efficiency without degrading the perceptual quality of reconstructed speech.

The parameter targeted in RCELP for improving the coding efficiency is the pitch period. In RCELP, the pitch delay is determined (and transmitted) only once every frame and is linearly interpolated between these updates to create a synthetic pitch contour. The creation of the synthetic pitch contour obviates the need for transmission of the pitch delays (adaptive codebook lag) every subframe, as is done in traditional CELP. Using the delays in the synthetic pitch contour, the adaptive codebook contribution for any subframe in the frame can be calculated. The adaptive codebook contribution is not optimal for the subframe, however, leading to a large mismatch between the original and synthesized residuals. A way to overcome this problem is to modify the original residual by time warping it such that its pitch contour matches the synthetic pitch contour. In RCELP, this modification is implemented by time shifting (rather than time warping) the original residual to match the adaptive codebook contribution optimally. This is done to reduce the computational burden. The modification of the pitch (within reasonable limits) is found not to degrade the perceived speech quality.

In the IS-127 coder, every frame is split into three subframes, and fixed codebook processing is done once every subframe. The fixed codebook used is an algebraic codebook. Variable-rate operation is obtained by using a 35-bit algebraic codebook for Rate 1 and a 10-bit algebraic codebook for Rate 1/2. For Rate 1/8, as in QCELP, the fixed codebook is replaced by a noise generator, and the adaptive codebook contribution is not used. Variable-rate transmission of LSP parameters is achieved by using split VQ, with the LSP parameters split into a different number of sets depending on the rate of coding used.

VII. Key Innovations and Advances

The current performance of wireless communications systems in terms of output speech quality at the relatively low rates is really quite good, especially considering the quality of the mobile channels. There are several key innovations and advances that have led to the current state of the art. Although most of these have been mentioned in other sections, they may have been masked by the myriad of other details, and hence we highlight them here. We are always "standing on the shoulders of those that have gone before," so we mention only the most recent ideas here.

Starting with CELP and the analysis-by-synthesis paradigm, the need in speech coding was to maintain or improve speech quality, reduce the rate, and control complexity. One important innovation that appears in some of the recent speech coders is fractional pitch. The use of fractional pitch lags increases the pitch resolution and allows better matches with past speech (or output) samples. Almost the same quality improvement can be obtained by employing multiple-tap pitch predictors; however, the multiple tap approach incurs a bit-rate penalty since the additional parameters must be coded and transmitted. Using fractional pitch lags requires increased signal processing but no increase in transmitted bit rate. Another innovation that affects both the complexity and the transmitted bit rate is the discovery that, in a multipulse-like excitation codebook, only a single gain is necessary. This reduces both the search effort and the number of gains to be transmitted. A closely related advance is the interleaved single pulse permutation search technique that enormously reduces the codebook search effort. All three of these innovations appear in G.729, which achieves an MOS of 4.2 at 8 kbits/s.

Modifications in the overall approach have also been advantageous. Silence removal had often been investigated for speech-coding applications over the years, and indeed was important in time alignment speech interpolation applications, but GSM's full-rate coder managed to combine voice activity detection (VAD) and insertion of comfort noise to produce good quality output speech. VAD and comfort noise are now staples in state-of-the-art speech coders. Of course, silence excision automatically implies variable-rate speech coding, and while researchers had investigated such techniques and were making great strides at the same time, the introduction and standardization of QCELP with a very simple energy-threshold-based rate control gave variable-rate speech coders the needed momentum. The energy threshold approach has proved fairly ineffective, but a second key idea in IS-96 is the matching of the variable-rate coding with the multiple access method. That is, since each user interferes with all other users in proportion to transmitted power, the combination of a variable-rate speech coder with CDMA was important to achieve the desired cell capacity.

The use of unequal error protection (UEP) has also been essential to achieving the necessary error protection to channel impairments while keeping the portion of the available transmitted bit rate allocated to error-control coding at a minimum. Again, UEP has become an accepted part of all current speech-coder designs.

VIII. Impairments

Achieving highly intelligible, high-quality speech at bit rates of 8 kbits/s and below with reasonable complexity is a sufficiently difficult problem in itself; however, the mobile or wireless communications problem has other characteristics that significantly add to the challenge. Two of these characteristics are the presence of background noise at the handset microphone input and the unreliability of the mobile channel. Both of these issues have also appeared earlier in the paper, but here we consider their implications in more detail.

A. Channel Effects

The information transmission theorem in information theory [12] says that source coding and channel coding can be performed separately without loss of optimality. The operative word here is "optimality" because if the channel is not being used optimally (that is, if we are using the channel at a nonzero error rate), then overall optimality is unachievable and the information transmission theorem no longer holds. Of course, the mobile radio channels that we use today are quite unreliable, and there is a nonzero bit error rate even after error-control coding. This implies that we should be pursuing joint source-channel coding designs.

At first glance, most of the mobile communications systems perform source and channel coding separately, since they basically split the bit rate between speech coding and error-control coding. The use of UEP at least couples these two efforts, but the speech coder is not jointly designed with the channel coder. A lesser known subtext of the information transmission theorem is that even though separate designs may be optimal in the information theoretic sense, this theorem says nothing about complexity. Hence, a joint source-channel coder that achieves optimal performance may be preferable to a separate design because of complexity.

A major drawback to achieving either optimality or joint source-channel coding designs is the time-varying nature of the wireless channel and the fact that a single end-to-end link may contain several channels of different quality. This latter fact implies that the designer should break down the end-to-end problem into its constituent channels and achieve the best possible error control on each individual channel [24].

Principally, the way mobile systems address the fact that error control is not perfect is really twofold. One way is to make the speech coder robust to channel errors so that channel errors do not produce catastrophic behavior in the reconstructed output. Another way is to use error-concealment techniques, such as replacing corrupted information in the current frame with more reliable (but less recent) information from a previously received frame. Both robustness and error concealment rely heavily on the designer's creativity, experience, and knowledge of human perceptual preferences.

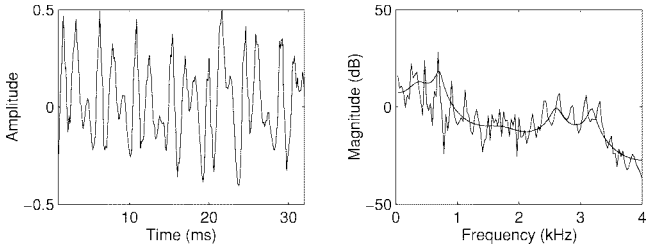

Fig. 4. Analysis of noisy speech.

B. Background Noise

A major challenge to designing low-rate speech coders for mobile radio applications is the presence of background noise. To develop speech coders that produce good quality, highly intelligible speech at bit rates below 16 kbits/s in a quiet environment, it has been necessary to incorporate more knowledge about the speech production model into the coder itself. Thus, the assumption is that, at the speech coder input, only clean speech and only the speech that one desires to be transmitted is present. When there are other sounds or speakers present, the coder treats them as if they are part of the desired speech signal and codes them subject to the assumed speech production model. This can often result in unnatural-sounding artifacts at the decoder output.

For example, simple additive white noise can be turned into harmonic sounds, and other more harmonically rich background sounds can cause wavering or warbling in the output speech. Users are fairly forgiving of natural-sounding background impairments, but artifacts such as those mentioned are unacceptable to the mobile phone user.

To illustrate the effect of additive distortion on speech coder performance, consider the time-domain waveform in Fig. 4(a), which is the speech segment from Fig. 3(a) with additive vehicular noise. Fig. 4(b) shows the spectrum of the noisy speech and the spectral envelope found using LP analysis. If the noisy speech segment in Fig. 4(a) is coded and reconstructed using G.726 at 32 kbits/s, we get the spectrum and LP spectral envelope shown in Fig. 5. If this result is contrasted with Fig. 3(c), which shows G.726 encoding of clean speech in Fig. 3(a), it is evident that the excitation harmonics are still fairly well preserved in Fig. 5.

If we pass the noisy speech in Fig. 4 through a CELP coder, we get the spectrum and LP envelope shown in Fig. 6. Comparing this figure to Figs. 4(b) and 5, there is clearly a loss in harmonic content, and there are additional

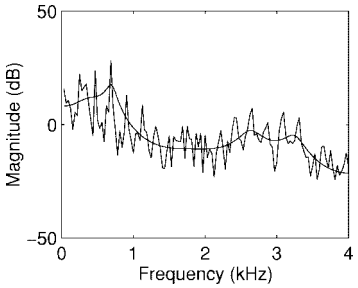

Fig. 5. ADPCM encoded noisy speech and LP analysis.

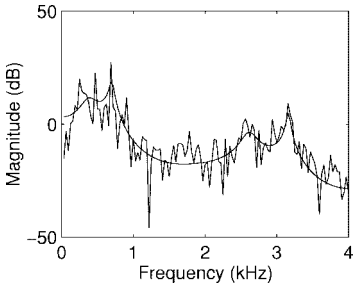

Fig. 6. CELP coded noisy speech and LP analysis.

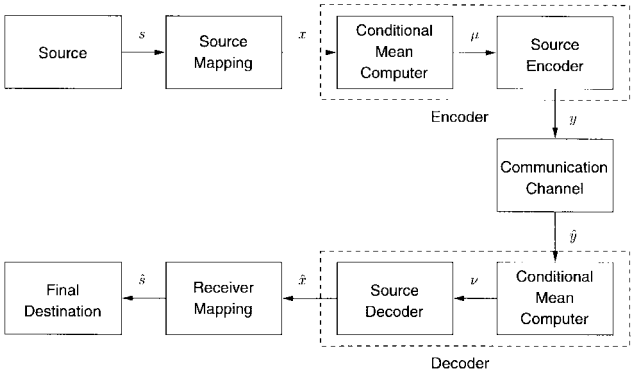

Fig. 7. Communication system for noisy inputs.

artifacts—note the enhancement of the highest formant in Fig. 6. The difficulties faced by the model-based CELP coders when coding noisy speech is that the estimation of the speech model parameters suffers significant degradation. Not only are the estimates of the linear prediction coefficients scaled and biased by the background noise, but inaccuracies in the estimates of the pitch gain and pitch lag may prove even more important.

One approach to reducing background-noise effects has been to utilize an adaptive filter at the speech-coder input, and another approach might be to use multiple microphones and noise cancellation. Wolf and Ziv [16] (see also Berger [12]) have examined the more general problem of coding a source that has been contaminated prior to reaching the source-coder input. The channel is assumed to be ideal, but the authors also allow for the possibility that there is distortion added after the source is decoded at the receiver. Their basic block diagram is shown in Fig. 7. For s and $\hat{s}$ elements of a separable Hilbert space with an inner product denoted as (a, b) and appropriate norm, we define the distortion function

$$d_N(s,\hat{s}) = \frac{1}{N}(\hat{s}-s, \hat{s}-s) = \frac{1}{N}\|\hat{s}-s\|^2 \qquad (1)$$

and the distortion measure as the expectation

$$D_N = E d_N(s,\hat{s}). \qquad (2)$$

The general result for distorted inputs, but no distortion on the output, is that (asymptotically in block length) the optimal encoder/decoder consists of an optimal estimator followed by optimal encoding of the resulting estimate. In the case of additive independent Gaussian background noise with variance σ_n^2 and a Gaussian source of variance σ_s^2, the minimum possible distortion is shifted away from zero by the amount

$$D_{\min} = \frac{\sigma_s^2 \sigma_n^2}{\sigma_s^2 + \sigma_n^2}. \qquad (3)$$

Hence, for a 0-dB signal-to-noise ratio at the source coder input, that is, noise power equal to speech power, the minimum average distortion is $\sigma_s^2/2$.

For the removal of additive white noise, the standard approaches have been spectral subtraction using Wiener filtering or Kalman filtering. Since the jointly optimal (here, minimum mean square error) estimation of parameters and filtering of the noisy signal is nonlinear, the joint filtering and parameter estimation problem is typically separated into the cascaded problem of parameter estimation on the noisy input followed by linear filtering using estimated parameters obtained in the first stage [13]–[15]. This process is then iterated to obtain a solution. Since we do not have perfect models for the speech signal and have limited knowledge of the speech model parameters and the noise power, the iteration process does not yield monotonically improving performance. Hence, after a few iterations, the speech may actually become more degraded rather than improved.

The Kalman filtering approach allows for time-domain noise models to be used, and a correlated noise model is used in [14]. A second-order version of one of these algorithms is the optional noise canceller in the half-rate JDC.

However, the basic difficulty for noise filtering or noise cancellation is that we lack a perfect model for the speech process, and we almost wholly lack even an approximate model for the noise. The IS-127 EVR coder improves on the familiar spectral subtraction method of noise suppression by taking a fairly sophisticated approach that uses an FFT and estimates the input signal-to-noise ratio at the various frequencies. This method is less model based than the time-domain approaches.

IX. New Directions and Conclusions

The rapid evolution of mobile voice communications based upon digital communications technologies has been facilitated by innovations and advances in speech coding. Linear predictive coding within the analysis-by-synthesis paradigm is the dominant technique for digital cellular applications today. Current challenges and limitations of these coders and their corresponding systems are background-noise impairments, channel errors, and coder complexity. As bit rates allocated to speech coding are pushed below 8 kbits/s, the challenge is to address each of these issues while achieving the clear channel performance demanded by the user. Joint source/channel coder designs, better front-end processing, and more accurate modeling of the vocal-tract excitation are promising research directions. Because of the time-varying nature of the speech production process and the time-varying channels inherent in mobile radio applications, perhaps the two most significant concepts for future mobile systems are variable-rate speech coding and adaptive signal processing. Indeed, adaptive signal processing algorithms for parameter estimation, noise suppression, channel equalization, diversity combining, and error control are already essential to achieving the overall performance required by today's users. It is expected that adaptive signal processing will play an even larger role in the next-generation systems. Furthermore, it is clear that as we move toward ubiquitous mobile communications, innovations in speech coding will be an essential element.

Acknowledgment

The authors gratefully acknowledge S. Gray, R. Levine, A. McCree, and P. Mermelstein for providing comments, insights, and suggestions that greatly enhanced the final version of this paper.

References

[1] N. S. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proc. IEEE*, vol. 81, pp. 1385–1422, Oct. 1993.
[2] A. Gersho, "Advances in speech and audio compression," *Proc. IEEE*, vol. 82, pp. 900–918, June 1994.
[3] A. S. Spanias, "Speech coding: A tutorial review," *Proc. IEEE*, vol. 82, pp. 1541–1582, Oct. 1994.
[4] R. V. Cox, "Three new speech coders from the ITU cover a range of applications," *IEEE Commun. Mag.*, vol. 35, pp. 40–47, Sept. 1997.
[5] A. M. Kondoz, *Digital Speech: Coding for Low Bit Rate Communications Systems*. Chichester, West Sussex, England: Wiley, 1994.
[6] W. B. Kleijn and K. K. Paliwal, Eds., *Speech Coding and Synthesis*. Amsterdam, The Netherlands: Elsevier, 1995.
[7] J. D. Gibson, T. Berger, T. Lookabaugh, D. Lindbergh, and R. L. Baker, *Digital Compression for Multimedia: Principles and Standards*. San Francisco, CA: Morgan Kaufmann, 1998.
[8] P. Kroon, "Time-domain coding of (near) toll quality speech at rates below 16 Kb/s," Ph.D. dissertation, Delft University of Technology, Delft, The Netherlands, May 1985.
[9] P. Kroon and E. F. Deprettere, "A class of analysis-by-synthesis predictive coders for high quality speech coding at rates between 4.8 and 16 Kbits/s," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 353–363, Feb. 1988.
[10] P. Kroon, E. F. Deprettere, and R. J. Sluyter, "Regular pulse excitation—A novel approach to effective and efficient multipulse coding of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 1054–1063, Oct. 1986.
[11] W. B. Kleijn and W. Granzow, "Methods for waveform interpolation in speech coding," *Digital Signal Processing*, vol. 1, no. 4, pp. 215–230, 1991.
[12] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.
[13] T. R. Fischer and J. D. Gibson, "Estimation and noisy source coding," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 23–34, Jan. 1990.
[14] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Signal Processing*, vol. 39, pp. 1732–1742, Aug. 1991.
[15] J. D. Gibson, "Adaptive prediction in speech differential encoding systems," *Proc. IEEE*, vol. 68, pp. 488–525, Apr.mj 1980.
[16] J. K. Wolf and J. Ziv, "Transmission of noisy information to a noisy receiver with minimum distortiion," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 406–411, July 1970.
[17] B. Koo, "Speech compression," in *The Communications Handbook*, J. D. Gibson, Ed. Boca Raton, FL: CRC Press, 1997, pp. 1424–1436.
[18] L. Hanzo, "The British cordless telephone standard: CT-2," in *The Communications Handbook*, J. D. Gibson, Ed. Boca Raton, FL: CRC Press, 1997, pp. 1289–1304.
[19] S. Asghar, "DECT (Digital European cordless telephone)," in *The Communications Handbook*, J. D. Gibson, Ed. Boca Raton, FL: CRC Press, 1997, pp. 1305–1326.
[20] P. Mermelstein, "The IS-54 digital cellular standard," in *The Communications Handbook*, J. D. Gibson, Ed. Boca Raton, FL: CRC Press, 1997, pp. 1246–1256.
[21] A. H. M. Ross and K. S. Gilhousen, "CDMA technology and the IS-95 North American standard," in *The Communications Handbook*, J. D. Gibson, Ed. Boca Raton, FL: CRC Press, 1997, pp. 1257–1275.
[22] K. Kinoshita and M. Nakagawa, "Japanese cellular standard," in *The Communications Handbook*, J. D. Gibson, Ed. Boca Raton, FL: CRC Press, 1997, pp. 1276–1288.
[23] W.-Y. Chan, I. Gerson, and T. Miki, "Half-rate standards," in *The Communications Handbook*, J. D. Gibson, Ed. Boca Raton, FL: CRC Press, 1997, pp. 1338–1352.
[24] P. Mermelstien, private communication, Dec. 1997.
[25] B. S. Atal and M. R. Schroeder, "Stochastic coding of speech at very low bit rates," in *Proc. Int. Conf. Commununications*, 1984, pp. 1610–1613.

Madhukar Budagavi received the B.E. degree (first class with distinction) in electronics and communications engineering from the Regional Engineering College, Trichy, India, in 1991, the M.Sc. (Eng.) degree in electrical engineering from the Indian Institute of Science, Bangalore, in 1993, and the Ph.D. degree in electrical engineering from Texas A&M University, College Station, in 1998.

During 1993–1995, he was first a Software Engineer and then a Senior Software Engineer with Motorola India Electronics, Ltd., primarily developing digital signal processing (DSP) software and algorithms for the Motorola DSP chips. He is now a Member of Technical Staff in the Video Technology Branch of the Media Technologies Lab in the Texas Instruments DSP Solutions R&D Center. His current research interests include video coding, speech coding, and wireless and Internet communications.

Jerry D. Gibson (Fellow, IEEE) currently is Chairman of the Department of Electrical Engineering at Southern Methodist University, Dallas, TX. He has been with General Dynamics–Fort Worth (1969–1972), the University of Notre Dame (1973–1974), and the University of Nebraska–Lincoln (1974–1976). During fall 1991, he was on sabbatical with the Information Systems Laboratory and the Telecommunications Program in the Department of Electrical Engineering, Stanford University, Stanford, CA. From 1987 to 1997, he held the J. W. Runyon, Jr. Professorship in the Department of Electrical Engineering, Texas A&M University, College Station. He is the coauthor of *Introduction to Nonparametric Detection with Applications* (New York: Academic, 1975 and IEEE Press, 1995) and author of *Principles of Digital and Analog Communications* (Englewood Cliffs, NJ: Prentice-Hall, 1993). He is Editor-in-Chief of *The Mobile Communications Handbook* (Boca Raton, FL: CRC Press, 1995) and of *The Communications Handbook* (Boca Raton, FL: CRC Press, 1996). His research interests include data, speech, image, and video compression, multimedia over networks, wireless communications, information theory, and digital signal processing.

Dr. Gibson received the Fredrick Emmons Terman Award from the American Society for Engineering Education in 1990. He was corecipient of the 1993 IEEE Signal Processing Society Senior Paper Award for the Speech Processing area. He was Associate Editor for Speech Processing for IEEE TRANSACTIONS ON COMMUNICATIONS from 1981 to 1985 and Associate Editor for Communications for IEEE TRANSACTIONS ON INFORMATION THEORY from 1988 to 1991. He was a Member of the Speech Technical Committee of the IEEE Signal Processing Society (1992–1995) and currently is a Member of the IEEE Information Theory Society Board of Governors (1990–1998). He was President of the IEEE Information Theory Society in 1996.