Speech Coding Methods, Standards, and Applications

Jerry D. Gibson

Department of Electrical & Computer Engineering
University of California, Santa Barbara
Santa Barbara, CA 93106-6065
gibson@ece.ucsb.edu

I. Introduction

Speech coding is fundamental to the operation of the public switched telephone network (PSTN), videoconferencing systems, digital cellular communications, and emerging voice over Internet protocol (VoIP) applications. The goal of speech coding is to represent speech in digital form with as few bits as possible while maintaining the intelligibility and quality required for the particular application. Interest in speech coding is motivated by the evolution to digital communications and the requirement to minimize bit rate, and hence, conserve bandwidth. There is always a tradeoff between lowering the bit rate and maintaining the delivered voice quality and intelligibility; however, depending on the application, many other constraints also must be considered, such as complexity, delay, and performance with bit errors or packet losses [1].

Two networks that have been developed primarily with voice communications in mind are the public switched telephone network (PSTN) and digital cellular networks. Additionally, with the pervasiveness of the Internet, voice over the Internet Protocol (VoIP) is growing rapidly and is expected to do so for the near future. A new and powerful development for data communications is the emergence of wireless local area networks (WLANs) in the embodiment of the 802.11 a, b, g standards, and the next generation, 802.11n, collectively referred to as Wi-Fi, and the on-going development of the IEEE 802.16 family of standards for wireless metropolitan area networks (WMAN), often referred to as WiMax. Because of the proliferation of these standards and the expected rapid expansion of these networks, considerable attention is now being turned to voice over Wi-Fi and WiMax, with some companies already offering proprietary networks, handsets, and solutions.

Based on these developments, it is possible today, and it is likely in the near future, that our day-to-day voice communications will involve multiple hops including heterogeneous networks. This is a considerable departure from the plain old telephone service (POTS) on the PSTN, and indeed, these future voice connections will differ greatly even from the digital cellular calls connected through the PSTN today. As the networks supporting our voice calls become less homogeneous and include more wireless links, many new challenges and opportunities emerge.

This paper addresses these challenges and opportunities by describing the basic issues in speech coder design, developing the important speech coding techniques and standards, discussing current and future applications, outlining techniques for evaluating speech coder performance, and identifying research directions. Before describing the contents of the paper, let us first state what this paper is not. This paper is not intended to be an historical survey of speech coding, a comprehensive description of standardized speech codecs, nor a complete development of speech coding methods. There are many excellent papers [2-4], books, and book chapters [1, 5-11] that address these topics very well, so in this paper, we focus on the principal issues raised by supporting conversational voice communications over the multihop tandem networks and wireless links we encounter today and that we will encounter in the future.

As we take speech codecs out of the lab and into networks, many issues arise in addition to bit rate versus quality and intelligibility. In Sec. II, we highlight these issues to put the later discussion into context. To understand why certain speech codecs are chosen for a particular application, or to understand why a specific speech codec would be a poor choice for an application, it is useful to have a basic idea of the commonly-used speech coding methods. Furthermore, there are many functionalities, such as bit rate scalability, diversity coding options, and bandwidth scalability, that are highly desirable for different applications. Basic speech coding methods and codec functionalities are outlined in Sec. III.

There was almost an exponential growth of speech coding standards in the 1990's for a wide range of networks and applications, including the PSTN, digital cellular, and multimedia streaming. We list the most prominent speech coding standards and their properties, such as performance, complexity, and coding delay, in Sec. IV. In this section, we examine the particular networks and applications for each standard, including those applications where a standard from one application has become a de facto standard in another network or application. Section V describes current challenges for effective voice communications and suggests possible research directions.

In order to compare the various speech coding methods and standards, it is necessary to have methods for establishing the quality and intelligibility produced by a speech coder. It is a difficult task to find objective measures of speech quality, and often, the only acceptable approach is to perform subjective listening tests. However, there have been some recent successes in developing objective quantities, experimental procedures, and mathematical expressions that have a good correlation with speech quality and intelligibility. We provide brief descriptions of these methods in Appendix A. Appendix B gives an overview of the organizations behind setting the many speech coding standards that are available. A brief summary and some conclusions are offered in Sec. VI.

We use the terms speech coding and voice coding interchangeably in this paper. Speech coding is a widely accepted terminology, and there seems to be no reason to distinguish between voice and speech, unless a restricted definition of one or the other is adopted. Generally, it is desired to reproduce the voice signal, since we are interested in not only knowing what was said, but also in being able to identify the speaker. However, no confusion should result, and we do not attempt to make any distinction here.

II. Basic Issues in Speech Coding

Speech and audio coding can be classified according to the bandwidth occupied by the input and the reproduced source. Narrowband or telephone bandwidth speech occupies the band from 200 to 3400 Hz, while wideband speech is contained in the range of 50 Hz to 7 kHz. High quality audio is generally taken to cover the range of 20 Hz to 20 kHz. The discussions in this paper address both narrowband and wideband speech, but we do not treat the quite different problem of high quality audio coding here.

Given a particular source, the classic tradeoff in lossy source compression is rate versus distortion--the higher the rate, the smaller the average distortion in the reproduced signal. Of course, since a higher bit rate implies a greater bandwidth requirement, the goal is always to minimize the rate required to satisfy the distortion constraint. For speech coding, we are interested in achieving a quality as close to the original speech as possible. Encompassed in the term quality are intelligibility, speaker identification, and naturalness. Absolute category rating tests are subjective tests of speech quality and involve listeners assigning a category and rating for each speech utterance according to the classifications, such as, Excellent (5), Good (4), Fair (3), Poor (2), and Bad (1). The average for each utterance over all listeners is the Mean Opinion Score (MOS) [5, 11]. More details on MOS are given in Appendix A.

Although important, the MOS values obtained by listening to isolated utterances do not capture the dynamics of conversational voice communications in the various network environments. It is intuitive that speech codecs should be tested, within the environment and while executing the tasks, for which they are designed. Thus, since we are interested in conversational (two-way) voice communications, a more realistic test would be conducted in this scenario. Recently, the perceptual evaluation of speech quality (PESQ) method was developed to provide an assessment of speech codec performance in conversational voice communications. The PESQ has been standardized by the ITU-T as P.862 and can be used to generate MOS values for both narrowband and wideband speech [12]. The narrowband PESQ performs fairly well for the situations that it has been qualified for, however, at the time of this writing, the wideband PESQ MOS values are much lower than might be expected. The wideband PESQ can be, and is, used to evaluate wideband codec performance, but it should only be interpreted as a relative ranking rather than a global result.

Throughout this paper, we quote available MOS values taken from many different sources for all of the codecs. Since MOS values can vary from test to test and across languages, these values should not be interpreted as an exact indicator of performance. Care has been taken only to present MOS values that are consistent with widely known performance results for each codec. For more details on subjective and objective measures for speech quality assessment, the reader is referred to Appendix A and the references.

Across codecs that achieve roughly the same quality at a specified bit rate, codec complexity can be a distinguishing feature. In standards and in the literature, the number of MIPs (million instructions per second) is often quoted as a broad indicator of implementation complexity, where MIPs numbers generally relate to implementations on DSP chips rather than CPUs. We follow that practice here, although qualifying statements are included as necessary.

An issue of considerable importance for voice codecs to be employed in conversational applications is the delay associated with the processing at the encoder and the decoder. This is because excessive delay can cause "talk over" and an interruption of the natural flow of the conversation. ITU-T Recommendation G.114 provides specifications for delay when echo is properly controlled [13]. In particular, one way transmission time (including processing and propagation delay) is categorized as: (a) 0 to 150 ms: Acceptable for most user applications; (b) 150 to 400 ms: Acceptable depending upon the transmission time impact; (c) Above 400 ms: Unacceptable for general network planning purposes. Figure 1 provides a broader view of the estimated impact of delay on user preferences as stated in the ITU-T G.114 standard. From this figure, it is seen that up to 150 ms represents the flat part of the curve, and acceptable delay is sometimes quoted as 200 ms, since delays up to 200 ms fall into the "Users Very Satisfied" category. Current thinking is pushing the acceptable delay upward to 250 ms, which is roughly near the middle of the "Users Satisfied" category.

Furthermore, guidelines are given in G.114 for calculating the delay based upon the codec frame size and look ahead and the application. For example, it is assumed that there is a delay of one frame before processing can begin and that the time required to process a frame of speech is the same as the frame length. As a result, if there is no additional delay at the interface of the codec and the network, the "mean" one-way delay of a codec is 2 × frame size + look-ahead. Another one frame delay is sometimes included to incorporate delay due to interfacing with the network or for the delay due to decoding. Thus, when consulting various references, the reader may encounter quite different delay values for the same codec. In this paper, we state the frame size and the look-ahead for each codec since these are the basic quantities that influence the encoding/decoding delay of speech codecs.

When different codecs are used in different, but interconnected networks, or when the speech signal must be decoded at network interfaces or switches, the problem of asynchronous tandem connections of speech codecs can arise. The term asynchronous tandeming originally referred to a series connection of speech coders that requires digital to analog conversion followed by re-sampling and re-encoding. Today, and within the context of this paper, it also refers to where the speech samples must be reconstructed and then re-encoded by the next codec. Asynchronous tandem connections of speech codecs can lead to a significant loss in quality, and of course, incur additional delays due to decoding and re-encoding. Another possible approach to tandem connections of different speech codecs is to map the parameters of one codec into the parameters of the following codec without reconstructing the speech samples themselves. This is usually referred to as transcoding. Transcoding produces some quality loss and delay as well. Asynchronous tandem connections of speech codecs are often tested as part of the overall codec performance evaluation process.

In many applications, errors can occur during transmission across networks and through wireless links. Errors can show up as individual bit errors or bursts of bit errors, or with the widespread movement toward IP-based protocols, errors can show up as packet losses. In error-prone environments or during error events, it is imperative that the decoded speech be as little affected as possible. Some speech codecs are designed to be robust to bit errors, but it is more common today that packet loss concealment methods are designed and associated with each speech codec. Depending upon the application, robustness to errors and/or a good packet loss concealment scheme may be a dominating requirement.

It is by now well known that speech codecs should be tested across a wide variety of speakers, but it is also important to note that speech codecs are tested across many different languages as well. It is difficult to incorporate multiple languages into a design philosophy, but suffice it to say, since the characteristics of spoken languages around the world can be so different, codec designs must be evaluated across a representative set of languages to reveal any shortcomings.

III. Speech Coding Techniques and Functionalities

The most common approaches to narrowband speech coding today center around two paradigms, namely, waveform-following coders and analysis-by-synthesis methods. Waveform-following coders attempt to reproduce the time domain speech waveform as accurately as possible, while analysis-by-synthesis methods utilize the linear prediction model and a perceptual distortion measure to reproduce only those characteristics of the input speech that are determined to be most important. Another approach to speech coding breaks the speech into separate frequency bands, called subbands, and then codes these subbands separately, perhaps using a waveform coder or analysis-by-synthesis coding, for reconstruction and recombination at the receiver. Extending the resolution of the frequency domain decomposition leads to transform coding, wherein a transform is performed on a frame of input speech and the resulting transform coefficients are quantized and transmitted to reconstruct the speech from the inverse transform. In this section, we provide some background details for each of these approaches to lay the groundwork for the later developments of speech coding standards based upon these principles. A discussion of the class of speech coders called vocoders or purely parametric coders is not included due to space limitations and their more limited range of applications today.

We also discuss the utility of postfiltering and some general approaches to variable rate speech coding. Further, we develop several important speech codec functionalities, such as bit rate (or SNR) scalable coding, bandwidth scalable coding, and diversity-based source coding structures.

Waveform Coding

Familiar waveform-following methods are logarithmic pulse code modulation (log-PCM) and adaptive differential pulse code modulation (ADPCM), and both have found widespread applications. Log PCM at 64 kilobits/sec (kbps) is the speech codec used in the long distance public switched telephone network at a rate of 64 kbps [11, 14]. It is a simple coder and it achieves what is called toll quality, which is the standard level of performance against which all other narrowband speech coders are judged. Log PCM uses a nonlinear quantizer to reproduce low amplitude signals, which are important to speech perception, well. There are two closely related types of log-PCM quantizer used in the World-- μ -law, which is used in North America and Japan, and A-law, which is used in Europe, Africa, Australia, and South America. Both achieve toll quality speech, and in terms of the MOS value, it is usually between 4.0 and 4.5 for log-PCM [11, 14].

ADPCM operates at 32 kbps or lower, and it achieves performance comparable to log-PCM by using a linear predictor to remove short-term redundancy in the speech signal before quantization as shown in Fig. 2. The most common form of ADPCM uses what is called backward adaptation of the predictors and quantizers to follow the waveform closely. Backward adaptation means that the predictor and quantizer are adapted based upon past reproduced values of the signal that are available at the encoder and decoder. No predictor or quantizer parameters are sent along with the quantized waveform values. By subtracting a predicted value from each input sample, the dynamic range of the signal to be quantized is reduced, and hence, good reproduction of the signal is possible with fewer bits.

The waveform coder paradigm for speech was extended by adding a long term predictor to the ADPCM structure to get what is sometimes called an Adaptive Predictive Coder (APC). An important embodiment of APC that served as a precursor to the analysis-by-synthesis schemes was APC with noise spectral shaping as shown in Fig. 3. The shaping of the noise spectrum by *F(z)* allowed the noise power to be redistributed across the frequency band so that the noise could be masked by the input speech spectrum. The structures for the noise spectral shaping filters were carried over to the perceptual weighting filters in analysis-by-synthesis methods. These same forms also motivated the formant filter shaping in postfilters [1].

Subband and Transform Methods

The process of breaking the input speech into subbands via bandpass filters and coding each band separately is called subband coding. To keep the number of samples to be coded at a minimum, the sampling rate for the signals in each band is reduced by decimation. Of course, since the bandpass filters are not ideal, there is some overlap between adjacent bands and aliasing occurs during decimation. Ignoring the distortion or noise due to compression, Quadrature mirror filter (QMF) banks allow the aliasing that occurs during filtering and subsampling at the encoder to be cancelled at the decoder. The codecs used in each band can be PCM, ADPCM, or even an analysis-by-synthesis method. The advantage of subband coding is that each band can be coded differently and that the coding error in each band can be controlled in relation to human perceptual characteristics.

Transform coding methods were first applied to still images but later investigated for speech. The basic principle is that a block of speech samples is operated on by a discrete unitary transform and the resulting transform coefficients are quantized and coded for transmission to the receiver. Low bit rates and good performance can be obtained because more bits can be allocated to the perceptually important coefficients, and for well-designed transforms, many coefficients need not be coded at all, but are simply discarded, and acceptable performance is still achieved.

Although classical transform coding has not had a major impact on narrowband speech coding and subband coding has fallen out of favor in recent years, filter bank and transform methods play a critical role in high quality audio coding, and at least one important standard for wideband speech coding (G.722.1) is based upon filter bank and transform methods. Although it is intuitive that subband filtering and discrete transforms are closely related, by the early 1990's, the relationships between filter bank methods and transforms were well-understood [15, 16]. Today, the distinction between transforms and filter bank methods is somewhat blurred, and the choice between a filter bank implementation and a transform method may simply be a design choice.

Analysis-by-Synthesis Methods

Analysis-by-synthesis (AbS) methods are a considerable departure from waveform-following techniques. The most common and most successful analysis-by-synthesis method is code-excited linear prediction (CELP). In CELP speech coders, a segment of speech (say, 5 ms) is synthesized using the linear prediction model along with a long-term redundancy predictor for all possible excitations in what is called a codebook. For each excitation, an error signal is calculated and passed through a perceptual weighting filter. This operation is represented in Fig. 4(a). The excitation that produces the minimum perceptually weighted coding error is selected for use at the decoder as shown in Fig. 4(b). Therefore, the best excitation out of all possible excitations for a given segment of speech is selected by synthesizing all possible representations at the encoder. Hence, the name analysis-by-synthesis. The predictor parameters and the excitation codeword are sent to the receiver to decode the speech. It is instructive to contrast the AbS method with waveform coders such as ADPCM, shown in Fig. 2, where each sample is coded as it arrives at the coder input.

The perceptual weighting is key to obtaining good speech coding performance, and the basic idea is that the coding error is spectrally shaped to fall below the envelope of the input speech across the frequency band of interest. Figure 5 illustrates the concept wherein the spectral envelope of a speech segment is shown, along with the coding error spectrum without perceptual weighting (unweighted denoted by short dashes) and the coding error spectrum with perceptual weighting (denoted by long dashes). The perceptually weighted coding error falls below the spectral envelope of the speech across most of the frequency band of interest, just crossing over around 3100 Hz. The coding error is thus masked by the speech signal itself. In contrast, the unweighted error spectrum is above the speech spectral envelope starting at around 1.6 kHz, which produces audible coding distortion for the same bit rate.

In recent years, it has become common to use an adaptive codebook structure to model the long term memory rather than a cascaded long term predictor. A decoder using the adaptive codebook approach is shown in Fig. 6. The analysis-by-synthesis procedure is computationally intensive, and it is fortunate that algebraic codebooks, which have mostly zero values and only a few nonzero pulses, have been discovered and work well.

Postfiltering

Although a perceptual weighting filter is used inside the search loop for the best excitation in the codebook for analysis-by-synthesis methods, there is often still some distortion in the reconstructed speech that is sometimes characterized as "roughness." This distortion is attributed to reconstruction or coding error as a function of frequency that is too high at regions between formants and between pitch harmonics. Several codecs thus employ a postfilter that operates on the reconstructed speech to de-emphasize the coding error between formants and between pitch harmonics. This is shown as "Post-processing" in Fig. 6. The general frequency response of the postfilter, illustrated in Fig. 7, has the form similar to the perceptual weighting filter with a pitch or long term postfilter added. There is also a spectral tilt correction, labeled as part of the short term postfilter, since the formant-based postfilter results in an increased low pass filter effect, and a gain correction term [17]. The postfilter is usually optimized for a single stage encoding (however, not always), so if multiple asynchronous tandems occur, the postfilter can cause a degradation in speech quality.

Variable Rate Coding

For more than 30 years, researchers have been interested in assigning network capacity only when a speaker is "active," as in TASI [14], or removing silent periods in speech to reduce the average bit rate. This was successfully accomplished for some digital cellular coders where silence is removed and coded with a short length code and then replaced at the decoder with "comfort noise." Comfort noise is needed because the background sounds for speech coders are seldom pure silence and inserting pure silence generates unwelcome artifacts at the decoder and can cause the impression that the call is lost. The result, of course, is a variable rate speech coder. Many codecs use voice activity detection to excise non-speech signals so that non-speech regions do not need to be coded explicitly. More sophisticated segmentation can also be performed so that different regions can be coded differently. For example, more bits may be allocated to coding strongly voiced segments and fewer allocated to unvoiced speech. Also, speech onset might be coded differently as well.

When used over fixed rate communications links, variable rate coders require the use of jitter buffers and can be sensitive to bit errors. However, packet switched networks reduce these problems and make variable rate coding attractive for packet networks.

SNR and Bandwidth Scalable Methods

SNR or bit rate scalability refers to the idea of coding the speech in stages such that the quality can be increased simply by appending an additional bit stream. For example, speech may be coded using a minimum bit rate stream that provides acceptable quality, often called a core layer, along with one or more incremental enhancement bit streams. When the first enhancement bit stream is combined with the core layer to reconstruct the speech, improved performance is obtained. SNR or bit rate scalable coding is inherently no better than single stage encoding at the combined rate, so the challenge is to design good

SNR scalable codecs that are as close to single stage encoding as possible. One attraction of SNR scalable coding is that the enhancement bit stream may be added or dropped, depending upon available transmission bit rate. Another option facilitated by SNR scalable coding is that the core bit stream may be subject to error control coding whereas the enhancement bit stream may not.

Bandwidth scalable coding is a method wherein speech with a narrower bandwidth is coded as a base layer bit stream, and an enhancement bit stream is produced that encodes frequencies above the base layer bandwidth. Particular applications of interest might be having a base layer bit stream that codes telephone bandwidth speech from 200 to 3400 Hz and an enhancement layer bit stream that codes speech in the band from 3400 Hz to 7 kHz. The goal is to allow flexibility in bit rate and still have a high quality representation of the speech at the low band and at the full band when the enhancement layer is combined with the base layer.

Multiple Descriptions Coding

Multiple descriptions coding is a source coding method wherein the total rate available is split into multiple side channels with bit rates that are a fraction of the full rate bit stream. For the case with two side channels, for example, the goal is to achieve good performance if either of the half rate side channels is received while still achieving near optimal full rate performance if both channels are received. This is a difficult design problem that has received much more attention for still images and video. However, there has been recent interest in multiple descriptions speech coding for various applications and some example scenarios are discussed in later sections.

IV. Networks and Standards for Voice Communications

In this section, we describe the relevant details of current standardized voice codecs in the context of the network for which they were designed or have been applied. In particular, we develop the voice codecs for the PSTN, digital cellular networks, voice over the Internet protocol (VoIP), and voice over wireless local area networks (voice over Wi-Fi). We also briefly mention standard codecs used in some secure and satellite telephony applications.

The Public Switched Telephone Network (PSTN)

The most familiar network for conversational voice communications is the PSTN, which consists of a wired, time division multiplexed (TDM), circuit-switched backbone network with (often) copper wire pair local loops [14]. The PSTN was designed, and evolved, with voice communications as the primary application. The voice codec most often used in the PSTN is 64 kilobits/sec. (kbps) logarithmic pulse code modulation (log-PCM), designated by the ITU-T as G.711, and which is taken as the standard for narrowband toll quality voice transmission. The TDM links in the PSTN are very reliable with bit error rates (BERs) of 10^{-6} to 10^{-9}. As a result, bit error concealment is not an issue for G.711 transmission over TDM PSTN links, even though at higher bit error rates, bit errors in G.711 encoded voice generate very irritating "pops" in the reconstructed speech. Furthermore, G.711 is designed with several asynchronous tandems in mind, since it was possible to encounter several analog switches during a long distance telephone call prior to the mid-1980's. Even 8 asynchronous tandems of G.711 with itself has been shown to still maintain a Mean Opinion Score greater than 4.0 when a single encoding is 4.4 to 4.5.

Other voice codecs have been standardized for the PSTN over the years. These include G.721 (now G.726), G.727, G.728, G.729, and G.729A for narrowband (telephone bandwidth) speech (200 to 3400 Hz) and G.722, G.722.1 [18], and G.722.2 [19] for wideband speech (50 Hz to 7 kHz). Table I summarizes the rate, performance, complexity, and algorithmic delay of each of the narrowband speech codec standards. It is emphasized that the MOS values shown are approximate since they are taken from several different sources. It is recognized that MOS values for a given codec can (and will) change from test to test, but the values given here provide an approximate ordering and indication of codec performance.

Table II presents some results concerning the tandem performance of the narrowband speech codecs. Note that asynchronous tandem connections of these codecs with themselves do not cause an unacceptable loss in performance compared to a single encoding, although the MOS for 4 tandems of G.726 and 3 tandems of G.729 drops considerably. The additional delay due to the decoding and re-encoding is not indicated in this table and the effect on quality of additional latency is not included in the MOS values shown.

Even though we are quite comfortable communicating using telephone bandwidth speech (200 to 3400 Hz), there has been considerable recent interest in compression methods for wideband speech covering the range of 50 Hz to 7 kHz. The primary reasons for this interest are that wideband speech improves intelligibility, naturalness, and speaker identifiability. Originally, the primary application of wideband speech coding was to videoconferencing, and the first standard, G.722, separated the speech into two subbands and used ADPCM to code each band. The G.722 codec is relatively simple and produces good quality speech at 64 kbps, and lower quality speech at the two other possible codec rates of 56 and 48 kbps. The G.722 speech codec is still widely available in the H.323 videoconferencing standard, and it is often provided as an option in VoIP systems.

Two recently developed wideband speech coding standards, designated as G.722.1 and G.722.2, utilize coding methods that are quite different from G.722, as well as completely different from each other. The G.722.1 standard employs a filter bank/transform decomposition called the modulated lapped transform (MLT) and operates at the rates of 24 and 32 kbps. A categorization procedure is used to determine the quantization step sizes and coding parameters for each region. The coder has an algorithmic delay of 40 msec, which does not include any computational delay. Since G.722.1 employs filter bank methods, it performs well for music and less well for speech. In one MOS test using British English, the G.722.1 coder at 24 kbps achieved an MOS value of 4.1. It is important to note that comparing MOS values for narrowband and wideband speech is not valid.

G.722.2 is actually an ITU-T designation for the adaptive multirate wideband (AMR-WB) speech coder standardized by the 3GPP. This coder operates at rates of 6.6, 8.85, 12.65, 14.25, 15,85, 18.25, 19.85, 23.05, and 23.85 kbps and is based upon an algebraic CELP (ACELP) analysis-by-synthesis codec. Since ACELP utilizes the linear prediction model, the coder works well for speech but less well for music, which does not fit the linear prediction model. G.722.2 achieves good speech quality at rates greater than 12.65 kbps and performance equivalent to G.722 at 64 kbps with a rate of 23.05 kbps and higher. For speech, one MOS test for the French language showed the G.722.2 codec achieving an MOS value of 4.5 at the 23.85 kbps rate and an MOS value of around 4.2 for the 12.65 kbps rate.

Table III summarizes the important properties of the G.722, G.722.1, and G.722.2 codecs.

Digital Cellular Networks

Digital cellular networks provide wireless voice connectivity by combining high quality voice codecs, error detection and concealment of uncorrected errors, and interleaving [20, 21]. Table IV contains the rate, performance, complexity, and algorithmic delay for selected digital cellular speech codecs. As in previous tables, the MOS values and complexity in MIPS are representative numbers taken from several sources. It is evident from Table IV that the voice codecs have good performance in ideal conditions without large algorithmic delay or excessive complexity.

For some digital cellular networks, the bits to be transmitted may be classified into categories for unequal error protection, as well as for enabling different modes of error concealment for errors detected after forward error correction. In Table IV, the rate is for the voice codec only without error correction/detection, and the frame size/look ahead numbers in the table do not include delay due to interleaving. Note that for variable rate coders, quality is often determined for a desired average data rate (ADR), where the average is computed by weighting each possible rate of the codec with the amount of time it is expected to use this rate. For example, if there is 60% active speech and 40% silence in a speech segment, and the codec uses one rate for each condition, then the ADR would be 0.6 times the rate for active speech and 0.4 times the rate for silence.

An important wideband coder for digital cellular applications is the AMR-WB codec described in the preceding section on the PSTN, since it was also standardized by the ITU-T as G.722.2 [19, 22]. The cdma2000® Variable-Rate Multimode Wideband (VMR-WB) speech coding standard shares the same core algorithm as AMR-WB, and was standardized by the 3GPP2 in March 2003 [23]. The codec has the five modes shown in Table V, which lists the average data rate and quality target for each mode. The rate can be varied by the speech signal characteristics, called source-controlled, or by the traffic conditions on the network, called network-controlled. Depending upon the traffic conditions and the desired quality of service (QoS), one of the five modes is used. The speech activity factor for the test database used to obtain the average data rates was about 60%. Mode 3 is fully interoperable with AMR-WB at 12.65, 8.85, and 6.6 kbps and since AMR-WB is standardized for GSM/WCDMA, this supports some cross-system applications. In addition to cdma2000®, the VMR-WB codec is targeted for a wide range of applications, including VoIP, packet-switched mobile-to-mobile calls, multimedia messaging, multimedia streaming, and instant messaging. If appropriate signaling is available, there is also the goal of end-to-end tandem free operation (TFO) or transcoder free operation (TrFO) for some mobile-to-mobile calls. The VMR-WB codec has a narrowband signal processing capability in all modes that is implemented by converting the narrowband input sampling rate of 8 kHz to the internal sampling frequency of 12.8 kHz for processing, and then after decoding, converting the sampling rate back to 8 kHz.

One of the first variable rate speech coders with more than a silence removal mode was the Qualcomm IS-96 QCELP coder, which operated at the rates of 0.8, 2, 4, and 8 kbps, depending upon the classification of the input speech. This coder was part of the CDMA standard in North America, but it did not have good performance even at the highest supported rate, achieving an MOS of about 3.3. A replacement coder for IS-96 QCELP is the IS-127 Enhanced Variable Rate Coder (EVRC) that has three possible bit rates of 0.8, 4, and 8.55 kbps, which are selected with a simple rate selection algorithm. The IS-127 EVRC coder achieves an MOS of about 3.8 at the highest rate of 8 kbps, but is not operated at lower average data rates because of low voice quality.

A more recent variable rate speech codec is the TIA-EIA/IS-893 cdma2000 standard Selectable Mode Vocoder (SMV), which has six different modes that produce different average data rates and voice quality. The highest quality mode, Mode 0, can achieve a higher MOS than the IS-127 EVRC at an average data rate of 3.744 kbit/s, and Mode 1 is theoretically equivalent to IS-127.

Table VI shows available results for multiple tandem encodings of the codecs in Table IV, including results from tandem connections of these codecs with the PSTN backbone codecs in Table I. It is clear that tandem encodings result in a drop in performance as seen in the lower MOS values. Furthermore, tandem encodings add to the end-to-end delay because of the algorithmic delays in decoding and re-encoding. Tandem encodings can occur today on mobile-to-mobile calls when the handsets support different voice codecs. Tandem connections with PSTN codecs are not discussed often within digital cellular applications since the codec for the backbone wireline network is often assumed to be G.711. However, it is recognized that tandem encodings with codecs other than G.711 can lead to a loss in performance and that tandem encodings constitute a significant problem for end-to-end voice quality. In particular, transcoding at network interfaces and source coding distortion accumulation due to repeated coding has been investigated with the goal of obtaining a transparent transition between certain speech codecs. Some system-wide approaches also have been developed for specific networks. The general tandeming/transcoding problem remains open.

Digital cellular networks perform exceedingly well given the difficulty of the task. The melding of voice coder design, forward error correction and detection, unequal error protection, and error concealment in digital cellular has important lessons for designing voice communications systems for VoIP and voice over Wi-Fi.

Wireline Packet Switched Networks

Voice over IP has been evolving for more than 10 years, but it is now projected finally to be taking off as a voice communications service [24]. Among the issues in developing good VoIP systems are voice quality, latency, jitter, packet loss performance, packetization, and the design of the network. Broadly speaking, the voice codec in VoIP systems should achieve toll or near toll quality, have as low a delay as possible, and have good resilience to packet loss. Interestingly, voice codecs used in prominent VoIP products are all ported from previous standards and other applications. Some of the earliest VoIP applications used G.723.1, which was originally intended for video telephony, but the relatively long frame size and look ahead and the somewhat lower quality lead developers to consider other alternatives. Today's VoIP product offerings typically include G.711, G.729, and G.722, in addition to G.723.1. See Table VII for a summary of the relevant properties offered by each coder. G.723.1 is often favored for videophone applications since the delay of the video coding, rather than the voice codec, sets the delay of the videophone operation. The AMR-WB codec (G.722.2) has packet loss concealment incorporated so it is also a good candidate for VoIP applications.

The coders in Table VII, as a set, offer a full range of alternatives in terms of rate, voice quality, complexity, and delay. What is not evident from this table is how effectively one can conceal packet losses with each of these coders. Packet loss concealment is particularly important since in order to reduce latency, retransmissions are not allowed in wireline VoIP.

Rather recently, packet loss concealment algorithms have been developed for G.711 [25, 26]. Based upon 10 ms packets and assuming the previous frame was received correctly, the method in [25] generates a synthesized or concealment waveform from the last pitch period with no attenuation. If the next packet is lost as well, the method uses multiple pitch periods with a linear atttentuation at a rate of 20% per 10 ms. After 60 ms, the synthesized signal is zero.

G.729 and G.723.1 suffer from the problem that the predictor parameters (line spectral frequencies) are predictively encoded from frame-to-frame. For G.729, the basic approach to packet loss concealment if a single 10 ms frame is erased is: (i) generate a replacement excitation based upon the classification of the previous frame as voiced or unvoiced, (ii) repeat the synthesis filter parameters from the previous frame, and (iii) attenuate the memory of the gain predictor [27]. It seems intuitive that a speech codec that allowed interpolation of erased frames would perform better than using only past information, and this is indeed true. However, the frame-to-frame predictive coding of the short-term predictor parameters precludes using interpolation. Note, however, that interpolation implies additional delay in reconstructing the speech, and so the performance improvement provided by interpolation schemes must include the effects of any additional delay.

A codec that has been developed with VoIP in mind is the iLBC from Global IP Sound. This codec is designed for narrowband speech and operates at 13.33 kbps for 30 ms frames and 15.2 kbps for 20 ms frames. The iLBC uses block-independent linear predictive coding (LPC) as the basic coding algorithm. This codec has been standardized by the IETF as "Experimental" RFC 3951 in December 2004 [43] and is specified as a mandatory codec in PacketCable 1.5 by CableLabs for VoIP over cable [44].

Another recently developed codec for VoIP is the BroadVoice16 (BV16) codec that codes narrowband speech at 16 kbps [45]. This codec was specifically targeted for VoIP and voice over cable/DSL applications, and was designed to have high quality, low delay, low complexity, and medium to low bit rate. The algorithmic details are not yet widely available, however, the BV16 codec has also been specified as a mandatory codec in PacketCable 1.5 by CableLabs [44].

End-to-end delay is critical to VoIP performance, and some measurements on real backbone network connections of major Internet Service Providers have revealed that delay variability is very high and that some complex loss events can keep the one way delay through the network well above 100 ms for time periods of tens of seconds [28]. While it is common to use packet loss rate and delay to characterize VoIP performance, as noted in Sec. II, an evaluation of delivered voice quality in terms of MOS values is important. How to specify a value for MOS during a VoIP connection, which reflects the user's experience, is difficult. It is usually considered sufficient to state the time average of the measured MOS values over successive short time intervals; however, it may also be important to include a term considering recent bad events or the total number of bad events.

The E-model (see Appendix A) has been used in several studies to obtain MOS values for some common voice codecs under conditions representative of VoIP connections and using measured delay traces obtained from backbone networks. The E-model incorporates terms due to the "intrinsic" quality produced by a voice codec, packet losses and packet loss concealment, and one way mouth-to-ear delay, among other quantities, to produce an R-value on a psychoacoustic scale. The R-value then can be mapped into an MOS value. Results show that generating an instantaneous MOS by calculating the delay and packet loss effects on short time intervals is more representative of user experience than an average MOS [28].

Wireless Local Area Networks (WLANs)

Wireless local area networks (WLANs), such as 802.11b, 802.11a, and 802.11g, are becoming extremely popular for use in businesses, homes, and public places. As a result, there is considerable interest in developing VoIP for Wi-Fi, which we designate here as voice over Wi-Fi. The issues involved for voice over Wi-Fi would seem to be very much the same as for VoIP over the Internet; and it is certainly true that voice quality, latency, jitter, packet loss performance, and packetization, all play a critical role. However, the physical link in Wi-Fi is wireless, and as a result, bit errors commonly occur and this, in turn, affects link protocol design and packet loss concealment.

The IEEE 802.11 MAC specifies two different mechanisms, namely the contention-based Distributed Coordination Function (DCF) and the polling-based Point Coordination Function (PCF), with the first being widely implemented [29]. The DCF uses a carrier sense multiple access with collision avoidance (CSMA/CA) scheme for medium access. DCF has two mechanisms for packet transmission, the default basic access mechanism and the optional request-to-send/clear-to-send mechanism (RTS/CTS). The basic DCF access mechanism is, if the channel is sensed to be idle, the node starts its transmissions. A CRC (cyclic redundancy check code) is computed over the entire received packet and an acknowledgment is sent if the packet is error-free. If even a single bit error is detected in the packet, a retransmission is requested. Up to 7 retransmissions for short packets and up to 4 retransmissions for large packets are allowed, but actual implementations vary. Implicit in this protocol is that for a fixed bit error rate, a longer packet has a higher probability of needing a retransmission. In the event of packet loss either by bit errors or collisions, a random backoff is initiated in order to reduce the probability of collisions. An explicit ACK transmission is required since these wireless transceivers are generally half-duplex, i.e. they do not transmit and receive at the same time and hence cannot detect a collision by listening to its own transmission. This access method clearly adds to latency and is in contrast to avoiding retransmissions altogether in wireline VoIP.

In addition to the basic access scheme, there is an optional four-way handshaking technique called the RTS/CTS mechanism. In the RTS/CTS mode, the transmitting station reserves the channel by transmitting a Request-To-Send (RTS) short frame. The RTS frame helps in reserving the channel for transmission as well as silencing any stations which hear it. The target station, on successful reception of the RTS frame, responds with a Clear-To-Send (CTS) frame, which also helps in silencing nearby stations. Once the RTS/CTS exchange is completed, the transmitted station resumes its packet transmission and the destination acknowledges a successful reception with an ACK. As collisions can only occur during the RTS/CTS exchange, the packet error rate caused by collisions is significantly reduced. In addition, the RTS/CTS scheme is also helpful in combating the hidden terminal problem. However, the RTS/CTS exchange has a higher latency and is used only in high-capacity environments with significant contention.

Since it is widely implemented and more appropriate for low delay applications such as conversational voice, most work has considered the basic DCF access scheme for voice over Wi-Fi investigations.

An important issue in using WLANs for voice communications is the fundamental inefficiency of CSMA/CA, particularly with the relatively long packet headers and the short packet payloads. The expected number of voice conversations supported by an 802.11b wireless access point with a transmission rate of 11 Mbits/s without collisions, without frame errors, and with no late packets varies from only 6 calls for both G.711 and G.729 at a delay budget of 20 ms, up to 25 calls for G.711 and 34 calls for G.729 if the delay budget is 90 ms [30]. When the channel BER is 10^{-4}, the number of supported calls drops to 4 for G.711 and 5 for G.729 with a 20 ms delay budget and to 11 for G.711 and 28 for G.729 with a 90 ms delay budget. Depending upon the particular test, the intrinsic MOS for G.729 may be lower than G.711. The sharp drop in the number of G.711 calls comes from the longer payload of G.711 compared to G.729.

The critical role of the wireless channel is also evident in [49] where the MOS performance of 802.11a for independent bit error rate channels and fading channels are compared. The MOS values are obtained by using the relationship between packet error rate and MOS specified in the E-model [31]. The results show that the independent BER assumption can prove overly optimistic concerning network performance, that the average MOS may not be a good performance indicator for all channel realizations, and that up to 3 retransmissions can improve the MOS performance.

Under the current protocol, numerous retransmissions may be requested due to bit errors or packet collisions. Since packet loss concealment algorithms can be fairly effective for many voice codecs, it may be better simply to discard a packet when the CRC fails and use packet loss concealment. Alternatively, perhaps it is not necessary to detect bit errors in all of the bits in a packet, but it is necessary to include only the perceptually significant bits in the CRC; if an error is detected, then a retransmission can be requested, or alternatively, concealment can be employed. These ideas can be studied through cross layer designs involving the MAC and Application layers, and much interesting work is being pursued along these lines. Furthermore, it is expected that involving the PHY layer by adapting packet size (say) in relationship to the quality of the channel will prove rewarding.

While voice over Wi-Fi is projected to be an exponentially growing market in the next five years, it is just now in its formative stages; however, there are some proprietary products and systems being offered. We mention only two here. First, Cisco announced their Wireless IP Phone 7920 for 802.11b [32]. The voice codecs available in this phone are G.711 and G.729A. Spectralink offers voice over Wi-Fi voice systems for businesses [33]. This system also operates over 802.11b and uses G.711 and G.729A as the candidate voice codecs. Additionally, Spectralink implements a proprietary protocol that gives voice priority over data within the confines of the 802.11 standard. In particular, their protocol specifies zero back-off after each packet transmission if the next packet is voice (This is in contrast to the requirement of random back-off after each transmission in 802.11, which results in variable delays of packets). Further, a priority queuing method (not specified in 802.11) is used to push voice packets to the head of the transmission queue.

One way to transmit voice over WLANs is to employ a reservation scheme that guarantees delay and bandwidth. A new IEEE standard, 802.11e, is designed to support delay-sensitive applications with multiple managed levels of QoS for data, voice, and video.

Streaming Media and the MPEG-4 Speech Coding Tools

The MPEG-4 audio coding standard specifies a complete toolbox of compression methods for everything including low bit rate narrowband speech, wideband speech, and high quality audio. It offers several functionalities not available in other standards as well, such as bit rate scalability (also called SNR scalability) and bandwidth scalability. We provide an overview here with an emphasis on narrowband and wideband speech coding. Tables VIII and IX summarize the many options available in the MPEG-4 toolbox [34].

The HVXC (Harmonic Vector Excitation Coder) tool performs a linear prediction analysis and calculates the prediction error signal. The prediction error is then transformed into the frequency domain where the pitch and envelope are analyzed. The envelope is quantized using vector quantization for voiced segments and a search for an excitation is performed for unvoiced speech.

The CELP coder in the MPEG-4 toolbox utilizes either a multipulse excitation (MPE) or a regular pulse excitation (RPE), which were both predecessors of code-excited systems. The MPE provides for better quality but it is more complicated than RPE. The coder also uses a long term pitch predictor rather than an adaptive codebook. Note from Table IX that there are 28 bit rates from 3850 bits/s to 12.2 kbps for narrowband speech, and 30 bit rates from 10.9 kbps to 23.8 kbps for wideband speech. The larger changes in rate come about by changes in the frame size, which, of course, leads to greater delay.

The functionalities built into the HVXC and CELP speech coding tools are impressive. The HVXC speech coder has a multi-bit rate capability and a bit rate scalability option. For multi-bit rate coding, the bit rate is chosen from a set of possible rates upon call setup. For HVXC, there are only two rates, 2 kbps and 4 kbps. For CELP, bit rates are selectable in increments as small as 200 bits/s. Bit rate scalability can be useful in multicasting as well as many other applications. The bit rate scalability options in MPEG-4 are many indeed. The HVXC and CELP coders can be used to generate core bit streams that are enhanced by their own coding method or by one of the other coding methods available in the MPEG-4 natural audio coding tool. In bit rate scalability, the enhancement layers can be added in rate increments of 2 kbps for narrowband speech, and in increments of 4 kbps for wideband speech. For bandwidth scalable coding, the enhancement bit stream increments depend upon the total coding rate. Table X summarizes the enhancement bit streams for bandwidth scalability in relation to the core bit stream rates.

The speech quality produced by the MPEG-4 natural audio coding tool is very good in comparison to other popular standards, especially considering the range of bit rates covered. For example, at 6 kbps the MPEG-4 tool produces an MOS comparable to G.723.1, and at 8.3 kbps, the MOS value achieved is comparable to G.729, while at 12 kbps, it performs as well as the GSM EFR at 12.2 kbps. The bit rate scalable modes perform slightly poorer than G.729 at 8 bits/s and the GSM EFR at 12 kbps. Thus, as should be expected, bit rate scalable functionality extracts a penalty in coder performance.

Satellite and Secure Telephony

Speech coder design for secure voice communications over wireline, HF radio, and satellite connections was an early driving force in speech coding research, and important secure voice applications exist today. A mixed excitation linear prediction (MELP) codec at 2.4 and 1.2 kbps was standardized by the U. S. Government's DoD Digital Voice Processing Consortium DDVPC) in 1996 [35]. This codec is based upon the familiar linear prediction model but it employs a mixed excitation with voicing determined for each of five frequency bands, and a frequency domain calculation of the first ten pitch harmonics. Many other refinements are included to improve the speech quality. An enhanced dual-rate (1.2/2.4 kbps) MELP coder, called MELPe, was selected for the NATO STANAG 4591 standard, which is for the universal secure voice system [36]. The MELPe codec achieves significant coding efficiency at 1.2 kbps by grouping the parameters from three 22.5 ms frames and jointly quantizing the parameters in this superframe.

Multiband excitation codecs do not rely on the linear prediction model and determine voicing for a relatively large number of frequency bands. The Improved Multiband Excitation (IMBE) codec has been standardized for INMARSAT-M at a rate of 4.15 kbps, and the advanced multiband excitation (AMBE) coder at 3.6 kbps has been standardized for INMARSAT mini-M [37]. MBE-based coders have been adopted as standards in other satellite and radio-based communications applications.

V. Research Directions and Challenges

Drawing on the preceding discussions, we briefly highlight some near term challenges in speech coding and some possible future research directions.

Tandem Connections

Earlier sections have discussed asynchronous tandem connections of different codecs in some detail; however, it is worthwhile to summarize the broad scenarios that are expected to persist. First, it is common for mobile-to-mobile calls to have asynchronous tandem connections of different codecs because cell phones may have different codecs. Second, there has been the implicit assumption that the backbone network for cellular calls will be the PSTN and that G.711 is the backbone codec of choice. However, with VoIP moving to the forefront, codecs other than G.711 may be relied upon for voice coding in the backbone network. Furthermore, it can be expected that packet-switched voice will also be carried over WLAN links, which again may not use G.711. To improve the performance of asynchronous tandems, temporary fixes such as removing or modifying any postfilters present should be pursued, plus additional work on parameter transcoding of different codecs is needed.

As noted in the section on Digital Cellular, provisions are being made for end-to-end packet connections with tandem free operation under some network scenarios. While such tandem free operation does eliminate asynchronous tandeming of speech codecs, it requires that the codec be negotiated between the two endpoints of the call and that both users support a common, desirable codec. Given the opportunities for innovation in VoIP and voice over WLANs, it seems that tandem free operation may be the exception rather than the rule.

Error Concealment

From previous discussions, it is evident that packet loss concealment is important and that considerable attention is being given to methods for packet loss concealment. Continued research on packet loss concealment is important to maintain conversational voice quality as our networks become more heterogeneous. Packet loss concealment methods need to be tailored to the specific codec and should use schemes that take advantage of everything that is known about the current state of the codec and the coded speech frame. This area offers considerable promise for performance improvements in the near future.

Functionalities

Other than the MPEG-4 Natural Coding Toolbox, relatively little effort has been put into developing SNR scalable or bandwidth scalable speech coders. Part of the reason for this relatively low level of effort is the lack of motivating applications. Motivation for SNR scalable coding includes the desire to cater to the different bandwidth connections of different users and to offer the possibility that the core layer can be better error-protected than the enhancement layers. Some digital cellular codecs already error protect certain sets of bits more than others, and since the newer variable rate digital cellular codecs can change rate and the coded band on a frame-by-frame basis, the scalability option again becomes less of an issue. It may eventually develop that pruning bit rates for wireless access points as the number of users increases might be advantageous, and it is certainly likely that for mobile ad hoc networks, unequal error protection and allowing intermediate nodes to prune bit rate and hence save battery power will be important.

Speech coding methods that offer a diversity advantage by using multiple transmitted bit streams will be critical for voice over mobile ad hoc networks, since nodes may leave at any time, and the re-establishment of a route can take at least 100 msec. The most commonly cited example of a diversity-based coding method is multiple descriptions coding. However, efficient, high quality multiple descriptions speech codecs are just being developed. Additionally, due to packet header overheads, the reduction in bit rate provided by multiple descriptions codes may not be significant, and therefore, simple repeated transmission of a single description bit stream over multiple routes may be preferable.

Multihop Wireless Networks

In addition to mobile ad hoc networks, discussed briefly in the preceding section, other multihop wireless connections, such as multiple cascaded WLAN access points or mesh networks, will appear and be used for conversational voice communications. For such multihop wireless connections, old issues will need to be reexamined and new issues will appear. For example, what would be the best speech codec for ad hoc networks or mesh networks, especially if the latter is used as a backbone for voice service? Further, in multihop networks, it seems obvious that decoding and re-encoding is undesirable, but it is less clear whether packet loss concealment of some kind should be attempted after each hop or just at the endpoints of the connection.

Speech Codec Design

From prior discussions, it is evident that taking a standard for one network or application and applying it to a different network or application is common, even though the two networks or applications may be quite different, e.g., the PSTN and voice over Wi-Fi. While creative additions to standards, such as adding packet loss concealment [25, 26] to the G.711 codec or comfort noise to the G.711, G.726, and G.728 codecs [38] help to extend their capabilities, it is intuitive that designing a voice codec with the particular network or application in mind would yield a better match in terms of voice quality, rate, latency, robustness to errors, and complexity. Voice codecs for the PSTN and digital cellular systems have been very successful with such an approach. With VoIP, voice over Wi-Fi, and voice over multihop networks of rapidly evolving interest, research on voice codecs with the particular characteristics of these networks/applications in mind offers unique opportunities. It is noted, of course, that such integrated designs can lead to other challenges in terms of asynchronous tandeming of different codecs. However, research on speech codecs which incorporate robustness to both errors and expected end-to-end network conditions (such as tandeming) could prove fruitful.

Codec and Network Performance Assessment

One of the key issues to be resolved in evolving voice communications networks is how to monitor voice quality in the network with the eventual goal of adapting codec/network parameters to maintain the quality desired by the user. Objective quality assessment of voice codec performance can be classified as intrusive, wherein the input clean speech is available, and non-intrusive, where only the degraded or processed speech is available. The ITU-T recently produced a standard algorithm for non-intrusive objective quality assessment designated P.563 [40]. We do not describe details of P.563 here, but we note that the development of such assessment methods is a critical research area, and more research in this area, such as in [41] is needed.

Noise Pre-Processors

The presence of unwanted sounds in the background of the voice signal to be compressed has long been recognized as a challenging problem. The basic problem is that as the codecs become more speech model based, this model can be forced upon the non-speech-like background sounds, thus creating artifacts in the coded speech. The background sounds can also redirect the codec signal processing efforts away from the desired signal. This latter effect is partially related to not having the appropriate input level for the desired speech with respect to the other sounds. The EVRC digital cellular codec has a front end processor and the Japanese Half-Rate Digital Cellular Standard provides for an optional noise canceller [46, 47]. Recent work on a preprocessor for noise reduction is reported in [48]. In comparing the approaches used in the cited preprocessors, it is evident that quite different methods have been used and that there is considerable room for further research.

VI. Summary and Conclusions

Speech coding is an integral part of our backbone communications services. As wireline VoIP becomes more important and voice over Wi-Fi is introduced, the end-to-end networks that support conversational voice will become much less homogeneous with respect to protocols, latency, physical layer characteristics, and voice codecs. Furthermore, it will be more difficult to develop and standardize optimal end-to-end designs that incorporate these disparate multihop connections. As a result, continued work that addresses asynchronous tandem connections of speech coders, latency, packet jitter, and packet loss concealment will be essential if we are to maintain the high quality of voice services that we have come to expect.

Cross layer designs involving the Application, Media Access, and Physical layers will be necessary to obtain the requisite quality and reliability and to improve the efficiency of the wireless links. This will require that protocol developers, speech codec designers, and physical layer engineers collaborate in establishing future voice communications solutions.

It is expected that mobile ad hoc networks and mesh networks will be used for voice communications as well. These networks will motivate the development of increased speech codec functionalities, such as bit rate scalability, bandwidth scalability, and diversity-oriented methods along the lines of multiple descriptions coding.

Finally, we note that voice is the preferred method of human communication. Although there have been times when it seemed that the voice communications problem was solved, such as when the PSTN was our primary network or later when digital cellular networks reached maturity, such is not the case today. Reflecting upon the issues and developments highlighted in this paper, it is evident that there is a diverse set of challenges and opportunities for research and innovation in speech coding and voice communications.

Appendix A: Speech Quality and Intelligibility

To compare the performance of two speech coders, it is necessary to have some indicator of the intelligibility and quality of the speech produced by each coder. The term intelligibility usually refers to whether the output speech is easily understandable, while the term quality is an indicator of how natural the speech sounds. It is possible for a coder to produce highly intelligible speech that is low quality in that the speech may sound very machine-like and the speaker is not identifiable. On the other hand, it is unlikely that unintelligible speech would be called high quality, but there are situations in which perceptually pleasing speech does not have high intelligibility. We briefly discuss here the most common measures of intelligibility and quality used in formal tests of speech coders. We also highlight some newer performance indicators that attempt to incorporate the effects of the network on speech coder performance in particular applications.

MOS

The Mean Opinion Score (MOS) is an often-used performance measure [5, Chap. 13; 11, Appendix F]. To establish a MOS for a coder, listeners are asked to classify the quality of the encoded speech in one of five categories: excellent (5), good (4), fair (3), poor (2), or bad (1). The numbers in parentheses are used to assign a numerical value to the subjective evaluations, and the numerical ratings of all listeners are averaged to produce a MOS for the coder. A MOS between 4.0 and 4.5 usually indicates high quality.

It is important to compute the variance of MOS values. A large variance, which indicates an unreliable test, can occur because participants do not know what categories such as good and bad imply. It is sometimes useful to present examples of good and bad speech to the listeners before the test to calibrate the 5-point scale. Sometimes the percentage of poor and bad votes may be used to predict the number of user complaints. MOS values can and will vary from test to test and so it is important not to put too much emphasis on particular numbers when comparing MOS values across different tests.

EMBSD

A relatively new objective measure that has a high correlation with MOS is the enhanced modified bark spectral distance (EMBSD) measure [39]. The EMBSD is based on the bark spectral distance measure that relates to perceptually significant auditory attributes. A value of zero for the EMBSD indicates no distortion and a higher value indicates increasing distortion. The G.729 codec has been tested to have an EMBSD of 0.9, indicating low distortion in the reconstructed speech. The EMBSD values are often mapped into MOS values, since acceptable MOS values are more readily known.

DRT

The diagnostic rhyme test (DRT) was devised to test the intelligibility of coders known to produce speech of lower quality. Rhyme tests are so named because the listener must determine which consonant was spoken when presented with a pair of rhyming words; that is, the listener is asked to distinguish between word pairs such as meat-beat, pool-tool, saw-thaw, and caught-taught. Each pair of words differs on only one of six phonemic attributes: voicing, nasality, sustention, sibilation, graveness, and compactness. Specifically, the listener is presented with one spoken word from a pair and asked to decide which word was spoken. The final DRT score is the percent responses computed according to $P = (R - W)\text{x}100/T$, where R is the number correctly chosen, W is the number of incorrect choices, and T is the total of word pairs tested. Usually, $75 \leq \text{DRT} \leq 95$, with a good being about 90.

DAM

The diagnostic acceptability measure (DAM) developed by Dynastat is an attempt to make the measurement of speech quality more systematic. For the DAM, it is critical that the listener crews be highly trained and repeatedly calibrated in order to get meaningful results. The listeners are each presented with encoded sentences taken from the Harvard 1965 list of phonetically balanced sentences, such as “Cats and dogs each hate the other” and “The pipe began to rust while new”. The listener is asked to assign a number between 0 and 100 to characteristics in three classifications—signal qualities, background qualities, and total effect. The ratings of each characteristic are weighted and used in a multiple nonlinear regression. Finally, adjustments are made to compensate for listener performance. A typical DAM score is 45 to 55%, with 50% corresponding to a good system.

PESQ

A new and important objective measure is the perceptual evaluation of speech quality (PESQ) method in ITU Recommendation P.862, which attempts to incorporate more than just speech codecs but also end-to-end network measurements [12]. The PESQ has been shown to have good accuracy for the factors listed in Table XI. It is clear that this is a very ambitious and promising testing method. There are parameters for which the PESQ is known to provide inaccurate predictions or is not intended to be used with, such as listening levels, loudness loss, effect of delay in conversational tests, talker echo, and two-way communications. The PESQ also has not been validated to test for packet loss and packet loss concealment with PCM codecs, temporal and amplitude clipping of speech, talker dependencies, music as input to a codec, CELP and hybrid codecs < 4 kbps, and MPEG-4 HVXC, among others. Also, as noted the wideband PESQ tends to underestimate the MOS by at least 0.5, and some work is being pursued in this area [42].

The E-Model

Another relatively recent objective method for speech quality evaluation is the E-Model in ITU Recommendation G.107 and G.108 [31]. The E-Model attempts to assess the "mouth to ear" quality of a telephone connection and is intended to be used in network planning. The E-Model has components for representing the effects of "equipment" and different types of impairments. The equipment effects include a term representing the "intrinsic" quality of the codec and a term due to the effects of packet loss concealment. There is also a term to model the loss in quality caused by delay. All of the components are summed to produce an R-value on a psychoacoustic scale, which maps to user satisfaction as shown in Fig. 1. Guidelines are given for mapping the R-values into MOS values as shown in Table XII. Some studies caution that the E-model may not sufficiently distinguish between various combinations of codec impairments and delay that are clearly perceived differently by users.

Appendix B: Standardization Bodies

There are many speech coding standards mentioned in this paper, and different standards bodies set the standards for different networks and applications. The Telecommunications Standardization Sector of the International Telecommunications Union (ITU-T) is responsible for setting the standards for telephony and video telephony, with the primary network being the PSTN. The standards bodies, such as the Group Speciale Mobile (GSM), the Third Generation Partnership (3GPP), and 3GPP2, set the speech coding standards for the various digital cellular applications. Standards for multimedia content distribution described herein are set by the Moving Pictures Experts Group (MPEG), which is Working Group 11 (WG11) of Subcommittee 29 of the International Organization for Standardization/International Electrotechnical Commission (ISO/IEC). Sometimes these groups come together to specify a joint standard such as the H.264/AVC video coding standard, which is both an ITU-T standard and an MPEG-4 standard. Standards for Internet applications are developed by the Internet Engineering Task Force (IETF). The IEEE has been active in the standardization of protocols and other computer network standards, such as the 802.11 WLAN series, but not in setting speech coding standards. The IETF involvement in specifying VoIP and voice over WLAN speech codecs is more to develop protocols that allow certain desired speech coders to be used in those applications. The IETF has not been involved in voice codec design thus far.

Acknowledgments

The author gratefully acknowledges the comments of the referees of this paper, the comprehensive comments provided by Rich Cox, Sassan Ahmadi, Wai Chu, Andreas Spanias, Juan Carlos de Martin, and

Antonio Servetti, and Juin-Whey Chen for the information concerning the Broadcom codecs. The author also acknowledges his students, J. Balam, S. Choudhury, J. Hu, J. Jung, and N. Shetty at UCSB, and B. Wei at SMU, and his colleagues, Professors E. Belding-Royer, K. Almeroth, and A. Gersho at UCSB, who have deepened the author's understanding of the field. This work was supported by the California Micro Program, Applied Signal Technology, Dolby Labs, Inc. and Qualcomm, Inc., by NSF Grant Nos. CCF-0429884 and CNS-0435527, and by the UC Discovery Grant Program and Nokia, Inc..

References **(In keeping with the goals of magazine articles, this is an abridged list of references. Cited references are not intended to attribute credit for ideas, concepts, algorithms, or codec designs, but are cited for the readers convenience.)**

[1] J. D. Gibson, T. Berger, T. Lookabaugh, D. Lindbergh, and R. L. Baker, *Digital Compression for Multimedia: Principles & Standards,* Morgan-Kaufmann, 1998.
[2] A. Gersho, "Advances in speech and audio compression," *Proceedings of the IEEE*, vol. 82, pp. 900-918, June 1994.
[3] A. S. Spanias, "Speech coding: A tutorial review," *Proceedings of the IEEE*, vol. 82, pp. 1541-1582, Oct. 1994.
[4] M. Budagavi and J. D. Gibson, "Speech Coding in Mobile Radio Communications, "*Proceedings of the IEEE*, vol. 86, pp. 1402-1412, July, 1998.
[5] W. B. Kleijn and K. K. Paliwal, eds., *Speech Coding and Synthesis*, Amsterdam, Holland: Elsevier, 1995.
[6] A. M. Kondoz, *Digital Speech: Coding for Low Bit Rate Communication Systems*, West Sussex, England: John Wiley & Sons, 2004.
[7] W. C. Chu, *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders*, Hoboken, NJ: Wiley Interscience, 2003.
[8] L. Hanzo, et al, *Voice Compression and Communications: Principles and Applications for Fixed and Wireless Channels*, New York: Wiley-IEEE Press, 2001.
[9] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Englewood Cliffs, NJ: Prentice-Hall, 2001.
[10] D. O'Shaughnessy, *Speech Communications: Human and Machine*, New York: IEEE Press, 2000.
[11] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*, Englewood Cliffs, NJ: Prentice-Hall, 1984.
[12] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Feb. 2001.
[13] ITU-T Recommendation G.114, One-Way Transmission Time, May 2000.
[14] J. C. Bellamy, *Digital Telephony*, John Wiley & Sons, 2000.
[15] T. Painter and A. Spanias, "Perceptual Coding of Digital Audio," *Proceedings of the IEEE*, vol. 88, pp. 451-513, April 2000.
[16] H. S. Malvar, *Signal Processing with Lapped Transforms*, Norwood, MA: Artech House, 1992.
[17] J.-H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Trans. Speech and Audio Processing,* vol. 3, pp. 59-71, Jan. 1995.
[18] ITU-T Recommendation G.722.1, Coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss, Sept. 1999.
[19] ITU-T Recommendation G.722.2, Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB), 2002.

[20] T. S. Rappaport, *Wireless Communications: Principles and Practice*, Prentice-Hall, second edition, 2002.
[21] R. Steele and L. Hanzo, eds., *Mobile Radio Communications*, second edition, John Wiley & Sons, 1999.
[22] B. Bessette, et al, "The adaptive multirate wideband speech codec (AMR-WB)," *IEEE Trans. on Speech and Audio Processing*, vol. 10, pp. 620-636, Nov. 2002.
[23] M. Jelinek, et al, "On the architecture of the cdma2000® variable-rate multimode wideband (VMR -WB) speech coding standard," *Proc. ICASSP*, 2004,pp. I-281--I-284.
[24] B. Goode, "Voice Over Internet Protocol (VoIP)," *Proceedings of the IEEE,* vol. 90, pp. 1495-1517, Sept. 2002.
[25] ITU-T, G.711, Appendix I: A high quality low-complexity algorithm for packet loss concealment with G.711," Sept. 1999.
[26] ANSI Standard T1.521a-2000, Packet loss Concealment for Use with ITU-T Recommendation G.711, June 7, 2000.
[27] R. Salami, et al, "Design and description of CS-ACELP: A toll quality 8 kb/s speech coder," *IEEE Trans. on Speech and Audio Processing*, vol. 6, pp. 116-130, March 1998.
[28] A. P. Markopoulou, F. A. Tobagi, and M. J. Karam, "Assessment of VoIP quality over Internet backbones," *Proc. IEEE INFOCOM 2002*, pp. 150-159.
[29] IEEE Std. 802.11a/D7.0-1999, Part 11, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: High Speed Physical Layer in the 5 GHz Band.
[30] D. P. Hole and F. A. Tobagi, "Capacity of an IEEE 802.11b wireless LAN supporting VoIP," *Proc. IEEE Int. Conf. on Communications*, 2004.
[31] ITU-T Recommendation G.107, The E-model, a computational model for use in transmission planning, March 2003.
[32] http://www.cisco.com
[33] http://www.spectralink.com
[34] K. Brandenburg, O. Kunz, and A. Sugiyama, "MPEG-4 natural audio coding," *Signal Processing: Image Communication,* Vol. 15, 2000, pp. 423-444.
[35] L. M. Supplee, R. P. Cohn, J. S. Collura, and A. V. McCree, "MELP: The new Federal standard at 2400 bps," *Proc. ICASSP*, pp. 1591-1594, 1997.
[36] T. Wang, K. Koishida, V. Cuperman, A. Gersho, and J. S. Collura, "A 1200/2400 bps coding suite based on MELP," 2002 IEEE Workshop on Speech Coding, Tsukuba, Ibaraki, Japan, Oct. 6-9.
[37] J. C. Hardwick and J. S. Lim, "The application of the IMBE speech coder to mobile communications," *Proc. ICASSP*, pp. 249-252, 1991.
[38] ITU-T Recommendation G.711, Appendix II: A comfort noise payload definition for ITU-T G.711 use in packet-based multimedia communication systems, Feb. 2000.
[39] W. Yang, M. Benbouchta, and R. Yantorno, "Performance of the modified bark spectral distortion as an objective speech quality measure," *Proc. ICASSP*, Seattle, 1998, pp. 541-544.
[40] ITU-T P.563, "Single ended method for objective speech quality assessment in narrow-band telephony applications," ITU, Geneva, Switzerland, May 2004.
[41] T. H. Falk, Q. Xu, and W.-Y. Chan, "Non-Intrusive GMM-Based speech quality measurement," *Proc. ICASSP 2005*, March 18-23, pp. I-125--I-128.
[42] C. Morioka, A. Kurashima, and A. Takahashi, "Proposal on objective speech quality asessment for wideband IP telephony, *Proc. ICASSP 2005,*March 18-23, pp. I-49--I-52.
[43] IETF RFC 3951, “Internet Low Bit Rate Codec (iLBC),” 2004.
[44] PacketCable™ 1.5 Specifications, Audio/Video Codecs, Jan. 28, 2005.
[45] Broadcom, "Enabling High Quality Voice over Internet Protocol (VoIP) Services with BroadVoice™ Speech Codecs and Wideband Telephony," March 2005.
[46] T. Ohya, H. Suda, and T. Miki, "5.6 kbits/s PSI-CELP of the half-rate PDC speech coding standard, *Proc. IEEE Veh. Tech. Conf.*, pp. 1680-1684, June 1994.

[47] B. Koo, J. D. Gibson, and S.D. Gray, “Filtering of Colored Noise for Speech Enhancement and Coding,” *IEEE Trans. Signal Processing,* vol. 39, pp. 1732-1742, Aug 1991.

[48] R. Martin, D. Malah, R. V. Cox, and A. J. Accardi, "A Noise Reduction Preprocessor for Mobile Voice Communication," *EURASIP J. Applied Signal Processing.*

[49] O. Awoniyi and F. A. Tobagi, "Effect of fading on the performance of VoIP in IEEE 802.11a WLANs," *Proc. ICC 2004*, June 20-24, pp. 3712-3717.

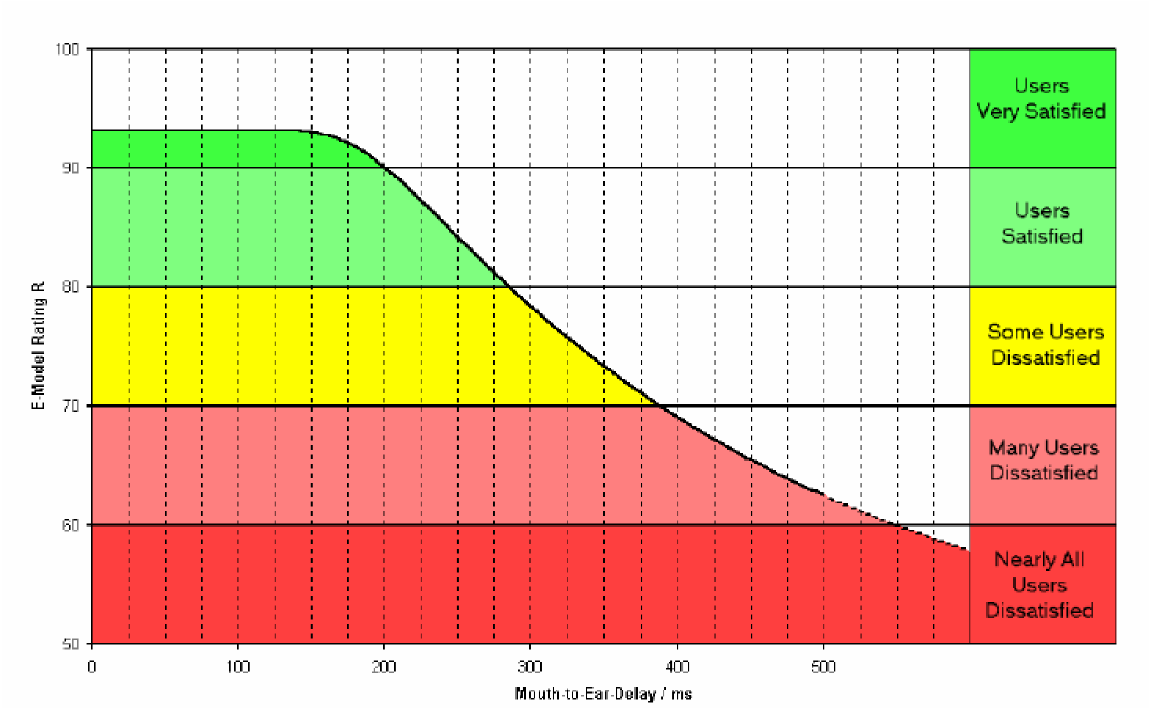

Figure 1. Effect of One-Way Delay on Speech Quality (G.114)

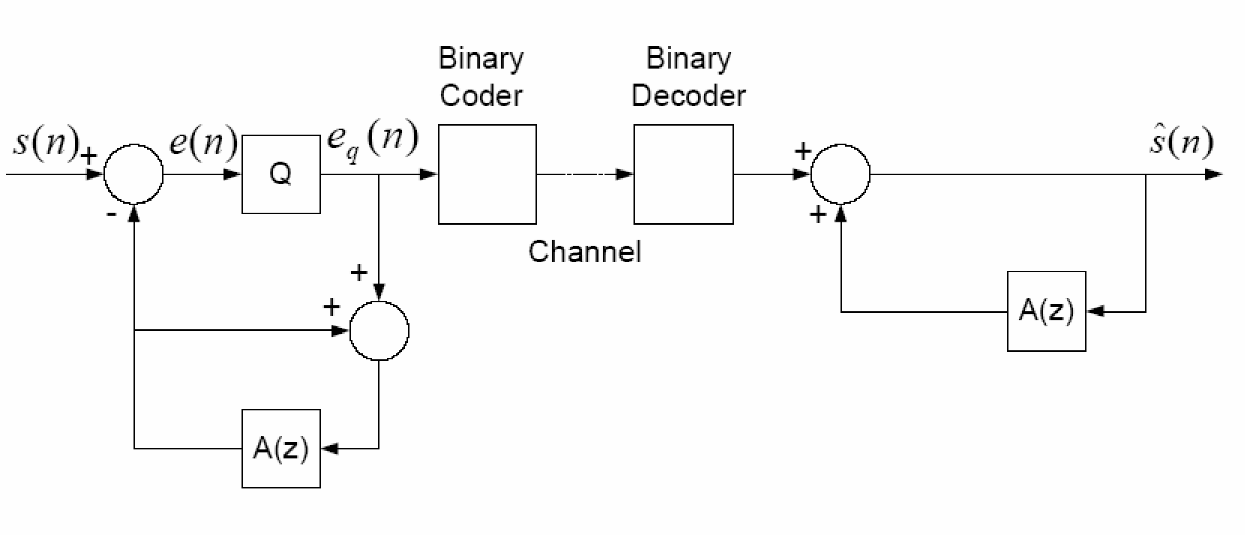

Figure 2. An ADPCM Speech Encoder and Decoder

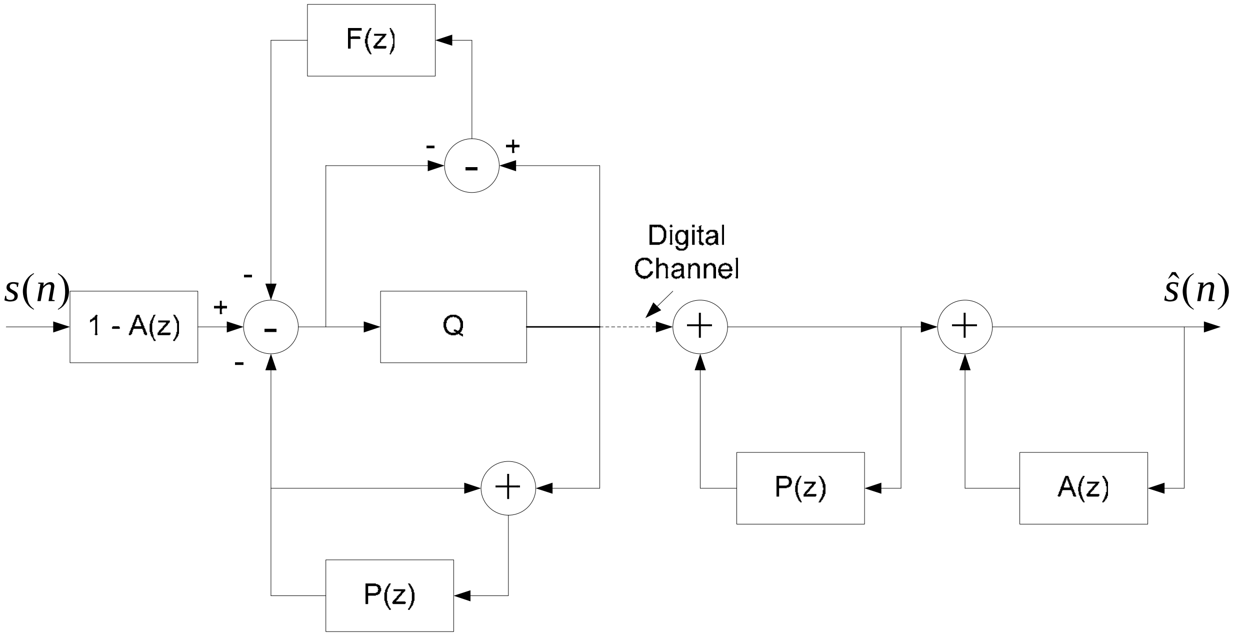

Figure 3. Adaptive Predictive Coding with a Long Term Predictor and Noise Spectral Shaping

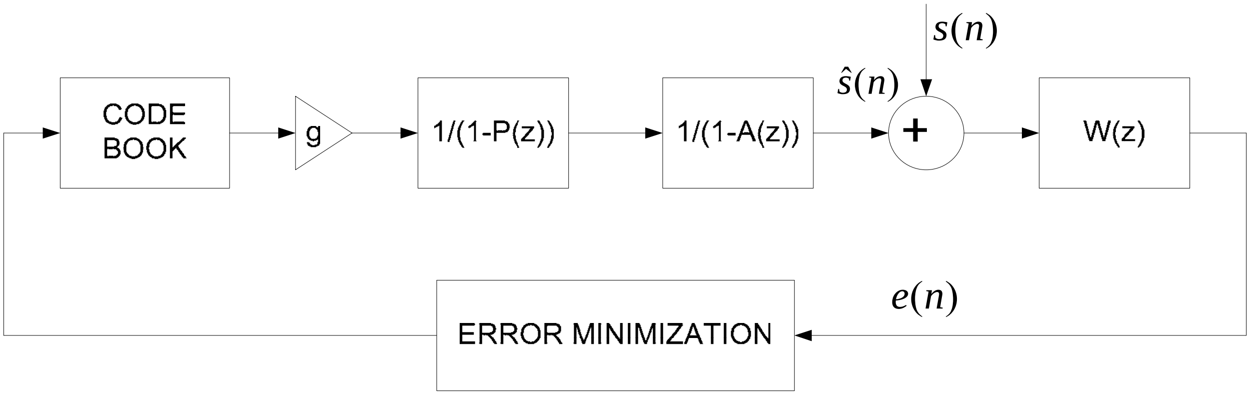

Figure 4(a). Encoder for Code-Excited Linear Predictive (CELP) Coding

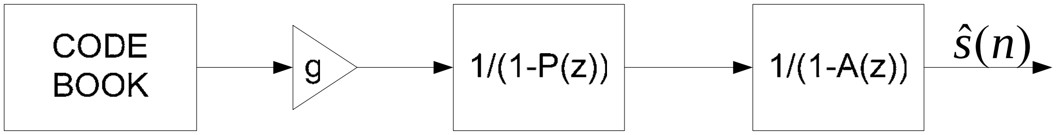

Figure 4(b). Decoder for CELP Coding

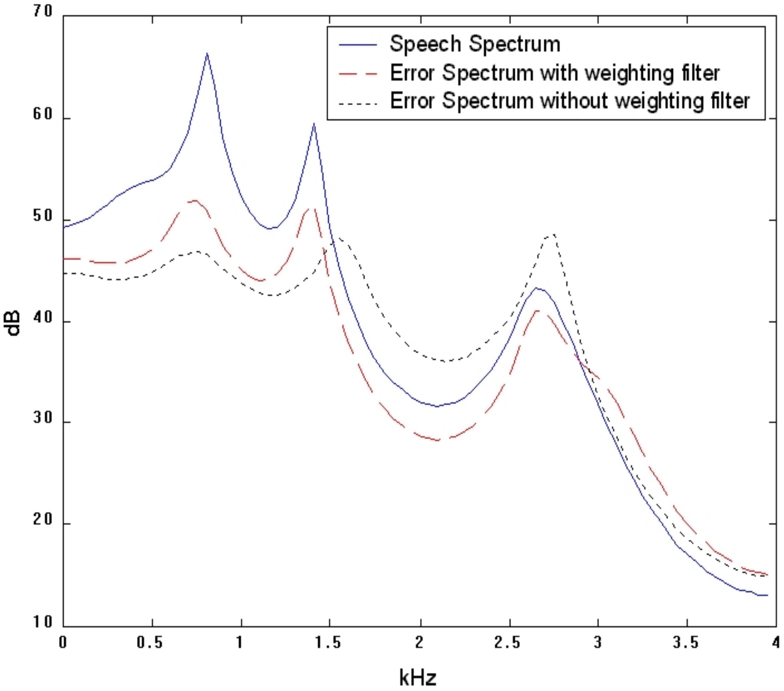

Figure 5. Perceptual Weighting of the Coding Error as a Function of Frequency

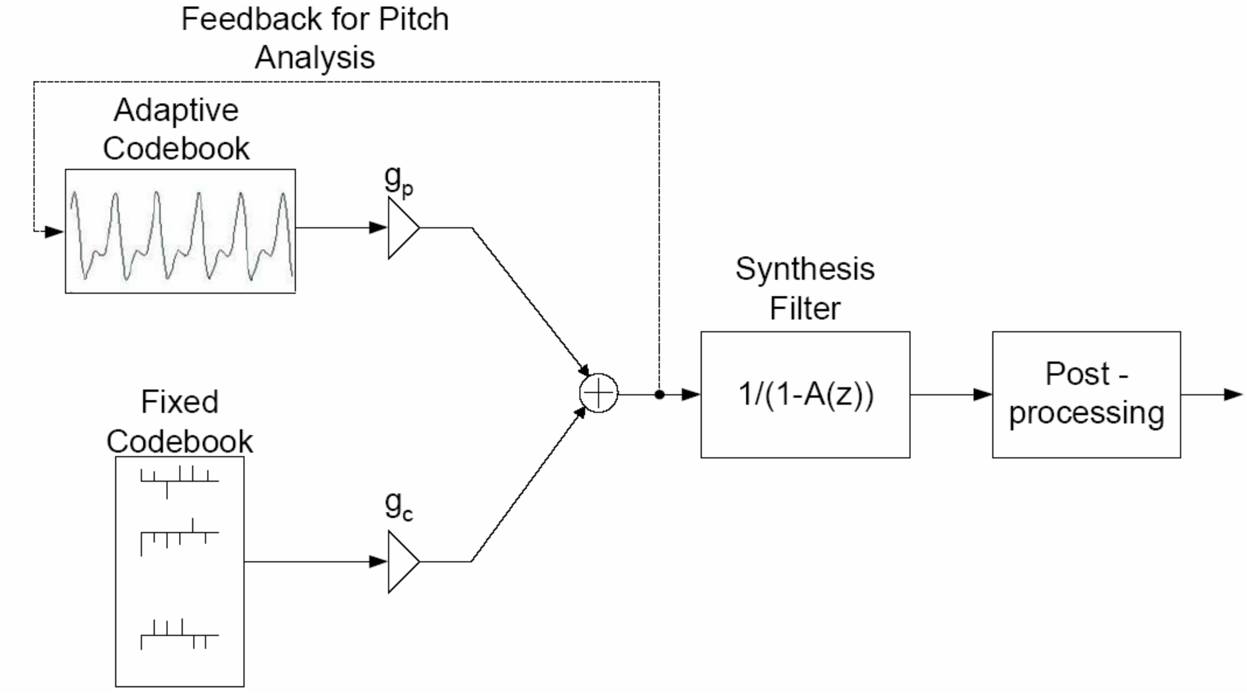

Figure 6. CELP Decoder with an Adaptive Codebook and Postfiltering

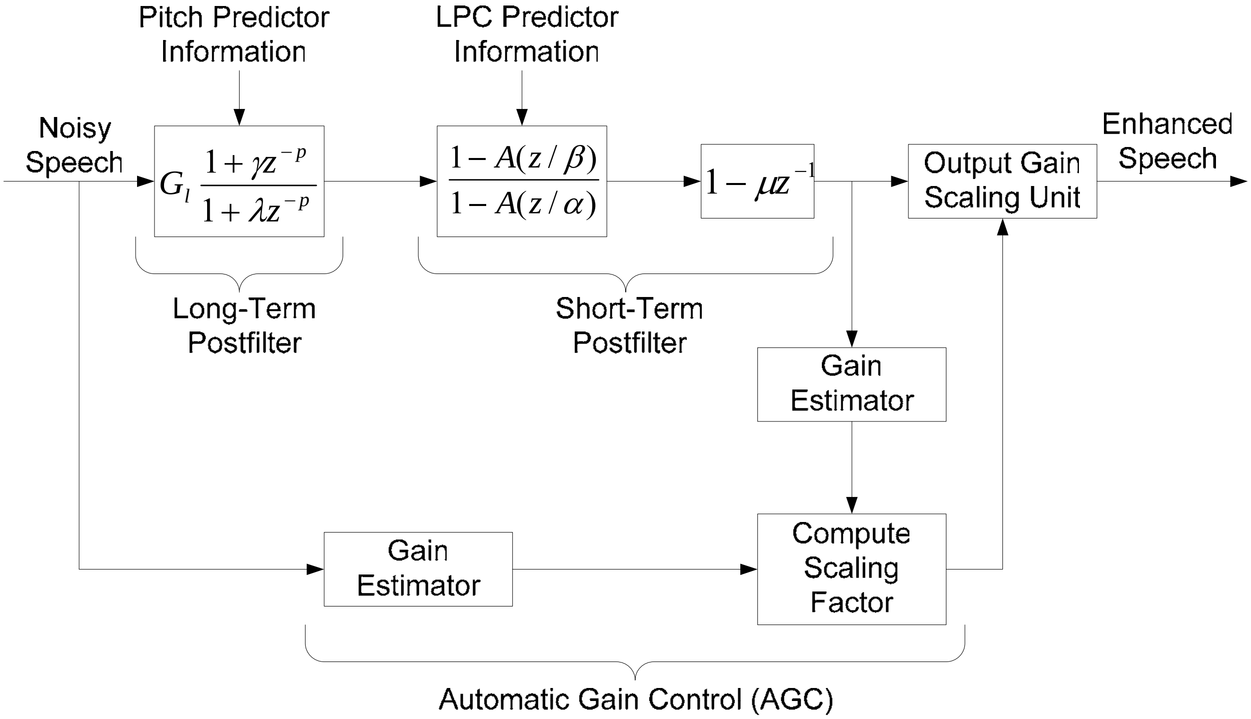
Pitch Predictor Information
LPC Predictor Information
Noisy Speech
$G_l \frac{1+\gamma z^{-p}}{1+\lambda z^{-p}}$
$\frac{1-A(z/\beta)}{1-A(z/\alpha)}$
$1-\mu z^{-1}$
Output Gain Scaling Unit
Enhanced Speech
Long-Term Postfilter
Short-Term Postfilter
Gain Estimator
Gain Estimator
Compute Scaling Factor
Automatic Gain Control (AGC)

Figure 7. Block Diagram of a General Postfilter Structure

Table I. Comparison of Voice Codecs for the PSTN

Codec	Rate (kbps)	MOS	Complexity (MIPS)	Frame Size/Look Ahead (ms)
G.711	64	4.0+	<<1	0.125
G.721/726	32	~4.0	1.25	0.125
G.728	16	3.9	30	0.625
G.729	8	4.0	20	10/5
G.729A	8	4.0	12	10/5
G.723.1	5.3/6.3	3.7/3.9	11	30/7.5

Table II. Representative Asynchronous Tandem Performance of Selected PSTN Codecs

Voice Codec	Mean Opinion Score (MOS)
G.711x4	>4.0
G.726x4	2.91
G.729x2	3.27
G.729x3	2.68
G.726+G.729	3.56
G.729+G.726	3.48

Table III. Characteristics of Some Wideband Speech Coding Standards

Standard	Bit Rate kbit/s	Coding Method	Quality	Frame Size/ Look ahead	Complexity
G.722	48, 56, 64	Subbband ADPCM	Commentary Grade	0.125/1.5	10 MIPS
G.722.1	24 and 32	Transform	Good Music Poorer for Speech	20/20	<15 MIPS
G.722.2	23.85	ACELP	Good Speech Poor Music	20/5	<40 MIPS

Table IV. Comparison of Selected Digital Cellular Voice Codecs

Codec	Rate (kbps)	MOS	Complexity (MIPS)	Frame Size/ (ms) Look Ahead
IS-641	7.4	4.09	14	20/5
IS-127-2 EVRC	8.55, 4.0, 0.8	3.82	20	20
GSM-EFR	12.2	4.26	14	20
NB-AMR	4.75-12.2	3.4-4.2	14	20/5
IS-893, cdma2000	8.5, 4.0, 2.0, 0.8	3.93 at 3.6 kbps ADR	18	20

Table V. Average Data Rates and Quality Targets for the VMR-WB Coder

Mode	Average Data Rate kbps	Quality Target
0	9.14	No worse than AMR-WB at 14.25 kbps
1	7.69	No worse than AMR-WB at 12.65 kbps
2	6.28	No worse than AMR-WB at 8.85 kbps
3	9.49	No worse than AMR-WB at 12.65 kbps
4	5.77	Same as Mode 2

Table VI. Representative Asynchronous Tandem Performance of Selected Digital Cellular Codecs

Voice Codec	Mean Opinion Score (MOS)
IS-641x2	3.62
GSM-EFRx2	4.13
IS-641+G.729	3.48
GSM-FR+G.729	3.05
GSM-EFR+G.729	3.53
GSM-EFR+G.729+G.729	3.21
IS-641+G.729+G.729	3.10

Table VII. Properties of Common VoIP Codecs

Codec	Relevant Properties
G.711	Low delay, toll quality, low complexity, higher rate
G.729	Toll quality, acceptable delay, low rate, acceptable complexity
G.723.1	Low rate, acceptable quality, relatively high delay
G.722	Wideband speech, low delay, low complexity, higher rate

Table VIII. Specifications of the HVXC Speech Coding Tool

Sampling Frequency	8 kHz
Bandwidth	300-3400 Hz
Bit Rate (bits/s)	2000 and 4000
Frame Size	20 msec
Delay	33.5-56 msec
Features	Multi-bit rate coding Bit rate scalability

Table IX. Specifications of the CELP Speech Coding Tool

Sampling Frequency	8 kHz	16 kHz
Bandwidth	300-3400 Hz	50-7000 Hz
Bit Rate (bits/s)	3850-12,200 (28 bit rates)	10,900-23,800 (30 bit rates)
Frame Size	10-40 msec	10-20 msec
Delay	15-45 msec	15-26.75 msec
Features	Multi-bit rate coding Bit rate scalability Bandwidth scalability	

Table X. Bandwidth Scalable Bit Rate Options

Core Bit Rate (bits/s)	Enhancement Layer Bit Rates (bits/s)
3850-4650	9200, 10400, 11600, 12400
4900-5500	9467, 10667, 11867, 12667
5700-10700	10000, 11200, 12400, 13200
11000-12200	11600, 12800, 14000, 14800

Table XI. Factors for which the PESQ has demonstrated acceptable accuracy

Test Factors
Speech input levels to a codec
Transmission channel errors
Packet loss and packet loss concealment with CELP codecs
Bit rates for multiple bit rate codecs
Transcodings
Environmental noise at the sending side
Effect of varying delay in listening only tests
Short-term time warping of the audio signal
Long-term time warping of the audio signal
Coding Technologies
Waveform codecs such as G.711, G.726, G.727
CELP and hybrid codecs at rates ≥ 4 kbps, such As G.728, G.729, G.723.1
Other codecs: GSM-FR, GSM-HR, GSM-EFR, GSM-AMR, CDMA-EVRC, TDMA-ACELP, TDMA-VSELP, TETRA
Applications
Codec Evaluation

Codec Selection
Live network testing using a digital or analog connection to the network
Testing of emulated and prototype networks

Table XII. Mapping of the E-Model R-Values into MOS

User Satisfaction	Very Satisfied	Satisfied	Some Dissatisfied	Many Dissatisfied	Nearly All Dissatisfied	Not Recommended
R	100-90	89-80	79-70	69-60	59-50	Below 50
MOS	4.5-4.3	4.3-4.0	4.0-3.6	3.6-3.1	3.1-2.6	Below 2.6