

Chapter 2

Challenges in Speech Coding Research

Jerry D. Gibson

Abstract Speech and audio coding underlie many of the products and services that we have come to rely on and enjoy today. In this chapter, we discuss speech and audio coding, including a concise background summary, key coding methods, and the latest standards, with an eye toward current limitations and possible future research directions.

2.1 Introduction

We distinguish between speech and audio coding according to the bandwidth occupied by the input source. *Narrowband* or telephone bandwidth speech occupies the band from 200 to 3,400 Hz, and is the band classically associated with telephone quality speech. The category of *wideband* speech covers the band 50 Hz–7 kHz. Audio is generally taken to cover the range of 20 Hz–20 kHz, and this bandwidth is sometimes referred to today as *fullband* audio [1, 2]. In recent years, quite a few other bandwidths have attracted attention, primarily for audio over the Internet applications, and the bandwidth of 50 Hz–14 kHz, designated as *superwideband*, has gotten considerable recent attention [3]. As the frequency bands being considered move upward from narrowband speech through wideband speech and superwideband audio, on up to fullband audio, the basic structures for digital processing and the quality expectations change substantially. In the following, we elaborate on these differences and highlight the challenges in combining the processing of this full range of bandwidths in single devices.

J.D. Gibson (✉)

Department of Electrical & Computer Engineering, University of California,
Santa Barbara, CA 93106-6065, USA

e-mail: gibson@ece.ucsb.edu

2.2 Speech Coding

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 [2]. Speech coding is a critical technology for videoconferencing systems, digital cellular communications, and voice over Internet protocol (VoIP) applications, while audio coding is essential for portable audio players, audio streaming, video streaming, and the storage and playback of movies.

The basic approaches for coding narrowband speech evolved over the years from waveform following codecs to code excited linear prediction (CELP) based codecs [2]. The process of this evolution was driven by applications that required lower bandwidth utilization and by advances in digital signal processing, which were facilitated by improvements in processor speeds that allowed more sophisticated processing to be incorporated. The reduction in bit rates was obtained by relaxing constraints on encoding delay and on complexity. This later relaxation of constraints, particularly on complexity, should be a lesson learned for future research; namely, complexity should not be a dominating concern at the beginning of a basic research effort.

Note that the basic speech coding problem for narrowband speech, in particular, follows the *distortion rate* paradigm; that is, given a rate constraint set by the application, the codec is designed to minimize distortion. The resulting distortion is not necessarily small or inaudible—just acceptable for the given constraints. The distortion rate structure should be contrasted with the *rate distortion* problem wherein the constraint is on allowable distortion and the rate required to achieve that distortion is minimized. Notice that for the rate distortion approach, a specified distortion is the goal and the rate is adjusted to obtain this level of distortion. Voice coding for digital cellular communications is an example of the distortion rate approach, since it has a rate constraint, while coding of fullband audio typically has the goal of transparent quality, and hence is an example of the rate distortion paradigm. We elaborate more on these ideas in the following.

We use the terms speech coding and voice coding interchangeably in this paper. 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.

Given a particular source such as voice, audio, or video, 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 channel or network bandwidth requirement, the goal is always either to minimize the rate required to satisfy the distortion constraint or minimize the distortion for a given rate constraint. For speech coding, we are interested in achieving a quality as close to the original speech as possible within the rate, complexity, latency, and any other constraints that might be imposed by the application of interest. Encompassed in the term quality are intelligibility, speaker identification, and naturalness. Absolute category rating (ACR) 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) [1].

Of course, listening tests involving human subjects are difficult to organize and perform, so the development of objective measures of speech quality is highly desirable. The perceptual evaluation of speech quality (PESQ) method, standardized by the ITU-T as P.862, was developed to provide an assessment of speech codec performance in conversational voice communications. The PESQ has been and can be used to generate MOS values for both narrowband and wideband speech [4, 5]. While no substitute for actual listening tests, the PESQ and its wideband version are widely used for initial codec evaluations and are highly useful. A newer objective measure, designated as P.863 POLQA (Perceptual Objective Listening Quality Assessment) has been developed but it has yet to receive widespread acceptance [6]. For a tutorial development of perceptual evaluation of speech quality, see [7].

More details on MOS and perceptual performance evaluation for voice codecs are provided in the references [1, 2, 7]. Later in the chapter, we discuss the relatively new nine point ACR ratings that are becoming popular as superwideband speech and audio become more prevalent in codec designs.

2.2.1 Speech Coding Methods

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 determined to be most important perceptually. 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 each subband, for reconstruction and recombination at the receiver. Extending the resolution of the frequency domain decomposition leads to transform coding and coding using filter banks, wherein a transform is performed on a frame of input speech/audio and the resulting transform coefficients are quantized and transmitted to reconstruct the speech/audio from the inverse transform. Subband decompositions and transform based decompositions are very closely related and combinations of the two are common in codecs that code bandwidths beyond narrowband speech.

2.2.1.1 Waveform Coding [2]

Familiar waveform-following methods are logarithmic pulse code modulation (log-PCM) and adaptive differential pulse code modulation (ADPCM), and both have

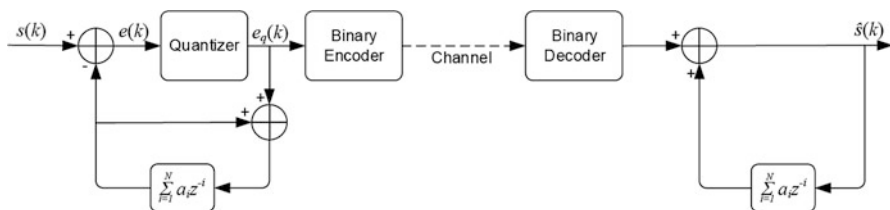


Fig. 2.1 An ADPCM encoder and decoder

found widespread applications. Log PCM at 64 kilobits/s (kbps) is the speech codec that was the work horse for decades in the long distance public switched telephone network at a rate of 64 kbps, and it is the most widely employed codec for VoIP applications. 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, South Korea and Japan, and A-law, which is used in the rest of the world. Both achieve toll quality speech, and which, in terms of the MOS value is usually between 4.0 and 4.5 for log-PCM. These quality levels are considered very good but not transparent.

ADPCM operates at 40 kbps or lower, and it achieves performance comparable to log-PCM by using an adaptive linear predictor to remove short-term redundancy in the speech signal before adaptive quantization of the prediction error. See Fig. 2.1. The reasoning behind differential coding like ADPCM is that 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 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 [2]. No predictor or quantizer parameters are sent along as side information with the quantized waveform values.

2.2.1.2 Subband and Transform Methods [2]

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 to a different accuracy 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 (with a slight recent resurgence such as the adoption of a subband codec optional for Bluetooth [8]), filter bank and transform methods play a critical role in high quality audio coding, and several important standards for wideband, superwideband, and fullband speech/audio coding are based upon filter bank and transform methods. Although it is intuitive that subband filtering and discrete transforms are closely related, by the early 1990s, the relationships between filter bank methods and transforms were well-understood [9]. 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. Often a combination of the two is the most efficient.

2.2.1.3 Analysis-by-Synthesis Methods [2, 10]

Analysis-by-synthesis (AbS) methods are a considerable departure from waveform-following techniques and from frequency domain methods as well, although they do build on linear prediction as used in ADPCM. 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, 5–10 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. 2.2a. The excitation that produces the minimum perceptually weighted coding error is selected for use at the decoder as shown in Fig. 2.2b. 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 (AbS). 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 where each sample is coded as it arrives at the coder input.

The perceptual weighting is key to obtaining good speech coding performance in CELP, and the basic idea is that the coding error is spectrally shaped to fall below

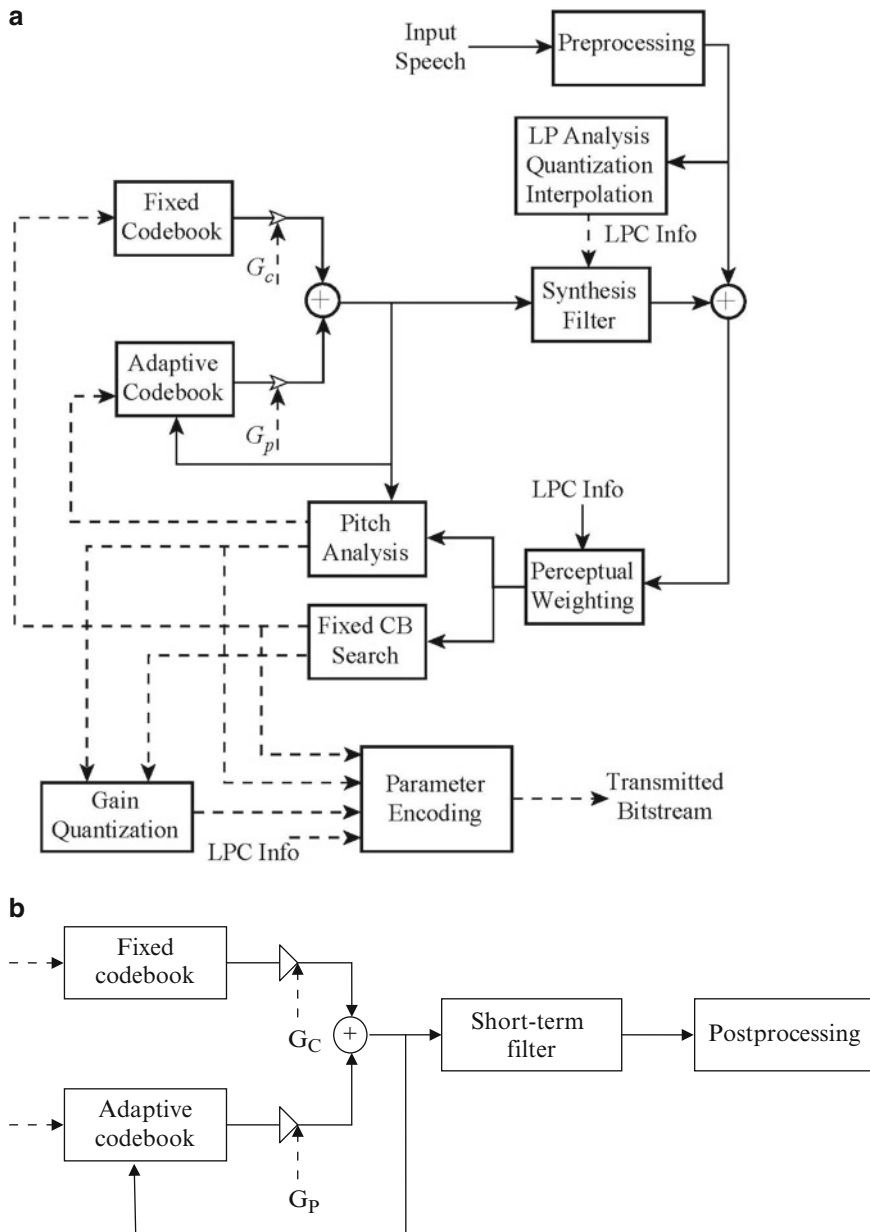


Fig. 2.2 (a) Encoder for code-excited linear predictive (CELP) coding with an adaptive codebook. (b) CELP decoder with an adaptive codebook and postfiltering

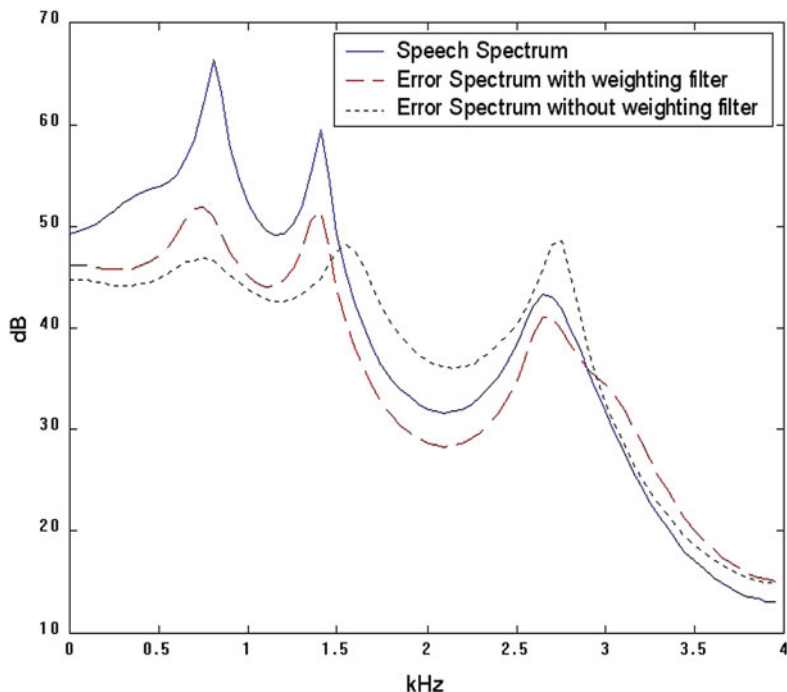


Fig. 2.3 Perceptual weighting of the coding error as a function of frequency

the envelope of the input speech across the frequency band of interest. Figure 2.3 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 3,100 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. The reader should note that if one analyzes each frame of a CELP-coded speech segment, the goal of pushing the error spectrum below that of the input speech is often not obtained across the entire band. This is because the perceptually shaping methods used are approximate and have not yet been refined to guarantee the desired result [2].

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. 2.2b. 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 for the fixed codebook [10].

2.2.1.4 Postfiltering [11]

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 some distortion remaining 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. 2.2b.

The general frequency response of the postfilter has the form similar to the perceptual weighting filter with a pitch or long term postfilter added. There is also a spectral tilt correction since the formant-based postfilter results in an increased low pass filter effect, and a gain correction term [2, 10, 11]. The postfilter is usually optimized for a single stage encoding (however, not always), so if multiple tandem connections of speech codecs occur, the postfilter can cause a degradation in speech quality.

2.2.1.5 Voice Activity Detection and Silence Coding

For many decades, researchers have been interested in assigning network capacity only when a speaker is “active,” by 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 [10].

Today, 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.

2.2.2 *Speech Coding Standards*

Different standardization bodies have adopted a host of codecs for rapidly evolving applications. For narrowband speech coding, the ITU-T and the several digital cellular standardization efforts are the dominant activities. There is a vast number of standards that have been set. We begin the discussion with ITU-T standardized codecs since some of those codecs have served as the basis for cellular codecs, and since some of these codecs have been adopted for VoIP applications.

Table 2.1 Comparison of ITU-T narrowband speech codecs

Standards body	ITU			
Recommendation	G.711	G.726	G.728	G.729
Coder type	Companded PCM	ADPCM	LD-CELP	CS-ACELP
Bit rate (kbps)	64	16–40	16	8
Complexity (MIPS)	$\ll 1$	~ 1	~ 30	≤ 20
Frame size (ms)	0.125	0.125	0.625	10
Lookahead (ms)	0	0	0	5
Codec delay (ms)	0.25	0.25	1.25	25

2.2.2.1 ITU-T Standards

Table 2.1 lists some of the narrowband voice codecs that have been standardized by the ITU-T over the years, including details concerning the codec technology, transmitted bit rate, performance, complexity, and algorithmic delay. Those shown include G.711, G.726, G.728, and G.729 for narrowband (telephone bandwidth) speech (200–3,400 Hz), where the first two codecs are waveform-following codecs, and the latter three are variations on code excited linear prediction.

G.711 at 64 kilobits/s (kbps) is the voice codec most often used in VoIP and many applications wherein somewhat higher bit rates are workable and very low complexity is desirable. This codec is based on a nonlinear scalar quantization method called logarithmic pulse code modulation (log-PCM), as discussed earlier. The G.726 waveform-following codec is based on ADPCM and operates at bit rates of 40, 32, 24, and 16 kbps. This codec achieves low delay and is still considered a low complexity codec. G.728 is a code-excited technique but when it was standardized, it was still desired to have a low encoding delay of 5 ms or less. G.728 is much more complex than either G.726 or G.711.

As the desired bit rate moved toward 8 kbps, the low delay requirement was relaxed. This allowed code-excited linear prediction methods to move to the forefront. The G.729 codec is an analysis-by-synthesis codec based on algebraic code excited linear prediction (ACELP), and it uses an adaptive codebook to incorporate the long term pitch periodicity [2, 10]. In addition to a lower complexity version of G.729, called G.729A, there is a higher rate codec based on G.729, designated G.729E, and a wideband version designated G.729.1 [12]. The G.729 codec structure has been very influential on subsequent voice coding standards for VoIP and digital cellular networks and this structure can be seen in most standardized voice codecs today.

Even though we are quite comfortable communicating using telephone bandwidth speech (200–3,400 Hz), there is considerable interest in compression methods for wideband speech covering the range of 50 Hz–7 kHz. The primary reasons for the interest in this band are that wideband speech (and wider bands) improves intelligibility, naturalness, and speaker identifiability. Table 2.2 lists codecs for

Table 2.2 ITU-T wideband and fullband speech coding standards

Recommendation	ITU-T G.722	ITU-T G.722.1	ITU-T G.722.2	ITU-T G.718	ITU-T G.719
Coder type	Sub-band ADPCM	MLT	3GPP AMR-WB ACELP	ACELP, MDCT	Adaptive resolution MDCT, FLVQ
Audio bandwidth (Hz)	50–7,000	50–7,000	50–7,000	50–7,000	20–20,000
Bitrate(s) (kbts/s)	48, 56, 64	24, 32	6.6, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05, 23.85	8, 12, 16, 24, 32 & 12.65 (G.722.2, AMR-WB, VMR-WB Interop Mode)	32 ... 128 steps of 4 kbps up to 96 kbps, steps of 8 kbps up to 128 kbps
Frame length (ms)	0.125	20	20	20	20
Algorithmic delay (ms)	1.625	40	25	32.875–43.875	40
Comp. complexity	10 MIPS	<5.5 WMOPS	27.2–39.0 WMOPS	57 WMOPS	15.39–21 WMOPS

wideband speech, including G.722, G.722.1 [13], and G.722.2 [14]. Also, shown in the table are ITU-T codecs G.718 for wideband speech (50 Hz–7 kHz) [15], and G.719 for fullband audio [16, 17].

The first 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 [2]. G.722 at 64 kbps is often employed as a benchmark for the performance of other wideband codecs.

Two additional 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. The coder has an algorithmic delay of 40 ms, 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. This codec structure for G.722.1 has much in common with the fullband audio codecs used for many music player products such as MP3.

G.722.2 is an ITU-T designation for the adaptive multirate wideband (AMR-WB) speech coder standardized by the cellular body 3GPP [14]. 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.

G.718 is a wideband speech codec that has an embedded codec structure and that operates at 8, 12, 16, 24, and 32 kbps, plus a special alternate coding mode that is bit stream compatible with AMR-WB. G.719 is a fullband audio codec that has relatively low complexity and low delay for a fullband audio codec, and the complexity is approximately evenly split between the encoder and decoder. This codec is targeted toward real-time communications such as in videoconferencing systems and the high definition telepresence applications.

2.2.2.2 Digital Cellular Standards

Digital cellular applications impose a stringent set of requirements on voice codecs in addition to rate and quality, such as complexity, robustness to background impairments, and the ability to perform well over wireless channels. Over the years, standards have been set by different bodies for different segmentations of the market, particularly according to geographic regions and wireless access technologies. More specifically, digital cellular standards were produced in the late

Table 2.3 Selected GSM voice codecs

Codec	Speech coding bit-rate (in kbit/s)	System/traffic channel	Speech coding algorithm	Complexity WMOPS
FR codec	13.0	GSM FR	Regular Pulse Excitation-Long Term Prediction (RPE-LTP)	3.0
HR codec	5.6	GSM HR	Vector-Sum Excited Linear Prediction (VSELP)	18.5
EFR codec	12.2	GSM FR	Algebraic Code Excited Linear Prediction (ACELP)	15.2
AMR codec	12.2, 10.2, 7.95, 7.4, 6.7, 5.9, 5.15, 4.75	GSM FR (all eight modes), GSM HR (six lowest modes), 3G WCDMA (all modes)	Algebraic Code Excited Linear Prediction (ACELP)	16.8
AMR-WB codec	23.85, 23.05, 19.85, 18.25, 15.85, 14.25, 12.65, 8.85, 6.60	GSM FR (seven lowest modes), EDGE (all modes), 3G WCDMA (all modes)	Algebraic Code Excited Linear Prediction (ACELP)	35.4

1980s and early 1990s in Europe, Japan, and North America. The competing North American standards then led to standards efforts more pointed toward each of the competing technologies.

The GSM standards developed in Europe were the basis of perhaps the first widely implemented digital cellular systems. Table 2.3 lists voice codecs standardized for GSM systems, wherein FR stands for “Full rate” and HR stands for “Half rate.” The terms FR and HR refer to the total transmitted bit rate for combined voice coding and error correction (or channel) coding, and FR is always 22.8 kbps and HR is always 11.4 kbps. By subtracting the rate of the voice codec from either 22.8 or 11.4, one obtains the bit rate allocated to error control coding.

The first GSM FR voice codec standardized in 1989 was not an analysis-by-synthesis codec but used a simpler regular pulse excited linear predictive structure with a long term predictor. As a result, the codec had to be operated at 13 kbps to achieve the needed voice quality, but it had very low complexity. An important and somewhat dominant voice codec in recent years is the Adaptive Multirate Codec, both narrowband and wideband versions. Note that AMR-NB has multiple rates and can be operated as a FR or HR codec, depending upon the rates. For GSM, the AMR-NB codec rates are not source-controlled as some prior codecs were, but the rates are switchable and usually adjusted by the network. The AMR codec maintains compatibility with other systems by incorporating the GSM EFR (Enhanced Full Rate) codec at 12.2 kbps and IS-641 at 7.4 kbps as two of its selectable rates. The AMR wideband codec, AMR-WB, also based upon ACELP is also a very important codec today; however, note how the complexity has grown.

2.2.2.3 VoIP Standards

VoIP for wireless access points involves many of the same issues as for wireline VoIP, such as voice quality, latency, jitter, packet loss performance, and packetization. One new challenge that arises is that since the physical link in Wi-Fi is wireless, bit errors commonly occur and this, in turn, affects link protocol design and packet loss concealment. A second challenge is that congestion can play a role, thus impacting real time voice communications. The way these two issues relate to voice codecs are that packet loss concealment methods are more critical and that codec delay should be more carefully managed for such wireless access points.

Turning our attention to the voice codecs normally implemented in VoIP solutions, we find that at this point in time, many VoIP codecs are borrowed from other standards bodies. Specifically, G.711, G.729, and G.722 are commonly offered in VoIP products. Additionally, AMR-NB and perhaps AMR-WB are optional voice codecs. All of these codecs have well-developed packet loss concealment methods, which makes them quite compatible with wireless applications. One thing to notice is that the AMR codecs are the only ones that are common with any digital cellular standards, and this can lead to tandem coding penalties when digital cellular and wireless VoIP are used for portions of the same connection for a voice call. The need to support multiple codecs can also be an issue as cell phones morph into smartphones that support both digital cellular and wireless access point connectivity.

There have also been voice codecs developed outside of standards bodies and offered as open source. Two such codecs are Speex [18] and iLBC (internet Low Bitrate Codec) [19]. Speex has become obsolete with the introduction of the Opus codec [20, 21], described in a later section. These codecs have been compared to other standardized codecs in several studies [22–24].

2.3 Audio Coding [25, 26]

The basic very successful paradigm for audio coding, meaning coding full band audio, in the past two decades has been the filter bank/transform based approach with noise masking using an iterative bit allocation. This technique does not lend itself to real time communications directly because of the iterative bit allocation method and because of complexity, and to a lesser degree, delay in the filter bank/transform/noise masking computations. As a result, the primary impact of high quality audio coding has been to audio players (decoders) such as MP3 and audio streaming applications.

A high level block diagram of an audio codec is shown in Fig. 2.4. In this diagram, two paths are shown for the sampled input audio signal, one path is through the filter bank/transform that performs the analysis/decomposition into spectral components to be coded, and the other path into the psychoacoustic analysis that computes the noise masking thresholds. The noise masking thresholds are then

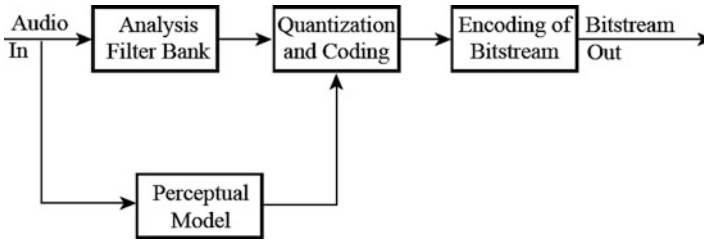


Fig. 2.4 Generic audio coding approach

used in the bit allocation that forms the basis for the quantization and coding in the analysis/decomposition path. All side information and parameters required for decoding are finally losslessly coded for storage or transmission.

The primary differences among the different audio coding schemes that have been standardized and/or found wide application are in the implementations of the time/frequency analysis/decomposition in terms of the types of filter banks/transforms used and their resolution in the frequency domain. Note that the frequency resolution of the psychoacoustic analysis is typically finer than the analysis/decomposition path since the perceptual noise masking is so critical for good quality. There are substantive differences in the other blocks as well, with many refinements over the years [2, 25, 26].

The strengths of the basic audio coding approach are that it is not model based, as in speech coding using linear prediction, and that the perceptual weighting is applied on a per-component basis, whereas in speech coding, the perceptual weighting relies on a spectral envelope shaping. A weakness in the current approaches to audio coding is that the noise masking theory that is the foundation of the many techniques is three decades old; further, the masking threshold for the entire frame is computed by adding the masking thresholds for each component. The psychoacoustic/audio theory behind this technique of adding masking thresholds has not been firmly established [25].

Other key ideas in the evolution of the full band audio coding methods have been pre- and post-masking and window switching to capture transients and steady state sounds. Details of the audio coding methods are left to the very comprehensive references cited [25, 26]. However, we will revisit full band audio coding in discussing the newer standards and when considering new research directions.

2.4 Newer Standards

The standardization processes continue to be vigorous in the classically active standards bodies such as ITU-T, the ISO, and the digital cellular community. Furthermore, there is considerable activity in developing alternative coding methods outside of the standards bodies that may be free of intellectual property claims.

A unifying thread in all of these efforts to develop new codecs is to have codecs that cover narrowband, wideband, superwideband, and full band in one specification. Differences in the codec development efforts revolve around whether the delay needed in coding is low enough to allow real time communications or whether the latency precludes most such real time applications. See [1, 27] for a more complete discussion of delay in voice codec design and its impact on voice coding applications.

Efforts in the last decade to design one codec that would cover all bands from narrowband speech to fullband audio have led to the approach of combining code excited linear prediction methods to cover narrowband and wideband speech with the filter bank methods to cover full band audio, using switching or mixing in between. Such a codec structure may be called an integrated codec, and examples of such integrated codecs include G.718 which combines ACELP and MDCT technologies, originally only covering 50 Hz–7 kHz but since extended to superwideband [28], and the recently standardized MPEG USAC (Unified Speech and Audio Coding) architecture shown in Fig. 2.5, which covers the entire range from 20 Hz to 20 kHz, with the goal of coding voice and fullband audio well [29]. The USAC codec utilizes signal classification and down mixes the stereo to mono for coding in the low band. There is a low pass/high pass decomposition, and enhanced spectral band replication (eSBR) is used to code the high band. There is both a baseline mode and an extension mode.

The applications targeted for the MPEG USAC codec are multimedia download to mobile devices, audio books, mobile TV, and digital radio. While there are strong targets for improving audio performance and it is designed to code speech, audio, and mixed content, there are no specifications on complexity or delay.

At first inspection, these integrated structures can be viewed as merely bolting together successful codecs for different bands, but the USAC effort notes that it is the handling of the transitions between different coding paradigms that requires innovation beyond a simple combination of known schemes. It is not difficult to understand how challenging it is to combine such different codec designs, and so the USAC and related codecs must be considered a substantial advance in the state-of-the-art.

A key limitation of the USAC effort is the lack of support for conversational services, which require low encoding delay and limitations on complexity. Several new and developing standards try to encompass these very important conversational applications.

For conversational speech, the ITU-T standardization efforts have already resulted in G.711.1 [30] and G.729.1 [12], both of which are extensions of existing standards to wider bands and different, higher rates. A superwideband version of G.722.1, designated G.722.1 Annex C [31], has also been standardized. The G.722.1C codec has a coding delay of 40 ms and a relatively modest complexity with transmitted bit rates of 24, 32 and 48 kbps. It codes voice, audio, and natural sounds well, and it is targeted for applications to videoconferencing, VoIP, and battery powered devices.

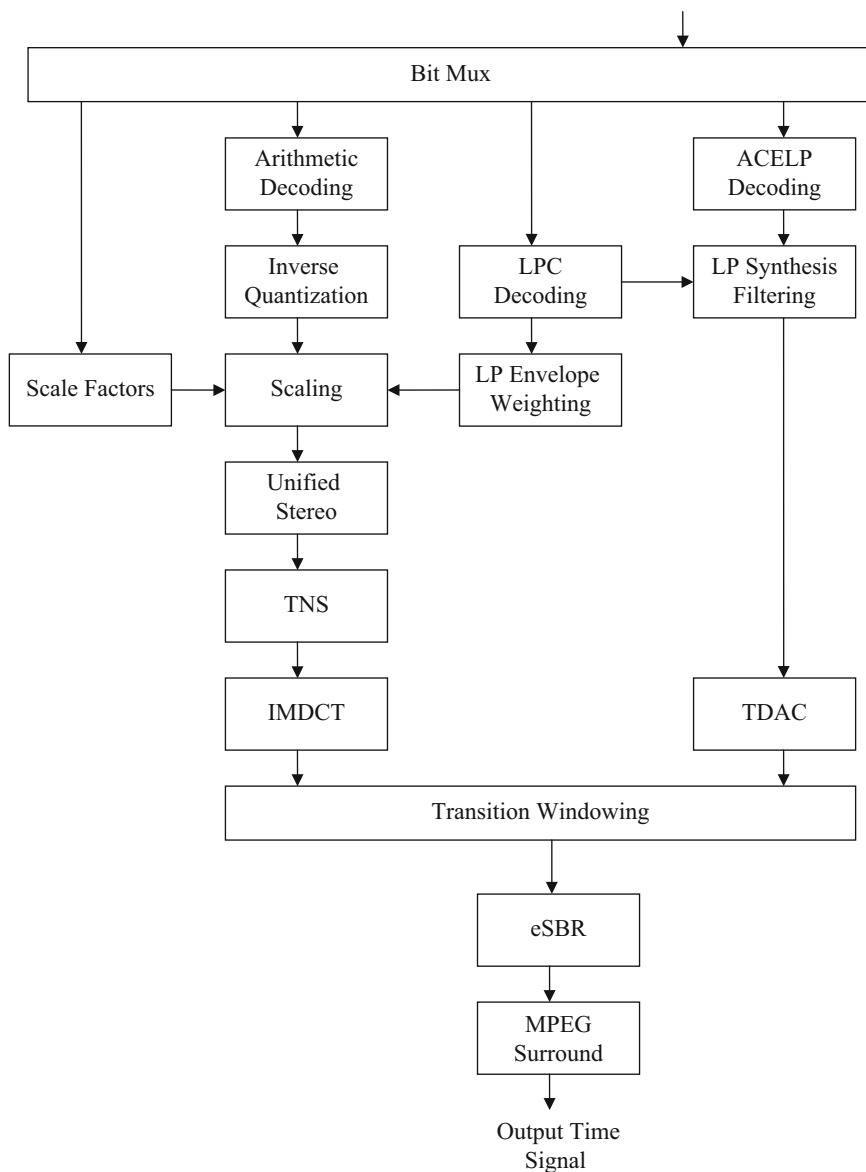


Fig. 2.5 USAC decoder structure

It is evident that digital cellular systems Worldwide will be based on 3GPP Long Term Evolution (LTE), which utilizes Orthogonal Frequency Division Multiple Access (OFDMA) in the downlink and Single Carrier Frequency Division Multiple Access (SC-FDMA) in the uplink. The initial releases of LTE rely upon the

Table 2.4 Objectives and features of the EVS codec

Enhanced quality and coding efficiency for narrowband (NB: 200–3,400 Hz) and wideband (WB: 50–7,000 Hz) speech
Enhanced quality by the introduction of super wideband (SWB: 50–14,000 Hz) speech
Enhanced quality for mixed content and music in conversational applications (e.g. in-call music)
Robustness to packet loss and delay jitter leading to optimized behavior in IP environments
Backward interoperability to AMR-WB by EVS modes supporting the AMR-WB codec format.
Source-controlled variable bit rate modes in addition to constant bit rate modes

AMR-NB/WB voice codecs for voice coding, but this is a stop gap effort while bodies work to develop a voice codec specifically for LTE for Enhanced Voice Services (EVS) [32]. Some objectives and features of the EVS voice codec for LTE are summarized in Table 2.4. Here we see that there is a desire to maintain interoperability with the AMR codecs while adding a superwideband capability and giving more attention to in-call music. As the EVS codec nears final characterization, there are some specific advances that will be widely deployed and used. First, new 5.9 kbps source controlled variable bit rate (VBR) modes for both narrowband and wideband speech that achieve the same quality as the AMR-NB/WB codec but at a lower average rate have been added to improve capacity. Further, there is better constant bit rate (CBR) coding of both WB and SWB music, and improved CBR coding of WB and SWB speech. Also included are optional full band and stereo modes for voice and music. Key design constraints are that codec delays up to 32 ms are allowed and a complexity up to twice that of AMR-WB, namely, 88 WMOPS.

High quality audio codecs for non-conversational services such as streaming, broadcasting, and multicasting also have been standardized earlier by 3GPP. These codecs are AMR-WB+ and aacPlus, but their high algorithmic delay restricts their importance for two-way conversational voice.

Another codec standardization effort had the goal of coding narrowband voice all the way up to fullband audio with the constraint of low delay. The Opus Audio Codec, standardized by the IETF, is designed for interactive voice and audio and has three modes [20, 21]: (a) A linear prediction based mode for low bit rate coding up to 8 kHz bandwidth, (b) A hybrid linear prediction and MDCT mode for fullband speech/audio at medium bit rates, and (c) an MDCT-only mode for very low latency coding of speech and audio. It has a wide range of bit rates from 6 kbps up to 510 kbps to support full band audio. Further, Opus has available frame sizes from 2.5 ms up to 60 ms and algorithmic delay in the range of 5–62.5 ms. Details of this codec can be found at [20, 21]. Speech quality tests indicate that the Opus codec produces excellent voice quality at medium rates of 20–40 kbps [23].

2.5 Emerging Topics

As the field of speech coding continues to evolve, new issues emerge, old challenges persist, and often old constraints are relaxed. It is clear from the prior discussions that extending the bandwidth covered by codecs is a high priority for essentially every standards activity. As stated earlier, wider bandwidths improve intelligibility, naturalness, and speaker identifiability. The advantages of incorporating wider bandwidths need to be elaborated further given the extraordinary efforts of the standards bodies to achieve ever-increasing bandwidth capabilities in codecs [33].

Consonants are key in the intelligibility of many words and phrases. The frequency content of consonants often occurs in the 4–14 kHz frequency range, which is partially encompassed by wideband speech, but much more so by the newer superwideband classification. There are other factors that arise in standard videoconferencing, VoIP, and even person-to-person calls that are addressed by wider bandwidths. In conference rooms and other such venues, there are natural reflections off walls and ceilings that can degrade communications, especially if a speaker moves away from a microphone or speakers are different distances from microphones. In person, a listener is able to use both ears, which greatly helps in alleviating misunderstandings, but for audio and video conference calls, often there is only one microphone and therefore only one channel being delivered to the other end. Experiments show that wider bandwidths aid greatly in reducing confusion and easing listener fatigue.

Another important point in this multinational business environment and with non-native English speakers routinely playing critical roles in organizations is accented speech. This point also holds for speakers within a country, such as the US, where there are quite different speech patterns. Speakers with accents will often have different pronunciations and different grammatical patterns. As a result, native listeners may not be able to correctly process sentences when there are different pronunciations because of the different grammatical structure. Increasing bandwidth provides considerable improvement in these situations.

Extending the lower end of the band is also of substantial value, since frequencies below 200 Hz add to listener comfort, warmth, and naturalness. It is thus very clear why the exceptional efforts to extend the bandwidths covered by codecs are being pursued.

Stereo audio is a new effort in communications applications. The capture of stereo, or more generally, multichannel audio, is simpler than it sounds, even for handheld devices. For example, there may be two microphones, one pointed toward the active speaker and the other outwardly to record the environment. There are many other microphone configurations that may be desirable as well [22]. As stereo audio capture and delivery becomes of interest, it is necessary to make decisions as to how to allocate bit rate; that is, if a choice must be made, is it preferable to send wider bandwidth speech/audio or stereo channels? Coupled to this question is how to evaluate the quality of the expanded bandwidths and additional multichannel audio when delivered to the user.

A recent addition to the perceptual quality evaluation area is the use of a nine point range for the MOS values [22, 32]. Unlike the five point scale, only the extreme values are given designations of Excellent (9) and Very Bad (1). It is shown that this scale allows the tests to be accomplished relatively quickly and that the various conditions are distinguished by the tests. Comparisons are given in [22] and [23], wherein the latter contains a performance evaluation of the Opus codec.

2.6 Conclusions and Future Research Directions

There are several clear trends in recent standardization efforts. First, single standard codecs that encompass the entire range of narrowband to fullband are highly desirable and the norm for the future. Second, while latency constraints have been relaxed for many applications, there is still a demand for lower latency codecs to be used in communications services. Third, increasing complexity is acceptable as long as the speech/audio quality is substantially improved. Fourth, there is a strong impetus to capture and code stereo channels for many applications, including handheld devices.

Another fact is also clear—the current standards still rely very heavily on the well-worn coding paradigms of code-excited linear prediction and transform/filter bank methods with noise masking. It is this fact that points to a great need for new research directions to try and identify new codec structures to continue the advance in speech/audio codec compression developments. Although standardization efforts have resulted in many new codec designs and the understanding of the basic structures, it is unlikely that given the time constraints and continuously competitive nature of codec standardization processes, these new research directions will be undertaken through the development of new standards.

Some suggested research directions are to incorporate increased adaptivity into the codec designs. For example, adapting the parameters of the perceptual weighting filters in CELP is one possible research direction. Another is to incorporate adaptive filter bank/transform structures such as adaptive band combining and adaptive band splitting. A third more difficult direction is to identify entirely new methods to incorporate perceptual constraints into codec structures.

It is hoped that the current chapter will motivate some of these new research efforts.

References

1. J.D. Gibson, Speech coding methods, standards, and applications. *IEEE Circuits Syst. Magazine* **5**, 30–49 (2005)
2. J.D. Gibson, T. Berger, T. Lookabaugh, D. Lindbergh, R.L. Baker, *Digital Compression for Multimedia: Principles and Standards* (Morgan-Kaufmann, San Francisco, 1998)

3. R. Cox, S.F. de Campos Neto, C. Lamblin, M.H. Sherif, ITU-T coders for wideband, superwideband, and fullband speech communication. *IEEE Commun. Magazine* **47**, 106–109 (2009)
4. 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 (2001)
5. ITU-T Recommendation P.862.2, Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs (2007)
6. ITU-T Recommendation P.863, Perceptual objective listening quality assessment (2011)
7. W.-Y. Chan, T.H. Falk, Machine assessment of speech communication quality, in *The Mobile Communications Handbook*, ed. by J.D. Gibson, 3rd edn. (CRC Press, BocaRaton, FL, 2012). Chapter 30
8. Advanced audio distribution profile (A2DP) specification version 1.2, Bluetooth SIG, Audio video WG, <http://www.bluetooth.org/>. April 2007
9. H.S. Malvar, *Signal Processing with Lapped Transforms* (Artech House, Norwood, 1992)
10. A.M. Kondoz, *Digital Speech: Coding for Low Bit Rate Communication Systems* (Wiley, West Sussex, 2004)
11. J.H. Chen, A. Gersho, Adaptive postfiltering for quality enhancement of coded speech. *IEEE Trans. Audio Process.* **3**, 59–70 (1995)
12. S. Ragot et al., ITU-T G.729.1: An 8-32 kbit/s scalable coder interoperable with G.729 for wideband telephony and Voice over IP, in *Proceedings of ICASSP*, Honolulu, April 2007
13. ITU-T Recommendation G.722.1, Coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss (1999)
14. ITU-T Recommendation G.722.2, Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB) (2002)
15. ITU-T Rec. G.718, Frame error robust narrowband and wideband embedded variable bit-rate coding of speech and audio from 8-32 kbit/s (2008)
16. ITU-T Rec. 719, Low-complexity, full-band audio coding for high-quality, conversational applications, June 2008
17. S. Karapetkov, G.719: the first ITU-T standard for full-band audio. Polycom white paper, April 2009
18. <http://www.speex.org/>
19. S.V. Andersen, W.B. Kleijn, R. Hagen, J. Linden, M.N. Murthi, J. Skoglund, iLBC – a linear predictive coder with robustness to packet losses, in *Proceedings of the IEEE Speech Coding Workshop*, October 2002, pp 23–25
20. IETF Opus Interactive Audio Codec, <http://opus-codec.org/> (2011)
21. RFC6716, Definition of the Opus Audio Codec, September 2012
22. A. Ramo, Voice quality evaluation of various codecs, in *ICASSP 2010*, Dallas, 14–19 March 2010
23. A. Ramo, H. Toukomaa, Voice quality characterization of the IETF Opus Codec, in *Proceedings of Interspeech 2011*, Florence (2011)
24. A. Ramo, H. Toukomaa, On comparing speech quality of various narrow- and wideband speech codecs, in *Proceeding of ISSPA*, Sydney (2005)
25. M. Bosi, R.E. Goldberg, *Introduction to Audio Coding and Standards* (Kluwer, Boston, 2003)
26. T. Painter, A. Spanias, Perceptual coding of digital audio. *Proc. IEEE* **88**, 451–512 (2000)
27. ITU-T Recommendation G.114, One-Way Transmission Time (2000)
28. ITU-T Rec. G.718 Amendment 2: New Annex B on superwideband scalable extension for ITU-T G.718 and corrections to main body fixed-point C-code and description text, March 2010
29. M. Neuendorf, P. Gournay, M. Multus, J. Lecomte, B. Bessette, R. Geiger, S. Bayer, G. Fuchs, J. Hilpert, N. Rettelbach, F. Nagel, J. Robilliard, R. Salami, G. Schuller, R. Lefebvre, B. Grill, A novel scheme for low bitrate unified speech and audio coding-MPEG RM0, in *Audio Engineering Society*, Convention Paper 7713, May 2009

30. Y. Hiwasaki et al., G.711.1: a wideband extension to ITU-T G.711. *EUSIPCO 2008*, Lausanne, 25–29 August 2008
31. M. Xie, D. Lindbergh, P. Chu, ITU-T G.722.1 Annex C: a new low-complexity 14 kHz audio coding standard, in *Proceedings of ICASSP*, Toulouse, May 2006
32. K. Jarvinen, I. Bouazizi, L. Laaksonen, P. Ojala, A. Ramo, Media coding for the next generation mobile system LTE. *Comput. Commun.* **33**, 1916–1927 (2010)
33. J. Rodman, The effect of bandwidth on speech intelligibility. Polycorn white paper, September 2006

Speech and Audio Processing for Coding, Enhancement
and Recognition

Ogunfunmi, T.; Togneri, R.; Narasimha, M.S. (Eds.)

2015, X, 345 p. 79 illus., 32 illus. in color., Hardcover

ISBN: 978-1-4939-1455-5