2

Coding Strategies and Standards

2.1 Introduction

The invention of Pulse Code Modulation (PCM) in 1938 by Alec H. Reeves was the beginning of digital speech communications. Unlike the analogue systems, PCM systems allow perfect signal reconstruction at the repeaters of the communication systems, which compensate for the attenuation provided that the channel noise level is insufficient to corrupt the transmitted bit stream. In the early 1960s, as digital system components became widely available, PCM was implemented in private and public switched telephone networks. Today, nearly all of the public switched telephone networks (PSTN) are based upon PCM, much of it using fibre optic technology which is particularly suited to the transmission of digital data. The additional advantages of PCM over analogue transmission include the availability of sophisticated digital hardware for various other processing, error correction, encryption, multiplexing, switching, and compression.

The main disadvantage of PCM is that the transmission bandwidth is greater than that required by the original analogue signal. This is not desirable when using expensive and bandwidth-restricted channels such as satellite and cellular mobile radio systems. This has prompted extensive research into the area of speech coding during the last two decades and as a result of this intense activity many strategies and approaches have been developed for speech coding. As these strategies and techniques matured, standardization followed with specific application targets. This chapter presents a brief review of speech coding techniques. Also, the requirements of the current generation of speech coding standards are discussed. The motivation behind the review is to highlight the advantages and disadvantages of various techniques. The success of the different coding techniques is revealed in the description of the

Digital Speech. A. Kondoz
 ISBN 0-470-87007-9 (HB)

many coding standards currently in active operation, ranging from 64 kb/s down to 2.4 kb/s.

2.2 Speech Coding Techniques

Major speech coders have been separated into two classes: waveform approximating coders and parametric coders. Kleijn [1] defines them as follows:

- **Waveform approximating coders:** Speech coders producing a reconstructed signal which converges towards the original signal with decreasing quantization error.
- **Parametric coders:** Speech coders producing a reconstructed signal which does not converge to the original signal with decreasing quantization error.

Typical performance curves for waveform approximating and parametric speech coders are shown in Figure 2.1. It is worth noting that, in the past, speech coders were grouped into three classes: waveform coders, vocoders and hybrid coders. Waveform coders included speech coders, such as PCM and ADPCM, and vocoders included very low bit-rate synthetic speech coders. Finally hybrid coders were those speech coders which used both of these methods, such as CELP, MBE etc. However currently all speech coders use some form of speech modelling whether their output converges to the

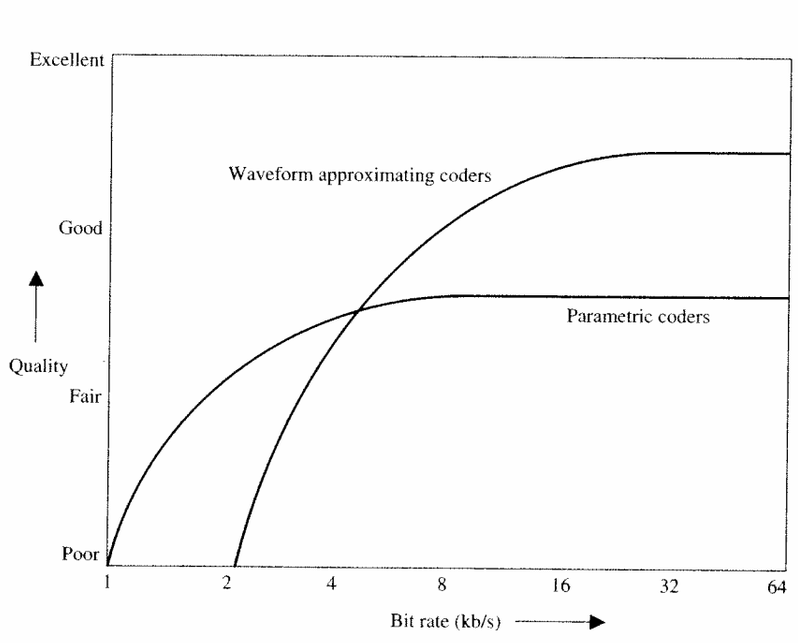

Figure 2.1 Quality vs bit rate for different speech coding techniques

original (with increasing bit rate) or not. It is therefore more appropriate to group speech coders into the above two groups as the old waveform coding terminology is no longer applicable. If required we can associate the name hybrid coding with coding types that may use more than one speech coding principle, which is switched in and out according to the input speech signal characteristics. For example, a waveform approximating coder, such as CELP, may combine in an advantageous way with a harmonic coder, which uses a parametric coding method, to form such a hybrid coder.

2.2.1 Parametric Coders

Parametric coders model the speech signal using a set of model parameters. The extracted parameters at the encoder are quantized and transmitted to the decoder. The decoder synthesizes speech according to the specified model. The speech production model does not account for the quantization noise or try to preserve the waveform similarity between the synthesized and the original speech signals. The model parameter estimation may be an open loop process with no feedback from the quantization or the speech synthesis. These coders only preserve the features included in the speech production model, e.g. spectral envelope, pitch and energy contour, etc. The speech quality of parametric coders do not converge towards the transparent quality of the original speech with better quantization of model parameters, see Figure 2.1. This is due to limitations of the speech production model used. Furthermore, they do not preserve the waveform similarity and the measurement of signal to noise ratio (SNR) is meaningless, as often the SNR becomes negative when expressed in dB (as the input and output waveforms may not have phase alignment). The SNR has no correlation with the synthesized speech quality and the quality should be assessed subjectively (or perceptually).

Linear Prediction Based Vocoders

Linear Prediction (LP) based vocoders are designed to emulate the human speech production mechanism [2]. The vocal tract is modelled by a linear prediction filter. The glottal pulses and turbulent air flow at the glottis are modelled by periodic pulses and Gaussian noise respectively, which form the excitation signal of the linear prediction filter. The LP filter coefficients, signal power, binary voicing decision (i.e. periodic pulses or noise excitation), and pitch period of the voiced segments are estimated for transmission to the decoder. The main weakness of LP based vocoders is the binary voicing decision of the excitation, which fails to model mixed signal types with both periodic and noisy components. By employing frequency domain voicing decision techniques, the performance of LP based vocoders can be improved [3].

Harmonic Coders

Harmonic or sinusoidal coding represents the speech signal as a sum of sinusoidal components. The model parameters, i.e. the amplitudes, frequencies and phases of sinusoids, are estimated at regular intervals from the speech spectrum. The frequency tracks are extracted from the peaks of the speech spectra, and the amplitudes and frequencies are interpolated in the synthesis process for smooth evolution [4]. The general sinusoidal model does not restrict the frequency tracks to be harmonics of the fundamental frequency. Increasing the parameter extraction rate converges the synthesized speech waveform towards the original, if the parameters are unquantized. However at low bit rates the phases are not transmitted and estimated at the decoder, and the frequency tracks are confined to be harmonics. Therefore point to point waveform similarity is not preserved.

2.2.2 Waveform-approximating Coders

Waveform coders minimize the error between the synthesized and the original speech waveforms. The early waveform coders such as companded Pulse Code Modulation (PCM) [5] and Adaptive Differential Pulse Code Modulation (ADPCM) [6] transmit a quantized value for each speech sample. However ADPCM employs an adaptive pole zero predictor and quantizes the error signal, with an adaptive quantizer step size. ADPCM predictor coefficients and the quantizer step size are backward adaptive and updated at the sampling rate.

The recent waveform-approximating coders based on time domain analysis by synthesis such as Code Excited Linear Prediction (CELP) [7], explicitly make use of the vocal tract model and the long term prediction to model the correlations present in the speech signal. CELP coders buffer the speech signal and perform block based analysis and transmit the prediction filter coefficients along with an index for the excitation vector. They also employ perceptual weighting so that the quantization noise spectrum is masked by the signal level.

2.2.3 Hybrid Coding of Speech

Almost all of the existing speech coders apply the same coding principle, regardless of the widely varying character of the speech signal, i.e. voiced, unvoiced, mixed, transitions etc. Examples include Adaptive Differential Pulse Code Modulation (ADPCM) [6], Code Excited Linear Prediction (CELP) [7, 8], and Improved Multi Band Excitation (IMBE) [9, 10]. When the bit rate is reduced, the perceived quality of these coders tends to degrade more for some speech segments while remaining adequate for others. This shows that the assumed coding principle is not adequate for all speech types. In order to circumvent this problem, hybrid coders that combine different

coding principles to encode different types of speech segments have been introduced [11, 12, 13].

A hybrid coder can switch between a set of predefined coding modes. Hence they are also referred to as multimode coders. A hybrid coder is an adaptive coder, which can change the coding technique or mode according to the source, selecting the best mode for the local character of the speech signal. Network or channel dependent mode decision [14] allows a coder to adapt to the network load or the channel error performance, by varying the modes and the bit rate, and changing the relative bit allocation of the source and channel coding [15].

In source dependent mode decision, the speech classification can be based on fixed or variable length frames. The number of bits allocated for frames of different modes can be the same or different. The overall bit rate of a hybrid coder can be fixed or variable. In fact variable rate coding can be seen as an extension of hybrid coding.

2.3 Algorithm Objectives and Requirements

The design of a particular algorithm is often dictated by the target application. Therefore, during the design of an algorithm the relative weighting of the influencing factors requires careful consideration in order to obtain a balanced compromise between the often conflicting objectives. Some of the factors which influence the choice of algorithm for the foreseeable network applications are listed below.

2.3.1 Quality and Capacity

Speech quality and bit rate are two factors that directly conflict with each other. Lowering the bit rate of the speech coder, i.e. using higher signal compression, causes degradation of quality to a certain extent (simple parametric vocoders). For systems that connect to the Public Switched Telephone Network (PSTN) and associated systems, the quality requirements are strict and must conform to constraints and guidelines imposed by the relevant regulatory bodies, e.g. ITU (previously CCITT). Such systems demand high quality (toll quality) coding. However, closed systems such as private commercial networks and military systems may compromise the quality to lower the capacity requirements. Although absolute quality is often specified, it is often compromised if other factors are allocated a higher overall rating. For instance, in a mobile radio system it is the overall average quality that is often the deciding factor. This average quality takes into account both good and bad transmission conditions.

2.3.2 Coding Delay

The coding delay of a speech transmission system is a factor closely related to the quality requirements. Coding delay may be algorithmic (the buffering of speech for analysis), computational (the time taken to process the stored speech samples) or due to transmission. Only the first two concern the speech coding subsystem, although very often the coding scheme is tailored such that transmission can be initiated even before the algorithm has completed processing all of the information in the analysis frame, e.g. in the pan-European digital mobile radio system (better known as GSM) [16] the encoder starts transmission of the spectral parameters as soon as they are available. Again, for PSTN applications, low delay is essential if the major problem of echo is to be minimized. For mobile system applications and satellite communication systems, echo cancellation is employed as substantial propagation delays already exist. However, in the case of the PSTN where there is very little delay, extra echo cancellers will be required if coders with long delays are introduced. The other problem of encoder/decoder delay is the purely subjective annoyance factor. Most low-rate algorithms introduce a substantial coding delay compared with the standard 64 kb/s PCM system. For instance, the GSM system's initial upper limit was 65 ms for a back-to-back configuration, whereas for the 16 kb/s G.728 specification [17], it was a maximum of 5 ms with an objective of 2 ms.

2.3.3 Channel and Background Noise Robustness

For many applications, the speech source coding rate typically occupies only a fraction of the total channel capacity, the rest being used for forward error correction (FEC) and signalling. For mobile connections, which suffer greatly from both random and burst errors, a coding scheme's built-in tolerance to channel errors is vital for an acceptable average overall performance, i.e. communication quality. By employing built-in robustness, less FEC can be used and higher source coding capacity is available to give better speech quality. This trade-off between speech quality and robustness is often a very difficult balance to obtain and is a requirement that necessitates consideration from the beginning of the speech coding algorithm design. For other applications employing less severe channels, e.g. fibre-optic links, the problems due to channel errors are reduced significantly and robustness can be ignored for higher clean channel speech quality. This is a major difference between the wireless mobile systems and those of the fixed link systems.

In addition to the channel noise, coders may need to operate in noisy background environments. As background noise can degrade the performance of speech parameter extraction, it is crucial that the coder is designed in such a way that it can maintain good performance at all times. As well as maintaining good speech quality under noisy conditions, good quality background noise

regeneration by the coder is also an important requirement (unless adaptive noise cancellation is used before speech coding).

2.3.4 Complexity and Cost

As ever more sophisticated algorithms are devised, the computational complexity is increased. The advent of Digital Signal Processor (DSP) chips [18] and custom Application Specific Integrated Circuit (ASIC) chips has enabled the cost of processing power to be considerably lowered. However, complexity/power consumption, and hence cost, is still a major problem especially in applications where hardware portability is a prime factor. One technique for overcoming power consumption whilst also improving channel efficiency is digital speech interpolation (DSI) [16]. DSI exploits the fact that only around half of speech conversation is actually active speech thus, during inactive periods, the channel can be used for other purposes, including limiting the transmitter activity, hence saving power. An important subsystem of DSI is the voice activity detector (VAD) which must operate efficiently and reliably to ensure that real speech is not mistaken for silence and vice versa. Obviously, a voice for silence mistake is tolerable, but the opposite can be very annoying.

2.3.5 Tandem Connection and Transcoding

As it is the end to end speech quality which is important to the end user, the ability of an algorithm to cope with tandeming with itself or with another coding system is important. Degradations introduced by tandeming are usually cumulative, and if an algorithm is heavily dependent on certain characteristics then severe degradations may result. This is a particularly urgent unresolved problem with current schemes which employ post-filtering in the output speech signal [17]. Transcoding into another format, usually PCM, also degrades the quality slightly and may introduce extra cost.

2.3.6 Voiceband Data Handling

As voice connections are regularly used for transmission of digital data, e.g. modem, facsimile, and other machine data, an important requirement is an algorithm's ability to transmit voiceband data. The waveform statistics and frequency spectrum of voiceband data signals are quite different from those of speech, therefore the algorithm must be capable of handling both types. The consideration of voiceband data handling is often left until the final stages of the algorithm development, which may be a mistake as end users expect nonvoice information to be adequately transported if the system is employed in the public network. Most of the latest low bit-rate speech coders are unable to pass voiceband data due to the fact they are too speech specific.

Other solutions are often used. A very common one is to detect the voiceband data and use an interface which bypasses the speech encoder/decoder.

2.4 Standard Speech Coders

Standardization is essential in removing the compatibility and conformability problems of implementations by various manufacturers. It allows for one manufacturer's speech coding equipment to work with that of others. In the following, standard speech coders, mostly developed for specific communication systems, are listed and briefly reviewed.

2.4.1 ITU-T Speech Coding Standard

Traditionally the International Telecommunication Union Telecommunication Standardization Sector (ITU-T, formerly CCITT) has standardized speech coding methods mainly for PSTN telephony with 3.4 kHz input speech bandwidth and 8 kHz sampling frequency, aiming to improve telecommunication network capacity by means of digital circuit multiplexing. Additionally, ITU-T has been conducting standardization for wideband speech coders to support 7 kHz input speech bandwidth with 16 kHz sampling frequency, mainly for ISDN applications.

In 1972, ITU-T released G.711 [19], an A/μ-Law PCM standard for 64 kb/s speech coding, which is designed on the basis of logarithmic scaling of each sampled pulse amplitude before digitization into eight bits. As the first digital telephony system, G.711 has been deployed in various PSTNs throughout the world. Since then, ITU-T has been actively involved in standardizing more complex speech coders, referenced as the G.72x series. ITU-T released G.721, the 32 kb/s adaptive differential pulse code modulation (ADPCM) coder, followed by the extended version (40/32/24/16 kb/s), G.726 [20]. The latest ADPCM version, G.726, superseded the former one. Each ITU-T speech coder except G.723.1 [21] was developed with a view to halving the bit rate of its predecessor. For example, the G.728 [22] and G.729 [23] speech coders, finalized in 1992 and 1996, were recommended at the rates of 16 kb/s and 8 kb/s, respectively. Additionally, ITU-T released G.723.1 [21], the 5.3/6.3 kb/s dual-rate speech coder, for video telephony systems. G.728, G.729, and G.723.1 principles are based on code excited linear prediction (CELP) technologies. For discontinuous transmission (DTX), ITU-T released the extended versions of G.729 and G.723.1, called G.729B [24] and G.723.1A [25], respectively. They are widely used in packet-based voice communications [26] due to their silence compression schemes. In the past few years there has been standardization activities at 4 kb/s. Currently there two coders competing for this standard but the process has been put on hold at the moment. One coder is based on the CELP model and the other

Table 2.1 ITU-T narrowband speech coding standards

Speech coder	Bit rate (kb/s)	VAD	Noise reduction	Delay (ms)	Quality	Year
G.711 (A/μ-Law PCM)	64	No	No	0	Toll	1972
G.726 (ADPCM)	40/32/24/16	No	No	0.25	Toll	1990
G.728 (LD-CELP)	16	No	No	1.25	Toll	1992
G.729 (CSA-CELP)	8	Yes	No	25	Toll	1996
G.723.1 (MP-MLQ/ACELP)	6.3/5.3	Yes	No	67.5	Toll/ Near-toll	1995
G.4k (to be determined)	4	–	Yes	~55	Toll	2001

is a hybrid model of CELP and sinusoidal speech coding principles [27, 28]. A summary of the narrowband speech coding standards recommended by ITU-T is given in Table 2.1.

In addition to the narrowband standards, ITU-T has released two wideband speech coders, G.722 [29] and G.722.1 [30], targeting mainly multimedia communications with higher voice quality. G.722 [29] supports three bit rates, 64, 56, and 48 kb/s based on subband ADPCM (SB-ADPCM). It decomposes the input signals into low and high subbands using the quadrature mirror filters, and then quantizes the band-pass filtered signals using ADPCM with variable step sizes depending on the subband. G.722.1 [30] operates at the rates of 32 and 24 kb/s and is based on the transform coding technique. Currently, a new wideband speech coder operating at 13/16/20/24 kb/s is undergoing standardization.

2.4.2 European Digital Cellular Telephony Standards

With the advent of digital cellular telephony there have been many speech coding standardization activities by the European Telecommunications Standards Institute (ETSI). The first release by ETSI was the GSM full rate (FR) speech coder operating at 13 kb/s [31]. Since then, ETSI has standardized 5.6 kb/s GSM half rate (HR) and 12.2 kb/s GSM enhanced full rate (EFR) speech coders [32, 33]. Following these, another ETSI standardization activity resulted in a new speech coder, called the adaptive multi-rate (AMR) coder [34], operating at eight bit rates from 12.2 to 4.75 kb/s (four rates for the full-rate and four for the half-rate channels). The AMR coder aims to provide enhanced speech quality based on optimal selection between the source and channel coding schemes (and rates). Under high radio interference, AMR is capable of allocating more bits for channel coding at the expense of reduced source coding rate and vice versa.

The ETSI speech coder standards are also capable of silence compression by way of voice activity detection [35–38], which facilitates channel

Table 2.2 ETSI speech coding standards for GSM mobile communications

Speech coder	Bit rate (kb/s)	VAD	Noise reduction	Delay (ms)	Quality	Year
FR (RPE-LTP)	13	Yes	No	40	Near-toll	1987
HR (VSELP)	5.6	Yes	No	45	Near-toll	1994
EFR (ACELP)	12.2	Yes	No	40	Toll	1998
AMR (ACELP)	12.2/10.2/7.95/ 7.4/6.7/5.9/ 5.15/4.75	Yes	No	40/45	Toll ~ Communi-cation	1999

interference reduction as well as battery life time extension for mobile communications. Standard speech coders for European mobile communications are summarized in Table 2.2.

2.4.3 North American Digital Cellular Telephony Standards

In North America, the Telecommunication Industries Association (TIA) of the Electronic Industries Association (EIA) has been standardizing mobile communication based on Code Division Multiple Access (CDMA) and Time Division Multiple Access (TDMA) technologies used in the USA. TIA/EIA adopted Qualcomm CELP (QCELP) [39] for Interim Standard-96-A (IS-96-A), operating at variable bit rates between 8 kb/s and 0.8 kb/s controlled by a rate determination algorithm. Subsequently, TIA/EIA released IS-127 [40], the enhanced variable rate coder, which features a novel function for noise reduction as a preprocessor to the speech compression module. Under noisy background conditions, noise reduction provides a more comfortable speech quality by enhancing noisy speech signals. For personal communication systems, TIA/EIA released IS-733 [41], which operates at variable bit rates between 14.4 and 1.8 kb/s. For North American TDMA standards, TIA/EIA released IS-54 and IS-641-A for full rate and enhanced full rate speech coding, respectively [42, 43]. Standard speech coders for North American mobile communications are summarized in Table 2.3.

2.4.4 Secure Communication Telephony

Speech coding is a crucial part of a secure communication system, where voice intelligibility is a major concern in order to deliver the exact voice commands in an emergency.

Standardization has mainly been organized by the Department of Defense (DoD) in the USA. The DoD released Federal Standard-1015 (FS-1015) and FS-1016, called 2.4 kb/s LPC-10e and 4.8 kb/s CELP coders, respectively [44–46]. The DoD also standardized a more recent 2.4 kb/s speech coder [47], based

Table 2.3 TIA/EIA speech coding standards for North American CDMA/TDMA mobile communications

Speech coder	Bit rate (kb/s)	VAD	Noise reduction	Delay (ms)	Quality	Year
IS-96-A (QCELP)	8.5/4/2/0.8	Yes	No	45	Near-toll	1993
IS-127 (EVRC)	8.5/4/2/0.8	Yes	Yes	45	Toll	1995
IS-733 (QCELP)	14.4/7.2/3.6/1.8	Yes	No	45	Toll	1998
IS-54 (VSELP)	7.95	Yes	No	45	Near-toll	1989
IS-641-A (ACELP)	7.4	Yes	No	45	Toll	1996

Table 2.4 DoD speech coding standards

Speech coder	Bit rate (kb/s)	VAD	Noise reduction	Delay (ms)	Quality	Year
FS-1015 (LPC-10e)	2.4	No	No	115	Intelligible	1984
FS-1016 (CELP)	4.8	No	No	67.5	Communication	1991
DoD 2.4 (MELP)	2.4	No	No	67.5	Communication	1996
STANAG (NATO) 2.4/1.2 (MELP)	2.4/1.2	No	Yes	>67.5	Communication	2001

on the mixed excitation linear prediction (MELP) vocoder [48] which is based on the sinusoidal speech coding model. The 2.4 kb/s DoD MELP speech coder gives better speech quality than the 4.8 kb/s FS-1016 coder at half the capacity. A modified and improved version of this coder, operating at dual rates of 2.4/1.2 kb/s and employing a noise preprocessor, has been selected as the new NATO standard. Parametric coders, such as MELP, have been widely used in secure communications due to their intelligible speech quality at very low bit rates. The DoD standard speech coders are summarized in Table 2.4.

2.4.5 Satellite Telephony

The international maritime satellite corporation (INMARSAT) has adopted two speech coders for satellite communications. INMARSAT has selected 4.15 kb/s improved multiband excitation (IMBE) [9] for INMARSAT M systems and 3.6 kb/s advanced multiband excitation (AMBE) vocoders for INMARSAT Mini-M systems (see Table 2.5).

2.4.6 Selection of a Speech Coder

Selecting the best speech coder for a given application may involve extensive testing under conditions representative of the target application. In general, lowering the bit rate results in a reduction in the quality of coded speech.

Table 2.5 INMARSAT speech coding standards

Speech coder	Bit rate (kb/s)	VAD	Noise reduction	Delay (ms)	Quality	Year
IMBE	4.15	No	No	120	Communication	1990
AMBE	3.6	No	No	–	–	–

Quality measurements based on SNR can be used to evaluate coders that preserve the waveform similarity, usually coders operating at bit rates above 16 kb/s. Low bit-rate parametric coders do not preserve the waveform similarity and SNR-based quality measures become meaningless. For parametric coders, perception-based subjective measures are more reliable. The Mean Opinion Score (MOS) [49] scale shown in Table 2.6 is a widely-used subjective quality measure.

Table 2.7 compares some of the most well-known speech coding standards in terms of their bit rate, algorithmic delay and Mean Opinion Scores and Figure 2.2 illustrates the performance of those standards in terms of speech quality against bit rate [50, 51].

Linear PCM at 128 kb/s offers transparent speech quality and its A-law companded 8 bits/sample (64 kb/s) version (which provides the standard for the best (narrowband) quality) has a MOS score higher than 4, which is described as Toll quality. In order to find the MOS score for a given

Figure 2.2 Performance of telephone band speech coding standards (only the top four points of the MOS scale have been used)

Table 2.6 Mean Opinion Score (MOS) scale

Grade (MOS)	Subjective opinion	Quality
5 Excellent	Imperceptible	Transparent
4 Good	Perceptible, but not annoying	Toll
3 Fair	Slightly annoying	Communication
2 Poor	Annoying	Synthetic
1 Bad	Very annoying	Bad

Table 2.7 Comparison of telephone band speech coding standards

Standard	Year	Algorithm	Bit rate (kb/s)	MOS*	Delay+
G.711	1972	Companded PCM	64	4.3	0.125
G.726	1991	VBR-ADPCM	16/24/32/40	toll	0.125
G.728	1994	LD-CELP	16	4	0.625
G.729	1995	CS-ACELP	8	4	15
G.723.1	1995	A/MP-MLQ CELP	5.3/6.3	toll	37.5
ITU 4	–	–	4	toll	25
GSM FR	1989	RPE-LTP	13	3.7	20
GSM EFR	1995	ACELP	12.2	4	20
GSM/2	1994	VSELP	5.6	3.5	24.375
IS54	1989	VSELP	7.95	3.6	20
IS96	1993	Q-CELP	0.8/2/4/8.5	3.5	20
JDC	1990	VSELP	6.7	commun.	20
JDC/2	1993	PSI-CELP	3.45	commun.	40
Inmarsat-M	1990	IMBE	4.15	3.4	78.75
FS1015	1984	LPC-10	2.4	synthetic	112.5
FS1016	1991	CELP	4.8	3	37.5
New FS 2.4	1997	MELP	2.4	3	45.5

* The MOS figures are obtained from formal subjective tests using varied test material (from the literature). These figures are therefore useful as a guide, but should not be taken as a definitive indication of codec performance.

+ Delay is the total algorithmic delay, i.e. the frame length and look ahead, and is given in milliseconds.

coder, extensive listening tests must be conducted. In these tests, as well as the 64 kb/s PCM reference, other representative coders are also used for calibration purposes. The cost of extensive listening tests is high and efforts have been made to produce simpler, less time-consuming, and hence cheaper, alternatives. These alternatives are based on objective measures with some subjective meanings. Objective measurements usually involve point to point comparison of systems under test. In some cases weighting may be used to

give priority to some system parameters over others. In early speech coders, which aimed at reproducing the input speech waveform as output, objective measurement in the form of signal to quantization noise ratio was used. Since the bit rate of early speech coders was 16 kb/s or greater (i.e. they incurred only a small amount of quantization noise) and they did not involve complicated signal processing algorithms which could change the shape of the speech waveform, the SNR measures were reasonably accurate. However at lower bit rates where the noise (the objective difference between the original input and the synthetic output) increases, the use of signal to quantization noise ratio may be misleading. Hence there is a need for a better objective measurement which has a good correlation with the perceptual quality of the synthetic speech. The ITU standardized a number of these methods, the most recent of which is P.862 (or Perceptual Evaluation of Speech Quality). In this standard, various alignments and perceptual measures are used to match the objective results to fairly accurate subjective MOS scores.

2.5 Summary

Existing speech coders can be divided into three groups: parametric coders, waveform approximating coders, and hybrid coders. Parametric coders are not expected to reproduce the original waveform; they reproduce the perception of the original. Waveform approximating coders, on the other hand, are expected to replicate the input speech waveform as the bit rate increases. Hybrid coding is a combination of two or more coders of any type for the best subjective (and perhaps objective) performance at a given bit rate.

The design process of a speech coder involves several trade-offs between conflicting requirements. These requirements include the target bit rate, quality, delay, complexity, channel error sensitivity, and sending of nonspeech signals. Various standardization bodies have been involved in speech coder standardization activities and as a result there have been many standard speech coders in the last decade. The bit rate of these coders ranges from 16 kb/s down to around 4 kb/s with target applications mainly in cellular mobile radio. The selection of a speech coder involves expensive testing under the expected typical operating conditions. The most popular testing method is subjective listening tests. However, as this is expensive and time-consuming, there has been some effort to produce simpler yet reliable objective measures. ITU P.862 is the latest effort in this direction.

Bibliography

[1] W. B. Kleijn and K. K. Paliwal (1995) 'An introduction to speech coding', in *Speech coding and synthesis* by W. B. Kleijn and K. K. Paliwal (Eds), pp. 1–47. Amsterdam: Elsevier Science

[2] D. O'Shaughnessy (1987) *Speech communication: human and machine*. Addison Wesley

[3] I. Atkinson, S. Yeldener, and A. Kondoz (1997) 'High quality split-band LPC vocoder operating at low bit rates', in *Proc. of Int. Conf. on Acoust., Speech and Signal Processing*, pp. 1559–62. May 1997. Munich

[4] R. J. McAulay and T. F. Quatieri (1986) 'Speech analysis/synthesis based on a sinusoidal representation', in *IEEE Trans. on Acoust., Speech and Signal Processing*, 34(4):744–54.

[5] ITU-T (1972) *CCITT Recommendation G.711: Pulse Code Modulation (PCM) of Voice Frequencies*. International Telecommunication Union.

[6] N. S. Jayant and P. Noll (1984) *Digital Coding of Waveforms: Principles and applications to speech and video*. New Jersey: Prentice-Hall

[7] B. S. Atal and M. R. Schroeder (1984) 'Stochastic coding of speech at very low bit rates', in *Proc. Int. Conf. Comm*, pp. 1610–13. Amsterdam

[8] M. Schroeder and B. Atal (1985) 'Code excited linear prediction (CELP): high quality speech at very low bit rates', in *Proc. of Int. Conf. on Acoust., Speech and Signal Processing*, pp. 937–40. Tampa, FL

[9] DVSI (1991) *INMARSAT-M Voice Codec*, Version 1.7. September 1991. Digital Voice Systems Inc.

[10] J. C. Hardwick and J. S. Lim (1991) 'The application of the IMBE speech coder to mobile communications', in *Proc. of Int. Conf. on Acoust., Speech and Signal Processing*, pp. 249–52.

[11] W. B. Kleijn (1993) 'Encoding speech using prototype waveforms', in *IEEE Trans. Speech and Audio Processing*, 1:386–99.

[12] E. Shlomot, V. Cuperman, and A. Gersho (1998) 'Combined harmonic and waveform coding of speech at low bit rates', in *Proc. of Int. Conf. on Acoust., Speech and Signal Processing*.

[13] J. Stachurski and A. McCree (2000) 'Combining parametric and waveform-matching coders for low bit-rate speech coding', in *X European Signal Processing Conf.*

[14] T. Kawashima, V. Sharama, and A. Gersho (1994) 'Network control of speech bit rate for enhanced cellular CDMA performance', in *Proc. IEE Int. Conf. on Commun.*, 3:1276.

[15] P. Ho, E. Yuen, and V. Cuperman (1994) 'Variable rate speech and channel coding for mobile communications', in *Proc. of Vehicular Technology Conf.*

[16] J. E. Natvig, S. Hansen, and J. de Brito (1989) 'Speech processing in the pan-European digital mobile radio system (GSM): System overview', in *Proc. of Globecom*, Section 29B.

[17] J. H. Chen (1990) 'High quality 16 kbit/s speech coding with a one-way delay less than 2 ms', in *Proc. of Int. Conf. on Acoust., Speech and Signal Processing*, pp. 453–6.