Speech Compression

2

This chapter presents an introduction to speech compression techniques, together with a detailed description of speech/audio compression standards including narrowband, wideband and fullband codecs. We will start with the fundamental concepts of speech signal digitisation, speech signal characteristics such as voiced speech and unvoiced speech and speech signal representation. We will then discuss three key speech compression techniques, namely waveform compression, parametric compression and hybrid compression methods. This is followed by a consideration of the concept of narrowband, wideband and fullband speech/audio compression. Key features of standards for narrowband, wideband and fullband codecs are then summarised. These include ITU-T, ETSI and IETF speech/audio codecs, such as G.726, G.728, G.729, G.723.1, G.722.1, G.719, GSM/AMR, iLBC and SILK codecs. Many of these codecs are widely used in VoIP applications and some have also been used in teleconferencing and telepresence applications. Understanding the principles of speech compression and main parameters of speech codecs such as frame size, codec delay, bitstream is important to gain a deeper understanding of the later chapters on Media Transport, Signalling and Quality of Experience (QoE) for VoIP applications.

2.1 Introduction

In VoIP applications, voice call is the mandatory service even when a video session is enabled. A VoIP tool (e.g., Skype, Google Talk and xLite) normally provides many voice codecs which can be selected or updated manually or automatically. Typical voice codecs used in VoIP include ITU-T standards such as 64 kb/s G.711 PCM, 8 kb/s G.729 and 5.3/6.3 kb/s G.723.1; ETSI standards such as AMR; open-source codecs such as iLBC and proprietary codecs such as Skype's SILK codec which has variable bit rates in the range of 6 to 40 kb/s and variable sampling frequencies from narrowband to super-wideband. Some codecs can only operate at a fixed bit rate, whereas many advanced codecs can have variable bit rates which may

L. Sun et al., *Guide to Voice and Video over IP*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4905-7_2, © Springer-Verlag London 2013

be used for adaptive VoIP applications to improve voice quality or QoE. Some VoIP tools can allow speech codecs used to be changed during a VoIP session, making it possible to select the most suitable codec for a given network condition.

Voice codecs or speech codecs are based on different speech compression techniques which aim to remove redundancy from the speech signal to achieve compression and to reduce transmission and storage costs. In practice, speech compression codecs are normally compared with the 64 kb/s PCM codec which is regarded as the reference for all speech codecs. Speech codecs with the lowest data rates (e.g., 2.4 or 1.2 kb/s Vocoder) are used mainly in secure communications. These codecs can achieve compression ratios of about 26.6 or 53.3 (compared to PCM) and still maintain intelligibility, but with speech quality that is somewhat ‘mechanical’. Most of speech codecs operate in the range of 4.8 kb/s to 16 kb/s and have good speech quality and reasonable compression ratio. These codecs are mainly used in bandwidth resource limited mobile/wireless applications. In general, the higher the speech bit rate, the higher the speech quality and the greater the bandwidth and storage requirements. In practice, it is always a trade-off between bandwidth utilisation and speech quality.

In this chapter, we first introduce briefly underpinning basics of speech compression, including speech signal digitisation, voice waveform, spectrum and spectrogram, and the concept of voiced and unvoiced speech. We then look at key techniques in speech compression coding which include waveform coding, parametric coding and hybrid coding (or Analysis-by-Synthesis coding). Finally, we present a number of key speech compression standards, from international standardisation body (ITU-T), regional standardisation bodies (Europe’s ETSI and North America’s TIA), together with some open source and proprietary codecs (such as GIP’s iLBC, now Google’s iLBC and Skype’s SILK codec).

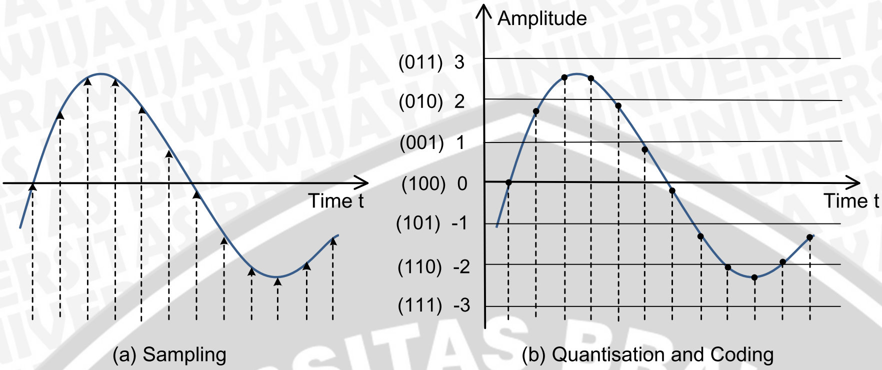
2.2 Speech Compression Basics

The purpose of speech compression is to reduce the number of bits required to represent speech signals (by reducing redundancy) in order to minimise the requirement for transmission bandwidth (e.g., for voice transmission over mobile channels with limited capacity) or to reduce the storage costs (e.g., for speech recording).

Before we start describing speech compression coding techniques, it is important to understand how speech signal is represented in its digital form, that is, the process of speech signal digitisation. We then need to understand what the key features of speech signal are (e.g., voiced and unvoiced speech) and their characteristics. In broad terms, speech compression techniques are mainly focused on removing short-term correlation (in the order of 1 ms) among speech samples and long-term correlation (in the order of 5 to 10 ms) among repeated pitch patterns. In this section, we will start with speech signal digitisation and then discuss speech signal features and speech representation (waveform, spectrum and spectrogram).

2.2.1 Speech Signal Digitisation

Speech signal digitisation is the process to convert speech from analog signal to digital signal in order for digital processing and transmission. The three main phases



(a) Sampling

(b) Quantisation and Coding

Fig. 2.1 Example of voice digitisation

in speech signal digitisation are sampling, quantisation and coding. As shown in Fig. 2.1, sampling is periodic measurement of an analog signal and changes a continuous-time signal into a discrete-time signal. For a narrow-band speech signal with a bandwidth limited to 300 to 3400 Hz (normally simplified to 0–4 kHz), the sampling rate is 8 kHz (i.e., 2 times the maximum signal bandwidth) in accordance with the sampling theorem. If the sampling rate is at least twice the Nyquist frequency (4 kHz for narrow-band voice), the analogue signal can be fully recovered from the samples [26]. If 8 kHz sampling rate is applied, the time difference between two consecutive samples is 0.125 milliseconds $(1/8000 = 0.125)$. Quantisation converts the signal from continuous-amplitude into discrete-amplitude signal. The coding process will further convert discrete-amplitude signal into a series of binary bits (or bitstream) for transmission and storage. For uniform quantisation, quantisation steps are kept the same for all signal amplitudes, see, for example, Fig. 2.1. In the figure, the amplitude space is evenly divided into 6 steps. For 6 different quantisation steps, three-bit binary codes can be used. Each sampled speech signal will be approximated by its closest available quantisation amplitude and then coded into binary bit streams through the coding process. For example, for the 1st sample, the quantised amplitude is zero and the coded bits are 100. For the 2nd sample in the figure, the quantised amplitude is 2 and the coded bits are 010. The difference between the quantised amplitude and actual signal amplitude is called “quantisation error”. Clearly, the more quantisation steps (fine quantisation) there are, the lower the quantisation error, but this requires more bits to represent the signal and the transmission bandwidth will also be greater. In practice, it is always a tradeoff between the desired quantisation error and the transmission bandwidth used.

Considering that speech signals have a non-uniform Probability Density Function (PDF) with lower level speech signal having a much higher PDF than high level speech signal, uniform quantisation will normally create higher quantisation error (or quantisation noise) for low speech signal and hence lower speech quality. Thus, non-uniform quantisation is normally used in speech compression coding. In non-uniform quantisation, fine quantisation is applied for low speech signal. As shown in Fig. 2.2, when uniform quantisation is applied, the quantisation step is kept the

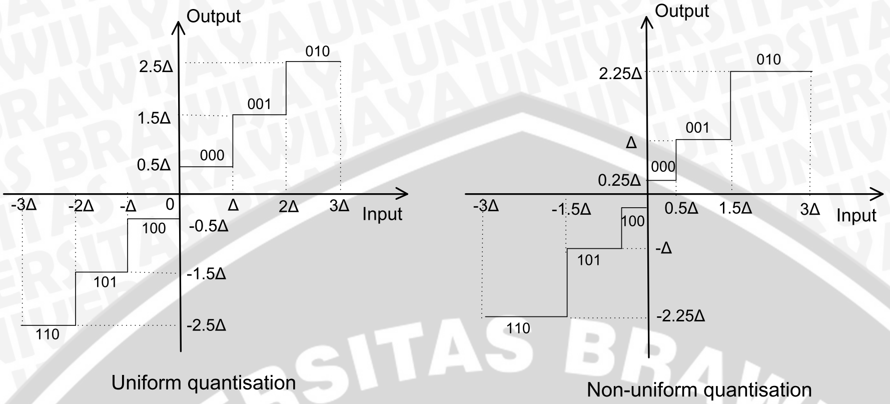


Fig. 2.2 Uniform quantisation and non-uniform quantisation

same (here the value of Δ) in the speech dynamic range considered. For a speech signal in the range of 0 to Δ (input), the output after quantisation will be represented by the quantised value of 0.5Δ with maximum quantisation error of 0.5Δ. When non-uniform quantisation is applied, different quantisation steps will be applied in the speech dynamic range. Due to the fact that speech has non-uniform PDFs, the quantisation step will be kept smaller in lower level signal. For example for speech signal in the range of 0 to 0.5Δ (input), the output will be represented by quantised value of 0.25Δ with maximum quantisation error of 0.25Δ (lower than that for uniform quantisation for low level signals). Similarly for higher level speech signal with lower PDF values, the quantisation step is set much bigger than that for uniform quantisation (coarse quantisation). As illustrated in the figure, for speech signal from 1.5Δ to 3Δ, the quantisation output will be 2.25Δ, with maximum quantisation error of 0.75Δ, much higher than that for uniform quantisation (0.5Δ), also higher than that for lower level speech signal (e.g., 0.25Δ for speech between 0 to 0.5Δ). As PDF of low level speech signal is much higher than that of high level speech signal. The overall performance (in terms of Signal-to-Noise Ratio (SNR)) will be better than that for uniform quantisation coding. In this example, for both uniform and non-uniform quantisation, same signal dynamic range is applied (i.e., from -3Δ to $+3\Delta$ for the input signal). Non-uniform quantisation has been applied in Pulse Coding Modulation (PCM), the most simple and commonly used speech codec. PCM explores non-uniform quantisation by using a logarithm companding method to provide fine quantisation for low speech and coarse quantisation for high speech signal.

After sampling, quantisation and coding, the analog speech signal is converted into a digitised speech signal which can be processed, transmitted or stored. Speech compression coding is normally carried out before digital transmission or storage in order to reduce the required transmission bandwidth or required storage space. For the PCM codec with 8000 sampling rate, each sample is represented by 8 bits, giving transmission bit rate of $8000 \times 8 = 64000$ bit/s (64 kb/s). Speech compression

UNIVERSITAS BRAWIJAYA

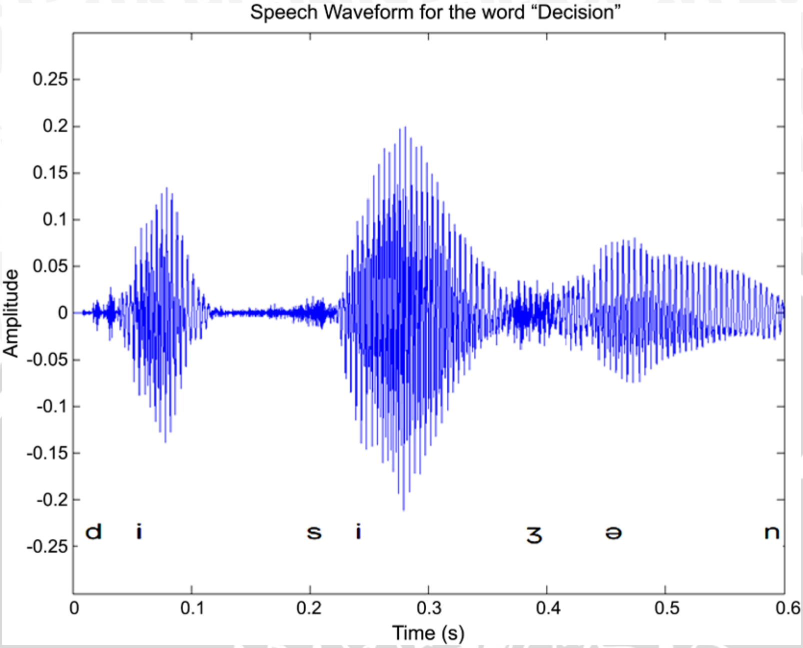


Fig. 2.3 Sample of speech waveform for the word 'Decision'

coding algorithms are normally compared with 64 kb/s PCM to obtain the compression ratio. Details of speech compression coding techniques will be discussed in Sect. 2.3.

2.2.2 Speech Waveform and Spectrum

Speech waveform is the time-domain representation of digitised speech signal. Speech spectrum is the representation of the speech signal in the frequency-domain. Figure 2.3 shows the speech waveform for the word 'Decision'. The speech waveform is normally formed up by voiced and unvoiced speech segments. This is mainly linked to the nature of how speech is produced (details will be discussed in the later Sect. 2.2.3). For voiced speech sounds (e.g., vowel sounds such as 'a', 'i'), these are essentially produced by the vibrations of the vocal cords, and are oscillatory in nature with repeatable patterns. Figure 2.4 illustrates a waveform for a voiced speech segment which has repetitive patterns and its spectrum which shows the basic frequency (pitch) and its harmonic frequencies. For unvoiced sounds, such as 's', 'sh', the signals are more noise-like and there are no repeatable patterns (see Fig. 2.5 for an example of a speech waveform and its spectrum for unvoiced speech segment).

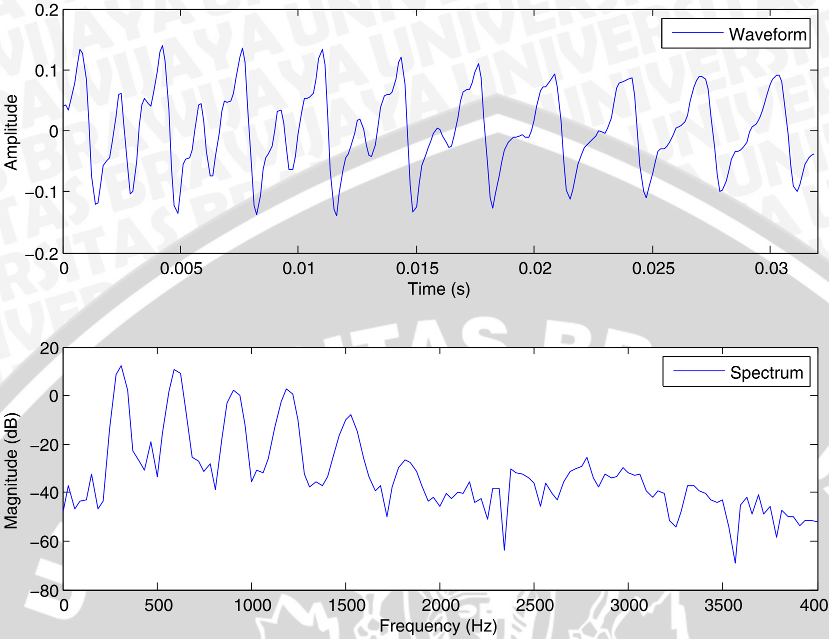


Fig. 2.4 Sample of voiced speech—waveform and spectrum

If we look more closely at the spectrum for voiced signal, it shows harmonic frequency components. For a normal male, the pitch is about 125 Hz and for a female the pitch is at about 250 Hz (Fig. 2.4 has a pitch of 285 Hz for a female sample) for voiced speech, whereas unvoiced signal does not have this feature (as can be seen from Fig. 2.5, the spectrum is almost flat and similar to the spectrum for white noise). The spectrum in Figs. 2.4 and 2.5 are obtained by using a Hamming window with 256 sample window length. The value of waveform amplitude has been normalised to −1 to +1. Spectrum magnitude is converted to dB value. For detailed function of Hamming window and roles of windows in speech signal frequency analysis, readers are recommended to read the book by Kondoz [26].

Figure 2.6 shows the speech waveform of a sentence, “Note closely the size of the gas tank” spoken by a female speaker and its spectrogram. The sentence is about 2.5 seconds long. Speech spectrogram displays the spectrum of the whole sentence of speech with the grade scale of grey for magnitude of the spectrum (the darker the color, the higher the spectrum energy). Pitch harmonised bars are also illustrated clearly in the spectrogram for voiced segments of speech. From the sentence, it is clearly shown that the percentage of the voiced speech segments (with pitch bars) are higher than that of the unvoiced ones (without pitch bars).

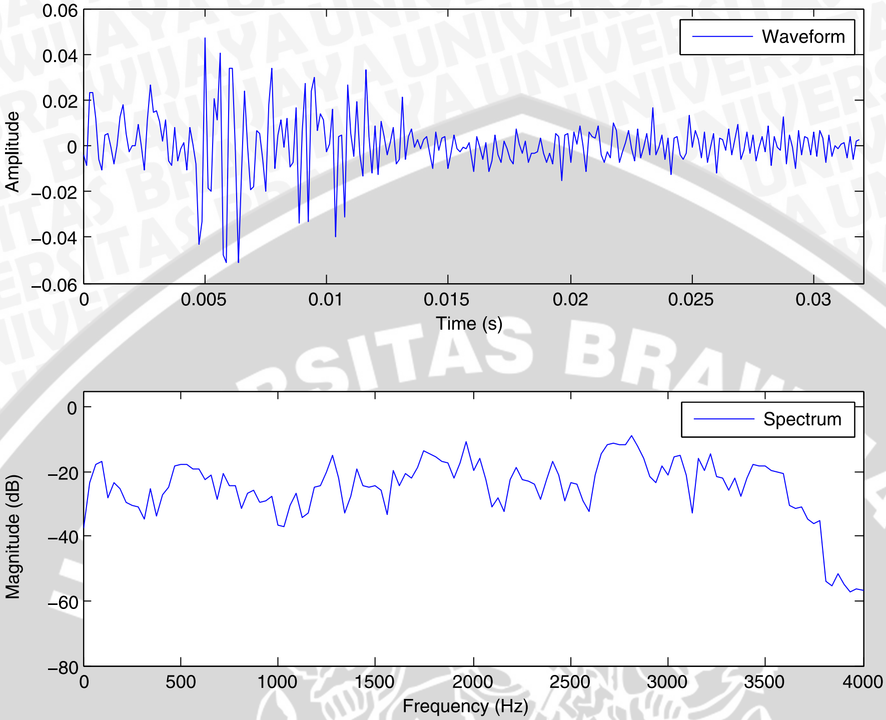


Fig. 2.5 Sample of unvoiced speech—waveform and spectrum

2.2.3 How Is Human Speech Produced?

Speech compression, especially at low bit rate speech compression, explores the nature of human speech production mechanism. In this section, we briefly explain how human speech is produced.

Figure 2.7 shows a conceptual diagram of human speech production physical model. When we speak, the air from lungs push through the vocal tract and out of the mouth to produce a sound. For some sounds for example,. a voiced sound, or vowel sounds of ‘a’, ‘i’ and ‘μ’, as shown in Fig. 2.4, the vocal cords vibrate (open and close) at a rate (fundamental frequency or pitch frequency) and the produced speech samples show a quasi-periodic pattern. For other sounds (e.g., certain fricatives as ‘s’ and ‘f’, and plosives as ‘p’, ‘t’ and ‘k’ , named as unvoiced sound as shown in Fig. 2.5) [28], the vocal cords do not vibrate and remain open during the sound production. The waveform of unvoiced sound is more like noise. The change of the shape of the vocal tract (in combination of the shape of nose and mouth cavities and the position of the tongue) produces different sound and the change of the shape is relatively slow (e.g., 10–100 ms). This forms the basis for the short-term stationary feature of speech signal used for all frame-based speech coding techniques which will be discussed in the next section.

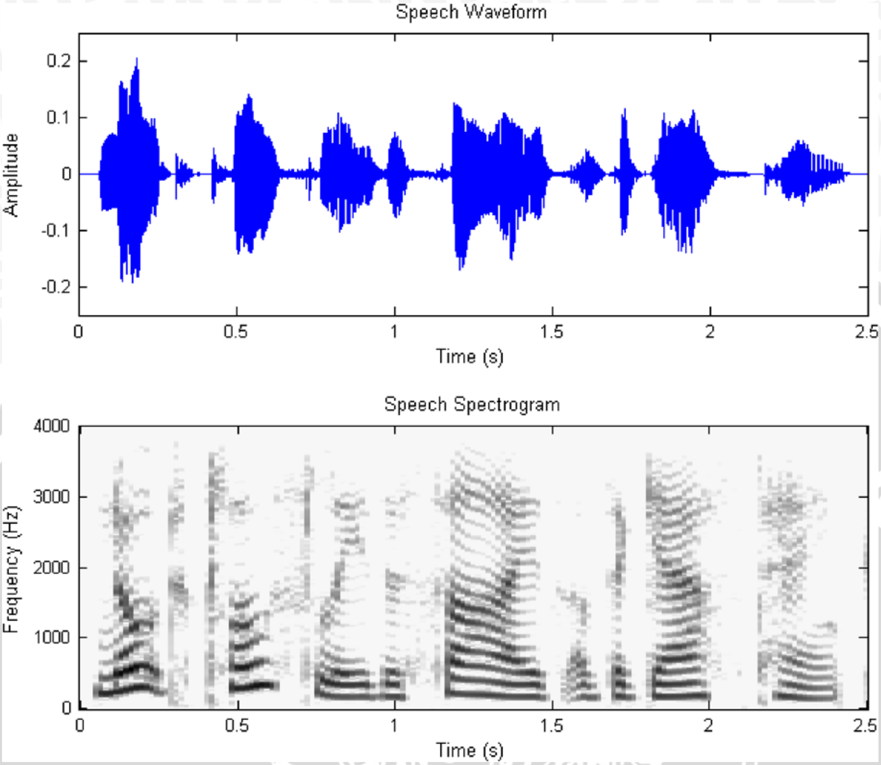
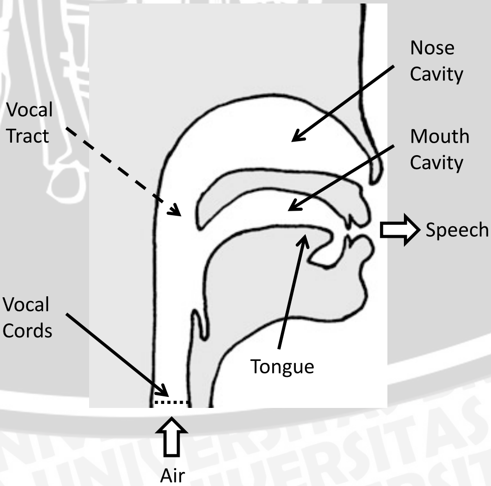


Fig. 2.6 Speech waveform and speech spectrogram

Fig. 2.7 Conceptual diagram of human speech production



UNIVERSITAS BRAWIJAYA

2.3 Speech Compression and Coding Techniques

Speech compression aims to remove redundancy in speech representation to reduce transmission bandwidth and storage space (and further to reduce cost). There are in general three basic speech compression techniques, which are waveform-based, parametric-based and hybrid coding techniques. As the name implied, waveform-based speech compression is mainly to remove redundancy in the speech waveform and to reconstruct the speech waveform at the decoder side as closely as possible to the original speech waveform. Waveform-based speech compression techniques are simple and normally low in implementation complexity, whereas their compression ratios are also low. The typical bit rate range for waveform-based speech compression coding is from 64 kb/s to 16 kb/s. At bit rate lower than 16 kb/s, the quantisation error for waveform-based speech compression coding is too high, and this results in lower speech quality. Typical waveform-based speech compression codecs are PCM and ADPCM (Adaptive Differential PCM) and these will be covered in Sect. 2.3.1.

Parametric-based speech coding is based on the principles of how speech is produced. It is based on the features that speech signal is stationary or the shape of the vocal tract is stable in short period of time (e.g., 20 ms). During this period of time, a speech segment can be classified as either a voiced or unvoiced speech segment. The spectral characteristics of the vocal tract can be represented by a time-varying digital filter. For each speech segment, the vocal tract filter parameters, voiced/unvoiced decision, pitch period and gain (signal energy) parameters are obtained via speech analysis at the encoder. These parameters are then coded into binary bitstream and sent to transmission channel. The decoder at the receiver side will reconstruct the speech (carry out speech synthesis) based on the received parameters. Compared to waveform-based codecs, parametric-based codecs are higher in implementation complexity, but can achieve better compression ratio. The quality of parametric-based speech codecs is low, with mechanic sound, but with reasonable intelligibility. A typical parametric codec is Linear Prediction Coding (LPC) vocoder which has a bit rate from 1.2 to 4.8 kb/s and is normally used in secure wireless communications systems when transmission bandwidth is very limited. The details of parametric-based speech coding will be discussed in Sect. 2.3.2.

As parametric-based codecs cannot achieve high speech quality because of the use of simple classification of speech segments into either voiced or unvoiced speech and simple representation of voiced speech with impulse period train, hybrid coding techniques were proposed to combine the features of both waveform-based and parametric-based coding (and hence the name of hybrid coding). It keeps the nature of parametric coding which includes vocal tract filter and pitch period analysis, and voiced/unvoiced decision. Instead of using an impulse period train to represent the excitation signal for voiced speech segment, it uses waveform-like excitation signal for voiced, unvoiced or transition (containing both voiced or unvoiced) speech segments. Many different techniques are explored to represent waveform-based excitation signals such as multi-pulse excitation, codebook excitation and vector quantisation. The most well known one, so called "Codebook Excitation Linear Prediction (CELP)" has created a huge success for hybrid speech codecs in the range of

UNIVERSITAS BRAWIJAYA

4.8 kb/s to 16 kb/s for mobile/wireless/satellite communications achieving toll quality (MOS over 4.0) or communications quality (MOS over 3.5). Almost all modern speech codecs (such as G.729, G.723.1, AMR, iLBC and SILK codecs) belong to the hybrid compression coding with majority of them based on CELP techniques. More details regarding hybrid speech coding will be presented in Sect. 2.3.3.

2.3.1 Waveform Compression Coding

Waveform-based codecs are intended to remove waveform correlation between speech samples to achieve speech compression. It aims to minimize the error between the reconstructed and the original speech waveforms. Typical ones are Pulse Code Modulation (PCM) and Adaptive Differential PCM (ADPCM).

For PCM, it uses non-uniform quantisation to have more fine quantisation steps for small speech signal and coarse quantisation steps for large speech signal (logarithmic compression). Statistics have shown that small speech signal has higher percentage in overall speech representations. Smaller quantisation steps will have lower quantisation error, thus better Signal-to-Noise Ratio (SNR) for PCM coding. There are two PCM codecs, namely PCM μ-law which is standardised for use in North America and Japan, and PCM A-law for use in Europe and the rest of the world. ITU-T G.711 was standardised by ITU-T for PCM codecs in 1988 [14].

For both PCM A-law and μ-law, each sample is coded using 8 bits (compressed from 16-bit linear PCM data per sample), this yields the PCM transmission rate of 64 kb/s when 8 kHz sample rate is applied (8000 samples/s $\times$ 8 bits/sample $=$ 64 kb/s). 64 kb/s PCM is normally used as a reference point for all other speech compression codecs.

ADPCM, proposed by Jayant in 1974 at Bell Labs [25], was developed to further compress PCM codec based on correlation between adjacent speech samples. Consisting of adaptive quantiser and adaptive predictor, a block diagram for ADPCM encoder and decoder (codec) is illustrated in Fig. 2.8.

At the encoder side, ADPCM first converts 8 bit PCM signal (A-law or μ-law) to 16 bit linear PCM signal (the conversion is not shown in the figure). The adaptive predictor will predict or estimate the current speech signal based on previously received (reconstructed) N speech signal samples $\tilde{s}(n)$ as given in Eq. (2.1).

$$\hat{s}(n) = \sum_{i=1}^{N} a_i(n)\tilde{s}(n-i) \tag{2.1}$$

where a_i, $i = 1, \ldots, N$ are the estimated predictor coefficients, and a typical N value is six.

Difference signal $e(n)$, also known as prediction error, is calculated from the speech signal $s(n)$ and the signal estimate $\hat{s}(n)$ and is given in Eq. (2.2). Only this difference signal (thus the name differential coding) is input to the adaptive quantiser for quantisation process. As the dynamic range of the prediction error, $e(n)$ is

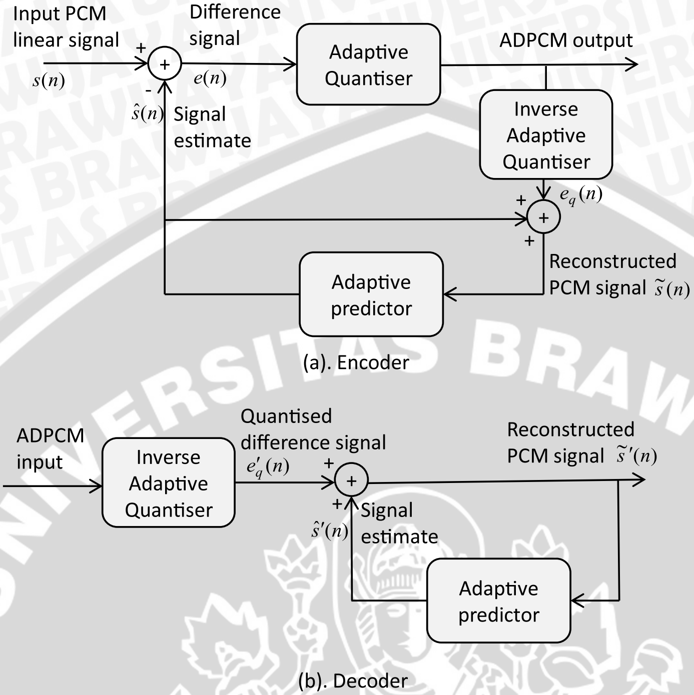


Fig. 2.8 Block diagram for ADPCM codec

smaller than that of the PCM input signal, less coding bits are needed to represent the ADPCM sample.

$$e(n) = s(n) - \hat{s}(n) \tag{2.2}$$

The difference between $e(n)$ and $e_q(n)$ is due to quantisation error ($n_q(n)$), as given in Eq. (2.3).

$$e(n) = e_q(n) + n_q(n) \tag{2.3}$$

The decoder at the receiver side will use the same prediction algorithm to reconstruct the speech sample. If we don't consider channel error, $e_q(n) = e'_q(n)$. The difference between the reconstructed PCM signal at the decoder ($\tilde{s}'(n)$) and the input linear PCM signal at the encoder ($s(n)$) will be just the quantisation error of $n_q(n)$. In this case, the Signal-to-Noise Ratio (SNR) for the ADPCM system will be mainly decided by the signal to quantisation noise ratio and the quality will be based on the performance of the adaptive quantiser.

If an ADPCM sample is coded into 4 bits, the produced ADPCM bit rate is $4 \times 8 = 32$ kb/s. This means that one PCM channel (at 64 kb/s) can transmit two ADPCM channels at 32 kb/s each. If an ADPCM sample is coded into 2 bits, then ADPCM bit rate is $2 \times 8 = 16$ kb/s. One PCM channel can transmit four ADPCM at 16 kb/s each. ITU-T G.726 [15] defines ADPCM bit rate at 40, 32, 24 and 16 kb/s which corresponds to 5, 4, 3, 2 bits of coding for each ADPCM sample. The higher the ADPCM bit rate, the higher the numbers of the quantisation levels, the lower the quantisation error, and thus the better the voice quality. This is why the quality for 40 kb/s ADPCM is better than that of 32 kb/s. The quality of 24 kb/s ADPCM is also better than that of 16 kb/s.

2.3.2 Parametric Compression Coding

Waveform-based coding aims to reduce redundancy among speech samples and to reconstruct speech as close as possible to the original speech waveform. Due to its nature of speech sample-based compression, waveform-based coding cannot achieve high compression ratio and normally operates at bit rate ranging from 64 kb/s to 16 kb/s.

In contrast, parametric-based compression methods are based on how speech is produced. In stead of transmitting speech waveform samples, parametric compression only sends relevant parameters related with speech production to the receiver side and reconstructs the speech from the speech production model. Thus, high compression ratio can be achieved. The most typical example of parametric compression is Linear Prediction Coding (LPC), proposed by Atal in 1971 [4] at Bell Labs. It was designed to emulate the human speech production mechanisms and the compression can reach the bit rate as lower as 800 bit/s (Compression Ratio reaches 80 when compared to 64 kb/s PCM). It normally operates at bit rates from 4.8 to 1.2 kb/s. The LPC based speech codecs can achieve high compression rate, however, the voice quality is also low, especially the natureness of the speech (i.e., can you recognise who is talking). The speech sound based on simple LPC model is more like mechanic or robotic sound, but can still achieve high intelligibility (i.e., understanding the meaning of a sentence). In this section, we will discuss briefly how human speech is generated and what a basic LPC model is.

Speech Generation Mathematic Model

Based on the nature of speech production, a speech generation mathematical model can be shown in Fig. 2.9. Depending on whether the speech signal is voiced or unvoiced, the speech excitation signal ($x(n)$) is switched between a period pulse train signal (controlled by the pitch period of T for the voiced signal) and random noise signal (for unvoiced speech). The excitation signal is amplified by Gain (G or energy of the signal) and then sent to the vocal tract filter or LPC filter.

The vocal tract filter can be modelled by a linear prediction coding (LPC) filter (a time-varying digital filter) and can be represented approximated by an all-pole filter as given by Eq. (2.4). The LPC filter mainly reflects the spectral envelope part of the speech segment.

$$H(z) = \frac{S(z)}{X(z)} = \frac{G}{1 - \sum_{j=1}^{p} a_j z^{-j}} \tag{2.4}$$

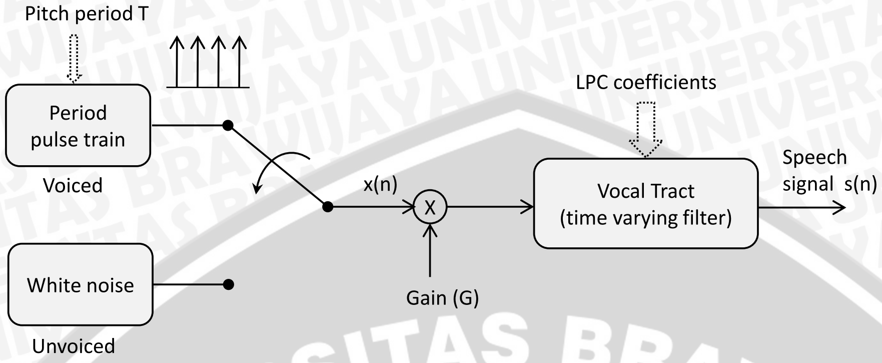


Fig. 2.9 Speech generation mathematical model

where a_j, $j = 1, \dots, p$, represents p-order LPC filter coefficients and the p value is normally ten for narrow-band speech (normally named as the ten-order LPC filter).

When converted to the time-domain, we can obtain the generated speech signal $s(n)$ from a difference equation (see Eq. (2.5)). This means that the output speech signal $s(n)$ can be predicted from the weighted sum of the past p speech output signal samples ($s(n-j)$, $j = 1, \dots, p$), or from the linear combination of previous speech outputs (thus, the name of Linear Prediction Coding, LPC), and the present excitation signal $x(n)$ and the gain (G). Equation (2.5) represents a general expression for LPC-based model which includes mainly two key elements, i.e. the excitation part and the LPC filter. In a basic LPC model, only impulse pulse train (for voiced) or white noise (for unvoiced) is used for the excitation signal. This simplified excitation model can achieve high compression efficiency (with bit rates normally between 800 bit/s to 2,400 bit/s), but with low perceived speech quality (due to mechanic sound) and reasonable intelligibility. They are mainly used in secure telephony communications.

$$s(n) = Gx(n) + \sum_{j=1}^{p} a_j s(n-j) \tag{2.5}$$

For more detailed explanation of speech signal generation model and LPC analysis, readers are recommended to read the reference book [26].

Linear Prediction Coding (LPC) Model

The LPC model, also known as the LPC vocoder (VOice enCODER), was proposed in 1960s and is based on the speech generation model presented in Fig. 2.9. The idea is that for a given segment of speech (e.g., 20 ms of speech, which corresponds to 160 samples at 8 kHz sampling rate), if we can detect whether it is voiced or unvoiced and estimate its LPC filter parameters, pitch period (for voiced signal) and its gain (power) via speech signal analysis, we can then just encode and send

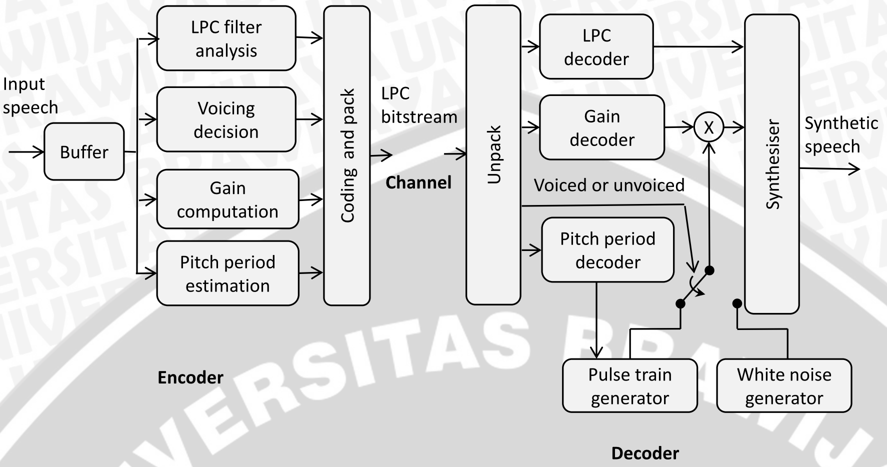


Fig. 2.10 The LPC model

these parameters to the channel/network and then synthesise the speech based on the received parameters at the decoder. For a continuous speech signal which is segmented for 20 ms speech frames, this process is repeated for each speech frame. The basic LPC model is illustrated in Fig. 2.10.

At the encoder, the key components are pitch estimation (to estimate the pitch period of the speech segment), voicing decision(to decide whether it is a voiced or unvoiced frame), gain calculation (to calculate the power of the speech segment) and LPC filter analysis (to predict the LPC filter coefficients for this segment of speech). These parameters/coefficients are quantised, coded and packetised appropriately (in the right order) before they are sent to the channel. The parameters and coded bits from the LPC encoder are listed below.

- Pitch period (T): for example, coded in 7 bits as in LPC-10 (together with voicing decision) [31].
- Voiced/unvoiced decision: to indicate whether it is voiced or unvoiced segment. For hard-decision, a binary bit is enough.
- Gain (G) or signal power: coded in 5 bits as in LPC-10.
- Vocal tract model coefficients: or LPC filter coefficients, normally in 10-order, i.e. $a_1, a_2, \ldots, a_{10}$, coded in 41 bits in LPC-10.

At the decoder, the packetised LPC-bitstream are unpacked and sent to the relevant decoder components (e.g., LPC decoder, pitch period decoder) to retrieve the LPC coefficients, pitch period and gain. The voicing detection bit will be used to control the voiced/unvoiced switch. The pitch period will control the impulse train sequence period when in a voiced segment. The synthesiser will synthesise the speech according to the received parameters/coefficients.

LPC-10 [31] is a standard specified by Department of Defence (DoD) Federal Standard (FS) 1015 in USA and is based on 10th order LP analysis. Its coded bits are 54 (including one bit for synchronisation) for one speech frame with 180 samples. For 8 kHz sampling rate, 180 samples per frame which is 22.5 ms per frame (180/8000 = 22.5 ms). For every 22.5 ms, 54 coded binary bits from the encoder are sent to the channel. The encoder bit rate is 2400 bit/s or 2.4 kb/s (54 bits/22.5 ms = 2.4 kb/s). The compression ratio is 26.7 when compared with 64 kb/s PCM (64/2.4). LPC-10 was mainly used in radio communications with secure voice transmissions. The quality of voice is low in its natureness (more mechanic sound), but with reasonable intelligibility. Some variants of LPC-10 explore different techniques (e.g., subsampling, silence detection, variable LP coded bits) to achieve bit rates from 2400 bit/s to 800 bit/s.

2.3.3 Hybrid Compression Coding—Analysis-by-Synthesis

Problems with Simple LPC Model

Using a sharp voiced/unvoiced decision to differentiate a speech frame as either voiced or unvoiced, and using a periodic impulse train to emulate voiced speech signal and noise for unvoiced speech are major limitations of LPC-based vocoders.

Now let's look at an example of the output from LPC analysis. Figure 2.11 shows the waveform, the LPC filter (vocal tract filter or spectral envelope) and the residual signal after removing the short-term LPC estimation from the original speech signal. From the residual signal, we can see that the signal energy is greatly less than that of the original speech and the period pattern is still there. This is because LPC filter can only remove short-term correlation between samples, but not long-term correlation between period pattern signals. This can also be shown from Fig. 2.12 with residual signal spectrum is more flat (with formants are removed via LPC filter). However, the pitch frequency and its harmonic frequencies are still there and this needs to be removed by Pitch filter, or the so-called Long-Term Prediction (LTP) filter which removes correlation between pitch period patterns.

From Fig. 2.11 for a voiced speech segment, we can see that LPC residual signal is not a simple period pulse signal. If we can find the best match of excitation signal which can represent as close as possible to this residual signal, then when this residual signal is passed through the LPC filter, a perfect reconstruction signal will be produced.

In order to find the best match of the excitation signal, a synthesiser (including LPC synthesiser and pitch synthesiser) is included at the encoder side and a closed-loop search is carried out in order to find the best match excitation signal (which results in a minimum perceptual error estimation between the original and the synthesised speech signal). This is the key concept of hybrid speech coding (combines the features of both waveform and parametric coding), also known as Analysis-by-Synthesis (AbS) method as shown in Fig. 2.13. The LPC synthesiser predicts the short-term vocal tract filter coefficients, whereas, the pitch synthesiser predicts the long-term pitch period and gain for the voiced segment. The parameters for the best match excitation signal, together with pitch period, gain and LPC filter coefficients

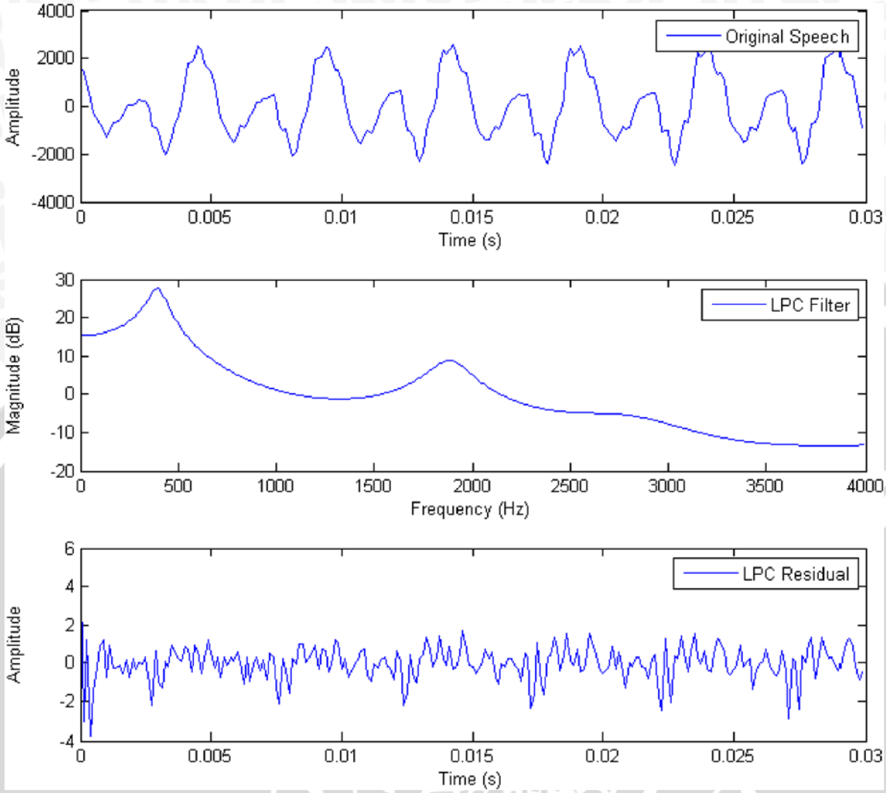


Fig. 2.11 Speech waveform, LPC filter, and residual signal

are transmitted to the receiver. The decoder will synthesise the speech signal based on the optimum excitation signal. The difference between the synthesised at the output of the decoder and the one estimated at the encoder is due to channel error. If there is no channel transmission error, the synthesised signals at the encoder and the decoder are the same.

In hybrid compression coding, the most successful one is Code-Excitation Linear Prediction (CELP) based AbS technique which was a major breakthrough at low bit rate speech compression coding in later 1980s. CELP-based coding normally contains a codebook with a size of 256 to 1024 at both sender and receiver. Each codebook entry contains a waveform-like excitation signal, or multi-pulse excitation signal [5] (instead of only periodic impulse train and noise in parametric coding). This resolves a major problem in the coding of a transition frame (or “onset” frame), for example, a frame contains transition from unvoiced to voiced, such as the phonetic sound at the beginning of the word “see” [si:] or “tea” [ti:] which is very important from perceptual quality point of view (affects the intelligibility of speech communications). The closed-loop search process will find the best match excitation from the codebook and only the index of the matched excitation of the codebook will be

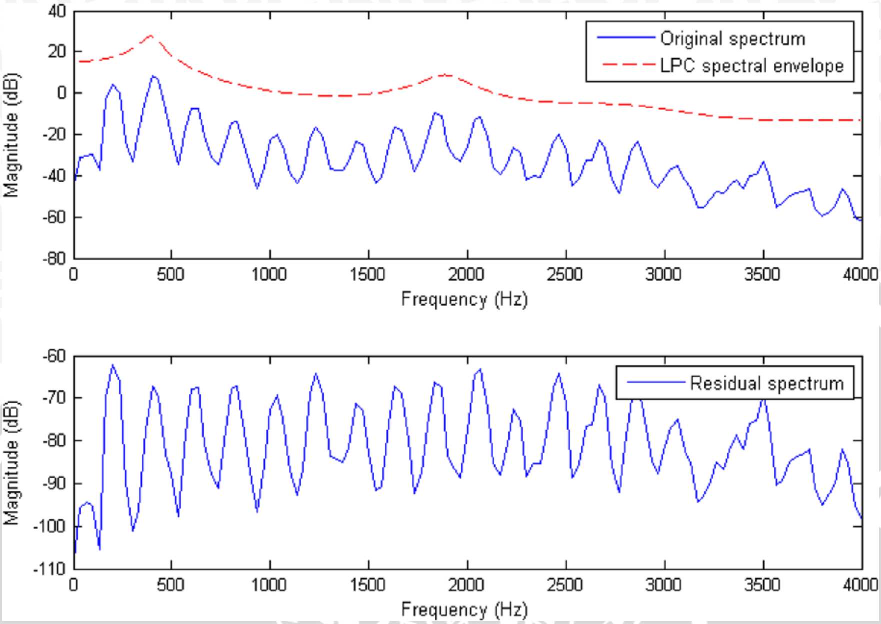
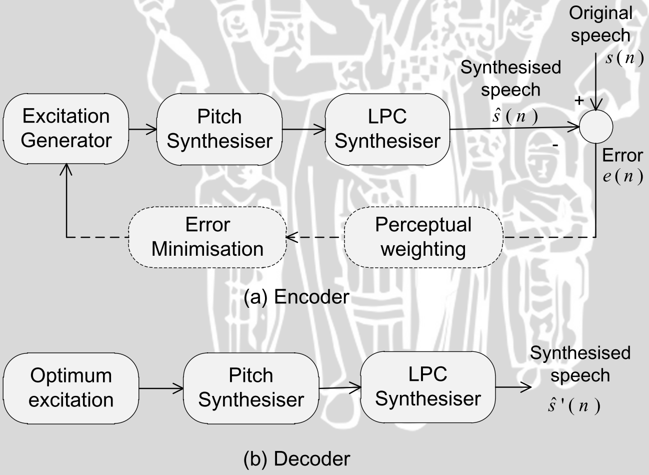


Fig. 2.12 Spectrum for original speech and residual signal



(a) Encoder

(b) Decoder

Fig. 2.13 Analysis-by-Synthesis LPC codec

coded and sent to the decoder at the receiver side. At the decoder side, the matched excitation signal will be retrieved from the same codebook and used to reconstruct the speech. For a codebook with the size of 256 to 1024, 8–10 bits can be used for the coding of codebook index. In order to achieve high efficiency in coding and low in

UNIVERSITAS BRAWIJAYA

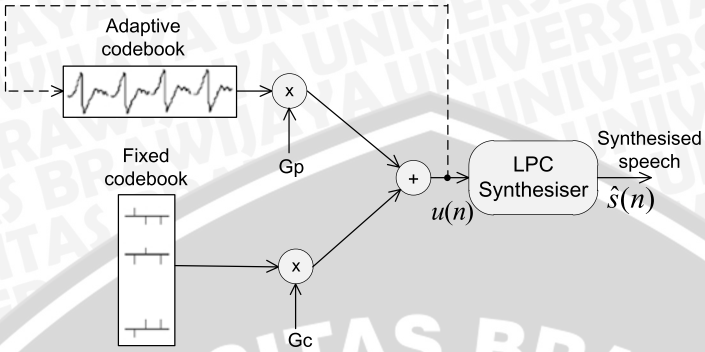


Fig. 2.14 Example of Code-Excitation Linear Prediction (CELP)

implementation complexity, a large codebook is normally split into several smaller codedbooks. Figure 2.14 shows an example of CELP used in the AMR codec [1] which includes two codebooks, an adaptive codebook to search for pitch excitation and a fixed codebook containing a set of fixed pulse train with preset pulse position and signs of the pulses. Pitch excitation codebook contains waveform-like excitation signals. Due to the successful use of the CELP techniques, the voice quality of hybrid compression coding has reached toll quality (MOS score over 4) or communications quality (MOS score over 3.5) at bit rates from 16 kb/s to 4.8 kb/s. This is impossible for waveform or parametric codecs to achieve high speech quality at this range of bit rates. The hybrid AbS-based codecs have been widely used in today's mobile, satellite, marine and secure communications.

In general, there are two categories in hybrid compression coding based on how the excitation signal is generated. One is based on excitation signal analysis and generation in the time-domain and aims to reconstruct the speech frame as close as possible on the speech waveform. Majority of CELP variants belong to this category, such as ACELP (Algebraic Code Excited Linear Prediction) and RELP (Residual pulse Excitation Linear Prediction). Another category is based on excitation signal analysis in the frequency-domain and aims to reconstruct the speech frame as close as possible from the speech spectrum point of view. Multiband Excitation (MBE) model proposed by Griffin and Lim in 1988 at MIT [10] is in this category. MBE divides the speech spectrum into several sub-bands (about 20) and a binary voiced/unvoiced parameter is allocated to each frequency band. This will make the spectrum of the reconstructed speech frame more close to the spectrum of the original speech frame and will produce better speech quality than traditional time-domain CELP at low bit rates, for example, 2.4 to 4.8 kb/s.

The typical hybrid compression codecs include the following from several standardisation bodies, such as the International Telecommunication Union, Telecommunication Standardisation Sector (ITU-T), European Telecommunication Standards Institute (ETSI), North America's Telecommunications Industry Association

(TIA) of the Electronic Industries Association (EIA), and the International Maritime Satellite Corporation (INMARSAT).

- LD-CELP: Low Delay CELP, used in ITU-T G.728 at 16 kb/s [16].
- CS-ACELP: Conjugate-Structure Algebraic-Code-Excited Linear Prediction, used in ITU-T G.729 [17] at 8 kb/s.
- RPE/LTP: Regular Pulse Excitation/Long Term Prediction, used in ETSI GSM Full-Rate (FR) at 13 kb/s [6].
- VSELP: Vector Sum Excited Linear Prediction: ETSI GSM Half-Rate (HR) at 5.6 kb/s [7].
- EVRC based on RCELP: Enhanced Variable Rate Codec [30], specified in TIA/EIA's Interim Standard TIA/EIA/IS-127 for use in the CDMA systems in North America, operating at bit rates of 8.5, 4 or 0.8 kb/s (full-rate, half-rate, eighth-rate at 20 ms speech frame) [30].
- ACELP: Algebraic CELP, used in ETSI GSM Enhanced Full-Rate (EFR) at 12.2 kb/s [9] and ETSI AMR from 4.75 to 12.2 kb/s [8].
- ACELP/MP-MLQ: Algebraic CELP/Multi Pulse—Maximum Likelihood Quantisation, used in ITU-T G.723.1 at 5.3/6.3 kb/s [18].
- IMBE: Improved Multiband Excitation Coding at 4.15 kb/s for INMARSAT-M.
- AMBE: Advanced Multiband Excitation Coding at 3.6 kb/s for INMARSAT-AMBE.

2.3.4 Narrowband to Fullband Speech Audio Compression

In the above sections, we mainly discussed Narrowband (NB) speech compression, aimed at speech spectrum from 0 to 4 kHz. Not only used in VoIP systems, this 0 to 4 kHz narrowband speech, expanded from speech frequency range of 300 Hz to 3400 Hz, has also been used in traditional digital telephony in the Public Switched Telephone Networks (PSTN) .

In VoIP and mobile applications, there is a trend in recent years to use Wideband (WB) speech to provide high fidelity speech transmission quality. For WB speech, the speech spectrum is expanded to 0–7 kHz, with sampling rate at 16 kHz. Compared to 0–4 kHz narrowband speech, wideband speech will have more higher frequency components and have high speech fidelity. The 0–7 kHz wideband speech frequency range is equivalent to general audio signal frequency range (e.g., music).

There are currently three wideband speech compression methods which have been used in different wideband speech codecs standardised by ITU-T or ETSI. They are:

- *Waveform compression based on sub-band* (*SB*) *ADPCM*: such as ITU-T G.722 [12].
- *Hybrid compression based on CELP*: such as AMR-WB or ITU-T G.722.2 [21].
- *Transform compression coding*: such as ITU-T G.722.1 [20].

Table 2.1 Summary of NB, WB, SWB and FB speech/audio compression coding

Mode	Signal bandwidth (Hz)	Sampling rate (kHz)	Bit-rate (kb/s)	Examples
Narrowband (NB)	300–3400	8	2.4–64	G.711, G.729, G.723.1, AMR, LPC-10
Wideband (WB)	50–7000	16	6.6–96	G.711.1, G.722, G.722.1, G.722.2
Super-wideband (SWB)	50–14000	32	24–48	G.722.1 (Annex C)
Fullband (FB)	20–20000	48	32–128	G.719

It needs to be mentioned that G.711.1 uses both waveform compression (for lower band signal based on PCM) and transform compression (for higher band signal based on Modified DCT).

Supter-wideband (SWB) is normally referred to speech compression coding for speech and audio frequency from 50 to 14 000 Hz.

Fullband (FB) speech/audio compression coding considers the full human auditory bandwidth from 20 Hz to 20 kHz to provide high quality, efficient compression for speech, music and general audio. The example is the latest ITU-T standard G.719 [23]. It is mainly used in teleconferencing and telepresence applications.

Table 2.1 summarizes the Narrowband, Wideband, Super-wideband and Fullband speech/audio compression coding basic information, including signal bandwidth, sampling rate, typical bit rate range and standards examples. Details regarding these standard codecs will be covered in the next section.

2.4 Standardised Narrowband to Fullband Speech/Audio Codecs

In this section, we will discuss some key standardised narrowband (NB), wideband (WB), super-wideband (SWB) and fullband (FB) speech/audio codecs from the International Telecommunication Union, Telecommunications Section (ITU-T) (e.g., G.729, G.723.1, G.722.2 and G.719), from the European Telecommunications Standards Institute (ETSI) (e.g., GSM, AMR, AMR-WB) and from the Internet Engineering Task Force (IETF) (e.g., iLBC and SILK) which are normally used in VoIP and conferencing systems.

2.4.1 ITU-T G.711 PCM and G.711.1 PCM-WB

G.711 for 64 kb/s Pulse Coding Modulation (PCM) was first adopted by ITU-T in 1972 and further amended in 1988 [14]. It is the first ITU-T speech compression coding standard for the ITU-T G-series for narrowband speech with a frequency range of 300–3400 Hz. Two logarithmic companding laws were defined due to historic reasons, with the μ-law for use in North America and Japan, and the A-law for use in Europe and the rest of the world. The G.711 encoder converts linear 14 bits uniform PCM code to 8 bits A-law or μ-law PCM (non-uniform quantisation, or logarithm companding) code per sample with fine quantisation for low level speech signal and coarse quantisation for high level speech signal. At the decoder side, decompanding process is applied to convert back to its uniform PCM signal. PCM operates at 64 kb/s and is sample-based coding, which means that the algorithmic delay for the encoder is only one sample of 0.125 ms at 8000 Hz sampling rate.

When PCM codec is used in VoIP applications, 20 ms of speech frame is normally formed up and packetised for transmission over the network. The original G.711 PCM standard did not contain packet loss concealment mechanism which is necessary for codecs for VoIP applications. G.711 Appendix I [19] was added in 1999 which contains a high quality low-complexity algorithm for packet loss concealment. This G.711 with packet loss concealment algorithm (PLC) is mandatory for all VoIP applications.

G.711.1 is the wideband extension for G.711 Pulse Code Modulation (PCM-WB) defined by ITU-T in 2008 [24]. It supports both narrowband and wideband speech coding. When it is applied for wideband speech coding, it can support speech and audio input signal frequency range from 50 to 7000 Hz. The encoder input signal, sampled at 16 kHz (in wideband coding case), is divided into 8 kHz sampled lower-band and higher-band signals with the lower-band using G.711-compatible coding, whereas the higher-band based on Modified Discrete Cosine Transform (MDCT) based on 5 ms speech frame. For the lower-band and higher-band signals, there are three layers of bitstreams as listed below.

- *Layer 0*: lower-band base bitstream at 64 kb/s PCM (base bitstream), 320 coded bits for 5 ms speech frame.
- *Layer 1*: lower-band enhancement bitstream at 16 kb/s, 80 coded bits for 5 ms speech frame.
- *Layer 2*: higher-band enhancement bitstream at 16 kb/s, 80 coded bits for 5 ms speech frame.

The overall bit rates for G.711.1 PCM-WB can be 64, 80 and 96 kb/s. With 5 ms speech frame, the coded bits are 320, 400, 480 bits, respectively. The algorithmic delay is 11.875 ms (5 ms speech frame, 5 ms look-ahead, and 1.875 ms for Quadrature-Mirror Filterbank (QMF) analysis/synthesis).

2.4.2 ITU-T G.726 ADPCM

G.726 [15], defined by ITU-T in 1990, is an ADPCM-based narrowband codec operating at bit rates of 40, 32, 24 and 16 kb/s. G.726 incorporates the previous ADPCM standards of G.721 [11] at 32 kb/s and G.723 [13] at 24 and 40 kb/s (both specified in 1988). For 40, 32, 24 and 16 kb/s bit rates, the corresponding ADPCM bits per sample are 5, 4, 3, and 2 bits. It operates at narrowband with the sampling rate of 8000 Hz. G.726 was originally proposed to be used for Digital Circuit Multiplication Equipment (DCME) to improve transmission efficiency for long distance speech transmission (e.g. one 64 kb/s PCM channel can hold two 32 kb/s ADPCM and four 16 kb/s ADPCM channels). The G.726 codec currently is also used for VoIP applications.

2.4.3 ITU-T G.728 LD-CELP

G.728 [16], a narrowband codec, was standardised by ITU-T in 1992, as a 16 kb/s speech compression coding standard based on low-delay code excited linear prediction (LD-CELP). It represented a major breakthrough in speech compression coding history and was the first speech standard based on Code Excited Linear Prediction (CELP) using an analysis-by-synthesis approach for codebook search. It achieves near toll quality (MOS score near 4) at 16 kb/s, similar quality as 32 kb/s ADPCM and 64 kb/s PCM. After G.728, many speech coding standards were proposed based on variants of CELP.

To achieve low delay, G.728 uses a small speech block for CELP coding. The speech block consists of only five consecutive speech samples, which has an algorithmic delay of 0.625 ms (0.125 ms $\times$ 5). It uses a size of 1024 vectors of codebook and is coded into 10 bits for codebook index ("code-vector"). Only these 10 bits for codebook index (or Vector Quantisation, VQ index) is sent to the receiver for each block of speech (equivalent to 10 bits/0.625 ms = 16 kb/s). Backward gain adaptation and backward predictor adaptation are used to derive the excitation gain and LPC synthesis filter coefficients at both encoder and decoder. These parameters are updated at every four consecutive blocks (every 20 speech samples or 2.5 ms). A 50th order (instead of 10th) LPC predictor is applied. To reduce codebook search complexity, two smaller codebooks are used instead of one 10-bit 1024-entry codebook (one 7-bit 128-entry "shape codebook" and one 3-bit 8-entry "gain codebook").

G.728 can further reduce its transmission bit rate to 12.8 and 9.6 kb/s which are defined at Annex H. The lower bit rate transmission is more efficient in DCME and VoIP applications.

2.4.4 ITU-T G.729 CS-ACELP

ITU-T G.729 [17], standardised in 1996, is based on CS-ACELP (Conjugate Structure- Algebraic Code Excited Linear Prediction) algorithm. It operates at 8 kb/s with 10 ms speech frame length, plus 5 ms look-ahead (a total algorithmic delay of 15 ms). Each 10 ms speech frame is formed up by two sub-frames with each of 5 ms. The LPC filter coefficients are estimated based on the analysis on the

UNIVERSITAS BRAWIJAYA

10 ms speech frame, whereas the excitation signal parameters (fixed and adaptive codebook indices and gains) are estimated based on the analysis of each subframe (5 ms). LPC filter coefficients are transformed to Line Spectrum Pairs (LSP) for stability and efficiency of transmission. For the G.729 encoder, every 10 ms speech frame (for 8 kHz sampling rate, it is equivalent to 80 speech samples) is analysed to obtain relevant parameters, which are then encoded to 80 bits and transmitted to the channel. The encoder bit rate is 8 kb/s (80 bits/10 ms = 8 kb/s). G.729 supports three speech frame types, which are normal speech frame (with 80 bits), Silence Insertion Description (SID) frame (with 15 bits, to indicate the features of background noise when voice activity detection (VAD) is enabled) and a null frame (with 0 bit). G.729 was designed for cellular and network applications. It has a built-in concealment mechanism to conceal a missing speech frame using interpolation techniques based on previous received speech frames. For detailed bit allocation of 80 bits to LPC filter coefficients and excitation codebooks, you can read ITU-T G.729 [17]. In the G.729 standard, it also defines G.729A (G.729 Annex A) for reduced complexity algorithm operating at 8 kb/s, Annex D for low-rate extension at 6.4 kb/s and Annex E for high-rate extension at 11.8 kb/s.

2.4.5 ITU-T G.723.1 MP-MLQ/ACELP

ITU-T G.723 [18], standardised in 1996, is based on Algebraic CELP (ACELP) for bit rate at 5.3 kb/s and Multi Pulse—Maximum Likelihood Quantisation (MP-MLQ) for bit rate at 6.3 kb/s. It was proposed for multimedia communications such as for very low bit rate visual telephony applications and provides dual rates for flexibility. The higher bit rate will have better speech quality. G.723.1 uses a 30 ms speech frame (240 samples for a frame for 8 kHz sampling rate). The switch between the two bit rates can be carried out at any frame boundary (30 ms). Each 30 ms speech frame is divided into four subframes (each 7.5 ms). The look-ahead of G.723.1 is 7.5 ms (one subframe length), this results in an algorithmic delay of 37.5 ms. The 10th order LPC analysis is applied for each subframe. Both open-loop and close-loop pitch period estimation/prediction are performed for every two subframes (120 samples). Two different excitation methods are used for the high and the low bit rate codecs (one on ACELP and one on MP-MLQ).

2.4.6 ETSI GSM

GSM (Global System for Mobile Communications), is a speech codec standard specified by ETSI for Pan-European Digital Mobile Radio Systems (2G mobile communications). GSM Rec 06.10 (1991) [6] defines full-rate GSM operating at 13 kb/s and is based on Regular Pulse Excitation/Long Term Prediction (RPE/LTP) Linear Prediction coder. The speech frame length is 20 ms (160 samples at 8 kHz sampling rate) and the encoded block is 260 bits. Each speech frame is divided into four subframes (5 ms each). LP analysis is carried out for each speech frame (20 ms). The Regular pulse excitation (RPE) analysis is based on the subframe, whereas Long Term Prediction (LTP) is based on the whole speech frame. The encoded block of 260 bits contains the parameters from LPC filter, RPE and LTP analysis. Detailed bits allocation can be found from [6].

GSM half rate (HR), known as GSM 06.20, was defined by ETSI in 1999 [7]. This codec is based on VSELP (Vector-Sum Excited Linear Prediction) operating at 5.6 kb/s. It uses vector-sum excited linear prediction codebook with each codebook vector is formed up by a linear combination of fixed basis vectors. The speech frame length is 20 ms and is divided into four subframes (5 ms each). The LPC filter is 10th order. The encoded block length is 112 bits containing parameters for LPC filter, codecbook indices and gain.

Enhanced Full Rate (EFR) GSM, known as GSM 06.60, was defined by ETSI in 2000 [9]. It is based on ACELP (Algebraic CELP) and operates at 12.2 kb/s, same as the highest rate in AMR (see the next section).

2.4.7 ETSI AMR

Adaptive Multi Rate (AMR) narrowband speech codec, based on ACELP (Algebraic Code Excited Linear Prediction), was defined by ETSI, Special Mobile Group (SMG), in 2000 [8]. It has been chosen by 3GPP (the 3rd Generation Partnership Project) as the mandatory codec for Universal Mobile Telecom Systems (UMTS) or the 3rd Generation Mobile Networks (3G). AMR is a multi-mode codec with 8 narrowband modes for bit rates of 4.75, 5.15, 5.9, 6.7, 7.4, 7.95, 10.2 and 12.2 kb/s. The speech frame length is 20 ms (160 speech samples at 8000 sampling rate). Mode switching can occur at the boundary of each speech frame (20 ms). For a speech frame, the speech signal is analysed in order to obtain the parameters of 10th LP coefficients, adaptive and fixed codebooks' indices and gains. The LP analysis is carried out twice for 12.2 kb/s AMR mode and only once for all other modes. Each 20 ms speech frame is divided into four subframes (5 ms each). Pitch analysis is based on every subframe, and adaptive and fixed codebooks parameters are transmitted for every subframe. The bit numbers for encoded blocks for the 8 modes from 4.75 to 12.2 kb/s are 95, 103, 118, 134, 148, 159, 204 and 244 bits, respectively. Here you can calculate and check the relevant bit rate based on bit numbers in an encoded block. For example, for 244 bits over a 20 ms speech frame, the bit rate is 12.2 kb/s (244 bits/20 ms = 12.2 kb/s). For detailed bit allocation for 8 modes AMR, the reader can follow the AMR ETSI specification [8].

The flexibility on bandwidth requirements and tolerance in bit errors for the AMR codec are not only beneficial for wireless links, but are also desirable for VoIP applications, e.g. in QoE management for mobile VoIP applications using automatic AMR bit rate adaptation in response to network congestions [27].

UNIVERSITAS BRAWIJAYA

2.4.8 IETF's iLBC

iLBC (Internet Low Bit Rate Codec), an open source speech codec, was proposed by Andersen et al. in 2002 [3] at Global IP Sound (GIP, acquired by Google Inc in 2011)[1] and was defined in IETF RFC 3951 [2] in 2004. It was aimed for Internet applications with robustness to packet loss. Based on block independent CELP (frame-independent long-term prediction), it can overcome the error propagation problem occurred in traditional CELP codecs and achieve better voice quality under packet loss conditions (when compared with other CELP codecs, such as G.729, G.723.1 and AMR) [29]. The frame length for iLBC is 20 ms (15.2 kb/s, with 304 bits per coded block) or 30 ms (13.33 kb/s, with 400 bits per coded block). Each speech frame is divided into four (for 20 ms frame with 160 samples) or six subframes (for 30 ms frame with 240 samples) with each subframe corresponding to 5 ms of speech (40 samples). For 30 ms frame, two LPC analyses are carried out, whereas for 20 ms frame, only one LPC analysis is required (both are based on 10th order LPC analysis). Codebook search is carried out for each subframe. Key techniques used in iLBC are LPC analysis, dynamic codebooks search, scalar quantization and perceptual weighting. The dynamic codebooks are used to code the residual signal only for the current speech block, without using the information based on previous speech frames, thus, eliminating the error propagation problem due to packet loss. This method enhances the packet loss concealment performance and results in better speech quality under packet loss conditions.

iLBC has been used in many VoIP tools such as Google Talk and Yahoo! Messenger.

2.4.9 Skype/IETF's SILK

SILK , the Super Wideband Audio Codec, is the recent codec used in Skype. It is designed and developed by Skype[2] as a speech codec for real-time and packet-based voice communications and was submitted to IETF in 2009 [32].

The SILK codec has four operating modes which are Narrowband (NB, 8 kHz sampling rate), Mediumband (MB, 8 or 12 kHz sampling rate), Wideband (WB, 8, 12 or 16 kHz sampling rate) and Super Wideband (SWB, 8, 12, 16 or 24 kHz sampling rate). Its basic speech frame is 20 ms (160 samples at 8 kHz sampling rate). The core Skype encoder uses similar AbS techniques which include pitch estimation (every 5 ms) and voicing decision (every 20 ms), short-term prediction (LPC) and long-term prediction (LTP), LTP scaling control, LPC transformed to LSF coefficients, together with noise shaping analysis.

The key scalability features of SILK codec can be categorized as following, as shown in Fig. 2.15.

[1] http://www.globalipsound.com

[2] https://developer.skype.com/silk/

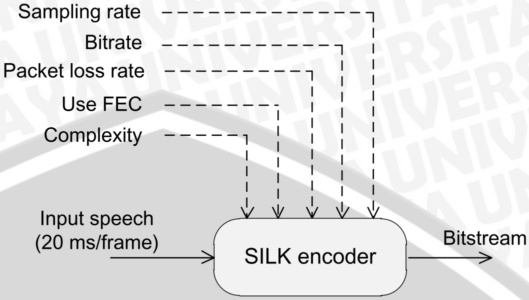


Fig. 2.15 Features for Skype codec

- *Sampling rate*: Skype supports the sampling rates of 8, 12, 16 or 24 kHz which can be updated in real-time to support NB, MB, WB and SWB voice applications.
- *Bit rate*: Skype supports bit rates from 6 to 40 kb/s. Bit rates can be adapted automatically according to network conditions.
- *Packet loss rate*: packet loss rate can be used as one of the control parameters for the Skype encoder to control its Forward Error Control (FEC) and packet loss concealment mechanisms.
- *Use FEC*: Forward Error Control (FEC) mechanism can be controlled whether to use or not depending on network conditions. Perceptually important packets for example, speech transition frames can be encoded at a lower bit rate and sent again over the channel. At the receiver side, if the main speech packet is lost, its lower bit rate packet can be used to recover the lost packet and to improve overall speech quality. However, FEC increases bandwidth usage as extra information is needed to be sent through the network.
- *Complexity*: There are three complexity settings provided in Skype which are high, medium and low. Appropriate complexity (CPU load) can be decided according to applications.

Other features such as changing packet size (e.g., one packet can contain 1, 2, up to 5 speech frames) and DTX (Discontinuous transmission) to stop transmitting packets in silence period are common features which can also be found for other speech codecs.

2.4.10 ITU-T G.722 ADPCM-WB

G.722 [12], defined by ITU-T in 1988, is a compression coding standard for 7 kHz audio at 16 kHz sampling rate. It is based on sub-band adaptive differential pulse code modulation (SB-ADPCM) with bit rates of 64, 56 or 48 kb/s (depending on the operation mode). When encoder bit rate is 56 or 48 kb/s, an auxiliary data channel of 8 or 16 kb/s bit rate can be added during transmission to form up a 64 kb/s data channel.

At the SB-ADPCM encoder, the input audio signal (0 to 8 kHz) at 16 kHz sampling rate is split into two sub-band signals, each at 8 kHz sampling rate. The lower sub-band is for the signal from 0 to 4 kHz (same frequency range as narrowband speech), and the higher sub-band is for signal from 4 to 8 kHz. Each sub-band signal is encoded based on ADPCM, a similar structure as illustrated in Fig. 2.8 including adaptive quantiser and adaptive predictor. The lower sub-band ADPCM applies an adaptive 60-level non-uniform quantisation which requires 6 bits coding for each ADPCM codeword, resulting in 48 kb/s bit rate. The higher sub-band ADPCM applies 4-level non-uniform quantisation using 2 bits coding and can achieve 16 kb/s transmission bit rate. Overall, 64 kb/s can be achieved for the SB-ADPCM coding. In the mode for 56 or 48 kb/s operation, 30-level or 15-level non-uniform quantisation is used, instead of 60-level quantisation, which results in a 5 or 4 bits coding for each ADPCM codeword for the lower-subband. 4-level quantisation for higher sub-band remains the same.

Due to the nature of ADPCM sample-based coding, G.722 ADPCM-WB is suitable for both wideband speech and music coding.

2.4.11 ITU-T G.722.1 Transform Coding

G.722.1 [20], approved by ITU-T in 1999, is for 7 kHz audio coding at 24 and 32 kb/s for hands-free applications, for example, conferencing systems. It can be used for both speech and music. Encoder input signal is sampled at 16 kHz sampling rate. The coding algorithm is based on transform coding, named as Modulated Lapped Transform (MLT). The audio coding frame is 20 ms (320 samples at 16 kHz sampling rate), with 20 ms look-ahead, resulting in coding algorithmic delay of 40 ms. For each 20 ms audio frame, it is transformed to 320 MLT coefficients independently, and then coded to 480 and 640 bits for the bit rate of 24 and 32 kb/s, respectively. This independent coding of MLT coefficients for each frame has a better resilience to frame loss as no error propagation exists in this coding algorithm. This is why G.722.1 is suitable for use in a conferencing system with low frame loss. Bit rate change for this codec can occur at the boundary of any 20 ms frames.

In the latest version of G.722.1 [22] (2005), it defines both the 7 kHz audio coding mode (in the main body) and the 14 kHz coding mode (in Annex C). The new 14 kHz audio coding mode further expands audio's frequency range from 7 kHz to 14 kHz, with sampling rate doubled from 16 to 32 kHz and samples doubled from 320 to 640 for each audio frame. The bit rates supported by Annex C are 24, 32 and 48 kb/s. The produced speech by the 14 kHz coding algorithm is normally referred to as "High Definition Voice" or "HD" voice. This codec has been used in video conference phones, and video streaming systems by Polycom.[3]

[3] http://www.polycom.com

2.4.12 ETSI AMR-WB and ITU-T G.722.2

Adaptive Multi-Rate Wideband (AMR-WB) has been defined by both 3GPP [1] in Technical Specification TS 26.190 and ITU-T G.722.2 [21]. It is for wideband application (7 kHz bandwidth speech signals) with 16 kHz sampling rate. It operates at a wide range of bit rates from 6.6 to 23.85 kb/s (6.60, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05 or 23.85 kb/s) with bit rate change at any 20 ms frame boundary. Same as AMR, AMB-WB is based on ACELP coding technique, but uses a 16th order linear prediction (LP) filter (or short-term prediction filter), instead of 10th LP as used in AMR narrowband. AMR-WB can provide high quality voice and is suitable for applications such as combined speech and music, and multi-party conferences.

2.4.13 ITU-T G.719 Fullband Audio Coding

G.719, approved by ITU-T in 2008, is the latest ITU-T standard for Fullband (FB) audio coding [23] with bit rates ranging from 32 to 128 kb/s and audio frequencies up to 20 kHz. It is a joint effort from Polycom and Ericsson,[4] and is aimed for high quality speech, music and general audio transmission and suitable for conversational applications such as teleconferencing and telepresence. The 20 Hz–20 kHz frequency range covers the full human auditory bandwidth and represents all frequency human ear can hear. The sample rate at the input of encoder and the output of the decoder is 48 kHz. The frame size is 20 ms, with 20 ms look-ahead, resulting in an algorithmic delay of 40 ms. The compression technique is based on Transform Coding. The features such as adaptive time-resolution, adaptive bit allocation and lattice vector quantization, make it flexible and efficient for incorporating different input signal characteristics of audio and to be able to provide a variable bit rate from 32 to 128 kb/s. The encoder detects each 20 ms input signal frame and classifies it as either a stationary frame (such as speech) or a non-stationary frame (such as music) and applies different transform coding techniques accordingly. For a stationary frame, the modified Discrete Cosine Transform (DCT) is applied, whereas for a non-stationary frame, a higher temporal resolution transform (in the range of 5 ms) is used. The spectral coefficients after transform coding are grouped into different bands, then quantised using lattice-vector quantisation and coded based on different bit allocation strategies to achieve different transmission bit rates from 32 to 128 kb/s. G.719 can be applied for high-end video conferencing and telepresence applications to provide high definition (HD) voice, in accompany with a HD video stream.

[4] http://www.ericsson.com

UNIVERSITAS BRAWIJAYA

UNIVERSITAS BRAWIJAYA

2.4.14 Summary of Narrowband to Fullband Speech Codecs

In the previous sections, we have discussed key narrowband to fullband speech compression codecs standardised by ITU-T, ETSI and IETF. We now summarize them in Table 2.2 which includes each codec's basic information such as which standardisation body was involved, which year was standardised, codec type, Narrowband (NB), Wideband (WB), Super-wideband (SWB) or Fullband (FB), bit rate (kb/s), length of speech frame (ms), bits per sample/frame (coded bits per sample or per frame), look-ahead time (ms), and coding's algorithmic delay (ms). From this table, you should be able to see the historic development of speech compression coding standards (from 64 kb/s, 32 kb/s, 16 kb/s, 8 kb/s to 6.4/5.3 kb/s) for achieving high compression efficiency, the mobile codecs development from GSM to AMR for 2G and 3G applications, the development from single rate codec, dual-rate codec, 8-mode codec to variable rate codec for achieving high application flexibility, and the trend from narrowband (NB) codecs to wideband codecs (WB) for achieving high speech quality (even for High Definition voice). This development has made speech compression codecs more efficient and more flexible for many different applications including VoIP. In the table, the columns on coded bits per sample/frame and speech frame for each codec will help you to understand payload size and to calculate VoIP bandwidth which will be covered in Chap. 4 on RTP transport protocol. The columns on look-ahead time and codec's algorithmic delay will help to understand codec delay and VoIP end-to-end delay, a key QoS metric, which will be discussed in detail in Chap. 6 on VoIP QoE.

It has to be mentioned that many VoIP phones (hardphones or softphones) have incorporated many different NB and even WB codecs. How to negotiate which codec to be used at each VoIP terminal and how to change the codec/mode/bit rate during a VoIP session on the fly will be discussed in Chap. 5 on SIP/SDP signalling.

2.5 Illustrative Worked Examples

2.5.1 Question 1

Determine the input and output data rates (in kb/s) and hence the compression ratio for a G.711 codec. Assume that the input speech signal is first sampled at 8 kHz and that each sample is then converted to 14-bit linear code before being compressed into 8-bit non-linear PCM by the G.711 codec.

SOLUTION: As the input speech signal is sampled at 8 kHz which means that there are 8000 samples per second. Then each sample is coded using 14-bit. Thus the input data rate is:

$$8000 \times 14 = 112{,}000 \text{ (bit/s)} = 112 \text{ (kb/s)}$$

Table 2.2 Summary of NB, WB, SWB and FB speech codecs

Codec	Standard Body/Year	Type	NB or WB or FB	Bit rate (kb/s)	Speech frame (ms)	Bits per sample/ frame	Look-ahead (ms)	Algor. delay (ms)
G.711	ITU/1972	PCM	NB	64	0.125	8	0	0.125
G.726	ITU/1990	ADPCM	NB	40	0.125	5	0	0.125
				32		4		
				24		3		
				16		2		
G.728	ITU/1992	LD-CELP	NB	16	0.625	10	0	0.625
G.729	ITU/1996	CS-ACELP	NB	8	10	80	5	15
G.723.1	ITU/1996	ACELP	NB	5.3	30	159	7.5	37.5
		MP-MLQ	NB	6.3		189		
GSM	ETSI/1991	(FR) RPE-LTP	NB	13	20	260	0	20
	ETSI/1999	(HR) VSELP	NB	5.6		112	0	20
	ETSI/2000	(EFR) ACELP	NB	12.2		244	0	20
AMR	ETSI/2000	ACELP	NB	4.75	20	95	5	25
				5.15		103		
				5.9		118		
				6.7		134		
				7.4		148		
				7.95		159		
				10.2		204		
				12.2		244	0	20
iLBC	IETF/2004	CELP	NB	15.2	20	304	0	20
				13.33	30	400		30
G.711.1	ITU/2008	PCM-WB (MDCT)	NB/WB	64	5	320	5	11.875
				80		400		
				96		480		
G.722	ITU/1988	SB-ADPCM	WB	64	0.125	8	0	0.125
				56		7		
				48		6		
G.722.1	ITU/1999	Transform Coding	WB	24	20	480	20	40
				32		640		
	ITU/2005		SWB	24/32/48		480–960		
G.719	ITU/2008	Transform Coding	FB	32–128	20	640–2560	20	40
AMR-WB (G.722.2)	ETSI/ITU /2003	ACELP	WB	6.6–23.85	20	132–477	0	20
SILK	IETF/2009	CELP	WB	6–40	20	120–800	0	20

UNIVERSITAS BRAWIJAYA

For the output data, each sample is coded using 8-bit, thus the output data rate is:

$$8000 \times 8 = 64{,}000 \text{ (bit/s)} = 64 \text{ (kb/s)}$$

The compression ratio for a G.711 codec is:

$$112/64 = 1.75$$

2.5.2 Question 2

The G.726 is the ITU-T standard codec based on ADPCM. Assume the codec's input speech signal is 16-bit linear PCM and the sampling rate is 8 kHz. The output of the G.726 ADPCM codec can operate at four possible data rates: 40 kb/s, 32 kb/s, 24 kb/s and 16 kb/s. Explain how these rates are obtained and what the compression ratios are when compared with 64 kb/s PCM.

SOLUTION: ADPCM codec uses speech signal waveform correlation to compress speech. For the ADPCM encoder, only the difference signal between the input PCM linear signal and the predicted signal is quantised and coded. The dynamic range of the difference signal is much smaller than that of the input PCM speech signal, thus less quantisation levels and coding bits are needed for the ADPCM coding.

For 40 kb/s ADPCM, let's assume the number of bits needed to code each quantised difference signal is x, then we have:

$$40 \text{ kb/s} = 8000 \text{ (samples/s)} \times x \text{ (bits/sample)}$$
$$x = 40 \times 1000/8000 = 5 \text{ (bits)}$$

Thus, using 5 bits to code each quantised difference signal will create an ADPCM bit steam operating at 40 kb/s.

Similarly, for 32, 24 and 16 kb/s, the required bits for each quantised difference signal is 4 bits, 3 bits and 2 bits, respectively. The lower the coding bits, the higher the quantisation error, thus, the lower the speech quality.

For the compression ratio for 40 kb/s ADPCM when compared with 64 kb/s PCM, it is $64/40 = 1.6$.

For 32, 24 and 16 kb/s ADPCM, the compression ratio is 2, 2.67, 4, respectively.

2.5.3 Question 3

For the G.723.1 codec, it is known that the transmission bit rates can operate at either 5.3 or 6.3 kb/s. What is the frame size for G.723.1 codec? How many speech samples are there within one speech frame? Determine the number of parameters bits coded for the G.723.1 encoding.

SOLUTION: For the G.723.1 codec, the frame size is 30 ms. As G.723.1 is narrowband codec, the sampling rate is 8 kHz. The number of speech samples in a speech frame is:

$$30\ (\text{ms}) \times 8000\ (\text{samples/s}) = 240\ (\text{samples})$$

So, there are 240 speech samples within one speech frame.

For 5.3 kb/s G.723.1, the number of parameters bits used is:

$$30\ (\text{ms}) \times 5.3\ (\text{kb/s}) = 159\ (\text{bits})$$

For 6.3 kb/s G.723.1, the number of parameters bits used is:

$$30\ (\text{ms}) \times 6.3\ (\text{kb/s}) = 189\ (\text{bits})$$

2.6 Summary

In this chapter, we discussed speech/audio compression techniques and summarised narrowband, wideband and fullband speech/audio compression standards from ITU-T, ETSI and IETF. We focused mainly on narrowband speech compression, but covered some wideband and the latest fullband speech/audio compression standards. We started the chapter from some fundamental concepts of speech, including speech signal digitisation (sampling, quantisation and coding), speech signal characteristics for voiced and unvoiced speech, and speech signal presentation including speech waveform and speech spectrum. We then presented three key speech compression techniques which are waveform compression, parametric compression and hybrid compression. For waveform compression, we mainly explained ADPCM which is widely used for both narrowband and wideband speech/audio compression. For parametric compression, we started from the speech production model and then explained the concept of parametric compression techniques, such as LPC-10. For hybrid compression, we started from the problems with waveform and parametric compression techniques, the need to develop high speech quality and high compression ratio speech codecs, and then discussed the revolutionary Analysis-by-Synthesis (AbS) and CELP (Code Excited Linear Prediction) approach. We also listed out major CELP variants used in mobile, satellite and secure communications systems.

In this chapter, we summarised major speech/audio compression standards for narrowband, wideband and fullband speech/audio compression coding from ITU-T, ETSI and IETF. We covered narrowband codecs including G.711, G.726, G.728, G.729, G.723.1, GSM, AMR and iLBC; wideband codecs including G.722, G.722.1, G.722.2/AMR-WB; and fullband codec (i.e., G.719). We explained the historic development of these codecs and the trend from narrowband, wideband to fullband speech/audio compression to provide high fidelity or “High Definition Voice” quality. Their applications cover VoIP, video call, video conferencing and telepresence.

This chapter, together with the next chapter on video compression, form the basis for other chapters in the book. We illustrated the concepts such as speech codec type, speech frame size, sampling rate, bit rate and coded bits for each speech frame. This will help you to understand the payload size and to calculate VoIP bandwidth which will be covered in Chap. 4 on the RTP transport protocol. The codec compression and algorithmic delay also affect overall VoIP quality which will be further discussed in Chap. 6 on VoIP QoE. How to negotiate and decide which codec to be used for a VoIP session and how to change the mode or codec type during a session will be discussed in Chap. 5 on the SIP/SDP signalling.

2.7 Problems

1. Describe the purpose of non-uniform quantisation.
2. What are the main differences between vocoder and hybrid coding?
3. What is the normal bit rate range for waveform speech codecs, vocoder and hybrid speech codecs?
4. From human speech production mechanism, explain the difference between ‘unvoiced’ speech and ‘voiced’ speech.
5. What is the LPC filter order used in modern codecs such as G.729, G.723.1 and AMR?
6. Based on the human speech production mechanism, illustrate and explain briefly the LPC model. What are the main reasons for LPC model achieving low bit rate, but with low speech quality (especially on fidelity and natureness of the speech). In which application areas, LPC-based vocoder is still used today?
7. What is the basis for speech compression for hybrid and parametric codings using 10 to 30 ms speech frames?
8. Describe the bit rate or bit rate ranges used in the following codecs, G.711, G.726, G.729.1, G.723.1 and AMR.
9. In an ADPCM system, it is known that 62-level non-linear quantiser is used. How many bits are required to code each ADPCM codeword (i.e. prediction error signal)? What is the bit rate of this ADPCM system?
10. Explain the reasons for CELP-based codecs to achieve better speech quality when compared with LPC vocoder.
11. Determine the compression ratio for LPC-10 at 2.4 kb/s when compared with G.711 PCM. Determine the number of parameters bits coded in LPC-10 for one speech frame with 180 speech samples.
12. Which ITU-T speech coding standard is the first ITU-T standard based on CELP technique? What is the size of codebook used in this standard? How many bits are required to transmit the codebook index? How do you calculate its bit rate?
13. Based on the bit rate sequence from high to low, list out for the following codecs: LPC-10, G.723.1 ACELP, G.711 PCM, G.728 LD-CELP, FR-GSM, G.729 CS-ACELP.

14. For G.722 ADPCM-WB, what is the sampling rate for signal at the input of the encoder? What is the sampling rate for the input at each sub-band ADPCM block?
15. Describe the speech/audio frequency range and sampling rate for narrowband, wideband, super-wideband and fullband speech/audio compression coding.
16. Describe the differences between G.711 and G.711.1.

References

1. 3GPP (2011) Adaptive Multi-Rate—Wideband (AMR-WB) speech codec, transcoding functions (Release 10). 3GPP TS 26.190 V10.0.0
2. Andersen S, Duric A, et al (2004) Internet Low Bit rate Codec (iLBC). IETF RFC 3951
3. Andersen SV, Kleijn WB, Hagen R, Linden J, Murthi MN, Skoglund J (2002) iLBC—a linear predictive coder with robustness to packet losses. In: Proceedings of IEEE 2002 workshop on speech coding, Tsukuba Ibaraki, Japan, pp 23–25
4. Atal BS, Hanauer SL (1971) Speech analysis and synthesis by linear prediction. J Acoust Soc Am 50:637–655
5. Atal BS, Remde JR (1982) A new model of LPC excitation for producing natural-sounding speech at low bit rates. In: Proc IEEE int conf acoust speech, signal processing, pp 614–617
6. ETSI (1991) GSM full rate speech transcoding. GSM Rec 06.10
7. ETSI (1999) Digital cellular telecommunications system (Phase 2+); half rate speech; half rate speech transcoding. ETSI-EN-300-969 V6.0.1
8. ETSI (2000) Digital cellular telecommunications system (Phase 2+); Adaptive Multi-Rate (AMR) speech transcoding. ETSI-EN-301-704 V7.2.1
9. ETSI (2000) digital cellular telecommunications system (phase 2+); Enhanced Full Rate (EFR) speech transcoding. ETSI-EN-300-726 V8.0.1
10. Griffin DW, Lim JS (1988) Multiband excitation vocoder. IEEE Trans Acoust Speech Signal Process 36:1223–1235
11. ITU-T (1988) 32 kbit/s adaptive differential pulse code modulation (ADPCM). ITU-T G.721
12. ITU-T (1988) 7 kHz audio-coding within 64 kbit/s. ITU-T Recommendation G.722
13. ITU-T (1988) Extensions of Recommendation G.721 adaptive differential pulse code modulation to 24 and 40 kbit/s for digital circuit multiplication equipment application. ITU-T G.723
14. ITU-T (1988) Pulse code modulation (PCM) of voice frequencies. ITU-T G.711
15. ITU-T (1990) 40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM). ITU-T G.726
16. ITU-T (1992) Coding of speech at 16 kbit/s using low-delay code excited linear prediction. ITU-T G.728
17. ITU-T (1996) Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP). ITU-T G.729
18. ITU-T (1996) Dual rate speech coder for multimedia communication transmitting at 5.3 and 6.3 kbit/s. ITU-T Recommendation G.723.1
19. ITU-T (1999) G.711: a high quality low-complexity algorithm for packet loss concealment with G.711. ITU-T G.711 Appendix I
20. ITU-T (1999) Coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss. ITU-T Recommendation G.722.1
21. ITU-T (2003) Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB). ITU-T Recommendation G.722.2
22. ITU-T (2005) Low-complexity coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss. ITU-T Recommendation G.722.1

23. ITU-T (2008) Low-complexity, full-band audio coding for high-quality, conversational applications. ITU-T Recommendation G.719. http://www.itu.int/rec/T-REC-G.719-200806-I
24. ITU-T (2008) Wideband embedded extension for G.711 pulse code modulation. ITU-T G.711.1
25. Jayant NS (1974) Digital coding of speech waveforms: PCM, DPCM and DM quantizers. Proc IEEE 62:611–632
26. Kondoz AM (2004) Digital speech: coding for low bit rate communication systems, 2nd ed. Wiley, New York. ISBN:0-470-87008-7
27. Mkwawa IH, Jammeh E, Sun L, Ifeachor E (2010) Feedback-free early VoIP quality adaptation scheme in next generation networks. In: Proceedings of IEEE Globecom 2010, Miami, Florida
28. Schroeder MR (1966) Vocoders: analysis and synthesis of speech. Proc IEEE 54:720–734
29. Sun L, Ifeachor E (2006) Voice quality prediction models and their applications in VoIP networks. IEEE Trans Multimed 8:809–820
30. TIA/EIA (1997) Enhanced Variable Rate Codec (EVRC). TIA-EIA-IS-127. http://www.3gpp2.org/public_html/specs/C.S0014-0_v1.0_revised.pdf
31. Tremain TE (1982) The government standard linear predictive coding algorithm: LPC-10. Speech Technol Mag 40–49
32. Vos K, Jensen S, et al (2009) SILK speech codec. IETF RFC draft-vos-silk-00

repository.ub.ac.id

Springer

http://www.springer.com/978-1-4471-4904-0

Guide to Voice and Video over IP
For Fixed and Mobile Networks
Sun, L.; Mkwawa, I.-H.; Jammeh, E.; Ifeachor, E.
2013, XII, 272 p., Hardcover

ISBN: 978-1-4471-4904-0

UNIVERSITAS BRAWIJAYA