Chapter 1

Speech Signals and an Introduction to Speech Coding

1.1 Motivation of Speech Compression

According to the lessons of information theory, the minimum bitrate at which the condition of distortionless transmission of any source signal is possible is determined by the entropy of the speech source message. Note, however, that in practical terms the source rate corresponding to the entropy is only asymptotically achievable as the encoding memory length or delay tends to infinity. Any further compression is associated with information loss or coding distortion. Many practical source compression techniques employ so-called 'lossy' coding, which typically guarantees further bitrate economy at the cost of nearly imperceptible speech, audio, video, etc, source representation degradation.

Note that the optimum Shannonian source encoder generates a perfectly uncorrelated source coded stream, where all the source redundancy has been removed, therefore the encoded source symbols – which are in most practical cases constituted by binary bits – are independent and each one has the same significance. Having the same significance implies that the corruption of any of the source encoded symbols results in identical source signal distortion over imperfect channels.

Under these conditions, according to Shannon's fundamental work [57–59], best protection against transmission errors is achieved, if source and channel coding are treated as separate entities. When using a block code of length N channel coded symbols in order to encode K source symbols with a coding rate of $R = K/N$, the symbol error rate can be rendered arbitrarily low if N tends to infinity and hence the coding rate to zero. This condition also implies an infinite coding delay. Based on the above considerations and on the assumption of additive white Gaussian noise (AWGN), channel source and channel coding have historically been separately optimised.

In designing a telecommunications system one of the most salient parameters is the number of subscribers that can be accommodated by the transmission media utilised. Whether it is a time division multiplex (TDM) or a frequency division multiplex (FDM) system,

Voice and Audio Compression for Wireless Communications Second Edition
L. Hanzo, F. C. A. Somerville and J. P. Woodard

whether it is analog or digital, the number of subscribers is limited by the channel capacity needed for one speech channel. If the channel capacity demand of the speech channels is halved, the total number of subscribers can be doubled. This gain becomes particularly important in applications like power- and band-limited satellite or mobile radio channels, where the urging demand for free channels overshadows the inevitable cost constraints imposed by a more complex low bitrate speech codec. In the framework of the basic limitations of state-of-the-art very large scale integrated (VLSI) technology the design of a speech codec is based on an optimum trade-off between lowest bitrate and highest quality, at the price of lowest complexity, cost and system delay. The analysis of these contradictory factors pervades all our forthcoming discussions.

1.2 Basic Characterisation of Speech Signals

In contrast to the so-called deterministic signals – random signals, such as speech, music, video, etc – information signals cannot be described with the help of analytical formulae. They are typically characterised with the help of a variety of statistical characteristics. The so-called power spectral density (PSD), auto-correlation function (ACF), cumulative distribution function (CDF) and probability density function (PDF) are some of the most frequent ones invoked, which will be exemplified during our further discourse.

Transmitting speech information is one of the fundamental aims of telecommunications and in this book we mainly concentrate on the efficient encoding of speech signals. The human vocal apparatus has been portrayed in many books dealing with human anatomy and has also been treated in references dealing with speech processing [5, 17, 22]. Hence, here we dispense with its portrayal and simply note that human speech is generated by emitting sound pressure waves, radiated primarily from the lips, although significant energy emanates in the case of some sounds also from the nostrils, throat, etc.

The air compressed by the lungs excites the vocal cords in two typical modes. Namely, when generating *voiced sounds*, the vocal cords vibrate and generate a high-energy quasi-periodic speech wave form, while in the case of lower energy *unvoiced sounds* the vocal cords do not participate in the voice production and the source behaves similar to a noise generator. In a somewhat simplistic approach the excitation signal denoted by $E(z)$ is then filtered through the vocal apparatus, which behaves like a spectral shaping filter with a transfer function of $H(z) = 1/A(z)$ that is constituted by the spectral shaping action of the glottis, which is defined as the opening between the vocal folds. Further spectral shaping is carried out by the vocal tract, lip radiation characteristics, etc. This simplified speech production model is shown in Figure 1.1.

Typical voiced and unvoiced speech waveform segments are shown in Figures 1.2 and 1.3, respectively, along with their corresponding power densities. Clearly, the unvoiced segment appears to have a significantly lower magnitude, which is also reflected by its PSD. Observe in Figure 1.3 that the low-energy, noise-like unvoiced signal has a rather flat PSD, which is similar to that of white noise. In general, the more flat the signal's spectrum, the more unpredictable it becomes and hence it is not amenable to signal compression or redundancy removal.

In contrast, the voiced segment shown in Figure 1.2 is quasi-periodic in the time-domain and it has an approximately 80-sample periodicity, identified by the positions of the

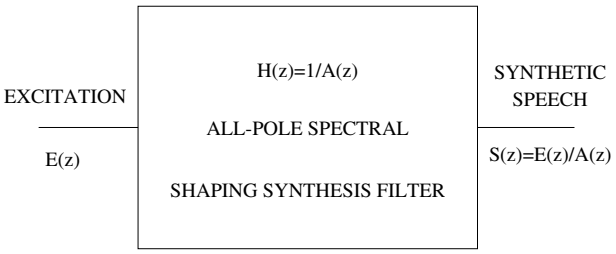

Figure 1.1: Linearly separable speech source model.

largest time-domain signal peaks, which corresponds to $80 \times 125\ \mu\text{s} = 10\,\text{ms}$. This interval is referred to as the *pitch period* and it is also often expressed in terms of the *pitch frequency* p, which is in this example $p = 1/10\ \text{ms} = 100\,\text{Hz}$. In the case of male speakers the typical pitch frequency range is between 40 and 120 Hz, while for females it can be as high as 300–400 Hz. Observe, furthermore, that within each pitch period there is a gradually decaying oscillation, which is associated with the excitation and gradually decaying vibration of the vocal cords.

A perfectly periodic time-domain signal would have a line spectrum, but since the voiced speech signal is quasi-periodic with a frequency of p – rather than being perfectly periodic – its spectrum in Figure 1.2 exhibits somewhat widened but distinctive spectral needles at frequencies of $n \times p$, rather than being perfectly periodic. As a second phenomenon, we can also observe three, sometimes four spectral envelope peaks. In our voiced spectrum of Figure 1.2 these so-called *formant frequencies* are observable around 500, 1500 and 2700 Hz and they are the manifestation of the resonances of the vocal tract at these frequencies. In contrast, the unvoiced segment of Figure 1.3 does not have a formant structure, it rather has a more dominant high-pass nature, exhibiting a peak around 2500 Hz. Observe, furthermore, that its energy is much lower than that of the voiced segment of Figure 1.2.

It is equally instructive to study the ACF of voiced and unvoiced segments, which are portrayed on an expanded scale in Figures 1.4 and 1.5, respectively. The voiced ACF shows a set of periodic peaks at displacements of about 20 samples, corresponding to $20 \times 125\ \mu\text{s} = 2.5\,\text{ms}$, which coincides with the positive quasi-periodic time-domain segments. Following four monotonously decaying peaks, there is a more dominant one around a displacement of 80 samples, which indicates the pitch periodicity. The periodic nature of the ACF can therefore be, for example, exploited to detect and measure the pitch periodicity in a range of applications, such as speech codecs, voice activity detectors, etc. Observe, however, that the first peak at a displacement of 20 samples is about as high as the one near 80 and hence a reliable pitch detector has to attempt to identify and rank all these peaks in order of prominence, exploiting also the *a priori* knowledge as to the expected range of pitch frequencies. Recall, furthermore, that, according to the Wiener–Khintshin Theorem, the ACF is the Fourier transform pair of the PSD of Figure 1.2.

By contrast, the unvoiced segment of Figure 1.5 has a much more rapidly decaying ACF, indicating no inherent correlation between adjacent samples and no long-term periodicity. Clearly, its sinc-function-like ACF is akin to that of band-limited white noise. The wider ACF of the voiced segment suggests predictability over a time-interval of some 3–400 μs.

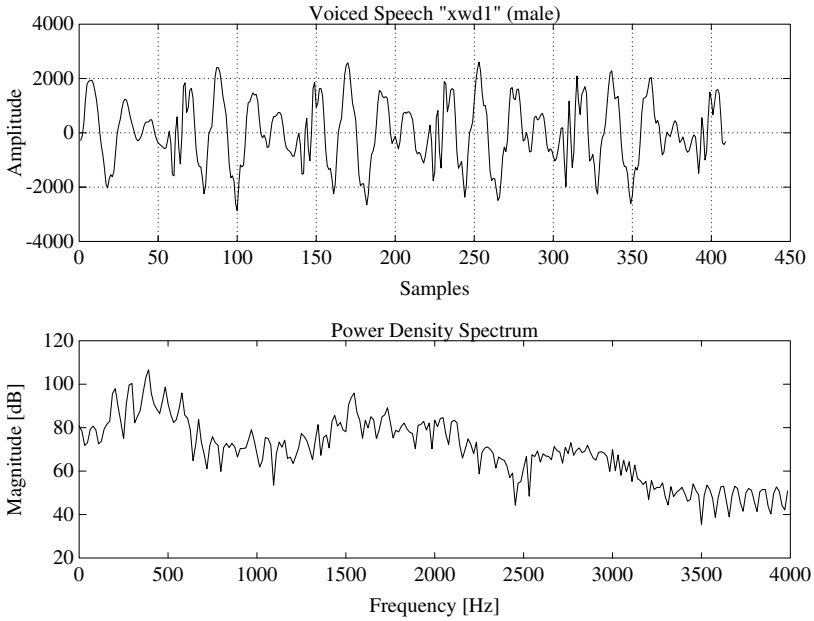

Figure 1.2: Typical voiced speech segment and its PSD for a male speaker.

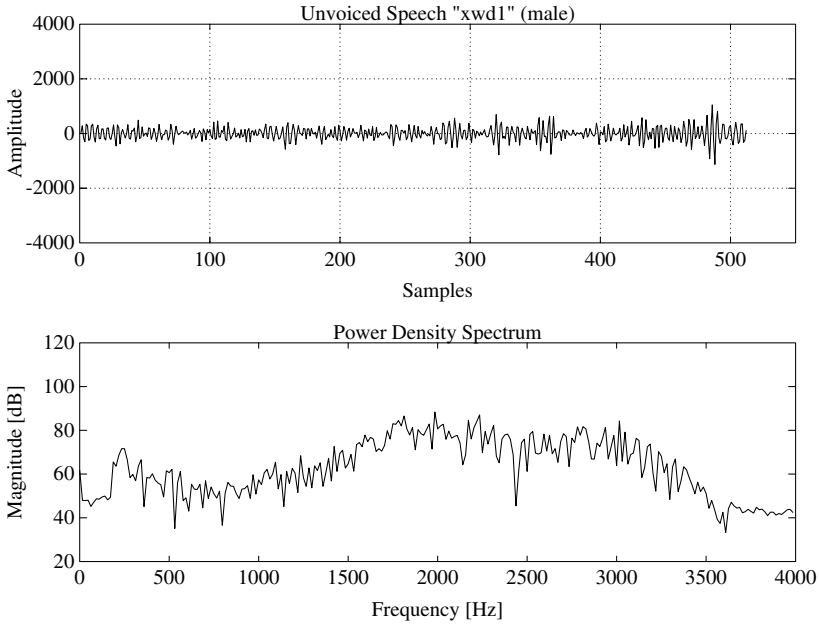

Figure 1.3: Typical unvoiced speech segment and its PSD for a male speaker.

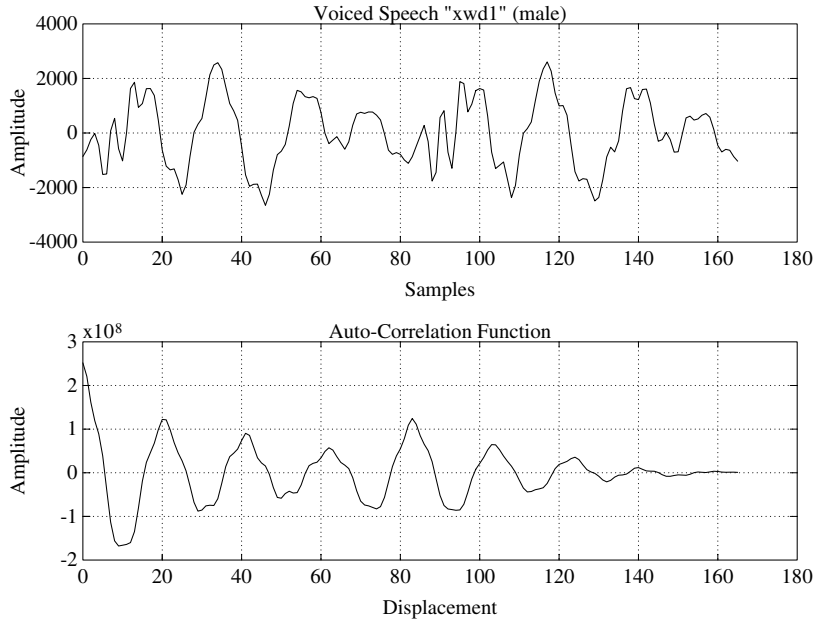

Figure 1.4: Typical voiced speech segment and its ACF for a male speaker.

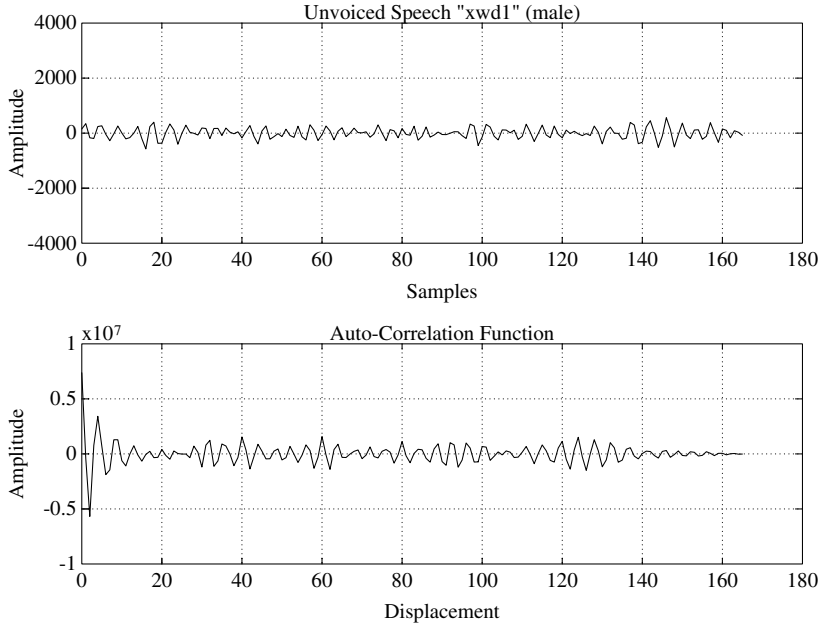

Figure 1.5: Typical unvoiced speech segment and its ACF for a male speaker.

Since human speech is voiced for about 2/3 of the time, redundancy can be removed from it using predictive techniques in order to reduce the bitrate required for its transmission.

Having characterised the basic features of speech signals, let us now focus our attention on their digital encoding. Intuitively it can be expected that the higher the encoder/decoder (codec) complexity, the lower the achievable bitrate and the higher the encoding delay. This is because more redundancy can be removed by considering longer speech segments and employing more sophisticated signal processing techniques.

1.3 Classification of Speech Codecs

Speech coding methods can be broadly categorised as *waveform coding, vocoding* and *hybrid coding*. The principle of these codecs will be considered later in this chapter, while the most prominent subclass of hybrid codecs referred to as analysis-by-synthesis schemes will be revisited in detail in Chapter 3 and will feature throughout this book. Their basic differences become explicit in Figure 1.6, where the speech quality versus bitrate performance of these codec families is portrayed in qualitative terms. The bitrate is plotted on a logarithmic axis and the speech quality classes 'poor to excellent' broadly correspond to the so-called five-point MOS scale values of 2–5 defined by the International Telegraph and Telephone Consultative Committee (CCITT), which was recently renamed as the International Telecommunications Union (ITU). We will refer to this diagram and to these codec families during our further discourse in order to allocate various codecs on this plane. Hence, here only a rudimentary interpretation is offered.

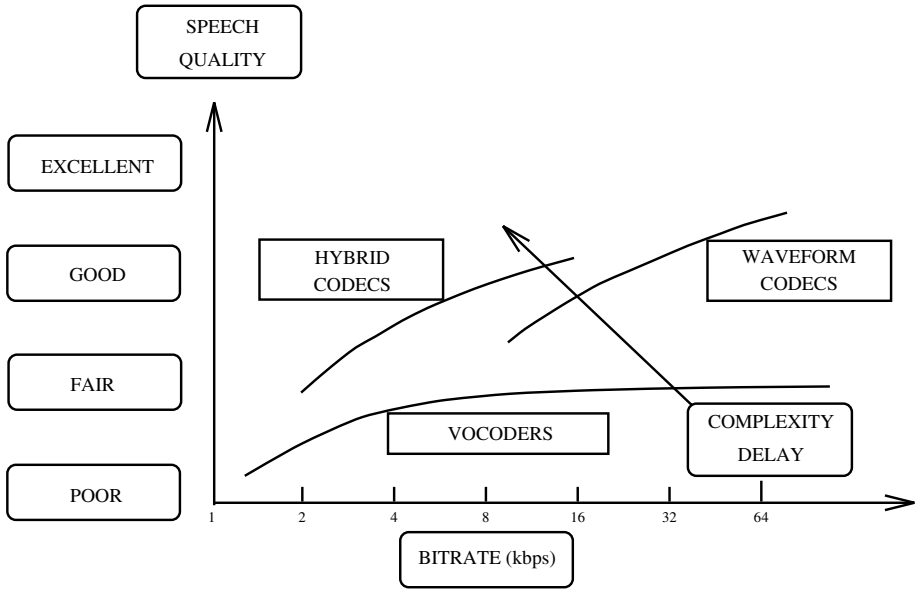

Figure 1.6: Speech quality versus bitrate classification of speech codecs.

1.3.1 Waveform Coding [10]

Waveform codecs have been comprehensively characterised by Jayant and Noll [10] and hence the spirit of virtually all treatises on the subject follows their approach. Our discourse in this section is no exception.

In general, waveform codecs are designed to be signal independent. They are designed to map the input waveform of the encoder into a facsimile-like replica of it at the output of the decoder. Due to this advantageous property they can also encode secondary types of information such as signalling tones, voice band data, or even music. Naturally, because of this signal 'transparency', their coding efficiency is usually quite modest. The coding efficiency can be improved by exploiting some statistical signal properties, if the codec parameters are optimised for the most likely categories of input signals, while still maintaining good quality for other types of signals as well. The waveform codecs can be further subdivided into time-domain waveform codecs and frequency-domain waveform codecs.

1.3.1.1 Time-domain Waveform Coding

The most well-known representative of signal independent time-domain waveform coding is the so-called A-law companded pulse code modulation (PCM) scheme, which has been standardised by the CCITT – now known as the International Telecommunications Union (ITU) – at 64 kbps, using nonlinear companding characteristics to result in near-constant signal-to-noise ratio (SNR) over the total input dynamic range. More explicitly, the nonlinear companding compresses large input samples and expands small ones. Upon quantising this companded signal, large input samples will tolerate higher quantisation noise than small samples.

Also well-known is the 32 kbps adaptive differential PCM (ADPCM) scheme standardised in ITU Recommendation G.721 – which will be the topic of Section 2.7 – and the so-called adaptive delta modulation (ADM) arrangement, where usually the most recent signal sample or a linear combination of the last few samples is used to form an estimate of the current one. Then their difference signal, the so-called prediction residual, is computed and encoded usually with a reduced number of bits, since it has a lower variance than the incoming signal. This estimation process is actually linear prediction with fixed coefficients. However, owing to the non-stationary statistics of speech, a fixed predictor cannot consistently characterise the changing spectral envelope of speech signals. Adaptive predictive coding (APC) schemes utilise, in general, two different time-varying predictors to describe speech signals more accurately. Namely, a so-called short-term predictor (STP) and a long-term predictor (LTP). During our further discourse we will show that the STP is utilised to model the speech spectral envelope, while the LTP is employed in order to model the line-spectrum-like fine-structure representing the voicing information due to quasi-periodic voiced speech.

All in all, time-domain waveform codecs treat the speech signal to be encoded as a full-band signal and attempt to map it into as close a replica of the input as possible. The difference amongst various coding schemes is in their degree and way of using prediction to reduce the variance of the signal to be encoded, so as to reduce the number of bits necessary to represent it.

1.3.1.2 Frequency-domain Waveform Coding

In frequency-domain waveform codecs the input signal undergoes a more or less accurate short-time spectral analysis. Clearly, the signal is split into a number of sub-bands, and the individual sub-band signals are then encoded by using different numbers of bits, to obey rate-distortion theory on the basis of their prominence. The various methods differ in their accuracies of spectral resolution and in the bit-allocation principle (fixed, adaptive, semi-adaptive). Two well-known representatives of this class are sub-band coding (SBC) and adaptive transform coding (ATC).

1.3.2 Vocoders

The philosophy of vocoders is based on our *a priori* knowledge about the way the speech signal to be encoded was generated at the signal source by a speaker, which was portrayed in Figure 1.1. The air compressed by the lungs excites the vocal cords in two typical modes. Namely, when generating voiced sounds they vibrate and generate a quasi-periodic speech wave form, while in the case of lower-energy unvoiced sounds they do not participate in the voice production and the source behaves similar to a noise generator. The excitation signal denoted by $E(z)$ in the z-domain is then filtered through the vocal apparatus, which behaves like a spectral shaping filter with a transfer function of $H(z) = 1/A(z)$ that is constituted by the spectral shaping action of the glotti, vocal tract, lip radiation characteristics, etc.

Accordingly, instead of attempting to produce a close replica of the input signal at output of the decoder, the appropriate set of source parameters is found, in order to characterise the input signal sufficiently closely for a given duration of time. First a decision must be made as to whether the current speech segment to be encoded is voiced or unvoiced. Then the corresponding source parameters must be specified. In the case of voiced sounds the source parameter is the time between periodic vocal tract excitation pulses, which is often referred to as the pitch p. In the case of unvoiced sounds the variance or power of the noise-like excitation must be determined. These parameters are quantised and transmitted to the decoder in order to synthesise a replica of the original signal.

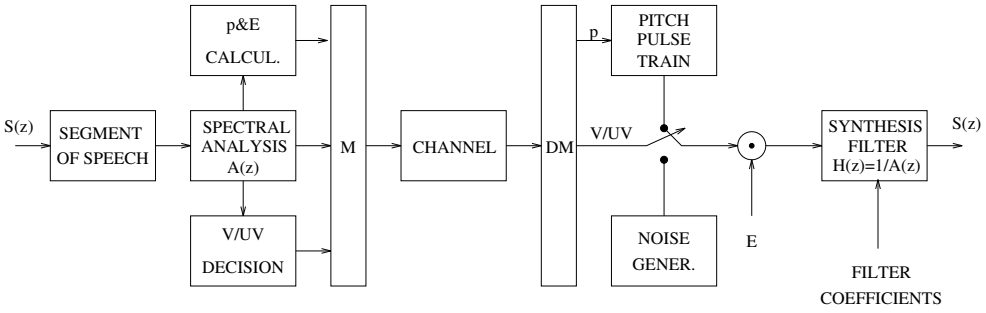

Figure 1.7: Vocoder schematic.

The simplest source codec arising from the above speech production model is depicted in Figure 1.7. The encoder is a simple speech analyser, determining the current source parameters. After initial speech segmentation it computes the linear predictive filter

coefficients a_i, $i = 1, \ldots, p$, which characterise the spectral shaping transfer function $H(z)$. A voiced/unvoiced decision is carried out and the corresponding pitch frequency and noise energy parameters are determined. These are then quantised, multiplexed and transmitted to the speech decoder, which is a speech synthesiser.

It is plausible that the associated speech quality of this type of system is predetermined by the adequacy of the source model, rather than by the accuracy of the quantisation of these parameters. This means that the speech quality of source codecs cannot simply be enhanced by increasing the accuracy of the quantisation, that is the bitrate, which is evidenced by the saturating MOS curve of Figure 1.6. Their speech quality is fundamentally limited by the fidelity of the model used. The main advantage of the above vocoding techniques is their low bitrate, with the penalty of relatively low, synthetic speech quality. A well-known representative of this class of vocoders is the 2400 bps American Military Standard LPC-10 codec.

In linear predictive coding (LPC) more complex excitation models are often used to describe the voice generating source. Once the vocal apparatus has been described with the help of its spectral domain transfer function $H(z)$, the central problem of coding is how to find the simplest adequate excitation for high-quality parametric speech representation. Strictly speaking this separable model represents a gross simplification of the vocal apparatus, but it provides the only practical approach to the problem. Vocoding techniques can also be categorised into frequency-domain and time domain sub-classes. However, frequency-domain vocoders are usually more effective than their time-domain counterparts.

1.3.3 Hybrid Coding

Hybrid coding methods constitute an attractive trade-off between waveform coding and source coding, both in terms of speech quality and transmission bitrate, although usually at the price of higher complexity. Every speech coding method, combining waveform and source coding methods in order to improve the speech quality and reduce the bitrate, falls into this broad category. However, adaptive predictive time-domain techniques used to describe the human spectral shaping tract combined with an accurate model of the excitation signal play the most prominent role in this category. The most important family of hybrid codecs, often referred to as *analysis-by-synthesis* (AbS) codecs, are ubiquitous at the time of writing and hence they will be treated in depth in a number of chapters after considering the conceptually more simple category of waveform codecs.

1.4 Waveform Coding [10]

1.4.1 Digitisation of Speech

The waveform coding of speech and video signals was comprehensively – in fact exhaustively – documented by Jayant and Noll in their classic monograph [10] and hence any treatise on the topic invariably follows a similar approach. Hence this section endeavours to provide a rudimentary overview of waveform coding following the spirit of Jayant and Noll [10]. In general, waveform codecs are designed to be signal independent. They are designed to map the input waveform of the encoder into a facsimile-like replica of it at the output of the decoder. Due to this advantageous property they can also encode secondary types of

information such as signalling tones, voice band data, or even music. Naturally, because of this transparency, their coding efficiency is usually quite modest. The coding efficiency can be improved by exploiting some statistical signal properties, if the codec parameters are optimised for the most likely categories of input signals, while still maintaining good quality for other types of signals.

The waveform codecs can be further subdivided into time-domain waveform codecs and frequency-domain waveform codecs. Let us initially consider the first category. The digitisation of analogue source signals, such as speech for example, requires the following steps, which are portrayed in Figure 1.8, while the corresponding waveforms are shown in Figure 1.9.

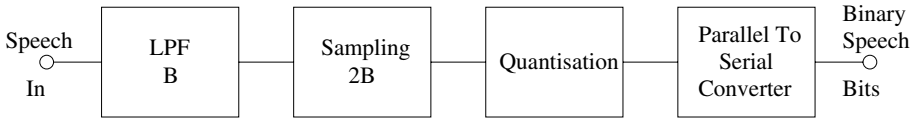

Figure 1.8: Digitisation of analogue speech signals.

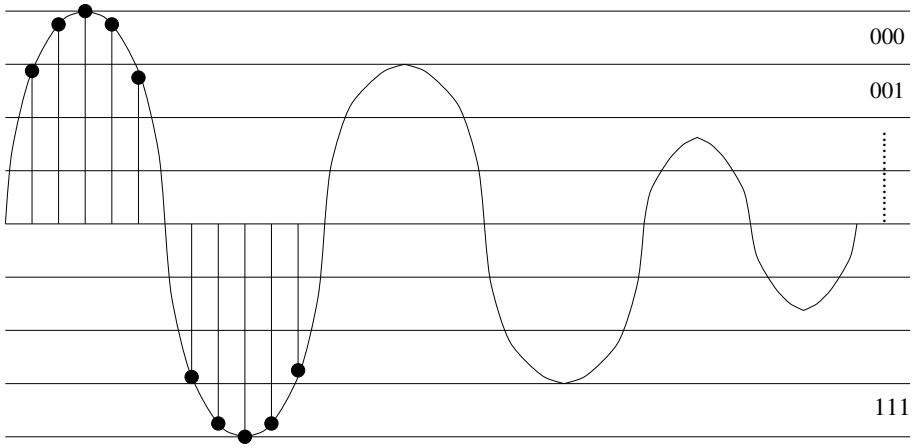

Figure 1.9: Sampled and quantised analogue speech signal.

- Anti-aliasing low-pass filtering (LPF) is necessary in order to bandlimit the signal to a bandwidth of B before sampling. In the case of speech signals about 1% of the energy resides above 4 kHz and only a negligible proportion above 7 kHz. Hence, so-called commentatory quality speech links, which are also often referred to as wideband speech systems, typically bandlimit the speech signal to 7–8 kHz. Conventional telephone systems usually employ a bandwidth limitation of 0.3–3.4 kHz, which results only in a minor speech degradation, hardly perceivable for the untrained listener.

- The bandlimited speech is sampled according to the Nyquist theorem, as seen in Figure 1.8, which requires a minimum sampling frequency of $f_{\text{Nyquist}} = 2 \cdot B$.

This process introduces time-discrete samples. Due to sampling, the original speech spectrum is replicated at multiples of the sampling frequency. This is why the previous bandlimitation was necessary, in order to prevent aliasing or frequency-domain overlapping of the spectral lobes. If this condition is met, the original analogue speech signal can be restored from its samples by passing the samples through a LPF having a bandwidth of B. In conventional speech systems, typically a sampling frequency of 8 kHz corresponding to a sampling interval of 125 μs is used.

- Lastly, amplitude discretisation or quantisation must be invoked, according to Figure 1.8, which requires an analogue-to-digital (A/D) converter. The out bits of the quantiser can be converted to a serial bitstream for transmission over digital links.

1.4.2 Quantisation Characteristics

It is clear from Figure 1.9 that the original speech signal is contaminated during the quantisation process by quantisation noise, which will be the subject of this section. The severity of contamination is a function of the signal's distribution, the quantiser's resolution and its transfer characteristic.

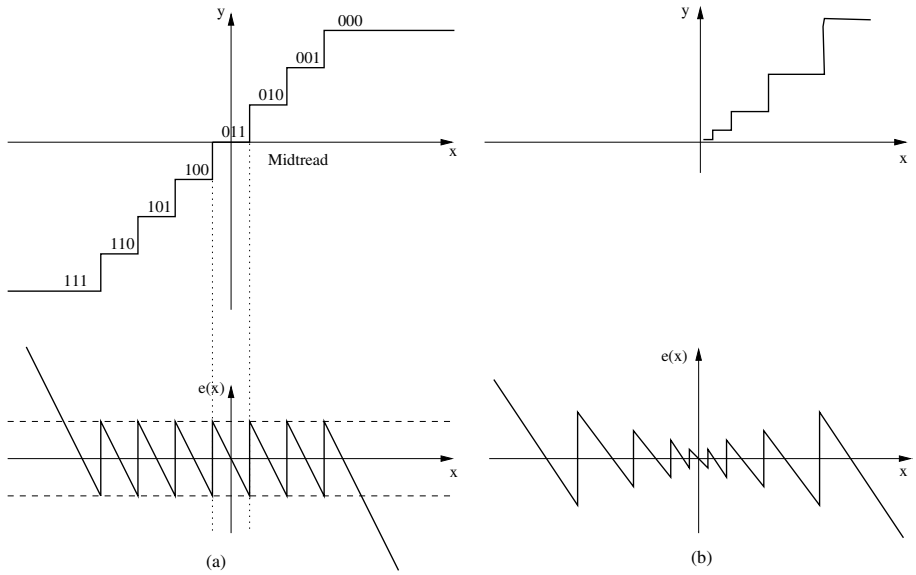

Figure 1.10: Linear quantisers and their quantisation errors: (a) midtread, (b) non-uniform.

The family of *linear quantisers* exhibits a linear transfer function within its dynamic range and saturation above that. They divide the input signal's dynamic range into a number of uniformly or non-uniformly spaced quantisation intervals, as seen in Figure 1.10, and assign an R-bit word to each so-called *reconstruction level*, which represent the legitimate output values. In Figure 1.10 according to $R = 3$ there are $2^3 = 8$ reconstruction levels which are labelled as 000, 001, -111 and a so-called *mid-tread quantiser* is featured, where the

quantiser's output is zero, if the input signal is zero. In the case of the so-called *mid-riser quantiser* the transfer function exhibits a level change at the abscissa value of zero. Note that the quantisation error characteristic of the quantisers is also shown in Figure 1.10. As expected, when the quantiser characteristic saturates at its maximum output level, the quantisation error increases without limit.

The difference between the *uniform* and *non-uniform quantiser* characteristics in Figure 1.10 is that the uniform quantiser maintains a constant maximum error across its total dynamic range, whereas the non-uniform quantiser employs unequal quantisation intervals (quantiles), in order to allow larger granular error, where the input signal is larger. Hence the non-uniform quantiser exhibits a near-constant SNR across its dynamic range. This may allow us to reduce the number of quantisation bits and the required transmission rate, while maintaining perceptually unimpaired speech quality.

In summary, linear quantisers are conceptually and implementationally simple and impose no restrictions on the analogue input signal's statistical characteristics, such as the PDF, etc. Clearly, they do not require *a priori* knowledge concerning the input signal. Note, however, that other PDF-dependent quantisers perform better in terms of overall quantisation noise power or SNR. These issues will be made more explicit during our further discourse.

1.4.3 Quantisation Noise and Rate-distortion Theory

Observe in Figure 1.10 that the instantaneous *quantisation error* $e(x)$ is dependent on the instantaneous input signal level. In other words, $e(x)$ is non-uniform across the quantiser's dynamic range and some amplitudes are represented without quantisation error, if they happen to be on a reconstruction level, while others are associated with larger errors. If the input signal's dynamic range exceeds the quantiser's linear range, the quantiser's output voltage saturates at its maximum level and the quantisation error may become arbitrarily high. Hence the knowledge of the input signal's statistical distribution is important for minimising the overall *granular* and *overload distortion*. The quantised version $\hat{x}(t)$ of the input signal $x(t)$ can be computed as

$$\hat{x}(t) = x(t) + e(t), \tag{1.1}$$

where $e(t)$ is the quantisation error.

It is plausible that if no amplitude discretisation is used for a source signal, a sampled analogue source has formally an infinite entropy, requiring an infinite transmission rate, which is underpinned by the formal application of Equation (1.2). If the analogue speech samples are quantised to R-bit accuracy, there are $q = 2^R$ different legitimate samples, each of which has a probability of occurrence $p_i, i = 1, 2, \ldots, q$. It is known from information theory that the above mentioned R bit/symbol channel capacity requirement can be further reduced using so-called entropy coding to the value of the source's entropy given by

$$H(x) = -\sum_{i=1}^{q} p_i \cdot \log_2 p_i, \tag{1.2}$$

without inflicting any further coding impairment, if an infinite delay entropy-coding scheme is acceptable. Since this is not the case in interactive speech conversations, we are more interested in quantifying the coding distortion, when using R bits per speech sample.

An important general result of information theory is the so-called *rate-distortion theorem*, which quantifies the minimum required average bitrate R_D in terms of [bpsample] in order to represent a random variable (rv) with less than D distortion. Explicitly, for a rv x with variance of σ_x^2 and quantised value $\hat{x}$ the distortion is defined as the mean squared error (MSE) expression given by

$$D = E\{(x - \hat{x})^2\} = E\{e^2(t)\}, \tag{1.3}$$

where E represents the expected value.

Observe that if $R_D = 0$ bits are used to quantise the quantity x, then the distortion is given by the signal's variance $D = \sigma_x^2$. If, however, more than zero bits are used, i.e. $R_D > 0$, then intuitively one additional bit is needed every time we want to halve the root mean squared (RMS) value of 'D', or quadruple the signal-to-noise ratio of $\mathrm{SNR} = \sigma_x^2/D$, which suggests a logarithmic relation between R_D and D. After Shannon and Gallager we can write

$$R_D = \frac{1}{2} \log_2 \frac{\sigma_x^2}{D} \quad \text{if } D \leq \sigma_x^2. \tag{1.4}$$

Upon combining $R_D = 0$ and $R_D > 0$ into one equation we arrive at

$$R_D = \begin{cases} \frac{1}{2} \log_2 \sigma_x^2/D & D < \sigma_x^2 \\ 0 & D \geq \sigma_x^2. \end{cases} \tag{1.5}$$

The qualitative or stylised relationship of D versus R_D inferred from Equation (1.5) is shown in Figure 1.11.

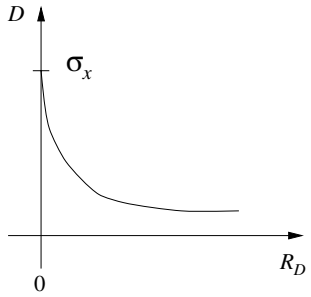

Figure 1.11: Stylised distortion (D) versus coding rate (R_D) curve.

In order to quantify the variance of the quantisation error it is reasonable to assume that if the quantisation interval q is small and no quantiser overload is incurred, then $e(t)$ is uniformly distributed in the interval $[-q/2, q/2]$. If the quantiser's linear dynamic range is limited to $[\pm V]$, then for a uniform quantiser the quantisation interval can be expressed with $q = 2V/2^{R_D}$, where R_D is the number of quantisation bits. The quantisation error variance can then be computed by squaring the instantaneous error magnitude e and weighting its contribution with its probability of occurrence expressed with the help of its PDF $p(e) = 1/q$

and finally integrating or averaging it over the range of $[-q/2, q/2]$ as follows:

$$\begin{aligned}\sigma_e^2 &= \int_{-q/2}^{q/2} e^2 p(e)\,\mathrm{d}e = \int_{-q/2}^{q/2} e^2 \frac{1}{q}\,\mathrm{d}e \\ &= \frac{1}{q}\left[\frac{e^3}{3}\right]_{-q/2}^{q/2} = \left(\frac{q^3}{8}\left(+\frac{q^3}{8}\right)\cdot 1/3q\right) = \frac{q^2}{12},\end{aligned} \tag{1.6}$$

which corresponds to an RMS quantiser noise of $q/\sqrt{12} \approx 0.3q$. In the case of uniform quantisers we can substitute $q = 2V/2^{R_D}$ into Equation (1.6) – where R_D is the number of bits used for encoding – giving the noise variance in the following form:

$$\sigma_q^2 = \frac{q^2}{12} = \frac{1}{12}\left(\frac{2V}{2_D^R}\right)^2 = \frac{1}{3}\frac{V^2}{2^{2R_D}}. \tag{1.7}$$

Similarly, assuming a *uniform signal PDF*, the signal's variance becomes

$$\sigma_x^2 = \int_{-\infty}^{\infty} x^2 p(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} x^2 \frac{1}{2V}\,\mathrm{d}x = \frac{1}{2V}\left[\frac{x^3}{3}\right]_{-V}^{V} = \frac{1}{6E}\cdot 2V^3 = \frac{E^2}{3}. \tag{1.8}$$

Then the SNR can be computed as

$$\mathrm{SNR} = \frac{\sigma_x^2}{\sigma_q^2} = \frac{V^2}{3}\cdot\frac{2^{2R_D}}{V^2}\cdot 3 = 2^{2R_D}, \tag{1.9}$$

which can be expressed in terms of dB as

$$\begin{aligned}\mathrm{SNR}_{\mathrm{dB}} &= 10\cdot\log_{10} 2^{2R} = 20R_D\cdot\log_{10} 2 \\ \mathrm{SNR}_{\mathrm{dB}} &\approx 6.02\cdot R_D\ [\mathrm{dB}].\end{aligned} \tag{1.10}$$

This simple result is useful for quick SNR estimates and it is also intuitively plausible, since every new bit used halves the quantisation error and hence doubles the SNR. In practice the speech PDF is highly non-uniform and hence the quantiser's dynamic range cannot be fully exploited in order to minimise the probability of quantiser characteristic overload error. Hence Equation (1.10) over-estimates the expected SNR.

1.4.4 Non-uniform Quantisation for a known PDF: Companding

If the input signal's PDF is known and can be considered stationary, higher SNR can be achieved by appropriately matched *non-uniform quantisation* (NUQ) than in the case of uniform quantisers. The input signal's dynamic range is partitioned into non-uniformly spaced segments as we have seen in Figure 1.10, where the quantisation intervals are more dense near the origin, in order to quantise the typically high-probability low-magnitude samples more accurately. In contrast, the lower-probability signal PDF tails are less accurately quantised. In contrast to uniform quantisation, where the maximum error was constant across

the quantiser's dynamic range, for non-uniform quantisers the SNR becomes more or less constant across the signal's dynamic range.

It is intuitively advantageous to render the width of the quantisation intervals or *quantiles* inversely proportional to the signal PDF, since a larger quantisation error is affordable in the case of infrequent signal samples and *vice versa*. Two different approaches have been proposed, for example, by Jayant and Noll [10] in order to minimise the total quantisation distortion in the case of non-uniform signal PDFs.

One of the possible system models is shown in Figure 1.12, where the input signal is first compressed using a so-called *nonlinear compander* characteristic and then uniformly quantised. The original signal can be recovered using an expander at the decoder, which exhibits an inverse characteristic with respect to that of the compander. This approach will be considered first, while the design of the minimum mean squared error (mmse) non-uniform quantiser using the so-called Lloyd–Max [60–62] algorithm will be portrayed during our further discussions.

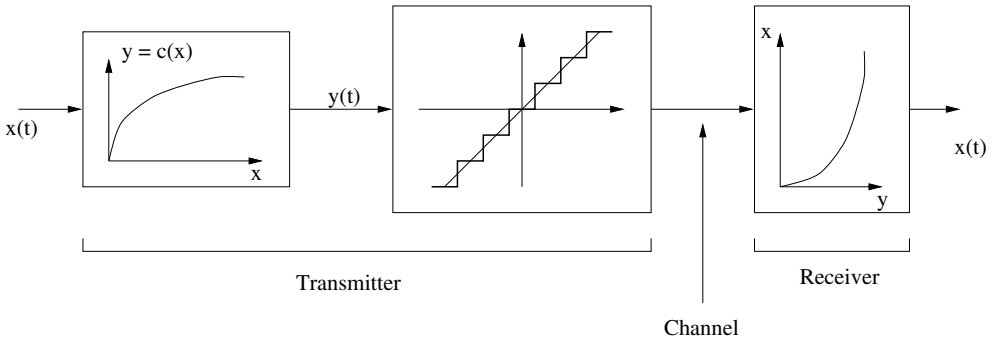

Figure 1.12: Stylised non-uniform quantiser model using companding, when the input signal's PDF is known.

The qualitative effect of nonlinear compression on the signal's PDF is portrayed in Figure 1.13, where it becomes explicit why the compressed PDF can be quantised by a uniform quantiser. Observe that the compander has a more gentle slope, where larger quantisation intervals are expected in the uncompressed signal's amplitude range and *vice versa*, implying that the compander's slope is proportional to the quantisation interval density and inversely proportional to the stepsize for any given input signal amplitude.

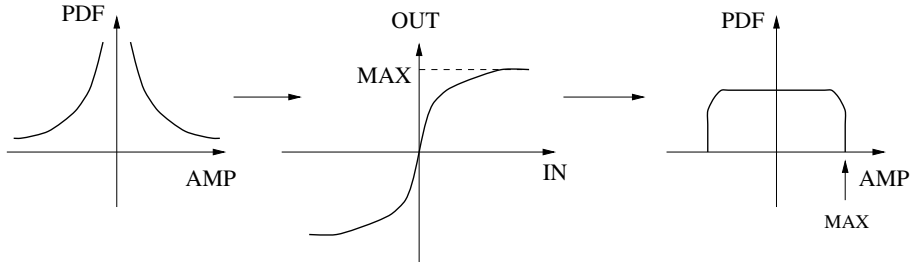

Figure 1.13: Qualitative effect of companding on a known input signal PDF shape.

Following Bennett's approach [63], Jayant and Noll [10] have shown that if the signal's PDF $p(x)$ is a smooth, known function and sufficiently fine quantisation is used – implying that $R \geq 6$ – then the quantisation error variance can be expressed as

$$\sigma_q^2 \approx \frac{q^2}{12} \int_{-x_{\max}}^{x_{\max}} \frac{p(x)}{|\dot{C}(x)|^2} \, \mathrm{d}x, \tag{1.11}$$

where $\dot{C}(x) = \mathrm{d}C(x)/\mathrm{d}x$ represents the slope of the compander's characteristic. It is instructive to note that where the input signal's PDF $p(x)$ is high, the σ_q^2 contributions are also high due to the high probability of occurrence of such signal amplitudes. This effect can be mitigated using a compander exhibiting a high gradient in this interval, since the factor $1/|\dot{C}(x)|^2$ de-weights the error contributions due to the highly peaked PDF near the origin. For an optimum compander characteristic $C(x)$ all quantiles give the same distortion contribution.

Jayant and Noll [10] have also shown that the minimum quantisation error variance is achieved by the compander characteristic given by

$$C(x) = x_{\max} \frac{\int_0^x \sqrt[3]{p(x)} \, \mathrm{d}x}{\int_0^{x_{\max}} \sqrt[3]{p(x)} \, \mathrm{d}x}, \tag{1.12}$$

where the denominator constitutes a normalising factor. Hence a simple practical compander design algorithm can be devised by evaluating the signal's histogram in order to estimate the PDF $p(x)$ and by graphically integrating $\sqrt[3]{p(x)}$ according to Equation (1.12) up to the abscissa value x, yielding the companding characteristic at the ordinate value $C(x)$, yielding the companding characteristic ordinate value $C(x)$.

Although this technique minimises the quantisation error variance or maximises the SNR in the case of a known signal PDF, if the input signal's PDF or variance is time-variant, the compander's performance degrades. In many practical scenarios this is the case and hence often it is advantageous to optimise the compander's characteristic to maximise the SNR independently of the shape of the PDF. Then no compander mismatch penalty is incurred. In order to achieve this, the quantisation error variance σ_e must be rendered proportional to the value of the input signal $x(t)$ across its dynamic range, implying that large signal samples will have larger quantisation error than small samples. This issue is the topic of the next section.

1.4.5 PDF-independent Quantisation using Logarithmic Compression

The input signal's variance is given in the case of an arbitrary PDF $p(x)$ as

$$\sigma_x^2 = \int_{-\infty}^{\infty} x^2 p(x) \, \mathrm{d}x. \tag{1.13}$$

Assuming zero saturation distortion, the SNR can be expressed from Equations (1.11) and (1.13) as

$$\mathrm{SNR} = \frac{\sigma_x^2}{\sigma_q^2} = \frac{\int_{-x_{\max}}^{x_{\max}} x^2 p(x) \, \mathrm{d}x}{\frac{q^2}{12} \int_{-x_{\max}}^{x_{\max}} (p(x)/|\dot{C}(x)|^2) \, \mathrm{d}x}. \tag{1.14}$$

In order to maintain an SNR value that is independent of the signal's PDF $p(x)$ the numerator of Equation (1.14) must be a constant times the denominator, which is equivalent to requiring that

$$|\dot{C}(x)|^2 \stackrel{!}{=} \left|\frac{K}{x}\right|^2, \tag{1.15}$$

or alternatively that

$$\dot{C}(x) = K/x \tag{1.16}$$

and hence

$$C(x) = \int_0^x \frac{K}{z}\,\mathrm{d}z = K \cdot \ln x + A. \tag{1.17}$$

This compander characteristic is shown in Figure 1.14(a) and it ensures a constant SNR across the signal's dynamic range, irrespective of the shape of the signal's PDF. Intuitively, large signals can have large errors, while small signal must maintain a low distortion, which gives a constant SNR for different input signal levels.

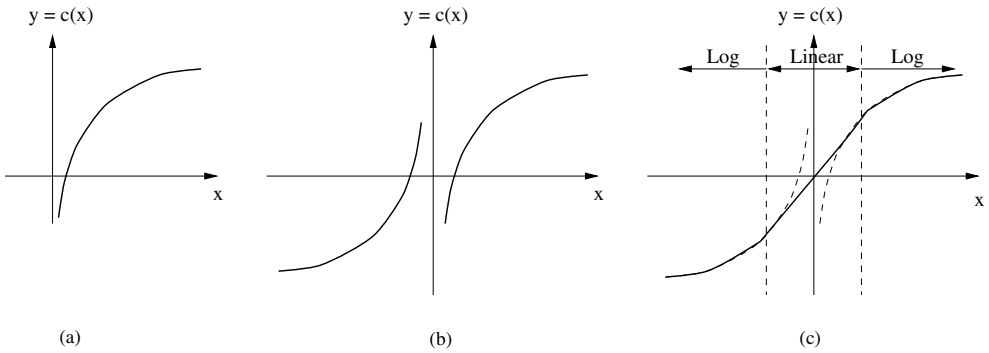

Figure 1.14: Stylised companding characteristic for a near-optimal quantiser.

Jayant and Noll also note that the constant A in Equation (1.17) allows for a vertical compander characteristic shift in order to satisfy the boundary condition of matching $x_{\max}$ and $y_{\max}$, yielding $y = y_{\max}$, when $x = x_{\max}$. Explicitly:

$$y_{\max} = C(x_{\max}) = K \cdot \ln x_{\max} + A. \tag{1.18}$$

Upon normalising Equation (1.17) to $y_{\max}$ we arrive at

$$\frac{y}{y_{\max}} = \frac{C(x)}{y_{\max}} = \frac{K \cdot \ln x + A}{K \cdot \ln x_{\max} + A}. \tag{1.19}$$

It is convenient to introduce an arbitrary constant B, in order to be able to express A as $A = K \cdot \ln B$, since then Equation (1.19) can be written as

$$\frac{y}{y_{\max}} = \frac{K \cdot \ln x + K \cdot \ln B}{K \cdot \ln x_{\max} + K \cdot \ln B} = \frac{\ln xB}{\ln x_{\max}B}. \tag{1.20}$$

Equation (1.20) can be further simplified upon rendering its denominator unity by stipulating $x_{\max} \cdot B = e^1$, which yields $B = e/x_{\max}$. Then Equation (1.20) simplifies to

$$\frac{y}{y_{\max}} = \frac{\ln xe/x_{\max}}{\ln e} = \ln\left(\frac{e \cdot x}{x_{\max}}\right), \tag{1.21}$$

which now gives $y = y_{\max}$, when $x = x_{\max}$. This logarithmic characteristic, which is shown in Figure 1.14(a), must be rendered symmetric with respect to the y-axis, which we achieve upon introducing the $\text{signum}(x) = \text{sgn}(x)$ function:

$$\frac{y}{y_{\max}} = \frac{C(x)}{y_{\max}} = \ln\left(\frac{e \cdot |x|}{x_{\max}}\right) \text{sgn}(x). \tag{1.22}$$

This symmetric function is displayed in Figure 1.14(b). However, a further problem is that the logarithmic function is non-continuous at zero. Hence around zero amplitude a linear section is introduced in order to ensure a seamless positive–negative transition in the compression characteristic.

Two practical logarithmic compander characteristics have emerged, which satisfy the above requirements. In the US the so-called μ-law compander was standardised [64–66], while in Europe the A-law compander was proposed [4]. The corresponding stylised logarithmic compander characteristic is depicted in Figure 1.14(c). Let us now consider the standard μ-law compander.

1.4.5.1 The μ-law Compander

This companding characteristic is given by

$$y = C(x) = y_{\max} \cdot \frac{\ln[1 + \mu \cdot (|x|/x_{\max})]}{\ln(1 + \mu)} \cdot \text{sgn}(x). \tag{1.23}$$

Upon inferring from the $\log(1 + z)$ function that

$$\log(1 + z) \approx z \text{ if } z \ll 1, \tag{1.24}$$

in the case of small and large signals, respectively, we have from Equation (1.23) that

$$y = C(x) = \begin{cases} y_{\max} \cdot \dfrac{\mu \cdot (|x|/x_{\max})}{\ln \mu} & \text{if } \mu \cdot \left(\dfrac{|x|}{x_{\max}}\right) \ll 1 \\ y_{\max} \cdot \dfrac{\ln[\mu \cdot (|x|/x_{\max})]}{\ln \mu} & \text{if } \mu \cdot \left(\dfrac{|x|}{x_{\max}}\right) \gg 1, \end{cases} \tag{1.25}$$

which is a linear function of the normalised input signal $x/x_{\max}$ for small signals and a logarithmic function for large signals. The $\mu \cdot |x|/x_{\max} = 1$ value can be considered to be the break-point between the small and large signal operation and the $|x| = x_{\max}/\mu$ is the corresponding abscissa value. In order to emphasise the logarithmic nature of the characteristic, μ must be large, which reduces the abscissa value of the beginning of the logarithmic section. It is plausible that the optimum value of μ is dependent on the quantiser

resolution R and for $R = 8$ the American standard so-called *pulse code modulation* (PCM) speech transmission system recommends $\mu = 255$.

Following the approach proposed by Jayant and Noll [10], the SNR of the μ-law compander can be derived upon substituting $y = C_\mu(x)$ from Equation (1.23) into the general SNR formula of Equation (1.14):

$$y = C_\mu(x) = y_{\max} \cdot \frac{\ln[1 + \mu(|x|/x_{\max})]}{\ln(1 + \mu)} \cdot \mathrm{sgn}(x) \tag{1.26}$$

$$\dot{C}_\mu(x) = \frac{y_{\max}}{\ln(1 + \mu)} \cdot \frac{1}{1 + \mu(|x|/x_{\max})} \cdot \mu \left(\frac{1}{x_{\max}}\right). \tag{1.27}$$

For large input signals we have $\mu(|x|/x_{\max}) \gg 1$, and hence

$$\dot{C}_\mu(x) \approx \frac{y_{\max}}{\ln \mu} \cdot \frac{1}{x}. \tag{1.28}$$

Upon substituting

$$\frac{1}{\dot{C}_\mu(x)} = \frac{\ln \mu}{y_{\max}} \cdot x \tag{1.29}$$

in Equation (1.14) we arrive at

$$\begin{aligned} \mathrm{SNR} &= \frac{\int_{-x_{\max}}^{x_{\max}} x^2 p(x)\, \mathrm{d}x}{(q^2/12) \int_{-x_{\max}}^{x_{\max}} (\ln \mu / y_{\max})^2 x^2 p(x)\, \mathrm{d}x} \\ &= \frac{1}{(q^2/12)(\ln \mu / y_{\max})^2} = 3 \left(\frac{2y_{\max}}{q}\right)^2 \cdot \left(\frac{1}{\ln \mu}\right)^2 \\ &= 3 \cdot 2^{2R} \cdot \left(\frac{1}{\ln \mu}\right)^2. \end{aligned} \tag{1.30}$$

Upon exploiting the fact that $2y_{\max}/q = 2^R$ represents the number of quantisation levels and expressing the above equation in terms of dB we get

$$\mathrm{SNR}_{\mathrm{dB}}^{\mu} = 6.02 \cdot R + 4.77 - 20 \log_{10}(\ln(1 + \mu)), \tag{1.31}$$

which gives an SNR of about 38 dB in the case of the American standard system using $R = 8$ and $\mu = 255$. Recall that under the assumption of no quantiser characteristic overload and a uniformly distributed input signal the corresponding SNR estimate would yield $6.02 \cdot 8 \approx 48$ dB. Note, however, that in practical terms this SNR is never achieved, since the input signal does not have a uniform distribution and saturation distortion is also often incurred.

1.4.5.2 The *A*-law Compander

Another practical logarithmic compander characteristic is the *A-Law Compander* [4] given below, which was standardised by the CCITT or ITU and which is used throughout Europe:

$$y = C(x) = \begin{cases} y_{\max} \cdot \dfrac{A(|x|/x_{\max})}{1+\ln A} \cdot \mathrm{sgn}(x) & 0 < \dfrac{|x|}{x_{\max}} < \dfrac{1}{A} \\ y_{\max} \cdot \dfrac{1+\ln[A(|x|/x_{\max})]}{1+\ln A} \cdot \mathrm{sgn}(x) & \dfrac{1}{A} < \dfrac{|x|}{x_{\max}} < 1, \end{cases} \tag{1.32}$$

where $A = 87.56$. Similar to the μ-law characteristic, it has a linear region near the origin and a logarithmic section above the break-point $|x| = x_{\max}/A$. Note, however, that in the case of $R = 8$ bits $A < \mu$, hence the A-law characteristic's linear–logarithmic break-point is at a higher input value than that of the μ-law characteristic.

Again, substituting

$$\frac{1}{\dot{C}_A(x)} = \frac{(1+\ln A)}{y_{\max}} \cdot x \tag{1.33}$$

into Equation (1.14) and exploiting the fact that $2y_{\max}/q = 2^R$ represents the number of quantisation levels, we have

$$\begin{aligned} \mathrm{SNR} &= \frac{\int_{-x_{\max}}^{x_{\max}} x^2 p(x)\,\mathrm{d}x}{(q^2/12)\int_{-x_{\max}}^{x_{\max}} ((1+\ln A)/y_{\max})^2 x^2 p(x)\,\mathrm{d}x} \\ &= \frac{1}{(q^2/12)((1+\ln A)/y_{\max})^2} = 3\left(\frac{2y_{\max}}{q}\right)^2 \cdot \left(\frac{1}{(1+\ln A)}\right)^2 \\ &= 3 \cdot 2^{2R} \cdot \left(\frac{1}{(1+\ln A)}\right)^2. \end{aligned} \tag{1.34}$$

Upon expressing the above equation in terms of dB we arrive at

$$\mathrm{SNR}_{\mathrm{dB}}^{A} = 6.02 \cdot R + 4.77 - 20\log_{10}(1+\ln A), \tag{1.35}$$

which, similar to the μ-law compander, gives an SNR of about 38 dB in the case of the European standard PCM speech transmission system using $R = 8$ and $A = 87.56$.

Further features of the European A-law standard system are that the characteristic given by Equation (1.32) is implemented in the form of a 16-segment piece-wise linear approximation, as seen in Figure 1.15. The segment retaining the lowest gradient of $1/4$ is at the top end of the input signal's dynamic range, which covers half of the positive dynamic range and it is divided into 16 uniformly spaced quantisation intervals. The second segment from the top covers a quarter of the positive dynamic range and doubles the top segment's steepness or gradient to $1/2$, etc. The bottom segment covers a 64th of the positive dynamic range, has the highest slope of 16 and the finest resolution. The first bit of each $R = 8$-bit PCM codeword represents the sign of the input signal, the next three bits specify which segment the input signal belongs to, while the last four bits divide a specific segment into 16 uniform-width quantisation intervals, as shown below:

$$\underbrace{b_7}_{\substack{\text{sign} \\ \text{(segment)}}} \quad \underbrace{b_6\ b_5\ b_4}_{\text{segments}} \quad \underbrace{b_3\ b_2\ b_1\ b_0}_{\substack{\text{uniform quant.} \\ \text{in each segment}}}$$

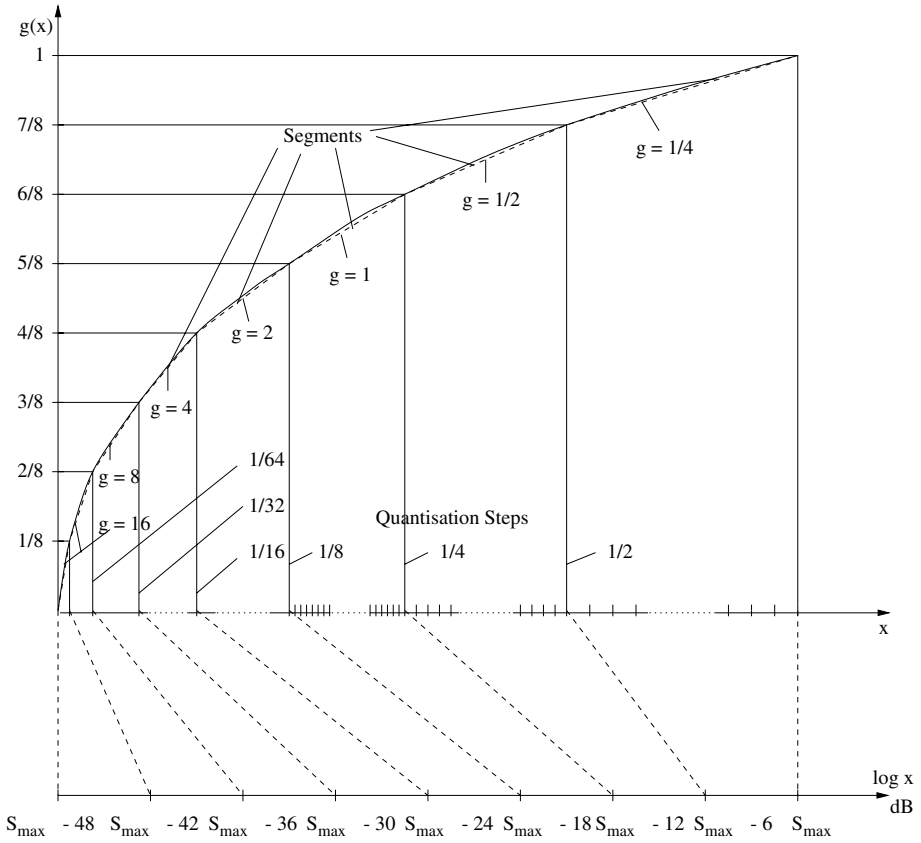

Figure 1.15: Stylised European A-law PCM standard characteristic.

This scheme was standardised by the *International Telegraph and Telephone Consultative Committee (CCITT)* as the G.711 Recommendation for the transmission of speech sampled at 8 kHz. Hence the transmission rate becomes $8 \times 8 = 64$ kbps (kbps). This results in perceptually unimpaired speech quality, which would require about 12 bits in the case of linear quantisation.

1.4.6 Optimum Non-uniform Quantisation

For non-uniform quantisers the quantisation error variance is given by

$$\sigma_q^2 = E\{|x - x_q|^2\} = \int_{-\infty}^{\infty} e^2(x) p(x)\, \mathrm{d}x, \tag{1.36}$$

which, again, corresponds to weighting and averaging the quantisation error energy over its magnitude range. Assuming an odd-symmetric quantiser transfer function and symmetric

PDF $p(x)$, the total quantisation distortion power σ_D^2 is

$$\sigma_D^2 = 2 \int_0^\infty e^2(x) p(x) \, \mathrm{d}x. \tag{1.37}$$

The total distortion can be expressed as the sum of the quantisation distortion in the quantiser's linear range, plus the saturation distortion in its nonlinear range

$$\sigma_D^2 = \underbrace{2 \int_0^V e^2(x) p(x) \, \mathrm{d}x}_{\sigma_q^2\text{: linear region}} + \underbrace{2 \int_V^\infty e^2(x) p(x) \, \mathrm{d}x}_{\sigma_s^2\text{: nonlinear region}} \tag{1.38}$$

or more simply as

$$\sigma_D^2 = \sigma_q^2 + \sigma_s^2. \tag{1.39}$$

In order to emphasise the fact that in the case of non-uniform quantisation each of the N quantisation intervals or so-called quantiles adds a different PDF-weighted contribution to the total quantisation distortion, we re-write the first term of Equation (1.38) as

$$\begin{aligned} \sigma_q^2 &= \sum_{n=1}^{N} \int_{x_n}^{x_{n+1}} e^2(x) p(x) \, \mathrm{d}x \\ &= \sum_{n=1}^{N} \int_{x_n}^{x_{n+1}} (x - x_q)^2 p(x) \, \mathrm{d}x \qquad (1.40) \\ &= \sum_{n=1}^{N} \int_{x_n}^{x_{n+1}} (x - r_n)^2 p(x) \, \mathrm{d}x, \qquad (1.41) \end{aligned}$$

where $x_q = r_n$ represents the so-called reconstruction levels.

Given a certain number of quantisation bits R and the PDF of the input signal, the optimum Lloyd–Max quantiser, which was independently invented by Lloyd [60, 61] and Max [62], determines the set of optimum quantiser decision levels and the corresponding set of quantisation levels.

Jayant and Noll [10] have provided a detailed discussion on two different methods of determining the mmse solution to the problem. One of the solutions is based on an iterative technique of rearranging the decision thresholds and reconstruction levels, while the other one is an approximate solution valid for fine quantisers using a high number of bits per sample. We first present the general approach to minimising the MSE by determining the set of optimum reconstruction levels r_n, $n = 1, \ldots, N$, and the corresponding decision threshold values t_n, $n = 1, \ldots, N$.

In general, it is a necessary but not sufficient condition for finding the global minimum of Equation (1.41) for its partial derivatives to become zero. However, if the PDF $p(s)$ is log-concave, that is the second derivative of its logarithm is negative, then the minimum found is a global one. For the frequently encountered uniform (U), Gaussian (G) and Laplacian (L) PDFs the log-concave condition is satisfied but, for example, for Gamma (Γ) PDFs is not.

Setting the partial derivatives of Equation (1.41) with respect to a specific r_n to zero, there is only one term in the sum which depends on the r_n value considered, hence we arrive at

$$\frac{\partial \sigma_q^2}{\partial r_n} = 2 \int_{t_n}^{t_{n+1}} (s - r_n) \cdot p(s)\, \mathrm{d}s = 0, \quad n = 1, \ldots, N, \tag{1.42}$$

which leads to

$$\int_{t_n^{\mathrm{opt}}}^{t_{n+1}^{\mathrm{opt}}} s \cdot p(s)\, \mathrm{d}s = r_n \int_{t_n^{\mathrm{opt}}}^{t_{n+1}^{\mathrm{opt}}} p(s)\, \mathrm{d}s, \tag{1.43}$$

yielding the optimum reconstruction level r_n^{opt} as

$$r_n^{\mathrm{opt}} = \frac{\int_{t_n^{\mathrm{opt}}}^{t_{n+1}^{\mathrm{opt}}} s \cdot p(s)\, \mathrm{d}s}{\int_{t_n^{\mathrm{opt}}}^{t_{n+1}^{\mathrm{opt}}} p(s)\, \mathrm{d}s}, \quad n = 1, \ldots, N. \tag{1.44}$$

Note that the above expression depends on the optimum quantisation interval thresholds t_n^{opt} and t_{n+1}^{opt}. Furthermore, for an arbitrary non-uniform PDF r_n^{opt} is given by the mean value or the 'centre of gravity' of s within the quantisation interval n, rather than by $(t_n^{\mathrm{opt}} + t_{n+1}^{\mathrm{opt}})/2$.

Similarly, when computing $\partial \sigma_q^2 / \partial t_n$, there are only two terms in Equation (1.41), which contain t_n, therefore we get

$$\frac{\partial \sigma_q^2}{\partial t_n} = (t_n - r_{n-1})^2 p(t_n) - (t_n - r_n)^2 p(t_n) = 0, \tag{1.45}$$

leading to

$$t_n^2 - 2t_n r_{n-1} + r_{n-1}^2 - t_n^2 + 2t_n r_n - r_n^2 = 0. \tag{1.46}$$

Hence the optimum decision threshold is given by

$$t_n^{\mathrm{opt}} = (r_n^{\mathrm{opt}} + r_{n-1}^{\mathrm{opt}})/2, \quad n = 2, \ldots, N,\ t_1^{\mathrm{opt}} = -\infty,\ t_N^{\mathrm{opt}} = \infty \tag{1.47}$$

which is half-way between the optimum reconstruction levels. Since these nonlinear equations are interdependent, they can only be solved by recursive iterations, starting from either a uniform quantiser or from a 'hand-crafted' initial non-uniform quantiser design.

Since most practical signals do not obey any analytically describable distribution, the signal's PDF typically has to be inferred from a sufficiently large and characteristic training set. Equations (1.44) and (1.47) will also have to be evaluated numerically for the training set. Below we provide a simple practical algorithm which can be easily implemented by the coding practitioner with the help of the flowchart of Figure 1.16.

Step 1: Input initial parameters such as the number of quantisation bits R, maximum number of iterations I, dynamic range minimum t_1 and maximum t_N.

Step 2: Generate the initial set of thresholds $t_1^0, \ldots, t_N^0$, where the superscript '0' represents the iteration index, either automatically creating a uniform quantiser between t_1 and t_N according to the required number of bits R, or by inputting a 'hand-crafted' initial design.

Step 3: While $t < T$, where T is the total number of training samples do:

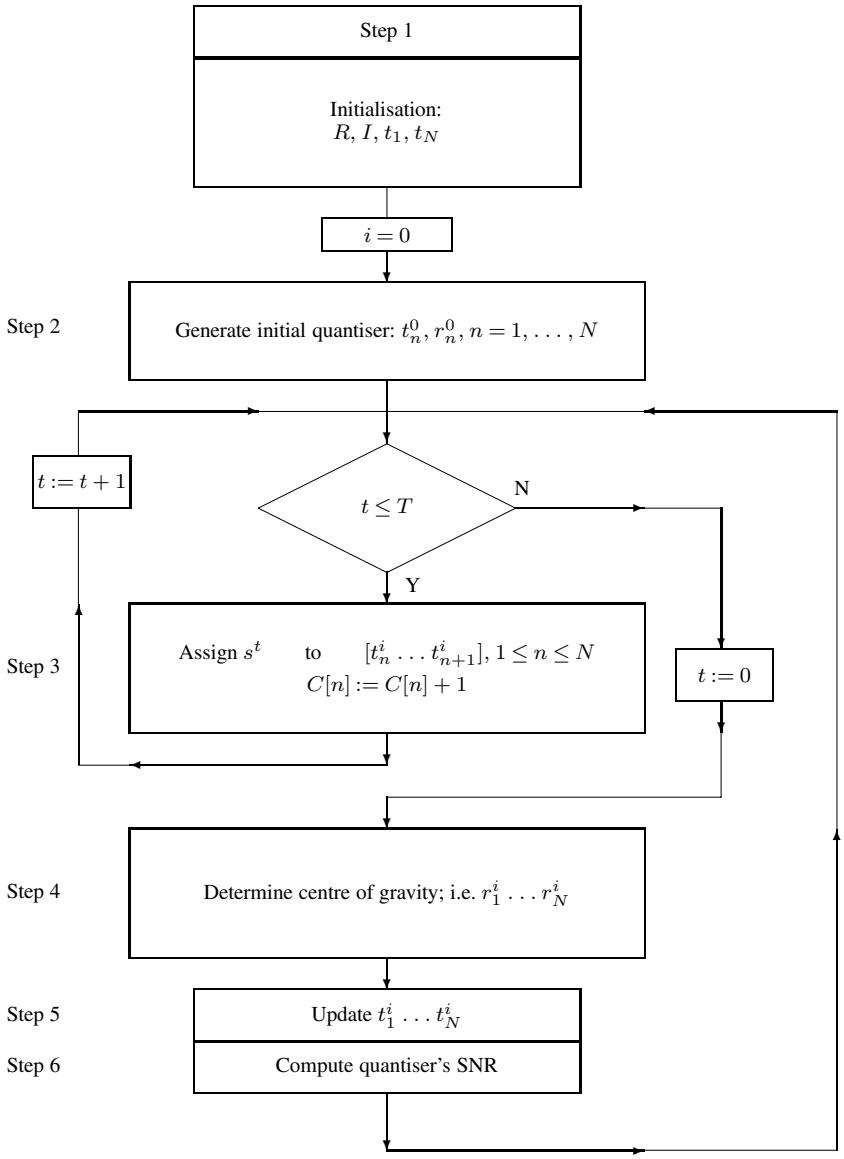

Figure 1.16: Lloyd–Max algorithm flowchart.

1. Assign the current training sample s^t, $t = 1, \ldots, T$ to the corresponding quantisation interval $[t_n^0 \ldots t_{n+1}^0]$ and increment the sample counter $C[n]$, $n = 1, \ldots, N$, holding the number of samples assigned to interval n. This corresponds to generating the histogram $p(s)$ of the training set.

2. Evaluate the MSE contribution due to assigning s^t to $\mathrm{bin}[n]$, that is $\mathrm{MSE}^t = (s^t - s_q^t)^2$ and the resultant total accumulated MSE, that is $\mathrm{MSE}^t = \mathrm{MSE}^{t-1} + \mathrm{MSE}^t$.

Step 4: Once all training samples have been assigned to their corresponding quantisation bins, that is the experimental PDF $p(s)$ is evaluated, the centre of gravity of each bin is computed by summing the training samples in each $\mathrm{bin}[n]$, $n = 1, \ldots, N$, and then dividing the sum by the number of training samples $C[n]$ in $\mathrm{bin}[n]$. This corresponds to the evaluation of Equation (1.44), yielding r_n.

Step 5: Rearrange the initial quantisation thresholds $t_1^0 \ldots t_N^0$ using Equation (1.47) by placing them half-way between the above computed initial reconstruction levels r_n^0, $n = 1, \ldots, N$, where again, the superscript '0' represents the iteration index. This step generates the updated set of quantisation thresholds $t_1^1 \ldots t_N^1$.

Step 6: Evaluate the performance of the current quantiser design in terms of

$$\mathrm{SNR} = 10 \log_{10} \left[\frac{\sum_{t=1}^{T} (s^t)^2}{\mathrm{MSE}^t} \right].$$

Recursion: Repeat Steps 3–6 by iteratively updating r_n^i, t_n^i for all bins $n = 1, \ldots, N$, until the iteration index i reaches its maximum I, while monitoring the quantiser SNR performance improvement given above.

Note that it is important to invoke the algorithm several times, while using a different initial quantiser, in order to ascertain its proper convergence to a global optimum. It is plausible from the inner workings of the algorithm that it will place the reconstruction levels and thresholds more sparsely, where the PDF $p(s)$ is low and *vice versa*. If the input signal's statistics obey a U, G, L or Γ distribution, the Lloyd–Max quantiser's SNR performance can be evaluated using Equations (1.44) and (1.47), and various authors have tabulated the achievable SNR values. Following Max [62], Noll and Zelinski [67] as well as Paez and Glisson [68], both Jayant and Noll [10] as well as Jain [69] collected these SNR values, which we have summarised in Table 1.1 for G and L distributions. Jayant and Noll [10] as well as Jain [69] also tabulated the corresponding t_n and r_n values for a variety of PDFs and R values.

Table 1.1: Maximum achievable SNR and MSE in the case of zero-mean, unit-variance input $[f(R)]$ for Gaussian (G) and Laplacian (L) PDFs for $R = 1, 2, \ldots, 7$. Copyright © Prentice Hall, Jayant-Noll [10] 1984, p. 135 and Jain [69] 1989, p. 104.

		$R=1$	$R=2$	$R=3$	$R=4$	$R=5$	$R=6$	$R=7$
G	SNR(dB)	4.40	9.30	14.62	20.22	26.01	31.89	37.81
	$f(R)$	0.3634	0.1175	0.0345	0.0095	0.0025	0.0006	0.0002
L	SNR(dB)	3.01	7.54	12.64	18.13	23.87	29.74	35.69
	$f(R)$	0.5	0.1762	0.0545	0.0154	0.0041	0.0011	0.0003

Note in Table 1.1 that apart from the achievable maximum SNR values the associated quantiser MSE $f(R)$ is also given as a function of the number of quantisation bits R. When designing a quantiser for an arbitrary non-unity input variance σ_s^2, the associated quantisation thresholds and reconstruction levels must be appropriately scaled by σ_s^2. It is plausible that in the case of a large input variance the reconstruction levels have to be sparsely spaced in

order to cater for the signal's expanded dynamic range. Hence the reconstruction MSE σ_q^2 must also be scaled by σ_s^2, giving

$$\sigma_q^2 = \sigma_s^2 \cdot f(R).$$

Here we curtail our discussion of *zero-memory quantisation* techniques, the interested reader is referred to the excellent in-depth reference [10] by Jayant and Noll for further details. Before we focus our attention on predictive coding techniques, the reader is reminded that in Section 1.2 we highlighted how redundancy is exhibited by both the time- and the frequency-domain features of the speech signal. In the next section we will endeavour to introduce a simple way of exploiting this redundancy in order to achieve better coding efficiency and reduce the required coding rate from 64 kbps to 32 kbps.

1.5 Chapter Summary

In this chapter we provided a rudimentary characterisation of voiced and unvoiced speech signals. It was shown that voice speech segments exhibit a quasi-periodic nature and convey significantly more energy than the more noise-like unvoiced segments. Due to their quasi-periodic nature voiced segments are more predictable, in other words they are more amenable to compression.

These discussions were followed by a brief introduction to the digitisation of speech and to basic waveform coding techniques. The basic principles of logarithmic compression were highlighted and the optimum non-uniform Lloyd–Max quantisation principle was introduced. In the next chapter we introduce the underlying principles of more efficient predictive speech coding techniques.