Chapter 2
Audio Acquisition, Representation and Storage

What the reader should know to understand this chapter

- Basic notions of physics.
- Basic notions of calculus (trigonometry, logarithms, exponentials, etc.)

What the reader should know after reading this chapter

- Human hearing and speaking physiology.
- Signal processing fundamentals.
- Representation techniques behind the main audio formats.
- Perceptual coding fundamentals.
- Audio sampling fundamentals.

2.1 Introduction

The goal of this chapter is to provide basic notions about *digital audio processing technologies*. These are applied in many everyday life products such as phones, radio and television, videogames, CD players, cellular phones, etc. However, although there is a wide spectrum of applications, the main problems to be addressed in order to manipulate digital sound are essentially three: *acquisition*, *representation* and *storage*. The acquisition is the process of converting the physical phenomenon we call sound into a form suitable for digital processing, the representation is the problem of extracting from the sound information necessary to perform a specific task, and the storage is the problem of reducing the number of bits necessary to encode the acoustic signals.

The chapter starts with a description of the sound as a physical phenomenon (Sect. 2.2). This shows that acoustic waves are completely determined by the energy distribution across different frequencies; thus, any sound processing approach must deal with such quantities. This is confirmed by an analysis of voicing and hearing

F. Camastra and A. Vinciarelli, *Machine Learning for Audio, Image and Video Analysis*, Advanced Information and Knowledge Processing,
DOI 10.1007/978-1-4471-6735-8_2

mechanisms in humans. In fact, the vocal apparatus determines frequency and energy content of the voice through the *vocal folds* and the *articulators*. Such organs are capable of changing the shape of the vocal tract like it happens in the cavity of a flute when the player acts on keys or holes. In the case of sound perception, the main task of the ears is to detect the frequencies present in an incoming sound and to transmit the corresponding information to the brain. Both production and perception mechanisms have an influence on audio processing algorithms.

The acquisition problem is presented in Sect. 2.3 through the description of the *analog-to-digital* (A/D) conversion, the process transforming any analog signal into a form suitable for computer processing. Such a process is performed by measuring at discrete time steps the physical effects of a signal. In the case of the sound, the effect is the displacement of an elastic membrane in a microphone due to the pressure variations determined by acoustic waves. Section 2.3 presents the two main issues involved in the acquisition process: the first is the *sampling*, i.e., the fact that the original signal is continuous in time, but the effect measurements are performed only at discrete-time steps. The second is the *quantization*, i.e., the fact that the physical measurements are continuous, but they must be quantized because only a finite number of bits is available on a computer.

The quantization plays an important role also in storage problems because the number of bits used to represent a signal affects the amount of memory space needed to store a recording. Section 2.4 presents the main techniques used to store audio signals by describing the most common audio *formats* (e.g. *WAV*, *MPEG*, *mp3*, etc.). The reason is that each format corresponds to a different *encoding* technique, i.e., to a different way of representing an audio signal. The goal of encoding approaches is to reduce the amount of bits necessary to represent a signal while keeping an acceptable perceptual quality. Section 2.4 shows that the pressure towards the reduction of the *bit-rate* (the amount of bits necessary to represent one second of sound) is due not only to the emergence of new applications characterized by tighter space and bandwidth constraints, but also by consumer preferences.

While acquisition and storage problems are solved with relatively few standard approaches, the representation issue is task dependent. For storage problems (see above), the goal of the representation is to preserve as much as possible the information of the acoustic waveforms, in prosody analysis or topic segmentation, it is necessary to detect the silences or the energy of the signal, in speaker recognition the main information is in the frequency content of the voice, and the list could continue. Section 2.5 presents some of the most important techniques analyzing the variations of the signal to extract useful information. The corpus of such techniques is called *time domain processing* in opposition to *frequency-domain* methods that work on the spectral representation of the signals and are shown in Appendix B and Chap. 12.

Most of the content of this chapter requires basic mathematical notions, but few points need familiarity with Fourier analysis. When this is the case, the text includes a warning and the parts that can be difficult for unexperienced readers can be skipped without any problem. An introduction to Fourier analysis and frequency domain techniques is available in Appendix B. Each section provides references to specialized books and tutorials presenting in more detail the different issues.

2.2 Sound Physics, Production and Perception

This section presents the sound from both a physical and physiological point of view. The description of the main acoustic waves properties shows that the sound can be fully described in terms of frequencies and related energies. This result is obtained by describing the propagation of a single frequency sine wave, an example unrealistically simple, but still representative of what happens in more realistic conditions. In the following, this section provides a general description of how the human beings interact with the sound. The description concerns the way the speech production mechanism determines the frequency content of the voice and the way our ears detect frequencies in incoming sounds.

For more detailed descriptions of the acoustic properties, the reader can refer to more extensive monographies [3, 16, 25] and tutorials [2, 11]. The psychophysiology of hearing is presented in [24, 31], while good introductions to speech production mechanisms are provided in [9, 17].

2.2.1 Acoustic Waves Physics

The physical phenomenon we call sound is originated by air molecule oscillations due to the mechanical energy emitted by an acoustic source. The displacement $s(t)$ with respect to the equilibrium position of each molecule can be modeled as a sinusoid:

$$s(t) = A\sin(2\pi f t + \phi) = A\sin\left(\frac{2\pi}{T}t + \phi\right) \tag{2.1}$$

where A is called amplitude and represents the maximum distance from the equilibrium position (typically measured in *nanometers*), ϕ is the phase, T is called period and it is the time interval length between two instants where $s(t)$ takes the same value, and $f = 1/T$ is the frequency measured in Hz, i.e., the number of times $s(t)$ completes a cycle per second. The function $s(t)$ is shown in the upper plot of Fig. 2.1. Since all air molecules in a certain region of the space oscillate together, the acoustic waves determine local variations of the density that correspond to periodic compressions and rarefactions. The result is that the pressure changes with the time following a sinusoid $p(t)$ with the same frequency as $s(t)$, but amplitude P and phase $\phi^* = \phi + \pi/2$:

$$p(t) = P\sin\left(2\pi f t + \phi + \frac{\pi}{2}\right) = P\sin\left(\frac{2\pi}{T}t + \phi + \frac{\pi}{2}\right). \tag{2.2}$$

The dashed sinusoid in the upper plot of Fig. 2.1 corresponds to $p(t)$ and it shows that the pressure variations have a delay of a quarter of period (due to the $\pi/2$ added to the phase) with respect to $s(t)$. The maximum pressure variations correspond, for

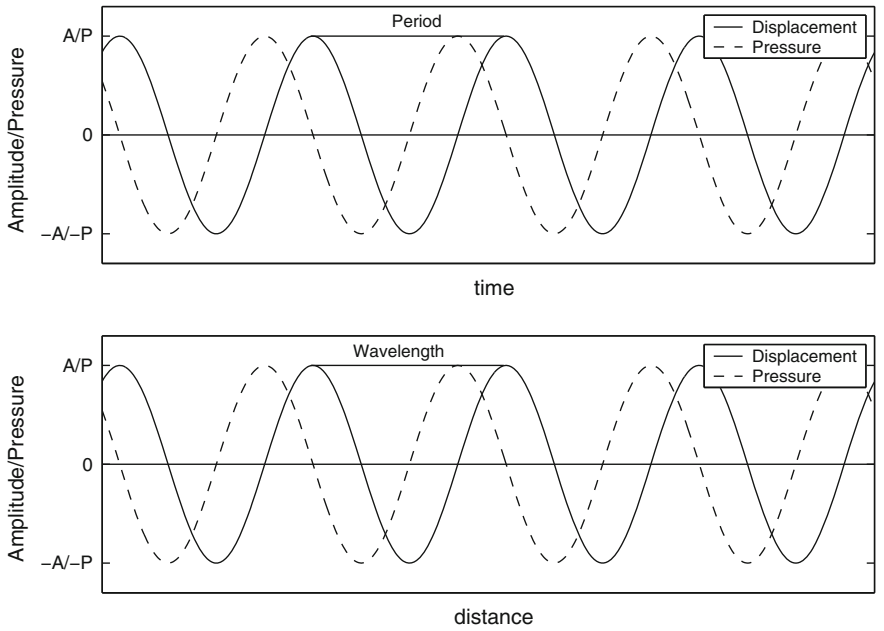

Fig. 2.1 Frequence and wavelength. The *upper* plot shows the displacement of air molecules with respect to their equilibrium position as a function of time. The *lower* plot shows the distribution of pressure values as a function of the distance from the sound source

the highest energy sounds in a common urban environment, to around 0.6 percent of the atmospheric pressure.

When the air molecules oscillate, they transfer part of their mechanical energy to surrounding particules through collisions. The molecules that receive energy start oscillating and, with the same mechanism, they transfer mechanic energy to further particles. In this way, the acoustic waves propagate through the air (or any other medium) and can reach listeners far away from the source. The important aspect of such a propagation mechanism is that there is no net flow of particles no matter is transported from the point where the sound is emitted to the point where a listener receives it. Sound propagation is actually due to energy transport that determines pressure variations and molecule oscillations at distance x from the source.

The lower plot of Fig. 2.1 shows the displacement $s(x)$ of air molecules as a function of the distance x from the audio source:

$$s(x) = A \sin\left(\frac{2\pi}{v} f x + \phi\right) = A \sin\left(\frac{2\pi}{\lambda} x + \phi\right) \tag{2.3}$$

where v is the sound speed in the medium and $\lambda = v/f$ is the *wavelength*, i.e., the distance between two points where $s(x)$ takes the same value (the meaning of the other symbols is the same as in Eq. (2.1). Each point along the horizontal axis of

the lower plot in Fig. 2.1 corresponds to a different molecule of which $s(x)$ gives the displacement. The pressure variation $p(x)$ follows the same sinusoidal function, but has a quarter of period delay like in the case of $p(t)$ (dashed curve in the lower plot of Fig. 2.1):

$$p(x) = P \sin\left(\frac{2\pi}{v} f x + \phi + \frac{\pi}{2}\right) = P \sin\left(\frac{2\pi}{\lambda} x + \phi + \frac{\pi}{2}\right). \tag{2.4}$$

The equations of this section assume that an acoustic wave is completely characterized by two parameters: the frequency f and the amplitude A. From a perceptual point of view, A is related to the *loudness* and f corresponds to the *pitch*. While two sounds with equal loudness can be distinguished based on their frequency, for a given frequency, two sounds with different amplitude are perceived as the same sound with different loudness. The value of f is measured in *Hertz* (Hz), i.e., the number of cycles per second. The measurement of A is performed through the physical effects that depend on the amplitude like pressure variations.

The amplitude is related to the energy of the acoustic source. In fact, the higher is the energy, the higher is the displacement and, correspondently, the perceived loudness of the sound. From an audio processing point of view, the important aspect is what happens for a listener at a distance R from the acoustic source. In order to find a relationship between the source energy and the distance R, it is possible to use the *intensity* I, i.e., the energy passing per time unit through a surface unit. If the medium around the acoustic source is *isotropic*, i.e., it has the same properties along all directions, the energy is distributed uniformly on spherical surfaces of radius R centered in the source. The intensity I can thus be expressed as follows:

$$I(R) = \frac{W}{4\pi R^2} \tag{2.5}$$

where $W = \Delta E / \Delta t$ is the source power, i.e., the amount of energy ΔE emitted in a time interval of duration Δt. The power is measured in watts (W) and the intensity in watts per square meter ($\mathrm{W/m^2}$). The relationship between I and A is as follows:

$$I = 2Z\pi^2 f^2 A^2 \tag{2.6}$$

where Z is a characteristic of the medium called *acoustic impedance*.

Since the only sounds that are interesting in audio applications are those that can be perceived by human beings, the intensities can be measured through their ratio I/I_0 to the *threshold of hearing* (THO) I_0, i.e., the minimum intensity detectable by human ears. However, this creates a problem because the value of I_0 corresponds to $10^{-12}\,\mathrm{W/m^2}$, while the maximum value of I that can be tolerated without permanent physiological damages is $I_{max} = 10^3\,\mathrm{W/m^2}$. The ratio I/I_0 can thus range across 15 orders of magnitude and this makes it difficult to manage different intensity values. For this reason, the ratio I/I_0 is measured using the *deciBel* (dB) scale:

$$I^* = 10 \log_{10} \left(\frac{I}{I_0} \right) \tag{2.7}$$

where I^* is the intensity measured in dB. In this way, the intensity values range between 0 ($I = I_0$) and 150 ($I = I_{max}$). Since the intensity is proportional to the square power of the maximum pressure variation P as follows:

$$I = \frac{P^2}{2Z}, \tag{2.8}$$

the value of I^* can be expressed also in terms of *db SPL* (sound pressure level):

$$I^* = 20 \log_{10} \left(\frac{P}{P_0} \right). \tag{2.9}$$

The numerical value of the intensity is the same when using dB or *db SPL*, but the latter unit allows one to link intensity and pressure. This is important because the pressure is a physical effect relatively easy to measure and the microphones rely on it (see Sect. 2.3).

Real sounds are never characterized by a single frequency f, but by an energy distribution across different frequencies. In intuitive terms, a sound can be thought of as a "sum of single frequency sounds," each characterized by a specific frequency and a specific energy (this aspect is developed rigorously in Appendix B). The important point of this section is that a sound can be fully characterized through frequency and energy measures and the next sections show how the human body interacts with sound using such informations.

2.2.2 Speech Production

Human voices are characterized, like any other acoustic signal, by the energy distribution across different frequencies. This section provides a high-level sketch of how the human vocal apparatus determines such characteristics. Deeper descriptions, especially from the anatomy point of view, can be found in specialized monographies [24, 31].

The voice mechanism starts when the diaphragm pushes air from lungs towards the oral and nasal cavities. The air flow has to pass through an organ called *glottis* that can be considered like a gate to the *vocal tract* (see Fig. 2.2). The glottis determines the frequency distribution of the voice, while the vocal tract (composed of larynx and oral cavity) is at the origin of the energy distribution across frequencies. The main components of the glottis are the vocal folds and the way they react with respect to air coming from the lungs enables to distinguish between the two main classes of sounds produced by human beings. When the vocal folds vibrate, the sounds are called *voiced*, otherwise they are called *unvoiced*. For a given language, all words

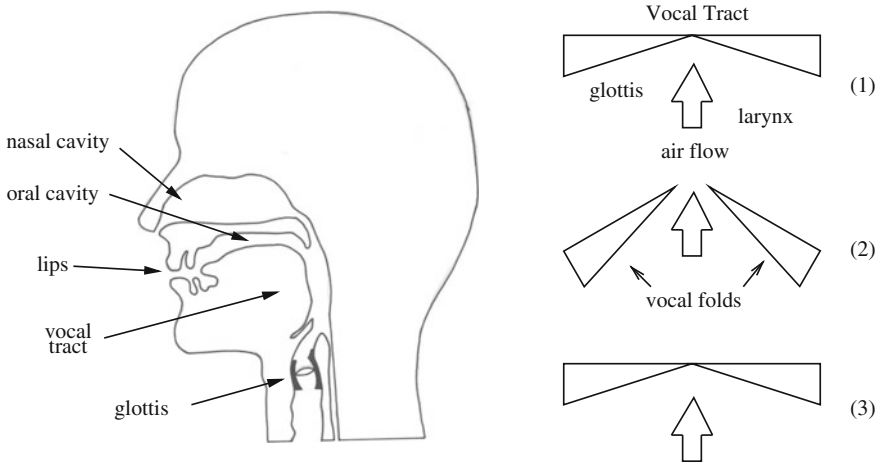

Fig. 2.2 Speech production. The *left* figure shows a sketch of the speech production apparatus (picture by Matthias Dolder); the *right* figure shows the glottal cycle: the air flows increases the pressure below the glottis (**1**), the vocal folds open to reequilibrate the pressure difference between larynx and vocal tract (**2**), once the equilibrium is achieved the vocal folds close again (**3**). The cycle is repeated as long as air is pushed by the lungs

can be considered like sequences of elementary sounds, called *phonemes*, belonging to a finite set that contains, for western languages, 35–40 elements on average and each phoneme is either voiced or unvoiced.

When a voiced phoneme is produced, the vocal folds vibrate following the cycle depicted in Fig. 2.2. When air arrives at the glottis, the pressure difference with respect to the vocal tract increases until the vocal folds are forced to open to reestablish the equilibrium. When this is reached, the vocal folds close again and the cycle is repeated as long as voiced phonemes are produced. The vibration frequency of the vocal folds is a characteristic specific of each individual and it is called *fundamental frequency* $F0$, the single factor that contributes more than anything else to the voice pitch. Moreover, most of the energy in human voices is distributed over the so-called *formants*, i.e. sound components with frequencies that are integer multiples of $F0$ and correspond to the resonances of the vocal tract. Typical $F0$ values range between 60 and 300 Hz for adult men and small children (or adult women) respectively. This means that the first 10–12 formants, on which most of the speech energy is distributed, correspond to less than 4000 Hz. This has important consequences on the human auditory system (see Sect. 2.2.3) as well as on the design of speech acquisition systems (see Sect. 2.3).

The production of unvoiced phonemes does not involve the vibration of the vocal folds. The consequence is that the frequency content of unvoiced phonemes is not as defined and stable as the one of voiced phonemes and that their energy is, on average, lower than that of the others. Examples of voiced phonemes are the vowels and the phonemes corresponding to the first sound in words like *milk* or *lag*, while unvoiced phonemes can be found at the beginning of words *six* and *stop*. As a further example

you can consider the words *son* and *zone* which have phonemes at the beginning where the vocal tract has the same configuration, but in the first case (*son*) the initial phoneme is unvoiced, while it is voiced in the second case. The presence of unvoiced phonemes at the beginning or the end of words can make it difficult to detect their boundaries.

The sounds produced at the glottis level must still pass through the vocal tract where several organs play as *articulators* (e.g. tongue, lips, velum, etc.). The position of such organs is defined *articulators configuration* and it changes the shape of the vocal tract. Depending on the shape, the energy is concentrated on certain frequencies rather than on others. This makes it possible to reconstruct the articulator configuration at a certain moment by detecting the frequencies with the highest energy. Since each phoneme is related to a specific articulator configuration, energy peak tracking, i.e. the detection of highest energy frequencies along a speech recording, enables, in principle, to reconstruct the voiced phoneme sequences and, since most speech phonemes are voiced, the corresponding words. This will be analyzed in more detail in Chap. 12.

2.2.3 Sound Perception

This section shows how the human auditory peripheral system (APS), i.e., what the common language defines as *ears*, detects the frequencies present in incoming sounds and how it reacts to their energies (see Fig. 2.3). The definition *peripheral* comes from the fact that no cognitive functions, performed in the brain, are carried out at its level and its only role is to acquire the information contained in the sounds and to transmit it to the brain. In machine learning terms, the ear is a basic *feature extractor* for the brain. The description provided here is just a sketch and more detailed introductions to the topic can be found in other texts [24, 31].

The APS is composed of three parts called *outer*, *middle* and *inner* ear. The outer ear is the *pinna* that can be observed at both sides of the head. Following recent experiments, the role of the outer ear, considered minor so far, seems to be important in the detection of the sound sources position. The middle ear consists of the auditory channel, roughly 1.3 cm long, which connects the external environment with the inner ear. Although it has such a simple structure, the middle ear has two important properties, the first is that it optimizes the transmission of frequencies between around 500 and 4000 Hz, the second is that it works as an impedance matching mechanism with respect to the inner ear. The first property is important because it makes the APS particularly effective in hearing human voices (see previous section), the second one is important because the inner ear has an acoustic impedance higher than air and all the sounds would be reflected at its entrance without an impedance matching mechanism.

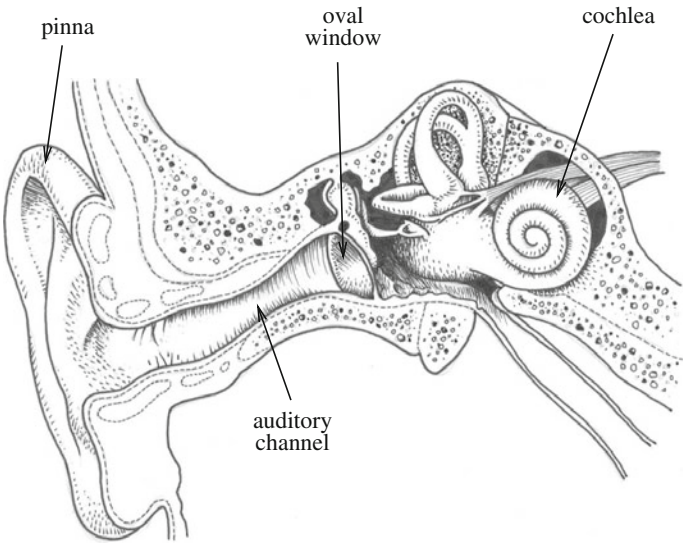

Fig. 2.3 Auditory peripheral system. The peripheral system can be divided into outer (the pinna is the ear part that can be seen on the sides of the head), middle (the channel bringing sounds toward the cochlea) and inner part (the cochlea and the hair cells). Picture by Matthias Dolder

The main organ of the inner ear is the *cochlea*, a bony spiral tube around 3.5 cm long that coils 2.6 times. Incoming sounds penetrate into the cochlea through the *oval window* and propagate along the *basilar membrane* (BM), an elastic membrane that follows the spiral tube from the *base* (in correspondence of the oval window) to the *apex* (at the opposite extreme of the tube). In the presence of incoming sounds, the BM vibrates with an amplitude that changes along the tube. At the base the amplitude is at its minimum and it increases constantly until a maximum is reached, after which point the amplitude decreases quickly so that no more vibrations are observed in the rest of the BM length. The important aspect of such a phenomenon is that the point where the maximum BM displacement is observed depends on the frequency. In other words, the cochlea operates a *frequency-to-place* conversion that associates each frequency f to a specific point of the BM. The frequency that determines a maximum displacement at a certain position is called the *characteristic frequency* for that place. The nerves connected to the external cochlea walls in correspondence of such a point are excited and the information about the presence of f is transmitted to the brain.

The frequency-to-place conversion is modeled in some popular speech processing algorithms through the *critical band analysis*. In such an approach, the cochlea is modeled as a bank of bandpass filters, i.e., as a device composed of several filters stopping all frequencies outside a predefined interval called *critical band* and centered around a *critical frequency* f_j. The problem of finding appropriate f_j values is addressed by selecting frequencies such that the perceptual difference between f_i and f_{i+1} is the same for all i. This condition can be achieved by mapping f onto an

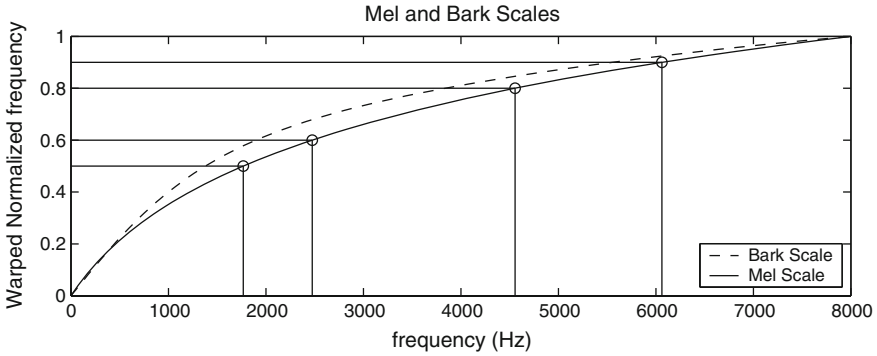

Fig. 2.4 Frequency normalization. Uniform sampling on the vertical axis induces on the horizontal axis frequency intervals more plausible from a perceptual point of view. Frequencies are sampled more densely when they are lower than 4 kHz, the region covered by the human auditory system

appropriate scale $T(f)$ and by selecting frequency values such that $T(f_{i+1}) - T(f_i)$ has the same values for every i. The most popular transforms are the *Bark scale*:

$$b(f) = 13 \cdot arctan(0.00076 f) + 3.5 \cdot arctan\left(\frac{f^2}{7500^2}\right), \tag{2.10}$$

and the *Mel scale*

$$B(f) = 1125 \cdot \ln\left(1 + \frac{f}{700}\right). \tag{2.11}$$

Both above functions are plotted in Fig. 2.4 and have finer resolution at lower frequencies. This means that ears are more sensitive to differences at low frequencies than at high frequencies.

2.3 Audio Acquisition

This section describes the audio *acquisition* process, i.e., the conversion of sound waves, presented in the previous section from a physical and physiological point of view, into a format suitable for machine processing. When the machine is a digital device, e.g. computers and *digital signal processors* (DSP), such a process is referred to as *analog-to-digital* (A/D) conversion because an analogic signal (see below for more details) is transformed into a digital object, e.g., a series of numbers. In general, the A/D conversion is performed by measuring one or more physical effects of a signal at discrete time steps. In the case of the acoustic waves, the physical effect that can be measured more easily is the pressure p in a certain point of the space. Section 2.2 shows that the signal $p(t)$ has the same frequency as the acoustic wave at its origin. Moreover, it shows that the square of the pressure is proportional to the

sound intensity I. In other words, the pressure variations capture the information necessary to fully characterize incoming sounds.

In order to do this, microphones contain an elastic membrane that vibrates when the pressure at its sides is different (this is similar to what happens in the ears where an organ called *eardrum* captures pressure variations). The displacement $s(t)$ at time t of a membrane point with respect to the equilibrium position is proportional to the pressure variations due to incoming sounds, thus it can be used as an indirect measure of p at the same instant t. The result is a signal $s(t)$ which is continuous in time and takes values over a continuous interval $S = [-S_{max}, S_{max}]$. On the other hand, the measurement of $s(t)$ can be performed only at specific instants t_i ($i = 0, 1, 2, \ldots, N$) and no information is available about what happens between t_i and t_{i+1}. Moreover, the displacement measures can be represented only with a finite number B of bits, thus only 2^B numbers are available to represent the non countable values of S. The above problems are called *sampling* and *quantization*, respectively, and have an important influence on the acquisition process. They can be studied separately and are introduced in the following sections.

Extensive descriptions of the acquisition problem can be found in signal processing [23, 29] and speech recognition [15] books.

2.3.1 Sampling and Aliasing

During the sampling process, the displacement of the membrane is measured at regular time steps. The number F of measurements per second is called *sampling frequency* or *sampling rate* and, correspondently, the length $T_c = 1/F$ of the time interval between two consecutive measurements is called *sampling period*. The relationship between the analog signal $s(t)$ and the sampled signal $s[n]$ is as follows:

$$s[n] = s(nT_c) \tag{2.12}$$

where the square brackets are used for sampled discrete-time signals and the parentheses are used for continuous signals (the same notation will be used throughout the rest of this chapter).

As an example, consider a sinusoid $s(t) = A\sin(2\pi ft + \phi)$. After the sampling process, the resulting digital signal is:

$$s[n] = A\sin(2\pi fnT_c + \phi) = A\sin(2\pi f_0 n + \phi) \tag{2.13}$$

where $f_0 = f/F$ is called *normalized frequency* and it corresponds to the number of sinusoid cycles per sampling period. Consider now the infinite set of continuous signals defined as follows:

$$s_k(t) = A\sin(2k\pi Ft + 2\pi ft + \phi) \tag{2.14}$$

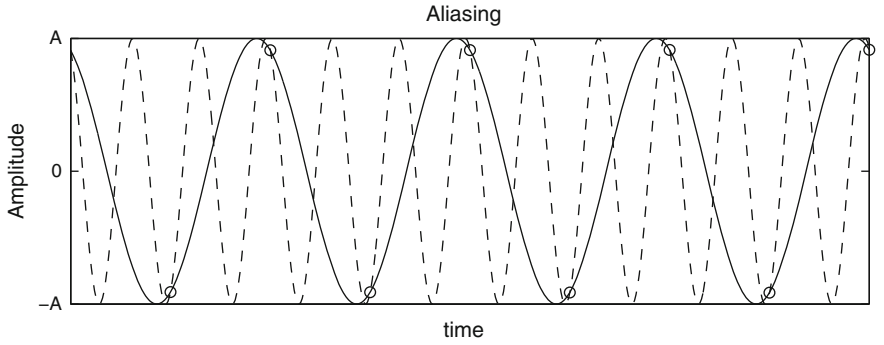

Fig. 2.5 Aliasing. Two sinusoidal signals are sampled at the same rate F and result in the same sequence of points (represented with circles)

where $k \in (0, 1, \ldots, \infty)$, and the corresponding digital signals sampled at frequence F:

$$s_k[n] = A \sin(2k\pi n + 2\pi f_0 n + \phi). \tag{2.15}$$

Since $\sin(\alpha + \beta) = \sin\alpha \cos\beta + \cos\alpha \sin\beta$, the sinus of a multiple of 2π is always null, and the cosine of a multiple of 2π is always 1, the last equation can be rewritten as follows:

$$s_k[n] = A \sin(2\pi f_0 n + \phi) = s[n] \tag{2.16}$$

where $k \in (0, 1, \ldots, \infty)$, then there are infinite sinusoidal functions that are transformed into the same digital signal $s[n]$ through an A/D conversion performed at the same rate F.

Such problem is called *aliasing* and it is depicted in Fig. 2.5 where two sinusoids are shown to pass through the same points at time instants $t_n = nT$. Since every signal emitted from a natural source can be represented as a sum of sinusoids, the aliasing can possibly affect the sampling of any signal $s(t)$. This is a major problem because does not allow a one-to-one mapping between incoming and sampled signals. In other words, different sounds recorded with a microphone can result, once they have been acquired and stored on a computer, into the same digital signal.

However, the problem can be solved by imposing a simple constraint on F. Any acoustic signal $s(t)$ can be represented as a superposition of sinusoidal waves with different frequencies. If f_{max} is the highest frequency represented in $s(t)$, the aliasing can be avoided if:

$$F > 2 f_{max} \tag{2.17}$$

where $2 f_{max}$ is called the *critical frequency*, *Nyquist frequency* or *Shannon frequency*. The inequality is strict; thus the aliasing can still affect the sampling process when $F = 2 f_{max}$. In practice, it is difficult to know the value of f_{max}, then the microphones

apply a low-pass filter that eliminates all frequencies below a certain threshold that corresponds to less than $F/2$. In this way the condition in Eq. (2.17) is met.[1]

The demonstration of the fact that the condition in Eq. (2.17) enables us to avoid the aliasing problem is given in the so-called *sampling theorem*, one of the foundations of signal processing. Its demonstration is given in the next subsection and it requires some deeper mathematical background. However, it is not necessary to know the demonstration to understand the rest of this chapter; thus unexperienced readers can go directly to Sect. 2.3.3 and continue the reading without problems.

2.3.2 The Sampling Theorem**

Aliasing is due to the effect of sampling in the frequency domain. In order to identify the conditions that enable to establish a one-to-one relationship between continuous signals $s(t)$ and corresponding digital sampled sequences $s[n]$, it is thus necessary to investigate the relationship between the Fourier transforms of $s(t)$ and $s[n]$ (see Appendix B).

The FT of $s(t)$ is given by:

$$S_a(j\omega) = \int_{-\infty}^{\infty} s(t)e^{-j\omega t}dt, \tag{2.18}$$

while the FT of the sampled signal is:

$$S_d(e^{j\omega}) = \sum_{n=-\infty}^{\infty} s[n]e^{-j\omega n}. \tag{2.19}$$

However, the above S_d form is not the most suitable to show the relationship with S_a, thus we need to find another expression. The sampling operation can be thought of as the product between the continuous signal $s(t)$ and a *periodic impulse train* (PIT) $p(t)$:

$$p(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT_c), \tag{2.20}$$

where T_c is the sampling period, and $\delta(k) = 1$ for $k = 0$ and $\delta(k) = 0$ otherwise. The result is a signal $s_p(t)$ that can be written as follows:

[1] Since the implementation of a low-pass filter that actually stops all frequencies above a certain threshold is not possible, it is more correct to say that the effects of the aliasing problem are reduced to a level that does not disturb human perception. See [15] for a more extensive description of this issue.

$$s_p(t) = s(t)p(t) = s(t) \sum_{n=-\infty}^{\infty} \delta(t - nT_c). \tag{2.21}$$

The PIT can be expressed as a Fourier series:

$$p(t) = \frac{1}{T_c} \sum_{k=-\infty}^{\infty} e^{j\frac{2\pi}{T_c}kt} = \frac{1}{T_c} \sum_{k=-\infty}^{\infty} e^{j\Omega_{T_c}kt} \tag{2.22}$$

and $s_p(t)$ can thus be reformulated as follows:

$$s_p(t) = \frac{s(t)}{T_c} \sum_{k=-\infty}^{\infty} e^{j\frac{2\pi}{T_c}kt} = \frac{s(t)}{T_c} \sum_{k=-\infty}^{\infty} e^{j\Omega_{T_c}kt}. \tag{2.23}$$

The FT of $s_p(t)$ is thus:

$$S_p(\Omega) = \frac{1}{T_c} \sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} s(t) e^{j\Omega_{T_c}kt - j\Omega t} dt \tag{2.24}$$

and this can be interpreted as an infinite sum of shifted and scaled replicas of the FT of $s(t)$:

$$S_p(j\Omega) = \frac{1}{T_c} \sum_{k=-\infty}^{\infty} S_a(j(\Omega - k\Omega_{T_c})), \tag{2.25}$$

where each term of the sum is shifted by integer multiples of Ω_{T_c} with respect to its neighbors.

The above situation is illustrated in Fig. 2.6. The sampling induces replications of $S_p(j\Omega)$ centered around integer multiples of Ω_{T_c}, in correspondence of the impulses of the PIT Fourier transform. Each replication is $2\Omega_{max}$ wide, where $\Omega_{max} = 2\pi f_{max}$ is the highest angular frequency represented in the original signal $s(t)$. The kth replication of $S_p(j\Omega)$ stops at $\Omega = k\Omega_{T_c} + \Omega_{max}$, while the $(k+1)$th one starts at $(k+1)\Omega_{T_c} - \Omega_{max}$. The condition to avoid overlapping between consecutive replications is thus:

$$\Omega_{T_c} > 2\Omega_{max}. \tag{2.26}$$

Since $\Omega = 2\pi f$, Eq. (2.26) corresponds to:

$$F > 2f_{max}. \tag{2.27}$$

This result is known as *sampling theorem*, and it is typically formulated as follows:

Theorem 2.1 *In order for a band-limited (i.e., one with a zero power spectrum for frequencies $f > f_{max}$) baseband ($f > 0$) signal to be reconstructed fully, it must be sampled at a rate $F \geq 2f_{max}$.*

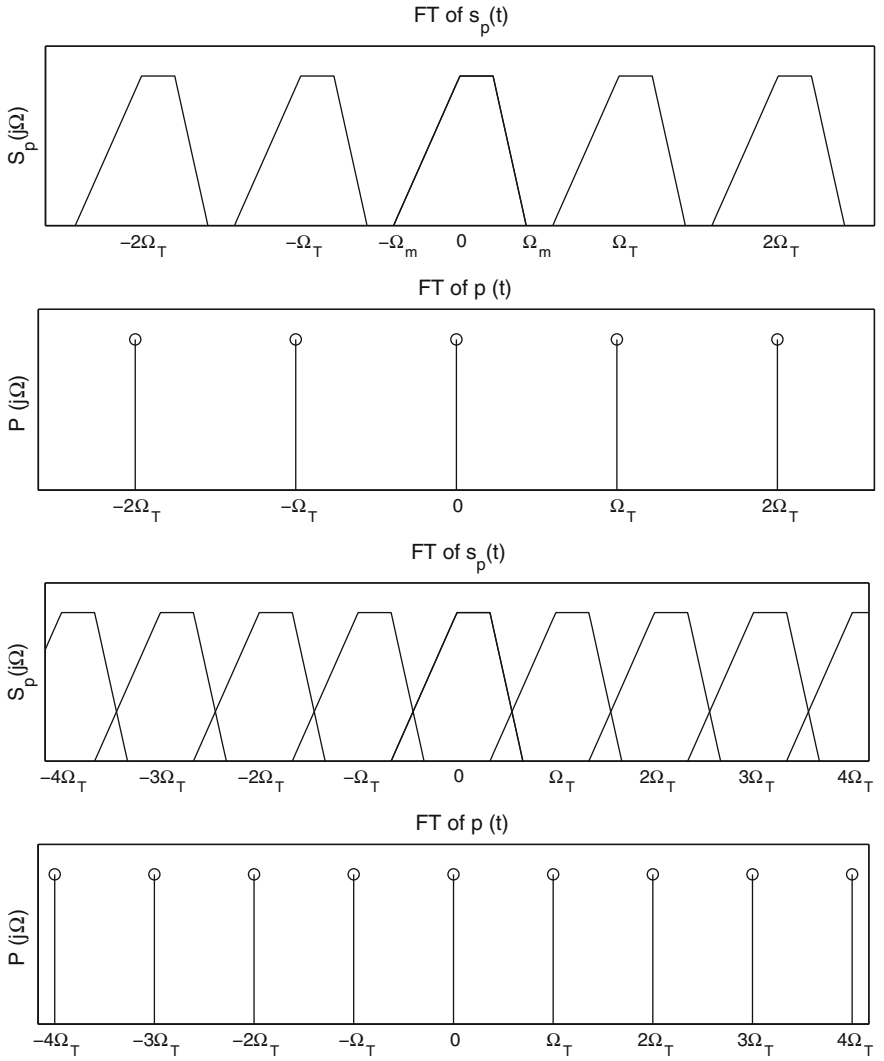

Fig. 2.6 Sampling effect in the frequency domain. The first two plots from above show the sampling effect when $\Omega_{T_c} > 2\Omega_m$. The replications of $S_p(j\Omega)$m, centered around the pulses in $P(j\Omega)$, are separated and the aliasing is avoided. In the third and fourth plot where the distance between the pulses in $P(j\Omega)$ is lower than $2\Omega_m$ and the aliasing takes place

Figure 2.6 shows what happens when the above condition is met (first and second plot from above) and when is not (third and fourth plot from above). Equation (2.26) is important because the overlapping between $S_p(\Omega)$ replications is the frequency domain effect of the aliasing. In other words, the aliasing can be avoided if signals are sampled at a rate F higher or equal than the double of the highest frequency f_{max}.

2.3.3 Linear Quantization

The second problem encountered in the acquisition process is the quantization, i.e., the approximation of a continuous interval of values by a relatively small set of discrete symbols or integer values. In fact, while the $s[n]$ measures range, in general, in a continuous interval $S = [-S_{max}, S_{max}]$, only 2^B discrete values are at disposition when B bits are available in a digital device. This section focuses on linear quantization methods, i.e., on quantization techniques that split the $s[n]$ range into 2^B intervals and represent all the $s[n]$ values lying in one of them with the same number. Other quantization techniques, called *vectorial*, will be described in Chap. 8.

The quantization can be thought of as a process that transforms a sequence of continuous values $s[n]$ into a sequence of discrete values $\hat{s}[n]$. The most straightforward method to perform such a task is the so-called *linear pulse code modulation* (PCM) [28]. The PCM splits the interval S into 2^B uniform intervals of length Δ:

$$\Delta = \frac{S_{max}}{2^{B-1}}. \tag{2.28}$$

Each interval is given a code corresponding to one of the 2^B numbers that can be described with B bits and $\hat{s}[n]$ is obtained in one of the following ways:

$$\begin{aligned} \hat{s}[n] &= sign(c[n])\frac{\Delta}{2} + c[n]\Delta \\ \hat{s}[n] &= c[n]\Delta \end{aligned} \tag{2.29}$$

where $c[n]$ is the code of the interval where $s[n]$ falls. The two equations correspond to the situation depicted in left (*mid-riser quantizer*) and right (*mid-tread quantizer*) plots of Fig. 2.7, respectively.

The use of $\hat{s}[n]$ to represent $s[n]$ introduces an error $\epsilon[n] = s[n] - \hat{s}[n]$. This leads to the use of the Signal to Noise Ratio (SNR) as a performance measure for quantization methods:

$$SNR = 10\log_{10}\left\{\frac{\sum_{n=0}^{M-1} s^2[n]}{\sum_{n=0}^{M-1}(s[n]-\hat{s}[n])^2}\right\} \tag{2.30}$$

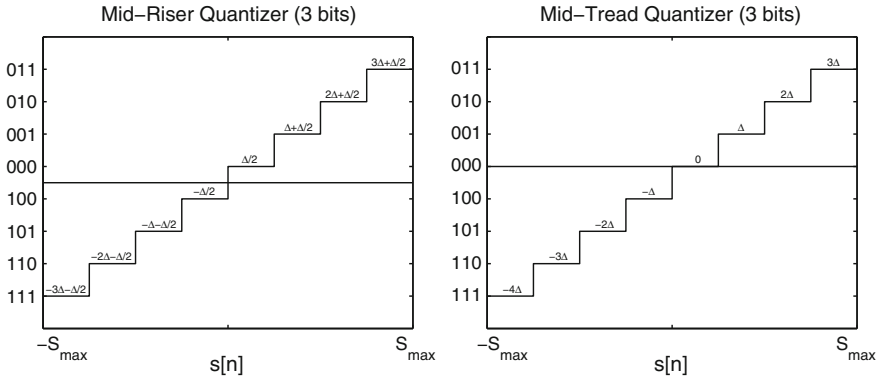

Fig. 2.7 Uniform quantization. The *left* plot shows a mid-riser quantizer, while the *right* plot shows a mid-tread quantizer

where M is the number of samples in the data. Since $\sum_n s^2[n]$ is the energy of a signal (see Sect. 2.5 for more details), the above equation is nothing but the ratio between the energy of the signal and the energy of the noise introduced by the quantization. The use of the logarithm (multiplied by 10) enables to use the dB as a measure unit (see Sect. 2.2). Higher SNR values correspond to better quantization performances because, for a given signal, the energy of the noise becomes smaller when the the values of the differences $s[n] - \hat{s}[n]$ decrease.

The main limit of the SNR is that it might hide temporal variations of the performance. Local deteriorations can be better detected by using short term SNR measures extracted from segments of predefinite length N. The average of local SNR values is called *segmental SNR* (SEGSNR) and it corresponds to the following expression:

$$SEGSNR = \frac{10}{L} \sum_{t=0}^{L-1} \log_{10} \left\{ \frac{\sum_{n=0}^{N-1} s^2[tN+n]}{\sum_{n=0}^{N-1} (s[tN+n] - \hat{s}[tN+n])^2} \right\} \tag{2.31}$$

where L is the number of N long segments spanning the M samples of the signal. The SEGSNR tends to penalize encoders with different performance for different signal energy and frequency ranges.

In the case of the PCM, the upper bound of $\epsilon[n]$ is Δ; in fact the maximum value that the difference $s[n] - \hat{s}[n]$ can assume is the length of the interval where $s[n]$ falls. The lower bound of the SNR is thus:

$$SNR_{PCM} = 10 \log_{10} \left\{ \frac{1}{\Delta^2} \sum_{n=0}^{M-1} s^2[n] \right\}. \tag{2.32}$$

The above expression shows the main limits of the PCM: if the SNR of lower energy signals decreases to a point that the perceptual quality of the quantized signal becomes unacceptable, the only way to improve the quantization performance is to reduce Δ,

i.e. to increase the number of bits B. On the other hand, it can happen that the same Δ value that makes unacceptable the perceptual quality for lower-energy signals can be tolerated in the case of higher-energy sounds. For the latter, an increase of B is thus not necessary and it leads to an improvement of the SNR that goes beyond the human ear sensibility. This is not desirable, because the number of bits must be kept as low as possible in order to reduce the amount of memory necessary to store the data as well as the amount of bits that must be transmitted through a line.

The solutions proposed to address such a problem are based on the fact that the SNR is a ratio and can be kept constant by adapting the quantization error $\epsilon[n]$ to the energy of the signal for any sample n. In other words, the SNR is kept at an acceptable level for all energy values by allowing higher quantization errors for higher-energy signals. Such an approach is used in differential PCM (DPCM), delta modulation (DM) and adaptive DPCM (ADPCM) [10]. However, satisfactory results can be obtained with two simple variants of the PCM that simply use a non uniform quantization interval. The variants, known as μ-law and A-law PCM, are currently applied in telecommunications and are described in the next section.

2.3.4 Nonuniform Scalar Quantization

The previous section has shown that the SNR value can be kept constant at different energies by adapting the quantization error $\epsilon[n]$ to the signal energy: the higher the energy of the signal, the higher the value of the quantization error that can be tolerated. This section shows how such a result can be obtained through functions called *logarithmic companders* and describes two quantization techniques based on such an approach and commonly applied in telecommunications: μ-law and A-law PCM.

A logarithmic compander is a function that uses a logarithm to compress part of the domain where it is defined:

$$y[n] = \ln(|s[n]|)sign(s[n]), \tag{2.33}$$

where $y[n] \in Y = [-\ln(S_{max}), \ln(S_{max})]$, $sign(x) = 1$ when $x \geq 0$ and $sign(x) = -1$ when $x < 0$ (see Sect. 2.3.3 for the meaning of symbols). If the uniform quantization is performed over Y (the vertical axis of Fig. 2.8), then $\hat{y}[n] - y[n] = \epsilon[n]$ and:

$$\hat{s}[n] = \exp(y[n])sign(s[n]) = s[n]\exp(\epsilon[n]) \tag{2.34}$$

Since Y is quantized uniformly, $\epsilon[n]$ can be approximated with the length Δ_Y of the quantization interval. When $\epsilon[n] \to 0$, the above equation can be rewritten as follows using a Taylor series expansion:

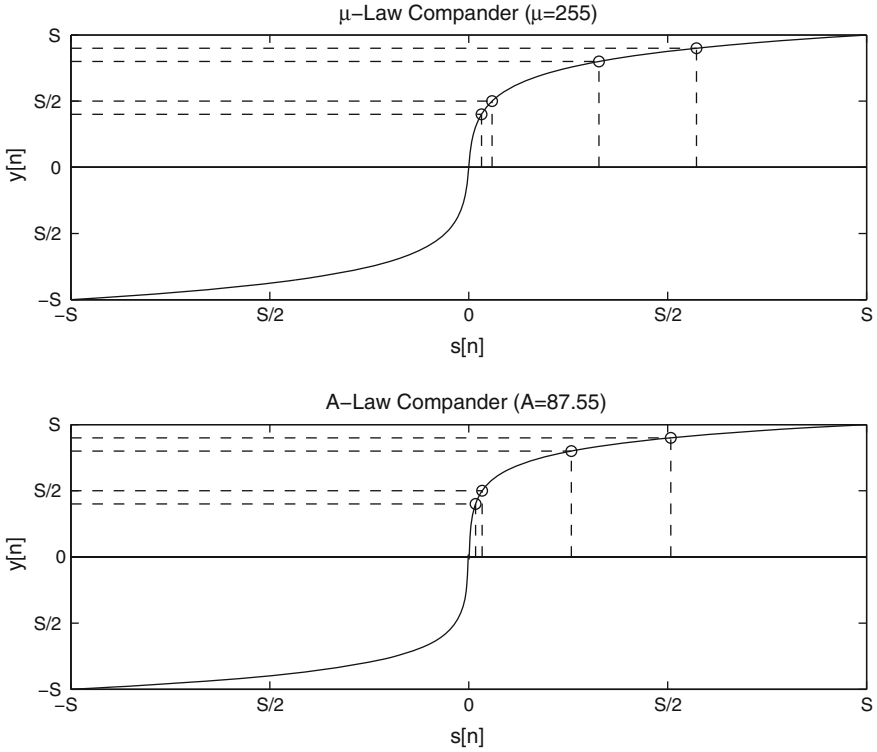

Fig. 2.8 Nonuniform quantization. The logarithmic companders induce finer quantization on lower-energy signals. Intervals with the same width on the vertical axis correspond to intervals with different width on the horizontal axis

$$\hat{s}[n] \simeq s[n](1 + \epsilon[n]) \tag{2.35}$$

and the expression of the SNR (see Eq. (2.30)) for the logarithmic compander corresponds to

$$SNR_{log} = \sum_{n=0}^{M-1} \frac{1}{\Delta_Y^2} = \frac{M}{\Delta_Y^2}; \tag{2.36}$$

thus, for a given signal length, SNR_{log} does not depend on the energy. This happens because the uniform quantization of Y induces a nonuniform quantization on S such that the quantization step is proportional to the signal energy. When the energy of the signal increases, the quantization error is increased as well and the SNR of Eq. (2.30) is kept constant.

The compander in Eq. (2.33) brings to the above effect only when $\epsilon[n] \to 0$, but this is not possible for real applications. For this reason two variants are used in real applications[2]:

$$y[n] = S_{max} \frac{\log\left(1 + \mu \frac{|s[n]|}{S_{max}}\right)}{\log(1+\mu)} sign(s[n]) \tag{2.37}$$

which is called the μ-law and

$$y[n] = \begin{cases} S_{max} \frac{A \frac{|s[n]|}{S_{max}}}{1+\log A} sign(s[n]); & 0 < \frac{|s[n]|}{S_{max}} < \frac{1}{A} \\ S_{max} \frac{1+\log\left(A \frac{|s[n]|}{S_{max}}\right)}{1+\log A} sign(s[n]); & \frac{1}{A} < \frac{|s[n]|}{S_{max}} < 1 \end{cases} \tag{2.38}$$

which is called the A-law. It can be demonstrated that both above quantizers lead to an SNR independent of the signal energy.

In telephone communications, an SNR of around 35 dB is considered acceptable. While a uniform quantizer requires 12 bits to guarantee such an SNR all over the energy spectrum, A-law and μ-law can achieve the same result by using only 8 bits [36]. For this reason, the above nonuniform quantization techniques are recommended by the *International Communications Union* and are applied to transmit speech through telephone networks [15].

2.4 Audio Encoding and Storage Formats

The number B of bits used to represent audio samples plays an important role in transmission and storage problems. In fact, the higher is B, the bigger is the amount of data to be transmitted through a channel and the larger is the memory space needed to store a recording. The amount of bits per time unit necessary to represent a signal is called *bit-rate* and it must be kept as low as possible to respect application constraints such as bandwidth and memory. On the other hand, a reduction of the bit-rate is likely to degradate the perceptual quality of the data and this, beyond a certain limit, is not tolerated by users (Sect. 2.3 shows that the reduction of B decreases the SNR of audio acquisition systems). The domain targeting techniques capable of reducing the bit-rate while still preserving a good perceptual quality is called *audio encoding*.

The main encoding methods result in audio *formats* (e.g. *MPEG*, *WAV*, *mp3*, etc.), i.e. into standardized ways of representing and organizing audio data inside files that can be used by computer applications. For this reason, this section presents not only encoding technologies, but also audio formats that make use of them. In particular, it

[2]There is no noticeable difference between the performance of the two companders, the A-law compander is used in Europe and other countries affiliated to the ITU (with $A = 87.56$), while the μ-law compander is mostly used in the USA (with $\mu = 255$).

will be shown how the development of new encoding methods and the definition of new formats is typically driven by two main factors: the first is the emergence of new applications that have bit-rate constraints tighter than the previous ones, the second is the expectation of users that accept different perceptual qualities depending on the applications.

The encoding problem is the subject of monographies [5] and tutorials [30, 36] that provide extensive introductions to the different algorithms and formats. For the MPEG audio format and coding technique, both tutorial level [4, 7, 27] articles and monographies [22] are available.

2.4.1 Linear PCM and Compact Discs

The earliest encoding approach is the linear PCM presented in Sect. 2.3. Although simple, such a technique is the most expensive in terms of bit-rate (see below) and the most effective for what concerns perceptual quality. Since it reproduces the whole information contained in the original waveform, the linear PCM is said *lossless*, in opposition to *lossy* approaches that discard selectively part of the original signal (see the rest of this section for more detail). In general, the samples are represented with $B = 16$ bits because this makes the quantization error small enough to be inaudible even by trained listeners (the so-called *golden ears* [30]). The sampling frequency commonly used for high-fidelity audio is $F = 44.1$ kHz and this leads to a bit rate of $2BF = 1,411,200$ bits per second. The factor 2 accounts for the two aural channels in a stereo recording.

Although high, such a bit-rate could be accommodated on the first supports capable of storing digital audio signals, i.e. digital audio tapes (DAT) and compact discs (CD). These last in particular started to spread in the early eighties, although invented in the sixties, and they are now, together with CD players, some of the most important consumer electronic products. One hour of high fidelity stereo sound at the 16-bit PCM rate requires roughly 635 MB. A CD can actually store around 750 MB, but the difference is needed for *error correction bits*, i.e., data required to recover acquisition errors. Since CDs have been used mainly to replace old vinyl recordings that were often shorter, the one-hour limit was largely accepted by users, and still is. For this reason, there was no pressure to decrease the PCM bit-rate in order to store more sound on CDs. At the same time, the perceptual improvement determined by the use of digital rather than analogic supports was so high, that the user expectations increased significantly and the CD-quality is currently used as a reference for any other encoding technique [27].

The linear PCM is the basis for several other formats that are used in conditions where the memory space is not a major problem: Windows *WAV*, Apple *AIFF* and Sun *AU*. In fact, such formats, with different values of B and F, are used to store sound on hard disks that are today large enough to contain hours of recordings and that promise to grow at a rate that makes the space constraint marginal.

The same does not apply to telephone communications where a high bit-rate results into an uneffective use of the lines. For this reason, the first efforts in reducing the bit-rate came from that domain. On the other hand, the development of encoding techniques for phone communications has an important advantage: since consumers are used to the fact that the so-called *telephone speech* is not as natural as in other applications (e.g. radio and television), their expectations are significantly lower and the bit-rate can be reduced with simple modifications of the linear PCM.

Section 2.3 shows that the main limit of the linear PCM is that the quantization error does not change with the signal energy. In this way, the parameter B must be kept at a level that leads to an SNR acceptable at low energies, but high beyond human earing sensibility at higher energies. In other words, there is a waste of bits at higher energies. The A-law and μ-law logarithmic companders address such a problem by adapting the quantization errors to the amplitude of the signals and reduce by roughly one third the bit-rate necessary to achieve a certain perceptual quality. For this reason the logarithmic companders are currently adviced by the *International Telecommunications Union* (ITU) and are widely applied with $A = 87.55$ and $\mu = 255$.

One of the most important lessons in the phone case, is that user expectations are not directed towards the highest possible quality, but simply at keeping constant the perceptual level in a given application. For this reason, the performance of an encoder is measured not only with the SNR, but also with the *mean opinion score* (MOS), a subjective test involving several *naïve* listeners, i.e., people that do not know encoding technologies (this might bias their evaluations). Each listener is asked to give a score between 1 (bad) and 5 (excellent) to a given encoded sound and the resulting MOS value is the average of all judgments given by the assessors. An MOS of 4.0 or more defines *good* or *toll* quality where the encoded signal cannot be distinguished from the original one. An MOS between 3.5 and 4.0 is considered acceptable for telephone communications [15]. The test can be performed unformally, but the results are accepted in the official organizations only if they respect the rigorous protocols given by the ITU [1].

2.4.2 MPEG Digital Audio Coding

Logarithmic companders and other approaches based on the adaptation of the noise to the signal energy (see Sect. 2.3) obtain significant reductions of the bit-rate. However, these are not sufficient to respect bandwidth and space constraints imposed by applications developed in the last years. Multimedia, streaming, online applications, content diffusion on cellular phones, wireless transmission, etc. require to go beyond the reduction by one-third achieved with A-law and μ-law encoding techniques. Moreover, user expectations correspond now to CD-like quality and any degradation with respect to such a perceptual level would not be accepted. For this reason, several efforts were made in the last decade to improve encoding approaches.

Table 2.1 MPEG audio layers. This table reports bit-rates (central column) and compression rates (right column), compared to CD bit-rate, achieved at different layers in the MPEG coding architecture

Layer	Bit-rate	Compression
I	384 kb/s	4
II	192 kb/s	8
III	128 kb/s	12

The compression rate is the ratio between CD and MPEG bit-rate at the same audio quality level

MPEG is the standard for multimedia (see Chap. 3), its digital audio coding technique is one of the major results in audio coding and it involves several major changes with respect to the linear PCM. The first is that the MPEG architecture is organized in *Layers* containing sets of algorithms of increasing complexity. Table 2.1 shows the bit-rates achieved at each layer and the corresponding compression rates with respect to the 16-bit linear PCM.

The second important change is the application of an *analysis and synthesis* approach implemented in layers I and II. This consists in representing the incoming signals with a set of compact parameters, in the case of sound frequencies, which can be extracted in the encoding phase and used to reconstruct the signal in the following decoding step (for a detailed description of the algorithms of the first two layers, see [30]). An average MOS of 4.7 and 4.8 has been reported for monaural layer I and II codecs operating at 192 and 128 kb/s [26].

The third major novelty is the application of psychoacoustic principles capable of identifying and discarding *perceptually irrelevant* frequencies in the signal. By perceptually irrelevant it is meant that a frequency cannot be perceived by human ears even if it is present in the signal, thus it can be discarded without degradation of the perceptual quality. Such an approach is called *perceptual coding* and, since part of the original signal is removed, the encoding approach is defined *lossy*. The application of the psychoacoustic principles is performed at layer III and it reduces by 12 the bit-rate of the linear PCM while achieving an average MOS between 3.1 and 3.7 [26]. The *MPEG* layer III is commonly called *mp3* and it is used extensively on the web because of its high compression rate (see Table 2.1). In fact, the good tradeoff between perceptual quality and size makes the *mp3* files easy to download and exchange. The format is now so popular that it gives the name to a new class of products, i.e. the *mp3 players*.

The main improvements of the *mp3* with respect to previous formats come from the application of perceptual coding. Section 2.4.4 provides a description of the main psychoacoustic phenomena used in *mp3*.

2.4.3 AAC Digital Audio Coding

The acronym AAC stands for *advanced audio coding* and the corresponding encoding technique is considered as the natural successor of the mp3 (see the previous

section) [30]. The structures of mp3 and AAC are similar, but the latter improves some of the algorithms included in the different layers.

AAC contains two major improvements with respect to mp3. The first is the higher adaptivity with respect to the characteristics of the audio. Different analysis windows (see Sect. 2.5) are used when the incoming sound has frequencies concentrated in a narrow interval or when strong components are separated by more than 220 Hz. The result is that the perceptual coding gain is maximized, i.e. most of the bits are allocated for perceptually relevant sound parts. The second improvement is the use of a predictor for the quantized spectrum. Some audio signals are relatively stationary and the same spectrum can be used for subsequent analysis frames (see Sect. 2.5). When several contiguous frames use the same spectrum, this must be encoded only the first time and, as a consequence, the bit-rate is reduced. The predictor is capable of deciding in advance wheather the next frame requires to compute a new spectrum or not.

In order to serve different needs, the AAC provides three profiles of decreasing complexity: the main profile offers the highest quality, the low-complexity profile does not include the predictor and the sampling-rate-scaleable profile has the lowest complexity (see [27] for details about each profile). The main profile AAC has shown higher performance the other formats in several comparisons[3]: at a bit-rate of 128 kb/s, listeners cannot distinguish between original and coded stereo sound. If the bit-rate is decreased at 96 kb/s, AAC has a quality higher than mp3 at 128 kb/s. On the other hand, if both AAC and mp3 have a bit-rate of 128 kb/s, the AAC shows a significantly superior performance.

2.4.4 Perceptual Coding

The main issue in perceptual coding is the identification of the frequencies that must be coded to preserve perceptual quality or, conversely, of the frequencies that can be discarded and for which no bits must be allocated. The selection, in both above senses, is based on three psychoacoustic phenomena: the existence of critical bands, the absolute threshold of hearing (TOH) and the masking. Critical band analysis has been introduced at the end of Sect. 2.2, the other two phenomena are briefly described in the following.

Section 2.2 defines the TOH as the lowest energy that a signal must carry to be heard by humans (corresponding to an intensity $I_0 = 10^{12}$ W/m). This suggests as a first frequency removal criterion that any spectral component with an energy lower than the TOH should not be coded. However, perceptual experiments have shown that the above TOH does not apply to any frequency and that the minimum audible energy is a function of f [12]:

[3]The results can be found on `www.apple.com/quicktime/technologies/aac/`.

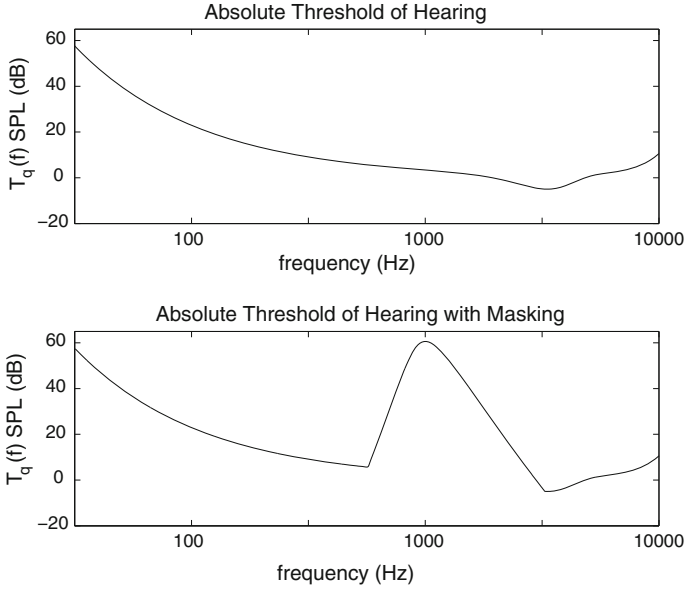

Fig. 2.9 Absolute TOH. The TOH is plotted on a logarithmic scale and shows how the energy necessary to hear frequencies between 50 and 4000 kHz is significantly lower than the energy needed for other frequencies

$$T_q(f) = 3.64\left(\frac{f}{10^3}\right)^{-0.8} - 6.5e^{-0.6(\frac{f}{10^3}-3.3)^2} + 10^{-3}\left(\frac{f}{10^3}\right)^4 (dBSPL). \quad (2.39)$$

The function $T_q(f)$ is referred to as *absolute* TOH and it enables to achieve better bit-rate reduction by removing any spectral component with energy $E_0 < T_q(f_0)$. Absolute TOH is plotted in Fig. 2.9, the lowest energy values correspond to frequencies ranging between 50 and 4000 Hz, not surprisingly those that propagate better through the middle ear (see Sect. 2.2). The main limit of the $T_q(f)$ introduced above is that it applies only to pure tones in noiseless environments, while sounds in everyday life have a more complex structure. In principle, it is possible to decompose any complex signal into a sum of waves with a single frequency f_0 and to remove those with energy lower than $T_q(f_0)$, but this does not take into account the fact that the perception of different frequencies is not independent.

In particular, components with a certain frequency can stop the perception of other frequencies in the auditory system. Such an effect is called *masking* and it modifies significantly the curve in Fig. 2.9. The waves with a given frequency f excite the auditory nerves in the region where they reach their maximum amplitude (the nerves are connected to the cochlea walls). When two waves of similar frequency occur together and their frequency is around the center of a critical band (see Sect. 2.2), the excitation induced by one of them can prevent from hearing the other. In other words, one of the two sounds (called *masker*) masks the other one (called *maskee*). From

an encoding point of view, this is important because no bits accounting for maskee frequencies need to be allocated in order to preserve good perceptual quality. The inclusion of masking in audio encoding is a complex process (see [30] for a detailed description for application in MPEG coding). For the sake of simplicity, we will show only how masker and maskee frequencies are identified in the two most common cases: tone masking noise (TMN) and noise masking tone (NMT).

The first step is to find tone and noise frequencies. The f values corresponding to masker tones are identified as peaks in the power spectrum with a difference of at least 7 Barks with respect to neighboring peaks. Noise maskers are detected through the geometric mean of frequencies represented between to consecutives tonal maskers. TMN takes place when noise masks tones with lower energy. Empirical models show that this happens when the difference between tone and noise energies is below a threshold $T_T(b)$ that can be calculated as follows:

$$T_T(b) = E_N - 6.025 - 0.275 \cdot g + S_m(b-g) \tag{2.40}$$

where b and g are the Bark frequencies of tone and noise, respectively, E_N is the noise energy and $S_m(h)$ is the *spread of masking* function given by

$$S_m(h) = 15.81 + 7.5 \cdot (h + 0.474) - 17.5\sqrt{1 + (h + 0.474)^2} \tag{2.41}$$

where h is the Bark frequency difference between noise and tone. The expression of the threshold for the NMT is similar:

$$T_N(b) = E_T - 2.025 - 0.175 \cdot g + S_m(b-g) \tag{2.42}$$

where E_T is the tone energy. Although Eqs. (2.40) and (2.42) seem to be symmetric, there is an important difference between TMN and NMT: in the first case only tones with signal-to-mask ratio (SMR) between –5 and 5 dB can be masked, while in the second case the SMR range where the masking takes place is between 21 and 28 dB. A tone can thus mask noise with energies roughly 100 to 1,000 times higher, while a noise can mask tones with energies from around one-third to three times its energy. The lower plot in Fig. 2.9 shows the effect of a masking tone noise of frequency 1 kHz and energy 69 dB. The energy necessary to hear frequencies close to 1 kHz is significantly higher than the corresponding TOH and this enables to reduce the number of bits necessary to encode the frequency region where masking takes place.

2.5 Time-Domain Audio Processing

The result of the acquisition process is a sequence of quantized physical measures $\{s[n]\} = (s[1], s[2], \ldots, s[N])$. Since both n and $s[n]$ are discrete, such sequences are referred to as *digital signals* and their form is particularly suitable for computer processing. This section presents some techniques that extract useful

information from the analysis of the variations across the sequences. The corpus of such techniques is called *time-domain audio processing* in opposition to *frequency-domain* techniques which operate on frequency distributions (see Appendix B for more details).

After presenting the fundamental notion of *system* and related properties, the rest of this section focuses on how to extract information related to energy and frequency. The subject of this section is covered in more detail in several speech and signal processing texts [15, 23, 33].

2.5.1 Linear and Time-Invariant Systems

Any operator T mapping a sequence $s[n]$ into another digital signal $y[n]$ is called *discrete-time system*:

$$y[n] = T\{s[n]\}, \tag{2.43}$$

the element $y[n]$ is a function of a single sample $s[n]$, of a subset of the samples of $\{s[n]\}$ or of the whole input digital signal $\{s[n]\}$. In the following, we show three examples corresponding to each of these situations: The *ideal delay* (function of a single sample), the *moving average* (function of a subset), and the *convolution* (function of the whole signal).

The *ideal delay* system is as follows:

$$y[n] = s[n - n_0] \tag{2.44}$$

where n_0 is an integer constant and $y[n]$ is function of the the only sample $s[n - n_0]$. The *moving average* is:

$$y[n] = \frac{1}{K_1 + K_2 + 1} \sum_{k=-K_1}^{K_2} s[k] \tag{2.45}$$

where K_1 and K_2 are two integer constants and $y[n] = T\{s[n]\}$ is function of the samples in the interval between $n - K_2$ and $n + K_1$. The expression of the convolution is:

$$y[n] = \sum_{k=-\infty}^{\infty} s[k]w[n - k] \tag{2.46}$$

where $w[n]$ is another digital signal and $y[n]$ is a function of the whole sequence $\{s[n]\}$.

A system is said *linear* when it has the following properties:

$$\begin{aligned} T\{s_1[n] + s_2[n]\} &= T\{s_1[n]\} + T\{s_2[n]\} \\ T\{as[n]\} &= aT\{s[n]\} \end{aligned} \tag{2.47}$$

where $s_1[n]$ and $s_2[n]$ are two different digital signals and a is a constant. The first property is called *additivity* and the second *homogeneity* or *scaling*. The two properties can be combined into the so-called *superposition principle*:

$$T\{as_1[n] + bs_2[n]\} = aT\{s_1[n]\} + bT\{s_2[n]\}. \tag{2.48}$$

Given a signal $\hat{s}[n] = s[n - n_0]$, a system is said to be *time invariant* when:

$$\hat{y}[n] = T\{\hat{s}[n]\} = y[n - n_0]. \tag{2.49}$$

The above equation means that a shift of the origin in the input digital signal determines the same shift in the output sequence. In other words, the effect of the system at a certain point of the sequence does not depend on the sample where T starts to operate.

When a system is LTI, i.e., both linear and time-invariant, the output sequence $y[n]$ can be obtained in a peculiar way. Consider the so-called *impulse*, i.e., a digital signal $\delta[n]$ such that $\delta[n] = 1$ for $k = 0$ and $\delta[n] = 0$ otherwise, the output of a system can be written as follows:

$$y[n] = T\left\{\sum_{k=-\infty}^{\infty} s[k]\delta[n-k]\right\} = \sum_{k=-\infty}^{\infty} s[k]T\{\delta[n-k]\}, \tag{2.50}$$

and the above equation can be rewritten as:

$$y[n] = \sum_{k=-\infty}^{\infty} s[k]h[n-k] \tag{2.51}$$

which corresponds to the convolution between the input signal $s[n]$ and $h[n-k]$, i.e., the response of the system to an impulse at time n. As a consequence, an LTI system is completely determined by its impulse response $h[n]$, in the sense that $h[n]$ can be used to obtain $y[n]$ for any other input signal $s[n]$ through a convolution operation $s[n] * h[n]$.[4]

2.5.2 Short-Term Analysis

Figure 2.14 shows a speech waveform sampled at 8 kHz. Such a value of F is common for spoken data because the highest formant frequencies in the human voice are

[4]The advantages of this property are particularly evident in the frequency domain. In fact, the Fourier transform of a convolution between two signals corresponds to the product between the Fourier transforms of the single signals, and this simplifies significantly the analysis of the effect of a system in the frequency domain.

around 4 kHz (see Sect. 2.2) and the lowest point of the absolute TOH curve for the human auditory system corresponds roughly to such frequency (see Fig. 2.9). Speech data are thus low-pass filtered at 4 kHz and sampled at 8 kHz to meet the sampling theorem conditions. The waveform of Fig. 2.14 shows two important aspects: the first is that different segments of the signal have different properties (e.g., speech and silence), the second is that the signal properties change relatively slowly, i.e. they are stable if an interval short enough is taken into account (e.g. 20–30 ms). Such assumptions underly the *short-term analysis*, an approach which takes into account segments short enough to be considered as sustained sounds with stable properties.

In mathematical terms this means that the value of the property $Q[n]$ at time nT, where $T = 1/F$ is the sampling period, can be expressed as follows:

$$Q[n] = \sum_{m=-\infty}^{\infty} K(s[m])w[n-m] \tag{2.52}$$

where K is a transform, either linear or nonlinear, possibly dependent upon a set of adjustable parameters, and $w[n]$ is the so-called analysis *window*. Two analysis windows are commonly applied: the first is called *rectangular* and the second is called *Hamming*. The latter has been introduced to avoid the main problems determined by the rectangular window, i.e., the presence of too high secondary lobes in the Fourier transform (see Appendix B). The *rectangular* window is defined as follows:

$$w[n] = \begin{cases} 1 : 0 \leq n \leq N-1 \\ 0 : n < 0 \\ 0 : n \geq N \end{cases}$$

and the Hamming window:

$$w[n] = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi n}{N-1}) & : 0 \leq n \leq N-1 \\ 0 & : n < 0 \\ 0 & : n \geq N. \end{cases}$$

In both above cases, as well as for any finite window, it is necessary to set the parameter N, the so-called *window length*. The value of N must be the tradeoff between two conflicting requirements: the first is that the window must be short enough to detect rapid changes of Q, the second is that it must be long enough to smooth local random fluctuations. Moreover, no window length gives satisfactory results for every application and different choices must be made for different tasks. In the case of spoken data, it is common to have a window corresponding to few fundamental periods $T0 = 1/F0$, where $F0$ is the fundamental frequency (see Sect. 2.2). In more general terms, the problem is addressed by observing that the variations of Q can be studied through the Fourier transform (FT) of $Q[n]$ (the unexperienced reader can move directly to Sect. 2.5.3). In this case high frequencies in the spectrum correspond to rapid Q variations, while low frequencies components are due to slow changes.

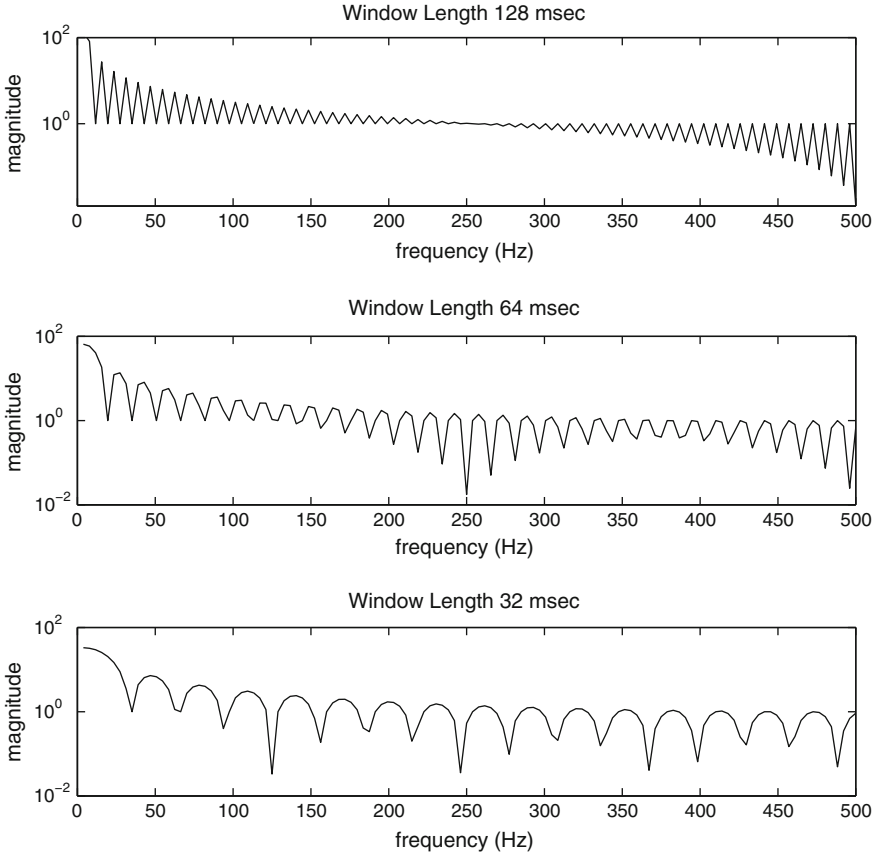

Fig. 2.10 Window effect in the frequency domain. The three plots show the spectrum of rectangular windows of length 128, 64 and 32 ms, respectively. All spectra show a first minimum in correspondence of $f_r = F/\Delta t$ Hz, where Δt is the length of the window. This means that variations of frequency higher than f_r are filtered and that longer windows tend to smooth higher frequency variations (and vice versa)

Since Eq. (2.52) can be interpreted as a discrete convolution, the FT of $Q[n]$ can be obtained as a product of the FT's of $K(s[n])$ and $w[n]$. The effect of N on the frequency with which Q changes can thus be evaluated through the FT of the window. Figure 2.10 shows the spectra of rectangular windows of different length. The windows act as a low-pass filters with cutoff frequencies $f_r = F/N$ ($f_h = 2F/N$ for the Hamming windows). The consequence is that the longer is the window, the narrower is the band of accepted frequencies. In other words, long windows tend to mask rapid changes and vice versa for short windows. In speech recognition (see Chap. 12) the window is typically 10–30 ms long. The reason is that physiological measurements performed using X-rays have shown that during such a time humans cannot significantly change the shape of the vocal tract.

2.5.3 Time-Domain Measures

This section presents the most important properties that can be extracted from a signal in the time domain. All of the properties are obtained with a short-term approach and provide a rough but meaningful representation of the audio signals (particular attention will be paid to speech data).

The first two properties are short-time *energy* and *average magnitude*. They carry the same kind of information, but the second one is less sensitive to local fluctuations. They are especially important to detect silences or to distinguish between voiced and unvoiced segments in spoken data, but they also play a role for the reduction of the bit-rate during the quantization. In fact, higher quantization errors can be allowed for higher energy signals (see Sect. 2.3). The short-time energy $E[n]$ of a signal can be extracted through the following convolution:

$$E[n] = \sum_{m=-\infty}^{\infty} s^2[n]w[n-m]. \tag{2.53}$$

The use of the square makes $E[n]$ too sensitive to the highest values of $s[n]$ that can be due to local random fluctuations. Moreover, the lowest energy parts of the signal tend to be suppressed as it can be observed in Fig. 2.14: the energy of the unvoiced phonemes at the end of the word *six* is so much lower than the other parts of the words that it can be difficult to distinguish them with respect to the silence. For this reason, $E[n]$ is often replaced with the short-term average magnitude $M[n]$:

$$M[n] = \sum_{m=-\infty}^{\infty} |s[n]w[n-m]|. \tag{2.54}$$

The dynamic range of $M[n]$ is smaller and the differences are smoother than in the $E[n]$ case. This can be seen at the end of the word *six* in Fig. 2.11 where the unvoiced phonemes have an average magnitude lower, but still comparable with the $M[n]$ value of voiced phonemes.

The length of the window should correspond more or less to a pitch period (see Sect. 2.2). Shorter windows detect uninteresting local fluctuations, while longer windows miss changes that should not be neglected. Since the pitch of human voices ranges between 50 (for male voices) and 400 kHz (for small children and women), no window length is optimal for any case, However, satisfactory results can be achieved, on average, with a 20–30 ms long analysis frame. Energy and magnitude are often used as features in speech recognition systems [15] as well as in multimedia content analysis where they have been applied to detect emotional states [18], to identify audio segments likely to attract the attention [20], to perform affective analysis [14].

Another important aspect of a signal is the frequency content. This is typically obtained through the Fourier transform (see Appendix B), but a simple time domain

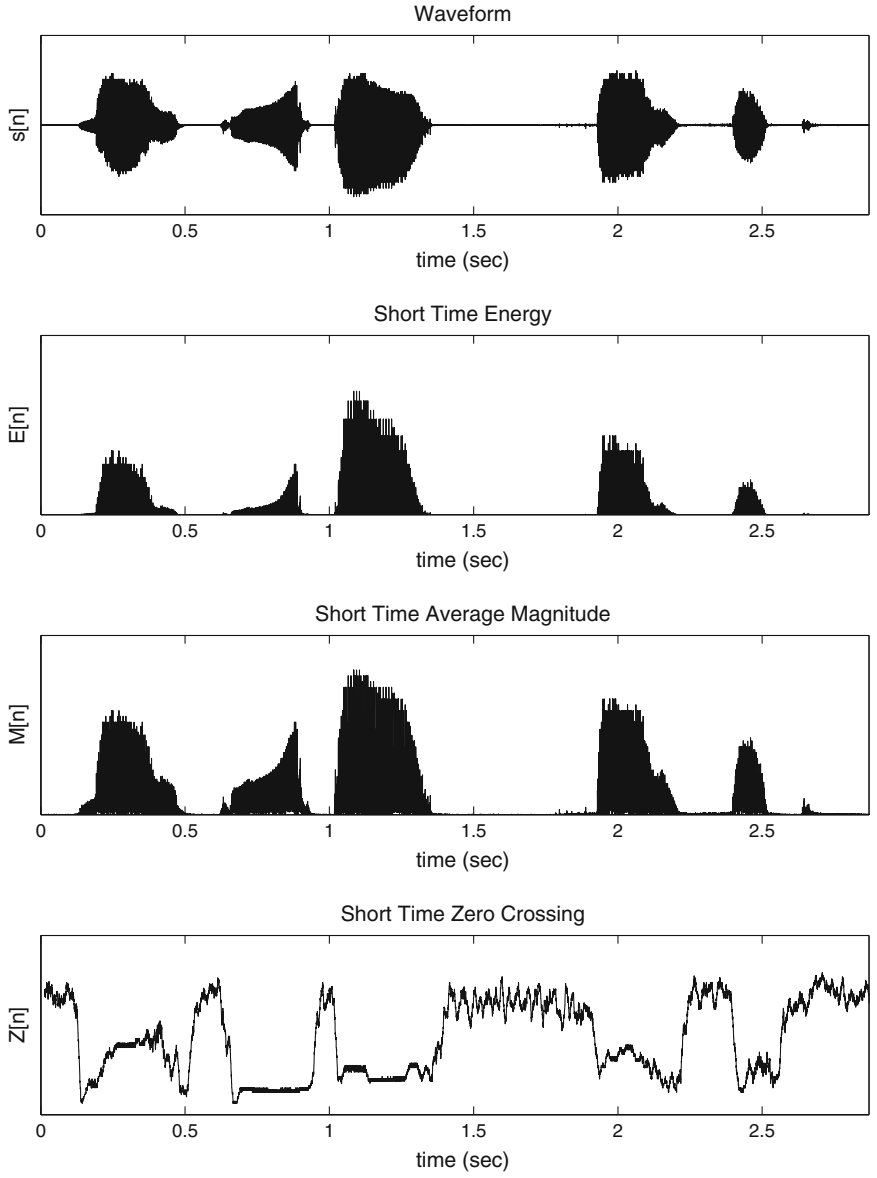

Fig. 2.11 Time domain processing. The plots show (from the *top* to the *bottom*) a waveform, the short-time energy, the short-time average magnitude, the short-time average zero crossing rate. The sampling rate is 8000 Hz and the window is 12.5 μs long

measure, called *short time average zero-crossing rate* ZCR, enables us to obtain a rough idea of the frequencies represented in the data. Such a measure can be obtained as follows:

$$Z[n] = \frac{1}{2N} \sum_{m=-\infty}^{\infty} |sign(s[m]) - sign(s[m-1])| w[n-m] \tag{2.55}$$

where $w(l)$ is a rectangular window of length N. If $s(t)$ is a sinusoid of frequency f, then there are two zero crossings every T seconds, where $T = 1/f$. If $s(t)$ is sampled at a rate $F > 2f$ for a time Δt corresponding to a high multiple of T, the average number of zero crossings Z can be obtained as follows:

$$Z \simeq \frac{2f}{F} \tag{2.56}$$

where f/F is nothing else than the number of sinusoid cycles per sampling period. For this reason, $Z[n]$ provides a rough description of the frequency content in $s[n]$. The lowest plot of Fig. 2.14 shows the value of $Z[n]$ for the spoken utterance used as example so far: on average, the $Z[n]$ value is between 0.1 and 0.2 in the spoken segments and this corresponds, using Eq. (2.56), to frequencies between 400 and 800 Hz. This is compatible with the fact that the speaker is a woman (and the fundamental frequencies are up to 300 Hz for women) and with the fact that the energy of the speech tends to concentrate below 3000 Hz. The value of $Z[n]$ in the silence segments is, on average, between 0.5 and 0.6 and this accounts for frequencies between 2000 and 2400 Hz. The reason is that the energy of nonspeech segments is concentrated on high-frequency noise. However, the above frequencies values must be considered indicative and must be used to discriminate rather than to describe different segments. The ZCR has been used in several audio processing technologies including the detection of word boundaries [34], speech-music discrimination [8, 35], audio classification [19].

The property examined next is the *autocorrelation function* $\phi[k]$ which has a different expression depending on the kind of signal under examination. For *finite energy* signals $\phi[k]$ is defined as follows:

$$\phi[k] = \sum_{m=-\infty}^{\infty} s[m]s[m+k]. \tag{2.57}$$

A signal is said to be finite energy when the following sum is finite:

$$E = \sum_{n=-\infty}^{\infty} s^2[n]. \tag{2.58}$$

for *constant power* signals the expression is:

$$\phi[k] = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{m=-N}^{N} s[m]s[m+k]. \tag{2.59}$$

A signal is said to be constant power when the following sum is constant:

$$P = \sum_{n=-T}^{T} s^2[n_0 - n] \tag{2.60}$$

for any n_0 and T. P can be thought of as the signal power, i.e., the average signal energy per time unit. The autocorrelation function has several important properties. The first is that if $s[n] = s[n + mp]$, where m is an integer number, then $\phi[k] = \phi[k+mp]$. in other words, the autocorrelation function of a periodic signal is periodic with the same period. The second is that $\phi[k] = \phi[-k]$, i.e., the autocorrelation function is even and it attains its maximum for $k = 0$:

$$|\phi[k]| \leq \phi[0] \qquad \forall k. \tag{2.61}$$

The value of $\phi[0]$ corresponds to the total energy of the signal which is thus a particular case of the autocorrelation function.

Equation (2.57) is valid for the signal as a whole, but in audio processing the analysis is performed, in general, on an analysis frame. This requires the definition of a *short-term autocorrelation function*:

$$R_n[k] = \sum_{m=-\infty}^{\infty} s[m]w[n-m]s[m+k]w[n-m-k]. \tag{2.62}$$

Such an expression corresponds to the value of $\phi[k]$ calculated over the intersection of two windows shifted by k sampling periods with respect to each other. If $k > N$ (where N is the window length), then $R_n[k] = 0$ because there is no intersection between the two windows.

The short-term properties considered so far (energy, average magnitude and average ZCR) provide a single value for each analysis frame identified by a specific position of the window. This is not the case of the short-time autocorrelation function which provides, for each analysis frame, a function of the *lag*. Figure 2.12 shows the short-term autocorrelation function obtained from a window of length $N = 401$ (corresponding to 50 ms). Upper and lower plots have been obtained over a speech ($t = 1.2$ s in Fig. 2.14) and a silence segment ($t = 1.5$ s in Fig. 2.14) respectively. In the first case there are clear peaks appearing roughly every 5 ms, and this corresponds to a fundamental frequency of around 200 Hz. In the second case no periodicity is observed and $R_n[k]$ looks rather like a high-frequency noise-like waveform. The autocorrelation function can thus be used as a further description of the frequency

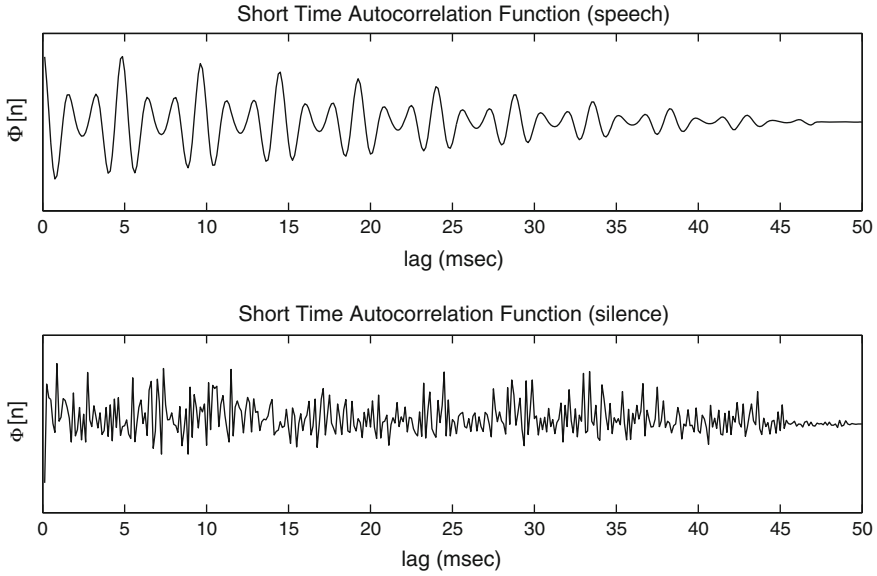

Fig. 2.12 Short term autocorrelation function. *Upper* and *lower* plots show the short term autocorrelation function for a speech and a silence point respectively. The plot in the silence case does not show any periodicity, while in the speech case there are peaks appearing roughly every 5 ms. This corresponds to a fundamental frequency of around 200 Hz, a value compatible with the ZCR measures made over the same signal and with the fact that the speaker is a woman

content that can help in discriminating different parts of the signal. Figure 2.12 shows that the amplitude of $R_n[k]$ decreases as the lag increases. The reason is that for higher values of k the intersection between the two windows decreases and there are less addends in the sum of Eq. (2.62).

The autocorrelation function has been used to detect the music meter [6], pitch detection [32], music and audio retrieval [13, 39], audio fingerprinting [38], and so on.

2.6 Linear Predictive Coding

Signals can be thought of as *temporal series*, i.e. sequences of values—typically measurements of an observable of interest—that follow and underlying dynamics. This means that samples close in time should not be independent, but correlated with one another. This is a major advantage when it comes to the possibility of *coding* a signal, i.e., of representing its properties and the information it conveys with a few parameters. By "a few" it is meant a number that is significantly smaller than the number of samples in the signal. In other words, the goal of coding a signal is to replace as many samples as possible with as a few numbers (the parameters) as possible. The advantages of coding are evident: On the one hand, transmission

and storage become easier because the signal requires much less band or space after having been coded. On the other hand, the value of the parameters gives an indication of the signal "content" (e.g., the type of sound a signal carries) and this makes it easier to compare different signals to verify whether they contain the same type of information or not (i.e., whether two signals are both human voices or not).

This section presents the *Linear Predictive Coding* (LPC) [21, 37], one of the most common and popular coding techniques. The general idea behind a coding approach is that a sample $s[k]$ can be represented as a function of a certain number of preceding samples:

$$s[k] = f(s[k-1], s[k-2], \ldots, s[k-p]). \tag{2.63}$$

The peculiarity of the LPC is that the function is a *linear combination* (hence its name):

$$s[k] = -\sum_{j=1}^{p} a_j s[k-j] + G \sum_{l=0}^{q} b_l u[k-l] \tag{2.64}$$

where the a_j and the b_l are the *predictor coefficients* (with $b_0 = 1$ by definition), G is the gain and $u[k]$ is an *unknown* input signal. The reason for such a formulation is that the LPC stems from the speech production model depicted in Fig. 2.13, a filter that produces voiced sounds (e.g., vowels and nasals) when the input signal is a quasi-periodic train of impulses and unvoiced sounds (e.g., fricatives like *sh*, *t*, *p*) when the input is random noise.

The z-transform $S(z)$ of a signal $s[k]$ can be obtained as follows (see Appendix B):

$$S(z) = \sum_{n=-\infty}^{\infty} s[n] z^{-n}. \tag{2.65}$$

By applying the z-transform to both sides of Eq. 2.64, it is possible to observe what happens in the frequency domain:

$$S(z) = -\sum_{j=1}^{p} a_j \sum_{k=-\infty}^{\infty} s[k-j] z^{-k} + G \sum_{l=0}^{q} b_l \sum_{k=-\infty}^{\infty} u[k-l] z^{-k}. \tag{2.66}$$

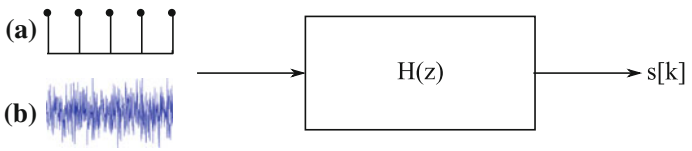

Fig. 2.13 Speech production model. The pixture depicts a speech production model where a filter produces voice sounds when excited by a train of quasi-periodic impulses (*input a*) and unvoiced sounds when excited by random noise (*input b*)

The last expression can be simplified by multiplying and dividing the first sum by z^j and the second sum by z^l:

$$S(z) = -\sum_{j=1}^{p} a_j z^{-j} \sum_{k=-\infty}^{\infty} s[k-j] z^{-(k-j)} + G \sum_{l=0}^{q} b_l z^{-l} \sum_{k=-\infty}^{\infty} u[k-l] z^{-(k-l)}$$

which corresponds to:

$$S(z) = -\sum_{j=1}^{p} a_j z^{-j} S(z) + G \sum_{l=0}^{q} b_l z^{-l} U(z). \tag{2.67}$$

Given the above, the transfer function $H(z)$ of the filter, the ratio of the output to the input, corresponds to the following:

$$H(z) = \frac{S(z)}{U(z)} = G \frac{\sum_{l=0}^{q} b_l z^{-l}}{1 + \sum_{j=1}^{p} a_j z^{-j}} = G \frac{1 + \sum_{l=1}^{q} b_l z^{-l}}{1 + \sum_{j=1}^{p} a_j z^{-j}} \tag{2.68}$$

where the last passage is possible because $b_0 = 1$ by definition (see above).

The main message of the last equation is that it is possible to obtain $S(z)$ by simply multiplying $H(z)$ by the z-transform of the input signal that we know to correspond to a train of impulses (when the signal carries a voiced sound) or random noise (when the signal carries an unvoiced sound). This explains why LPC *codes* the signal. Once the a_j's and the b_l's are known, it is not longer necessary to store or transmit the entire signal. It is sufficient to store or tansmit the parameters and then to use Eq. 2.68 to obtain $S(z)$ and, hence, the original signal $s[k]$. Needless to say, this is an advantage as long as the number of parameters is significantly lower than the number of samples in the signal.

Of all possible filters that can be obtained by chaning the parameter values in Eq. 2.68, two are of particular interest:

- the *all-zero model*: $a_j = 0$ for $j = 1, \ldots, p$
- the *all-pole method*: $b_l = 0$ for $l = 1, \ldots, q$.

The all-zero model is typically referred to as the *Moving Average* and it expresses $s[k]$ as a weighted sum of the last q input samples:

$$s[k] = G \sum_{l=1}^{q} b_l u[k-l]. \tag{2.69}$$

The reason for the name is that the case where $b_l = 1/q$ for all l values corresponds to the Moving Average defined in Sect. 2.5.

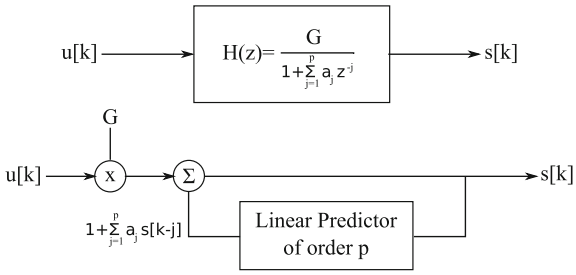

Fig. 2.14 Speech production model. The pixture depicts a speech production model where a filter produces voice sounds when excited by a train of quasi-periodic impulses (*input a*) and unvoiced sounds when excited by random noise (*input b*)

The all-pole model is called *Auto-Regression Model* (AR) and it expresses sample $s[k]$ as a linear combination of p past samples and the current input sample:

$$s[k] = -\sum_{j=1}^{p} a_j s[k-j] + Gu[k]. \tag{2.70}$$

The rest of this section focuses on the all-pole because it is one of the most commonly applied models and it has been the subject of extensive investigation.

2.6.1 Parameter Estimation

According to Eq. (2.64), the estimate of $\hat{s}_N[k]$ of sample k can be expressed as follows if the all pole model is adopted:

$$\hat{s}_N[k] = -\sum_{j=1}^{N} a_j s[k-j], \tag{2.71}$$

where N is the order of the model. Correspondingly, the error $e_N[k] = s[k] - \hat{s}_N[k]$ can be expressed in the following terms:

$$e_N[k] = s[k] + \sum_{j=1}^{N} a_j s[k-j]. \tag{2.72}$$

The *mean square error* $\mathcal{E}$ is therefore the average of $e_N[k]$ over the entire signal:

$$\mathcal{E} = \frac{1}{T}\sum_{k=1}^{T} |e_N[k]|^2 = E\left[|e_N[k]|^2\right], \tag{2.73}$$

where $E[.]$ denotes the expectation and T is the total number of samples in the signal. Since the goal of LPC is to reconstruct the signal as accurately as possible, i.e., to minimize the value of $e_N[k]$, a reasonable choice for the parameters a_j is to use the values that minimize the mean square error above. Such values can be found by minimizing the mean square error with respect to the parameters, i.e. by solving the following system of N equations with N unknown variables:

$$\begin{cases} \frac{\partial \mathcal{E}}{\partial a_1} = 0 \\ \frac{\partial \mathcal{E}}{\partial a_2} = 0 \\ \cdots \\ \frac{\partial \mathcal{E}}{\partial a_N} = 0 \end{cases} \tag{2.74}$$

where the ith equation can be obtained as follows:

$$\frac{\partial \mathcal{E}}{\partial a_i} = \frac{\partial}{\partial a_i} \sum_{k=1}^{T} \frac{1}{T} \left(s[k] + \sum_{j=1}^{N} a_j s[k-j] \right)^2 = 0 \tag{2.75}$$

that boils down to:

$$\frac{\partial \mathcal{E}}{\partial a_i} = \sum_{k=1}^{T} \frac{2}{T} \left(s[k] + \sum_{j=1}^{N} a_j s[k-j] \right) s[k-i] = 0 \tag{2.76}$$

and, finally, to:

$$\sum_{k=1}^{T} \frac{2}{T} s[k]s[k-i] + \sum_{j=1}^{N} a_j \sum_{k=1}^{T} \frac{2}{T} s[k-j]s[k-i] = 0. \tag{2.77}$$

Given that

$$\sum_{k=1}^{T} \frac{2}{T} s[k]s[k-i] = E\left(s[k]s[k-i]\right), \tag{2.78}$$

and

$$\sum_{k=1}^{T} \frac{2}{T} s[k-j]s[k-i] = E\left(s[k-j]s[k-i]\right) \tag{2.79}$$

Equation (2.77) can be further reformulated in the following final form:

$$\sum_{j=1}^{N} a_j E\left(s[k-j]s[k-i]\right) = -E\left(s[k]s[k-i]\right). \tag{2.80}$$

According to Eq. (2.62), the expectation of the product $s[k]s[k-i]$ can be thought of as the autocorrelation of he signal with lag i in the case $w[k] = 1\ \forall k$, i.e., the analysis window is infinite and its samples are all equal to 1:

$$T \cdot E\left(s[k]s[k-i]\right) = R(i), \tag{2.81}$$

where there is no need to use the subscript k because the value $R(i)$ is estimated using the entire signal. Similarly, for the expectation of $s[k-j]s[k-i]$:

$$T \cdot E\left(s[k-j]s[k-i]\right) = R(i-j). \tag{2.82}$$

Given that $R(-l) = R(l)$, this means that Eq. (2.77) can be written in the following form:

$$\sum_{j=1}^{N} a_j R(i-j) = -R(i) \tag{2.83}$$

and the systems of equations above can be interpreted as a product of matrices:

$$\begin{bmatrix} R(0) & R(1) & \dots & R(N-1) \\ R(1) & R(0) & \dots & R(N-2) \\ \dots & \dots & \dots & \dots \\ R(N-1) & R(N-2) & \dots & R(0) \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_N \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \dots \\ R(N) \end{bmatrix}. \tag{2.84}$$

The expression above is known as *Normal Equation* and can be solved like any other system of linear equations.

2.7 Conclusions

This chapter has provided an overview of the main problems revolving around the processing of audio signals with computers. After showing how the human phonatory apparatus and ears work, the chapter has shown how audio signals can be represented and stored. Furthermore, it has shown how it is possible to extract information about their content by using simple time-domain processing techniques. Chapter 16 how the analysis of speech signals can be used to redict the personality traits that people attribute to speakers.

Problems

2.1 Consider a sound of intensity $I = 5\,\text{dB}$. Calculate the energy emitted by its source in a time interval of length $\Delta t = 22.1\,\text{s}$. Given the air acoustic impedance $Z = 410\,\text{Pa} \cdot \text{s} \cdot \text{m}^{-1}$, calculate the pressure corresponding to the maximum compression determined by the same sound wave.

2.2 Human ears are particularly sensitive to frequencies between 50 and 4000 Hz. Given the speed of sound in air ($v \simeq 331.4\,\mathrm{m} \cdot \mathrm{s}^{-1}$), calculate the wavelengths corresponding to such frequencies.

2.3 Consider a sum of N sinusoids with frequencies $f_0, 3f_0, \ldots, (2N+1)f_0$:

$$f(t) = \sum_{n=0}^{N} \frac{1}{2n+1} sin[2\pi f_0(2n+1)t] \tag{2.85}$$

Plot $f(t)$ in the range $[0, 10]$ for $f_0 = 1$ and $N = 1, 2, \ldots, 100$ and observe the signal $f(t)$ converges to.

2.4 The Mel scale (see Sect. 2.2.3) maps frequencies f into values $B(f)$ that are more meaningful from a perceptual point of view. Segment the $B(f)$ interval $[0, 3375]$ into 20 intervals of the same length and find the frequencies f corresponding to their limits.

2.5 Extract the waveform from an audio file using *HTK* (see Chap. 12 for a description of the HTK software package) and calculate the number of bits N necessary to represent the sample values. Perform a uniform quantization of the waveform using a number of bits n ranging from 2 to $N-1$ and calculate, for each n, the signal-to-noise ratio (SNR). Plot the SNR as a function of n.

2.6 Calculate sampling frequency and bit-rate of the audio file used in Problem 2.5.

2.7 Plot the TOH in presence of a masking tone noise of frequency 200 Hz and intensity 50 dB.

2.8 Consider the system known as *moving average* (see Sect. 2.5). Demonstrate that such system is linear and time invariant.

2.9 Consider an audio file including both speech and silence and extract the waveform it contains. Obtain magnitude and zero crossing rate as a function of time using a rectangular analysis window 30 ms long. A pair $(M[n], Z[n])$ is available for each sample $s[n]$ and can be plotted on a plane where the axes are magnitude and ZCR. Do sound and speech samples form separate clusters (see Chap. 6)?

2.10 Demonstrate that the autocorrelation function $R_n[k]$ corresponds to the short time energy when $k = 0$ and that $|R_n[k]| < R_n[0]$ for $k > 0$.

References

1. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. Technical report, International Telecommunication Union, 1997.
2. L.L. Beranek. Concert hall acoustics. *The Journal of the Acoustical Society of America*, 92(1): 1–39, 1992.

3. D.T. Blackstock. *Fundamentals of Physical Acoustics*. John Wiley and Sons, 2000.
4. J. Bormans, J. Gelissen, and A. Perkis. MPEG-21: The 21st century multimedia framework. *IEEE Signal Processing Magazine*, 20(2):53–62, 2003.
5. M. Bosi and R.E. Goldberg. *Introduction to Digital Audio Coding and Standards*. Kluwer, 2003.
6. J.C. Brown. Determination of the meter of musical scores by autocorrelation. *The Journal of the Acoustical Society of America*, 94(4):1953–1957, 1993.
7. R. Burnett, I. and van de Walle, K. Hill, J. Bormans, and F. Pereira. MPEG-21: Goals and achievements. *IEEE Multimedia*, 10(4):60–70, 2003.
8. M.J. Carey, E.S. Parris, and H. Lloyd-Thomas. A comparison of features for speech-music discrimination. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, pages 149–152, 1999.
9. J.C. Catford. *Theoretical Acoustics*. Oxford University Press, 2002.
10. P. Cummiskey. Adaptive quantization in differential PCM coding of speech. *Bell Systems Technical Journal*, 7:1105, 1973.
11. T.F.W. Embleton. Tutorial on sound propagation outdoors. *The Journal of the Acoustical Society of America*, 100(1):31–48, 1996.
12. H. Fletcher. Auditory patterns. *Review of Modern Physics*, pages 47–65, 1940.
13. A. Ghias, J. Logan, D. Chamberlin, and B.C. Smith. Query by humming: musical information retrieval in audio database. In *Proceedings of the ACM Conference on Multimedia*, pages 231–236, 1995.
14. A. Hanjalic and L.-Q. Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, 2005.
15. X. Huang, A. Acero, and H.-W. Hon. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice-Hall, 2001.
16. L.E. Kinsler, A.R. Frey, A.B. Coppens, and J.V. Sanders. *Fundamentals of Acoustics*. John Wiley and Sons, New York, 2000.
17. P. Ladefoged. *Vowels and consonants*. Blackwell Publishing, 2001.
18. C.M. Lee and S.S. Narayanan. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Multimedia*, 13(2):293–303, 2005.
19. L. Lu, H. Jiang, and H.J. Zhang. A robust audio classification and segmentation method. In *Proceedings of the ACM Conference on Multimedia*, pages 203–211, 2001.
20. Y.-F. Ma, X.-S Hua, L. Lu, and H.-J. Zhang. A generic framework for user attention model and its application in video summarization. *IEEE Transactions on Multimedia*, 7(5):907–919, 2005.
21. J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975.
22. B.S. Manjunath, P. Salembier, and T. Sikora, editors. *Introduction to MPEG-7*. John Wiley and Sons, Chichester, UK, 2002.
23. S.K. Mitra. *Digital Signal Processing - A Computer Based Approach*. McGraw-Hill, 1998.
24. B.C.J. Moore. *An Introduction to the Psychology of Hearing*. Academic Press, 1997.
25. P.M. Morse and K. Ingard. *Theoretical Acoustics*. McGraw-Hill, 1968.
26. P. Noll. Wideband speech and audio coding. *IEEE Communications Magazine*, (11):34–44, november 1993.
27. P. Noll. MPEG digital audio coding. *IEEE Signal Processing Magazine*, 14(5):59–81, 1997.
28. B.M. Oliver, J. Pierce, and C.E. Shannon. The philosophy of PCM. *Proceedings of IEEE*, 36:1324–1331, 1948.
29. A.V. Oppenheim and R.W. Schafer. *Discrete-Time Signal Processing*. Prentice-Hall, 1989.
30. T. Painter and A. Spanias. Perceptual coding of digital audio. *Proceedings of IEEE*, 88(4):451–513, 2000.
31. J.O. Pickles. *An Introduction to the Physiology of Hearing*. Academic Press, 1988.
32. L. Rabiner. On the use of autocorrelation analysis for pitch detection. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 25(1):24–33, 1977.

33. L.R. Rabiner and R.W. Schafer, editors. *Digital Processing of Speech Signals*. Prentice-Hall, 1978.
34. L.R. R Rabiner and M.R. Sambur. An algorithm for determining the endpoints of isolated utterances. *Bell System Technical Journal*, 54(2):297–315, 1975.
35. E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, pages 1331–1334, 1997.
36. A. Spanias. Speech coding: a tutorial review. *Proceedings of IEEE*, 82(10):1541–1582, 1994.
37. A.S. Spanias. Speech coding: A tutorial review. *Proceedings of the IEEE*, 82(10):1541–1582, 1994.
38. S. Sukittanon and L.E. Atlas. Modulation frequency features for audio fingerprinting. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, pages 1773–1776, 2002.
39. E. Wold, T. Blum, D. Keislar, and J. Wheaten. Content-based classification, search, and retrieval of audio. *IEEE MultiMedia*, 3(3):27–36, 1996.

http://www.springer.com/978-1-4471-6734-1

Machine Learning for Audio, Image and Video Analysis
Theory and Applications
Camastra, F.; Vinciarelli, A.
2015, XVI, 561 p. 119 illus., Hardcover
ISBN: 978-1-4471-6734-1