3

Fundamentals

In this chapter we give a short overview of the major types of signals one has to deal with in acoustic echo and noise control systems. Furthermore, we describe the basic mechanism of the origin of acoustic echoes and justify the use of transversal filters for echo suppression. We also briefly review standards formulated by the International Telecommunication Union (ITU) and by the European Telecommunication Standards Institute (ETSI).

3.1 SIGNALS

The two main classes of signals involved in acoustic echo and noise control systems are *speech* and *background noise*. For someone who deals with (adaptive) algorithms, speech certainly belongs to the most difficult signals she or he has to handle. Its properties are far away from the properties of white noise that is popular for designing and analyzing algorithms. As background noise we will discuss noise in a car and noise in an office. Speech signals are treated in the literature in more detail (e.g.,by Rabiner and Juang [193]) than background noise. Therefore, we will be brief on speech and elaborate on background noise.

3.1.1 Speech

Speech might be characterized by nearly periodic (voiced) segments, by noiselike (unvoiced) segments, and by pauses (silence). The envelope of a speech signal exhibits a very high dynamics. If a sampling frequency of 8 kHz (telephony) is used, the

mean spectral envelope may range over more than 40 dB [126]. If higher sampling rates are implemented (e.g., for teleconferencing systems), the variations increase even further. Parameters derived from a speech signal may be considered valid only during an interval of about 20 ms [49, 193]. Properties of a speech signal are illustrated in Fig. 3.1. It shows a 5-second time sequence sampled at 8 kHz (upper diagram), the average power spectral density (center), and the result of a time–frequency analysis (lower diagram).

Due to the frequent changes of the properties of speech signals, the (long-term) power spectral density is not sufficient to characterize the spectral behavior of speech. The time–frequency—also called "visible speech"—diagram clearly shows the sequences of voiced and unvoiced sections and of pauses. During voiced intervals the *fundamental* or *pitch frequency* and the harmonics bare the major part of the energy.

A spectral feature of vowels are peaks in the spectral envelope caused by resonances in the vocal tract. These resonance frequencies shape ("form") the spectral envelope. Thus they are called *formant frequencies*. Their change over time characterizes the *intonation* of a word or sentence. The pitch frequency depends on the speaker. Mean pitch frequencies are about 120 Hz for male speakers and about 215 Hz for female speakers. In Fig. 3.2 three 60-ms segments of speech (as well as the entire sequence) are depicted in the time and in the frequency domain. The two diagrams in the center show two vowels: *a* from *wonderful* and *i:* from *be*. In the last diagram the sibilant sequence *sh* from *she* is depicted. The terms *a*, *i:*, and *sh* are written in phonetic transcription.

The differences between voiced and unvoiced speech are also clearly visible in the power spectral density. While voiced speech has a comblike spectrum, unvoiced speech exhibits a nonharmonic spectral structure. Furthermore, the energy of unvoiced segments is located at higher frequencies. The rapidly changing spectral characteristics of speech motivate the utilization of signal processing in subbands or in the frequency domain. These processing structures allow a frequency-selective power normalization leading to a smaller eigenvalue spread [111] and therefore to faster convergence of adaptive filters excited and controlled by such signals.

Sampling frequencies range from 8 kHz in telephone systems up to about 96 kHz in high-fidelity systems. Even in the case of 8 kHz sampling frequency, consecutive samples are highly correlated. The normalized autocorrelation coefficient

$$a_1 = \frac{\mathrm{E}\{\, s(n)\, s(n+1)\,\}}{\mathrm{E}\{\, s^2(n)\,\}} = \frac{s_{ss}(1)}{s_{ss}(0)} \tag{3.1}$$

of neighboring samples reaches values in the range of 0.8–0.95. Because of the periodic components and the pauses, short-time autocorrelation matrices very often become singular. Therefore, speech signals belong to the class of nonpersistent signals. Thus, special precautions are necessary to prevent instability of algorithms that use—directly or indirectly—the inverse of the autocorrelation matrix.

The *probability density function* $f_s(s)$ of the amplitudes of speech signals $s(n)$ is characterized by a marked peak for zero amplitudes and an exponential decay for large amplitudes. Voiced sounds mainly contribute to large values, whereas unvoiced sounds are responsible for the peak at zero amplitude [48]. Analytic approximations

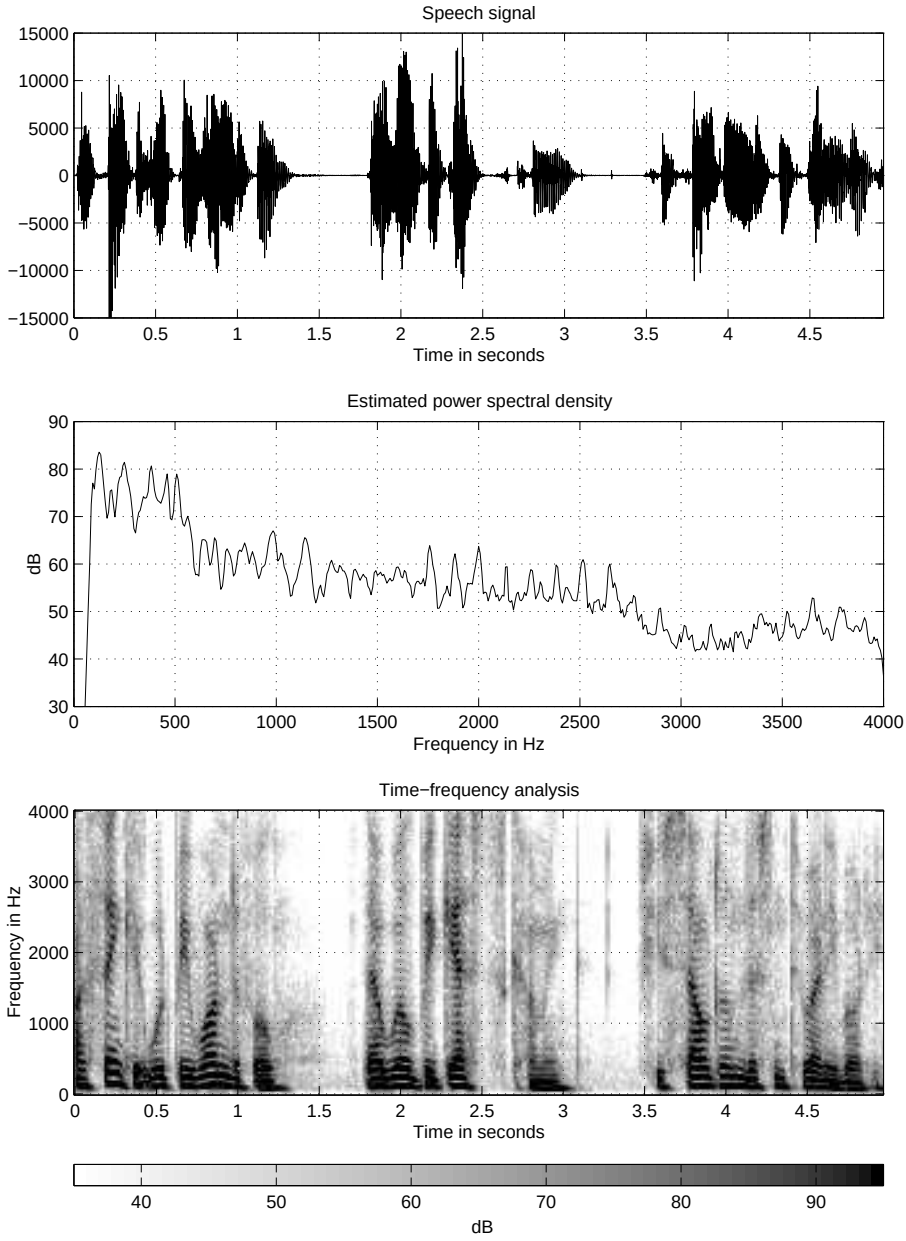

Fig. 3.1 Example of a speech sequence. In the upper part a 5-second sequence of a speech signal is depicted. The signal was sampled at 8 kHz. The second diagram shows the mean power spectral density of the entire sequence (periodogram averaging). In the lowest part a time–frequency analysis is depicted. Dark colors represent areas with high energy; light colors mark areas with low energy.

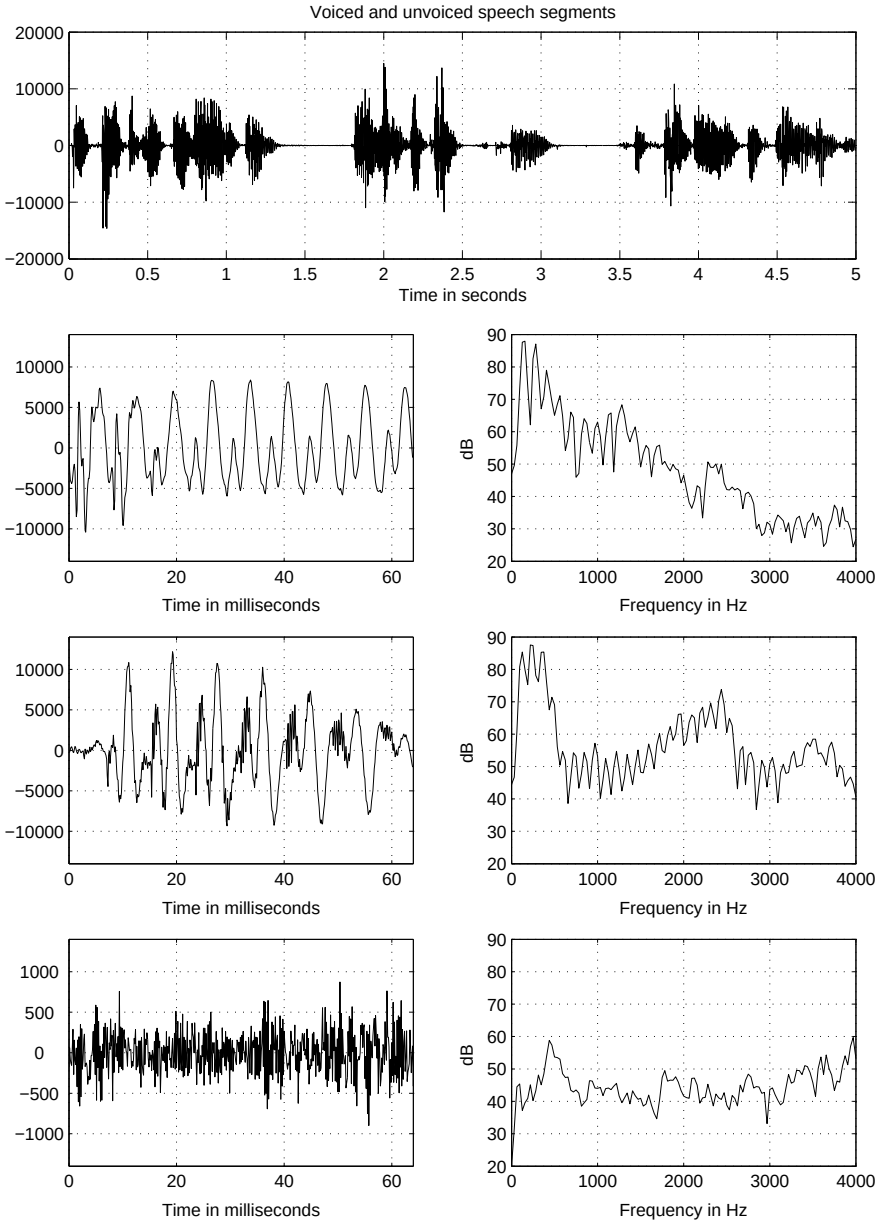

Fig. 3.2 Voiced and unvoiced speech segments. In the upper part a 5-second sequence of a speech signal is depicted. The following three diagram pairs show 60-ms sequences of the speech signal (left) as well as the squared spectrum of the sequences (right). First two vowels (*a* from *wonderful* and *i:* from *be*) are depicted. Finally, an unvoiced (sibilant) sequence (*sh* from *she*) is shown.

for the probability density function are the *Laplacian probability density function*

$$f_{s,1}(s) = \frac{1}{\sqrt{2}\sigma_s}\, e^{-\sqrt{2}\,\frac{|s|}{\sigma_s}}\,, \tag{3.2}$$

and the (two-sided) *Gamma probability density function*

$$f_{s,2}(s) = \frac{\sqrt[4]{3}}{2\sqrt{2\,\pi\,\sigma_s}}\,\frac{1}{\sqrt{|s|}}\, e^{-\frac{\sqrt{3}}{2}\,\frac{|s|}{\sigma_s}}\,, \tag{3.3}$$

where σ_s is the standard deviation of the signal, that is assumed to be stationary and zero mean in this context [23, 196, 242, 243]. It depends on the speaker which approximation provides the better fit [242].

In order to compare these functions with the well-known Gaussian density

$$f_g(s) = \frac{1}{\sqrt{2\,\pi}\,\sigma_s}\, e^{-\frac{s^2}{2\,\sigma_s^2}}\,, \tag{3.4}$$

the densities are transformed using

$$\bar{s} = \frac{s}{\sigma_s} \tag{3.5}$$

as normalized amplitude. It follows for Eq. 3.2

$$f_{\bar{s},1}(\bar{s}) = \frac{1}{\sqrt{2}}\, e^{-\sqrt{2}\,|\bar{s}|}\,, \tag{3.6}$$

and for Eq. 3.3 that

$$f_{\bar{s},2}(\bar{s}) = \frac{\sqrt[4]{3}}{2\sqrt{2\,\pi}}\,\frac{1}{\sqrt{|\bar{s}|}}\, e^{-\frac{\sqrt{3}}{2}\,|\bar{s}|}\,. \tag{3.7}$$

The Gaussian density for normalized amplitudes reads

$$f_{\bar{g}}(\bar{s}) = \frac{1}{\sqrt{2\,\pi}}\, e^{-\frac{\bar{s}^2}{2}}\,. \tag{3.8}$$

Figure 3.3 shows the three functions. It becomes obvious that especially for small amplitudes the approximations of the density of speech differ considerably from the Gaussian density.

In order to be able to handle the wide range of speech amplitudes the densities in Eqs. 3.2 and 3.3 are transformed into functions of normalized logarithmic amplitudes

$$\bar{s}_{\log} = 20\,\log_{10}|\bar{s}| = 20\,\log_{10}\frac{|s|}{\sigma_s}\,. \tag{3.9}$$

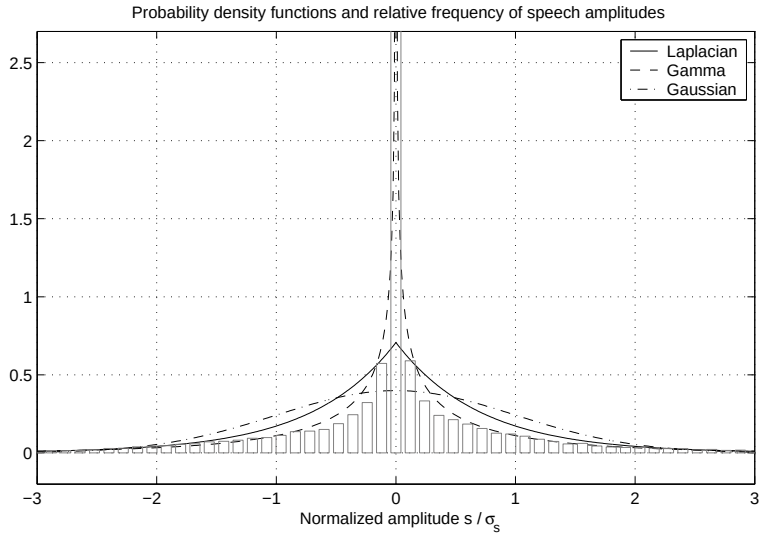

Fig. 3.3 Normalized Gaussian probability density function and normalized approximations for speech signals. The bars represent the measured relative frequency of the amplitudes of a speech sequence.

After a few manipulations the transformed probability density functions read

$$f_{\bar{s}_{\log},1}(\bar{s}_{\log}) = \frac{\sqrt{2}\,\ln 10}{20}\,10^{\frac{\bar{s}_{\log}}{20}}\;e^{-\sqrt{2}\,10^{\frac{\bar{s}_{\log}}{20}}} \tag{3.10}$$

and

$$f_{\bar{s}_{\log},2}(\bar{s}_{\log}) = \frac{\sqrt[4]{3}}{\sqrt{2\,\pi}}\,\frac{\ln 10}{20}\,10^{\frac{\bar{s}_{\log}}{40}}\;e^{-\frac{\sqrt{3}}{2}\,10^{\frac{\bar{s}_{\log}}{20}}}\;. \tag{3.11}$$

Both hold for $\bar{s}_{\log} > 0$. Figure 3.4 shows these functions compared with a measured histogram of speech amplitudes.

The assumption of a Gaussian probability density function turns out to be an applicative model for the *logarithm* of the normalized power spectral density $\Theta_{ss}(\Omega)$:

$$\Theta_{ss}(\Omega) = 10\,\log_{10}\frac{S_{ss}(\Omega)}{\sigma_s^2}\,, \tag{3.12}$$

where $S_{ss}(\Omega)$ is the power spectral density of the speech signal at frequency Ω, $S_{ss}(\Omega) > 0$ and σ_s^2 denotes the average power of this signal.[1] With this notation the

[1]If $S_{ss}(\Omega)$ describes a "short-term" power spectral density, this quantity also has a time-index n: $S_{ss}(\Omega, n)$. The time index is omitted here.

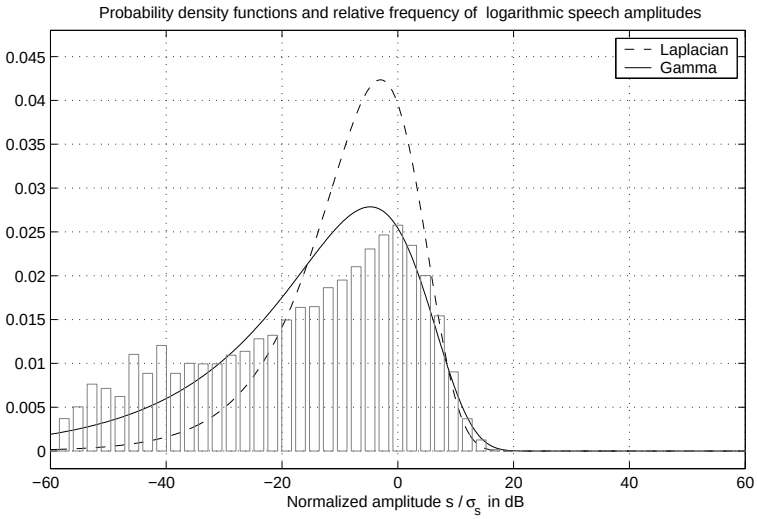

Fig. 3.4 Approximations of the probability density function for logarithmic normalized speech amplitudes and histogram of logarithmic normalized measured speech signal.

assumed density function reads

$$f_{\Theta_{ss}(\Omega)}(\Theta) = \frac{1}{\sqrt{2\,\pi}\,\kappa_{\Theta_{ss}(\Omega)}}\,\exp\left[-\frac{(\Theta - m_{\Theta_{ss}(\Omega)})^2}{2\,\kappa^2_{\Theta_{ss}(\Omega)}}\right], \tag{3.13}$$

where $m_{\Theta_{ss}(\Omega)}$ stands for the mean of the logarithmic normalized power spectral density at frequency Ω and $\kappa_{\Theta_{ss}(\Omega)}$ denotes the standard deviation of $\Theta_{ss}(\Omega)$. If we normalize once more using

$$\bar{\Theta}_{ss}(\Omega) = \frac{\Theta_{ss}(\Omega)}{\kappa_{\Theta_{ss}(\Omega)}}, \tag{3.14}$$

the probability density function finally reads

$$f_{\bar{\Theta}_{ss}(\Omega)}(\bar{\Theta}) = \frac{1}{\sqrt{2\,\pi}}\,e^{-\frac{(\bar{\Theta} - m_{\bar{\Theta}_{ss}(\Omega)})^2}{2}}, \tag{3.15}$$

with

$$m_{\bar{\Theta}_{ss}(\Omega)} = \frac{m_{\Theta_{ss}(\Omega)}}{\kappa_{\Theta_{ss}(\Omega)}}. \tag{3.16}$$

In Fig. 3.5 we show the probability density function (see Eq. 3.13) of the logarithmic normalized power spectral density of speech and histograms calculated from measured signals for selected frequencies.

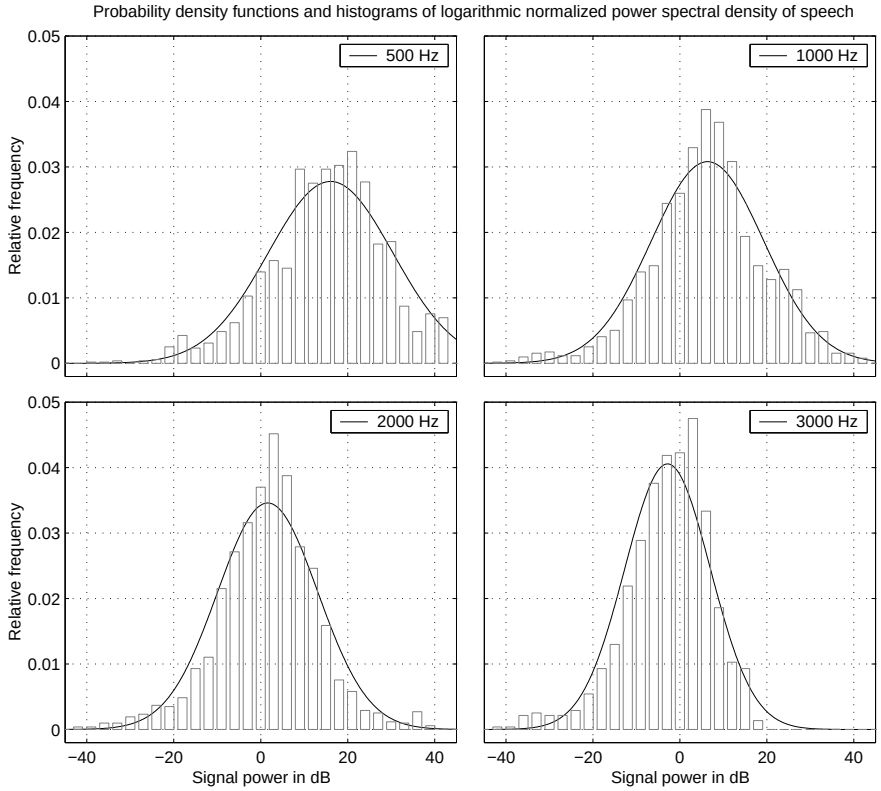

Fig. 3.5 Probability density function of the logarithmic normalized power spectral density of speech and histograms calculated from measured signals for selected frequencies.

Recent results demonstrate that the Laplacian (Eq. 3.2) or the gamma (Eq. 3.3) probability density provide close approximations also of the densities of the real and the imaginary parts of the DFT coefficients of speech signals calculated from frames shorter than 100 ms. Therefore, noise reduction procedures based on these densities prove to be superior to those assuming the Gaussian density [162].

3.1.2 Noise

Among the many types of acoustical noise in human environments, office and car noise are of special interest in connection with echo and noise control. Figure 3.6 shows estimates of power spectral densities of noises measured in a car at various speeds and in different offices. The sound level in a car can vary over 40 dB depending on the speed of the car. In an office the loudness varies depending on the size of the office, the furniture, and the equipment that is in use. However, the noise level in an office is comparable only with the level in cars driving at low speeds.

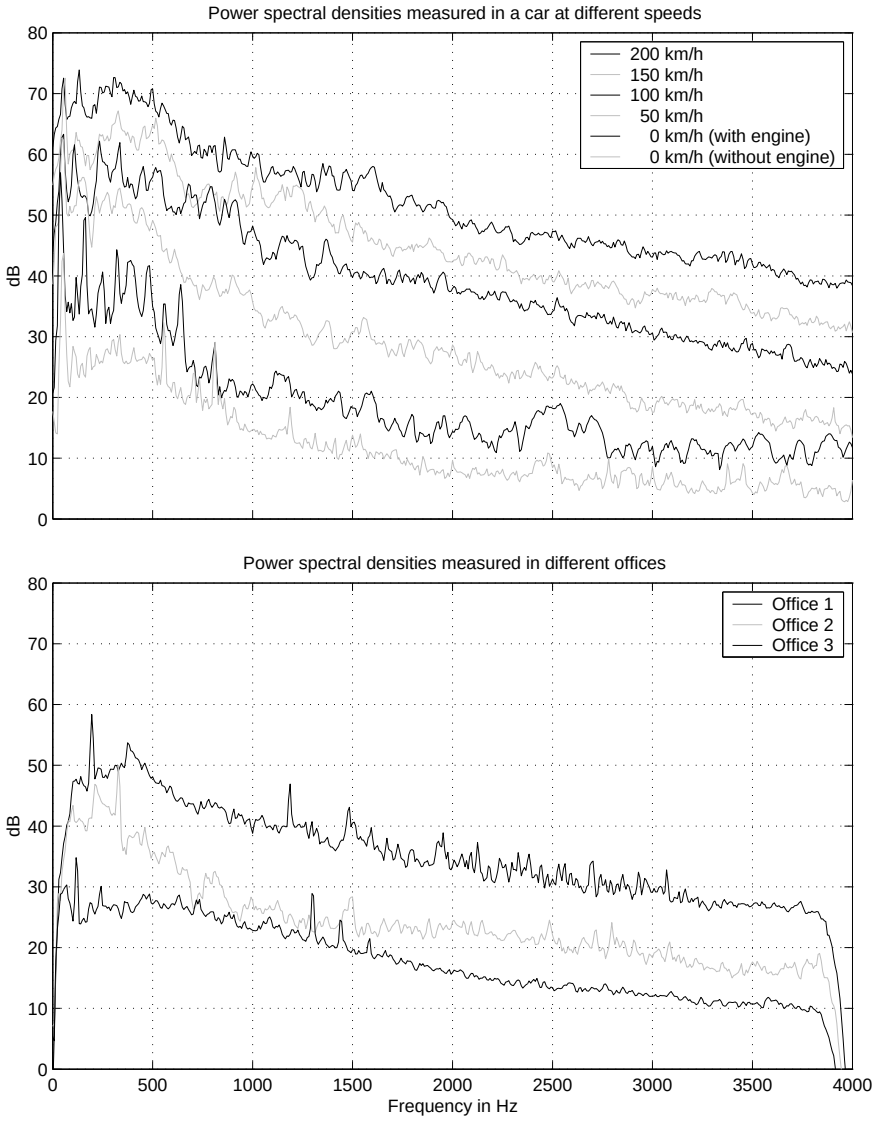

Fig. 3.6 Estimates of the power spectral densities of noises measured in a car and in offices.

3.1.2.1 Office Noise

The main sources of noise in an office are on one hand the ventilators of computers, printers, or facsimile machines and on the other hand (telephone) conversations of colleagues sharing the office.

The first category can be considered as (long-term) stationary. Figure 3.7 gives an example of the noise produced by a computer fan. It shows the time signal, the estimated power spectral density, and the time–frequency analysis. The latter confirms the stationarity of this type of noise. It also shows the harmonic components. With respect to the level of this noise, one has to be aware that its sources are often very close to microphones. This holds especially in IP-based[2] applications that are constantly gaining importance. The slow decay of the power spectral density with increasing frequency may be considered as typical for this type of office noise.

Conversational noise, of course, has speech character and is therefore—even if its level is lower than that of ventilator noise—much harder to attenuate by a noise reduction unit.

3.1.2.2 Car Noise

Car noise results from a large number of sources. The main components are engine noise, wind noise, rolling noise, and noise from various devices (e.g., ventilators) inside the passenger compartment. We will discuss these components in a little bit more detail, because their specific properties will become important later in this book when we present methods for noise reduction.

3.1.2.2.1 Engine Noise

The main energy of the noise produced by the engine of the car is concentrated at frequencies below 500 Hz. The power spectral density shows maxima at half of the engine speed and at integer multiples of it. This holds for 4-stroke engines. Figure 3.8 gives examples of engine noise. The upper diagram shows the result of a time–frequency analysis of an engine during acceleration and during deceleration. The center and lower diagrams depict the time–frequency analyses and the time signals of noises from engines running at constant speed: a gasoline engine at 2100 revolutions per minute (rpm) (center) and a diesel engine at 3500 rpm (lower diagram). In the first case the harmonics are 17.5 Hz apart from each other, in the second case their distance is 29.2 Hz. If it comes to the application of noise reduction methods it is important that the information of the exact engine speed can be read from the data bus system of the car.

3.1.2.2.2 Wind Noise

The wind noise depends on the design of the body of a car. Modern designs reduce the aerodynamic resistance as much as possible in order to lower the gasoline consumption of the car. Therefore wind noise contributes remarkably only at high speeds. Figure 3.9 shows two results of time–frequency analyses and power spectral densities of wind noise at two different speeds of a car. The measurements for these

[2] IP stands for Internet Protocol.

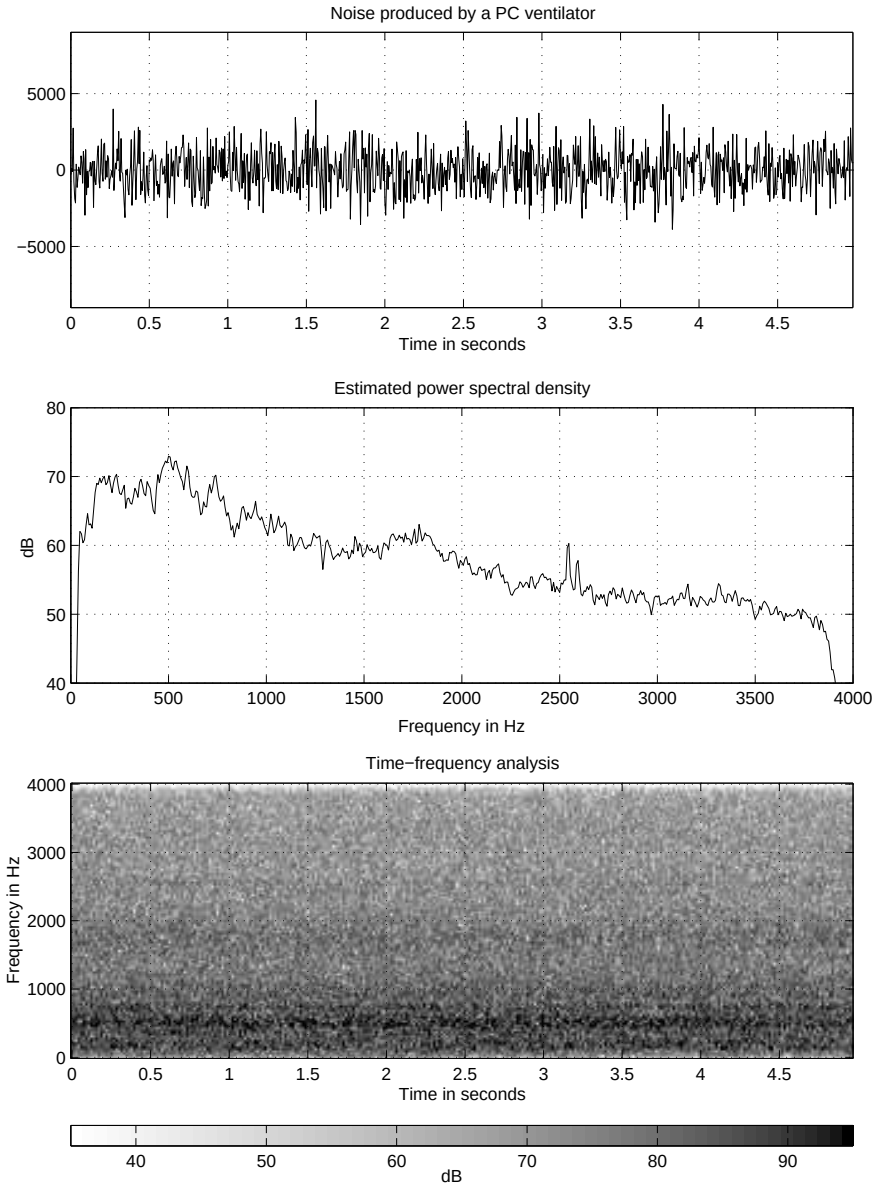

Fig. 3.7 Noise produced by a PC ventilator: time signal (upper diagram), estimated power spectral density (center), and time–frequency analysis (lower diagram). The sampling frequency is 8 kHz.

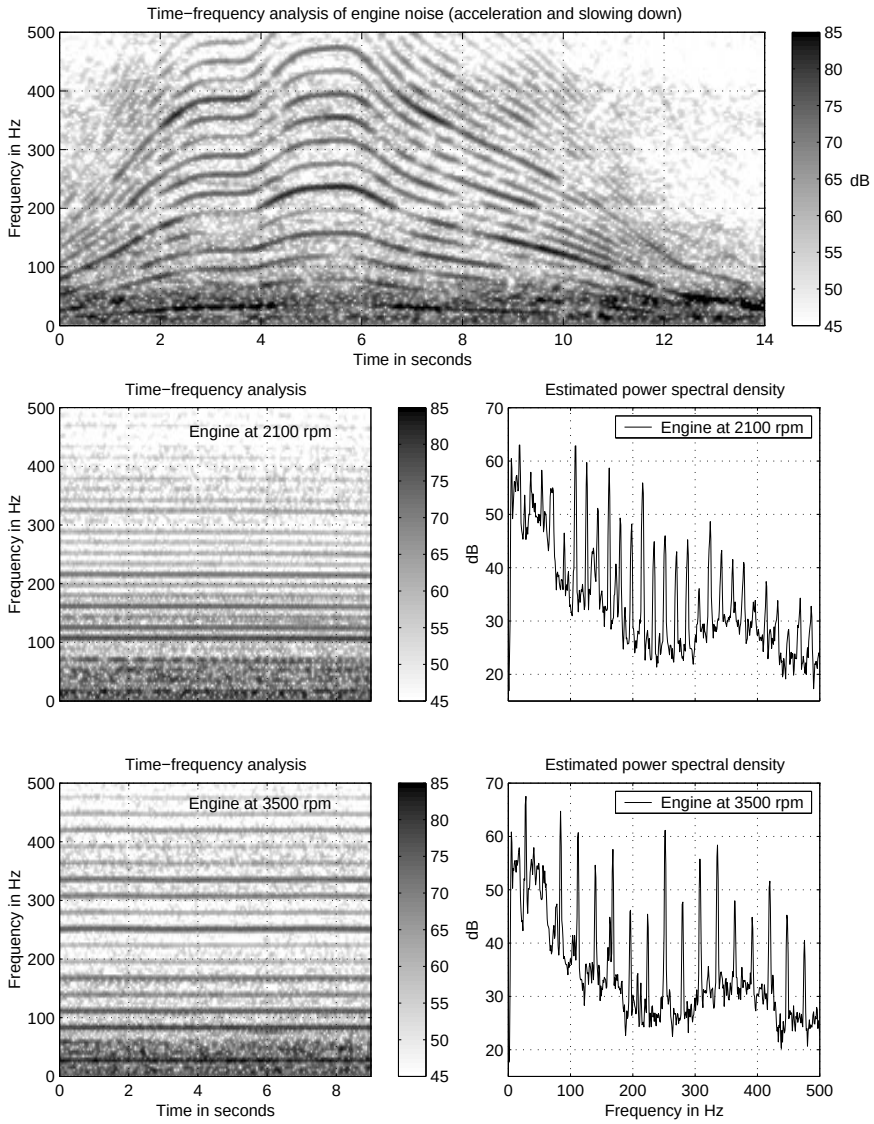

Fig. 3.8 Examples of engine noise: time–frequency analysis during acceleration and during deceleration (upper diagram), time–frequency analyses, and the time signals of noises from engines running at constant speed; gasoline engine at 2100 rpm (center) and diesel engine at 3500 rpm (lower diagram).

diagrams have been performed in a wind tunnel. Thus, the noise was free of noise from other sources.

Wind noise occupies a wider frequency range than engine noise. Since changes in the speed of a car are much slower than that of the engine wind noise is “more stationary” than engine or rolling noise.

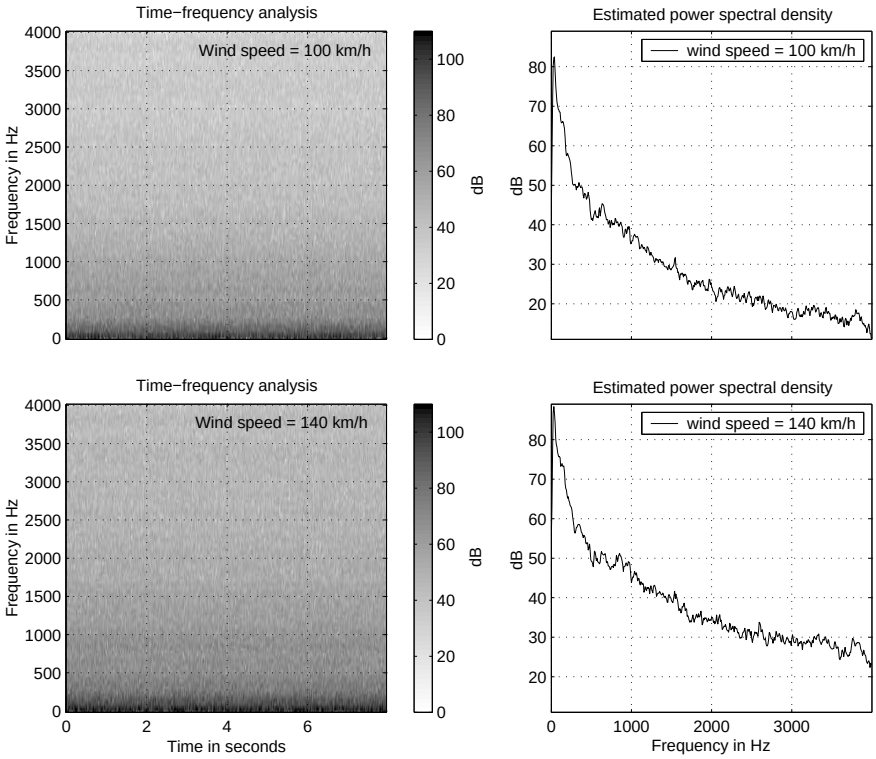

Fig. 3.9 Time–frequency analyses and power spectral densities of wind noise at two different speeds of a car: 100 km/h (upper diagrams) and 140 km/h (lower diagrams).

3.1.2.2.3 Rolling Noise

Rolling noise arises when the tires contact the road surface. It strongly depends on the type of the tires and on the road surface. The tread patterns are designed such that periodic noise at higher frequencies is avoided. Changes of the road surface cause sudden noise changes as shown in Fig. 3.10. During the measurements of these diagrams the car engine was idling to minimize the engine noise. Changes in the power spectral density are up to 15 dB and are mainly below 1000 Hz.

As very coarse approximation one can assume that—in case of nonchanging road surface—the power of the rolling noise increases exponentially with the speed of the car [189]:

$$P_{\mathrm{RN}}(v) = P_{\mathrm{RN}}(v_0)\, e^{K_{\mathrm{v}}\,(v-v_0)}\,, \tag{3.17}$$

where v is the speed of the car and K_{v} a constant depending on the type of the tires and the surface of the road. Plotted in decibels (dB), this results in a linear increase.

3.1.2.2.4 Internal Noise Sources

Besides the sources for car noise already discussed in this chapter, there are a number of additional sources in the passenger compartment. As an example, we mention ventilators. In unfavorable situations they may blow directly toward close-by microphones.

We show an example of ventilator noise in Fig. 3.11. The upper diagram depicts the time signal when the fan is switched from "low" to "high" through intermediate steps. The result of the time–frequency analysis of the noise is given in the center diagram, whereas the power spectral densities for "low" and "high" are shown at the bottom diagram. At position "low" the ventilator noise may be masked by other noise sources. At "high," however, ventilator noise cannot be neglected.

3.1.3 Probability Density of Spectral Amplitudes of Car Noise

For the real part $R(e^{j\Omega})$ and the imaginary part $I(e^{j\Omega})$ of the (short-term) Fourier transform $B(e^{j\Omega})$ of car noise $b(n)$

$$B\left(e^{j\Omega}\right) = R\left(e^{j\Omega}\right) + j\, I\left(e^{j\Omega}\right)\,, \tag{3.18}$$

Gaussian probability density functions prove to be good approximations [233]:

$$f_{R(\Omega)}(a) \;=\; \frac{1}{\sqrt{2\,\pi}\,\sigma_{R(e^{j\Omega})}}\; e^{-\dfrac{a^2}{2\,\sigma^2_{R(e^{j\Omega})}}}\,, \tag{3.19}$$

$$f_{I(\Omega)}(b) \;=\; \frac{1}{\sqrt{2\,\pi}\,\sigma_{I(e^{j\Omega})}}\; e^{-\dfrac{b^2}{2\,\sigma^2_{I(e^{j\Omega})}}}\,, \tag{3.20}$$

where $\sigma_{R(e^{j\Omega})}$ and $\sigma_{I(e^{j\Omega})}$ are the standard deviations of $R(e^{j\Omega})$ and $I(e^{j\Omega})$, respectively. Consequently, for

$$\sigma_{R(e^{j\Omega})} = \sigma_{I(e^{j\Omega})}\,, \tag{3.21}$$

and

$$|B\left(e^{j\Omega}\right)|^2 \;=\; R^2\left(e^{j\Omega}\right) + I^2\left(e^{j\Omega}\right)\,, \tag{3.22}$$

$$\sigma^2_{|B(e^{j\Omega})|^2} \;=\; 2\,\sigma^2_{R(e^{j\Omega})} \;=\; 2\,\sigma^2_{I(e^{j\Omega})}\,, \tag{3.23}$$

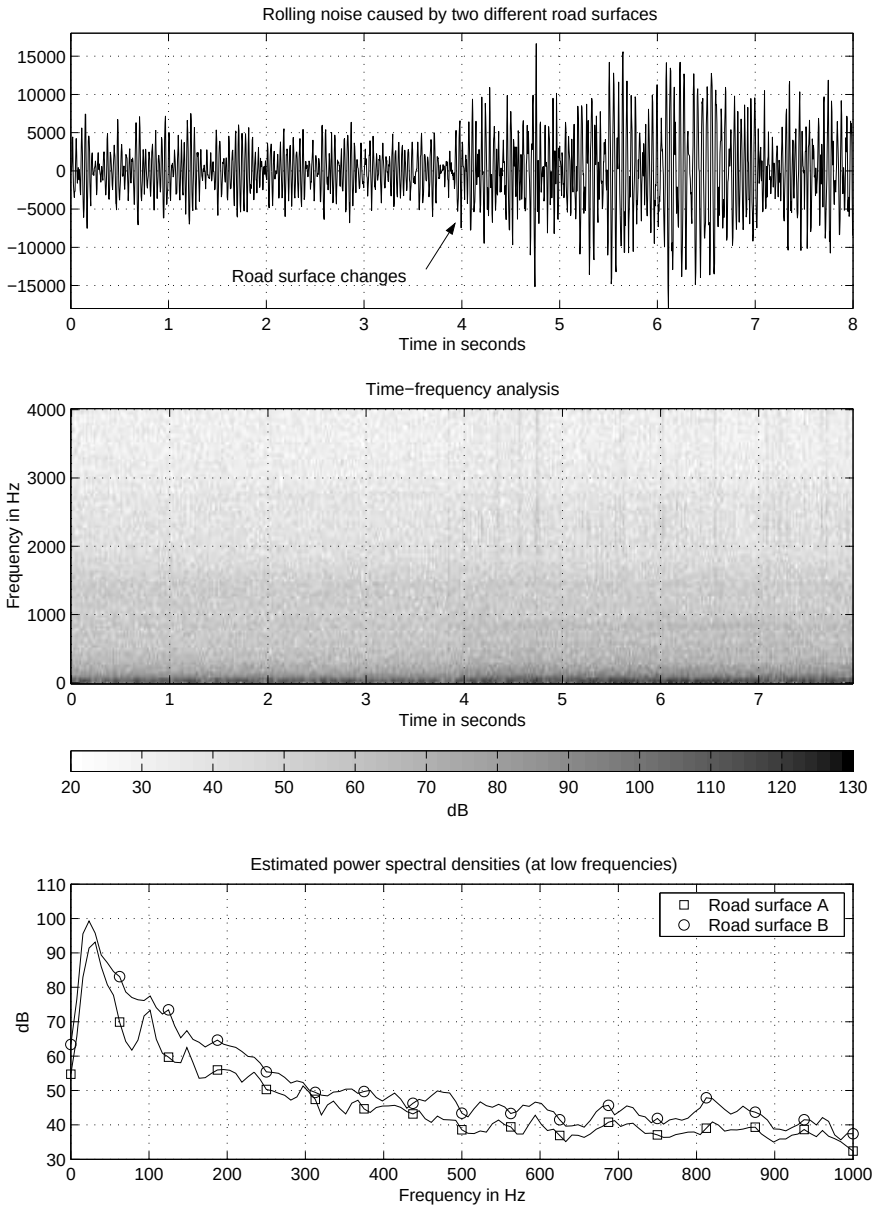

Fig. 3.10 Noise during a change in road surface: time function (upper diagram), time–frequency analysis (center), and power spectral densities (lower diagram).

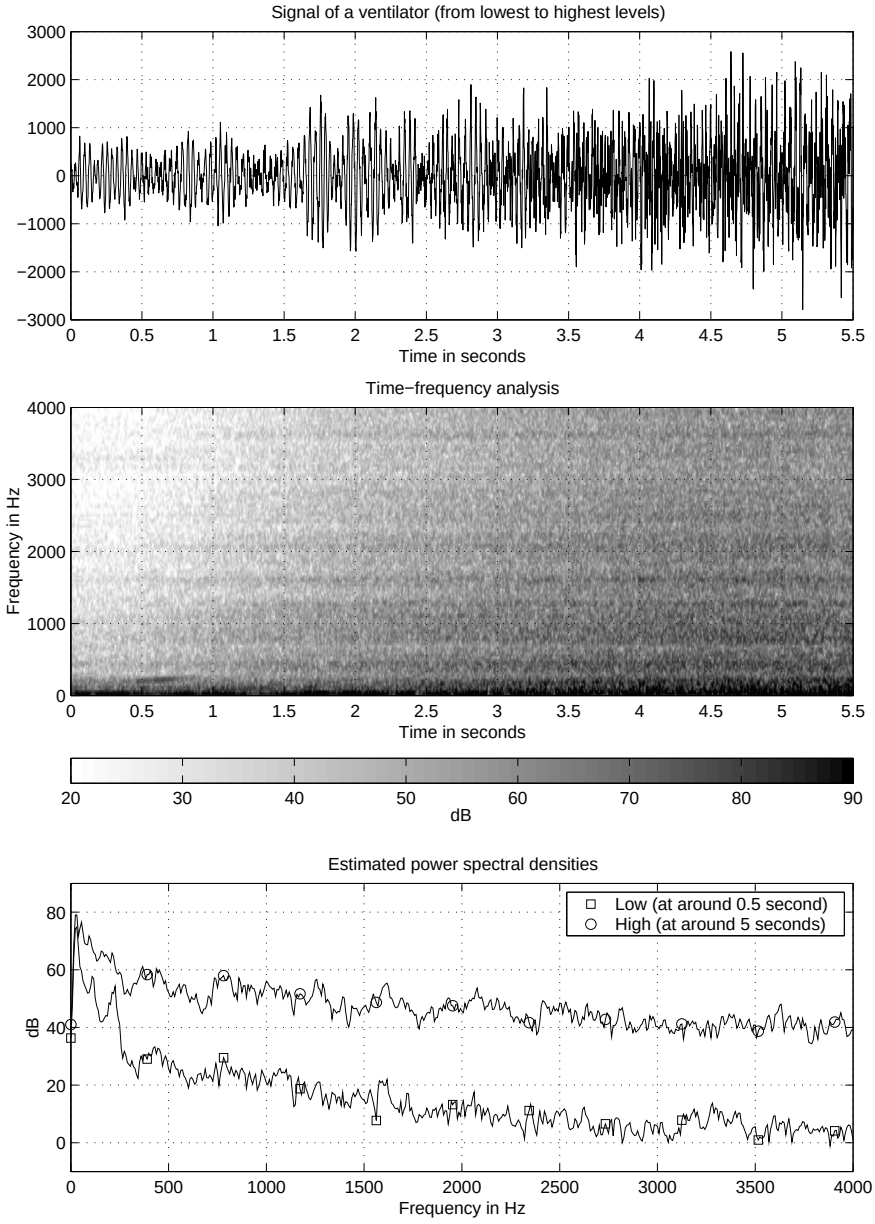

Fig. 3.11 Noise caused by a ventilator switched in steps from “low” to “high”: time signal (upper diagram), time–frequency analysis (center diagram), and power spectral densities (bottom diagram).

the probability density function of the absolute value $|B(e^{j\Omega})|$ of the noise spectrum is approximated by a *Rayleigh density*:

$$f_{|B(e^{j\Omega})|}(b) = \begin{cases} \dfrac{2\,b}{\sigma^2_{|B(e^{j\Omega})|^2}}\; e^{-\dfrac{b^2}{\sigma^2_{|B(e^{j\Omega})|^2}}}, & \text{for } b \geq 0\,, \\ 0\,, & \text{else}\,. \end{cases} \tag{3.24}$$

The power spectral density $S_{bb}(\Omega)$—estimated by $S_{bb}(\Omega) = |B(e^{j\Omega})|^2$—then obeys a χ^2 *density* with 2 degrees of freedom:

$$f_{S_{bb}(\Omega)}(S) = \begin{cases} \dfrac{1}{\sigma^2_{|B(e^{j\Omega})|^2}}\; e^{-\dfrac{S}{\sigma^2_{|B(e^{j\Omega})|^2}}}, & \text{for } S \geq 0\,, \\ 0\,, & \text{else}\,. \end{cases} \tag{3.25}$$

With respect to the wide range of amplitudes of the power spectral density a logarithmic transformation may be useful:

$$\Theta_{bb}(\Omega) = 10\,\log_{10}\frac{S_{bb}(\Omega)}{\sigma^2_{|B(e^{j\Omega})|^2}}\,. \tag{3.26}$$

Then, the probability density function transforms to

$$f_{\Theta_{nn}(l)}(\Theta) = \frac{\ln 10}{10}\,10^{\Theta/10}\;e^{-10^{\frac{\Theta}{10}}}\,. \tag{3.27}$$

In Fig. 3.12 the probability density function (see Eq. 3.27) of the logarithmic normalized power spectral density of car noise and histograms calculated from measured noise signals for selected frequencies are shown.

As a summary of the discussion of noise in a car one can state that besides stationary components it also contains highly nonstationary elements. Consequently, a noise control method can be efficient only if it can cope with both types of noise signals.

3.2 ACOUSTIC ECHOES

Loudspeaker(s) and microphone(s) in the same enclosure are connected by an acoustical path. In the following section we give a few facts from room acoustics about this path and we discuss how to model it by an electronic filter.

3.2.1 Origin of Acoustic Echoes

In a loudspeaker–enclosure–microphone (LEM) system the loudspeaker and the microphone are connected by an acoustical path formed by a *direct connection* (if both

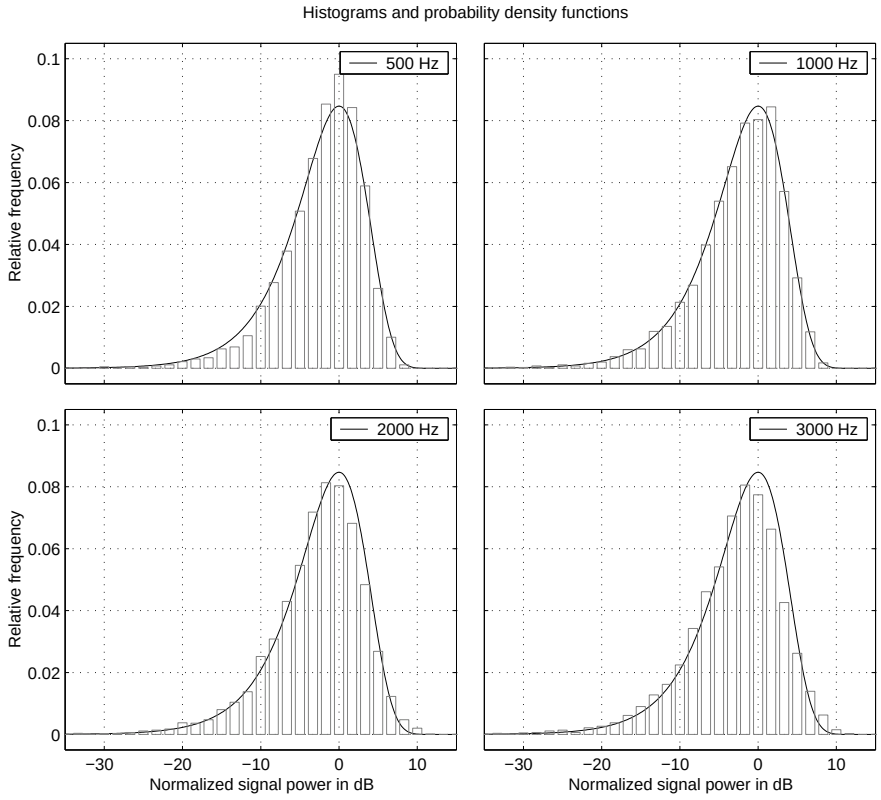

Fig. 3.12 Probability density functions of the logarithmic normalized power spectral density of car noise and histograms calculated from measured noise signals for selected frequencies.

can "see" each other) and in general a large number of *reflections* at the boundaries of the enclosure. For low sound pressure and no overload of the converters, this system may be modeled with sufficient accuracy as a linear system. The echo signal $d(n)$ can be described as the output of a convolution of the (causal) time-variant impulse response of the LEM system $h_i(n)$ and the excitation signal $x(n)$ (see Fig. 3.13):

$$d(n) = \sum_{i=0}^{\infty} h_i(n)\, x(n-i)\,, \tag{3.28}$$

where we assume for the moment that the impulse response is real.[3] The reason for having a time and a coefficient index for the LEM impulse response will become clearer at the end of this section. Signals from local speakers $s(n)$ and background

[3]For complex impulse responses, see Eq. 5.2 and the remarks in surrounding text there.

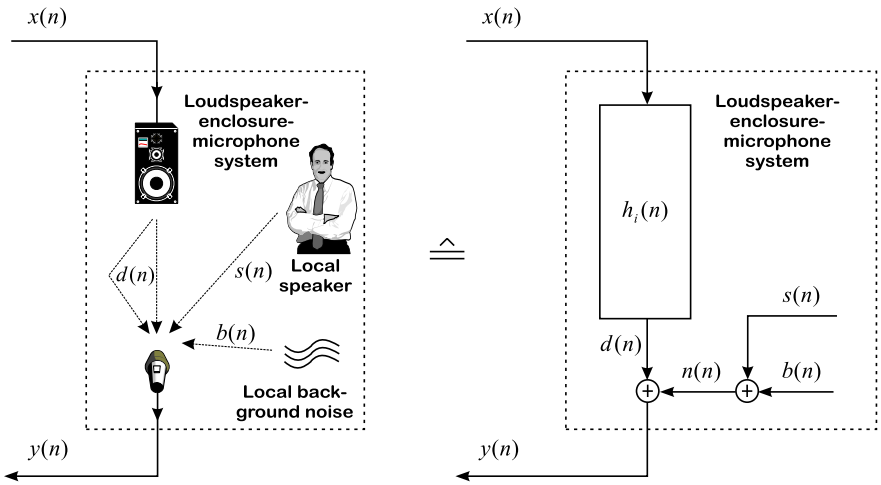

Fig. 3.13 Model of the loudspeaker–enclosure–microphone (LEM) system (see also Fig. 2.1).

noise $b(n)$ are combined to a local signal

$$n(n) = s(n) + b(n). \tag{3.29}$$

Adding the echo signal $d(n)$ and the local signals results in the microphone signal, given by

$$y(n) = d(n) + s(n) + b(n) = d(n) + n(n). \tag{3.30}$$

The time-varying impulse response $h_i(n)$ of an LEM system can be described by a sequence of delta impulses delayed proportionally to the geometrical length of the related path and the inverse of the sound velocity. The amplitudes of the impulses depend on the reflection coefficients of the boundaries and on the inverse of the path lengths. As a first order approximation one can assume that the impulse response decays exponentially. A measure for the degree of this decay is the *reverberation time* T_{60}. It specifies the time necessary for the sound energy to drop by 60 dB after the sound source has been switched off [136]. Depending on the application, it may be possible to design the boundaries of the enclosure such that the reverberation time is small, resulting in a short impulse response. Examples are telecommunication studios. For ordinary offices the reverberation time T_{60} is typically in the order of a few hundred milliseconds. For the interior of a passenger car this quantity is a few tens of milliseconds long.

To give an example, Fig. 3.14 shows the floor plan of the office used for most of the measurements and experiments described in this book. Its size measures about 18 m^2 and its volume is about 50 m^3. The floor is covered by carpet. There are

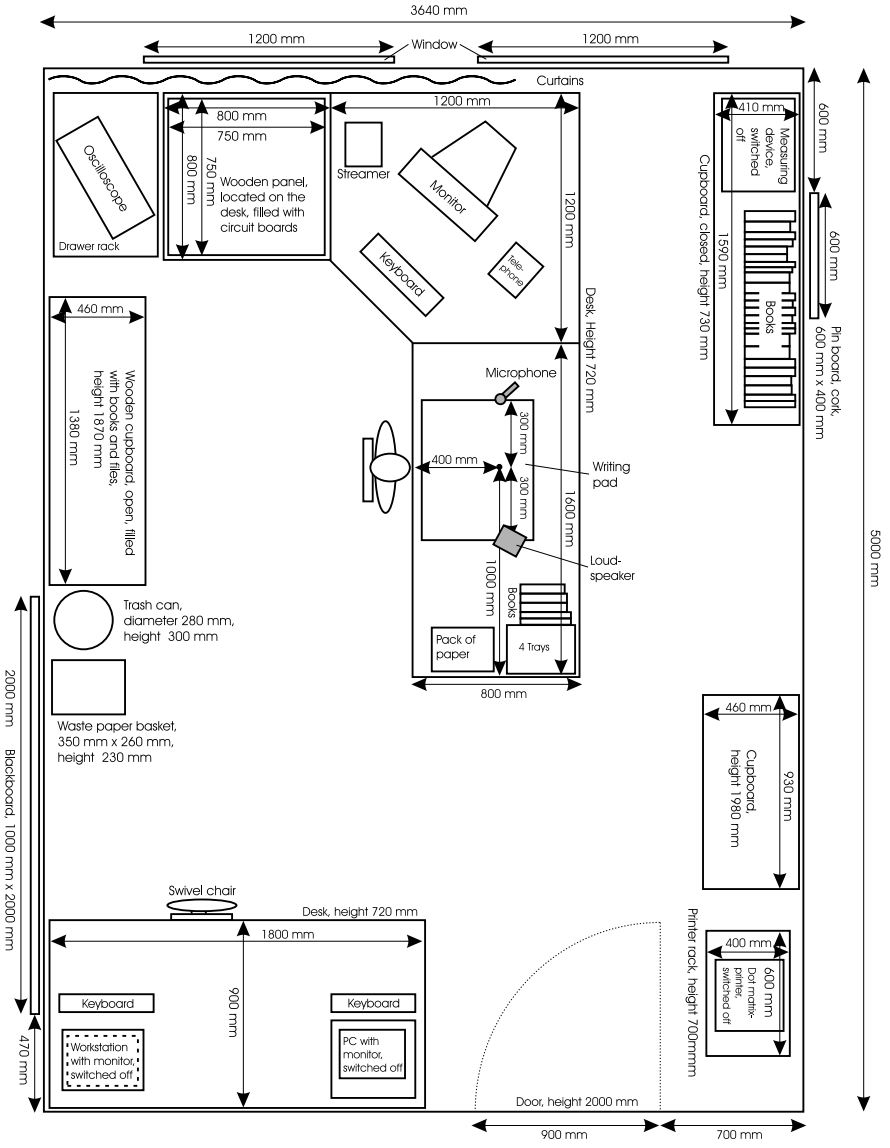

Fig. 3.14 Floorplan of an office with 300 ms reverberation time.

curtains in front of the windows. The ceiling consists of plaster board. All these materials exhibit attenuations that increase with frequency. The walls of the office are acoustically hard. The reverberation time of this office is approximately 300 ms. The microphone and the loudspeaker on the main desk are placed according to the ITU-T recommendations [120].

The impulse responses of LEM systems are highly sensitive to any changes such as the movement of a person within it. This is explained by the fact that, assuming a sound velocity of 343 m/s and 8 kHz sampling frequency, the distance traveled between two sampling instants is 4.3-cm. Therefore, a 4.3-cm change in the length of an echo path, the move of a person by only a few centimeters, shifts the related impulse by one sampling interval. Thus, the impulse response of an LEM system is timevariant. For this reason, we have chosen the notation $h_i(n)$ with filter coefficient index i and time index n for the impulse response in Eq. 3.28.

3.2.2 Electronic Replica of LEM Systems

From a control engineering point of view, acoustic echo cancellation constitutes a system identification problem. However, the system to be identified—the LEM system—is highly complex: Its impulse response exhibits up to several thousand sample values noticeably different from zero and it is time varying at a speed mainly according to human movements. The question of the optimal structure of the model of an LEM system and therefore also the question of the structure of the echo cancellation filter has been discussed intensively. Since a long impulse response has to be modeled by the echo cancellation filter, a recursive (IIR) filter seems best suited at first glance. At second glance, however, the impulse response exhibits a highly detailed and irregular shape. An office impulse response and the absolute value of its frequency response are shown in Fig. 3.15. To achieve a sufficiently good match, the model must offer a large number of adjustable parameters. Therefore, an IIR filter does not show an advantage over a nonrecursive (FIR) filter [142, 164]. The even more important argument in favor of an FIR filter is its guaranteed stability during adaptation.

In the following we assume the echo cancellation filter to have an FIR structure of order $N-1$. In this case the output of the echo cancellation filter $\widehat{d}(n)$, which is an estimate of the echo signal $d(n)$, can be described as a vector product of the real impulse response of the adaptive filter and the excitation vector:

$$\widehat{d}(n) = \sum_{i=0}^{N-1} \hat{h}_i(n)\, x(n-i) = \hat{\boldsymbol{h}}^{\mathrm{T}}(n)\, \boldsymbol{x}(n). \tag{3.31}$$

The vector $\boldsymbol{x}(n)$ consists of the last N samples of the excitation signal

$$\boldsymbol{x}(n) = [x(n),\, x(n-1),\, \ldots,\, x(n-N+1)]^{\mathrm{T}}, \tag{3.32}$$

and the filter coefficients $\hat{h}_i(n)$ have been combined to a column vector

$$\hat{\boldsymbol{h}}(n) = \left[\hat{h}_0(n),\, \hat{h}_1(n),\, \ldots,\, \hat{h}_{N-1}(n)\right]^{\mathrm{T}}. \tag{3.33}$$

A measure to express the effect of an echo cancellation filter is the *echo-return loss enhancement* (ERLE):

$$ERLE(n) = \frac{\mathrm{E}\left\{d^2(n)\right\}}{\mathrm{E}\left\{(d(n)-\widehat{d}(n))^2\right\}}, \tag{3.34}$$

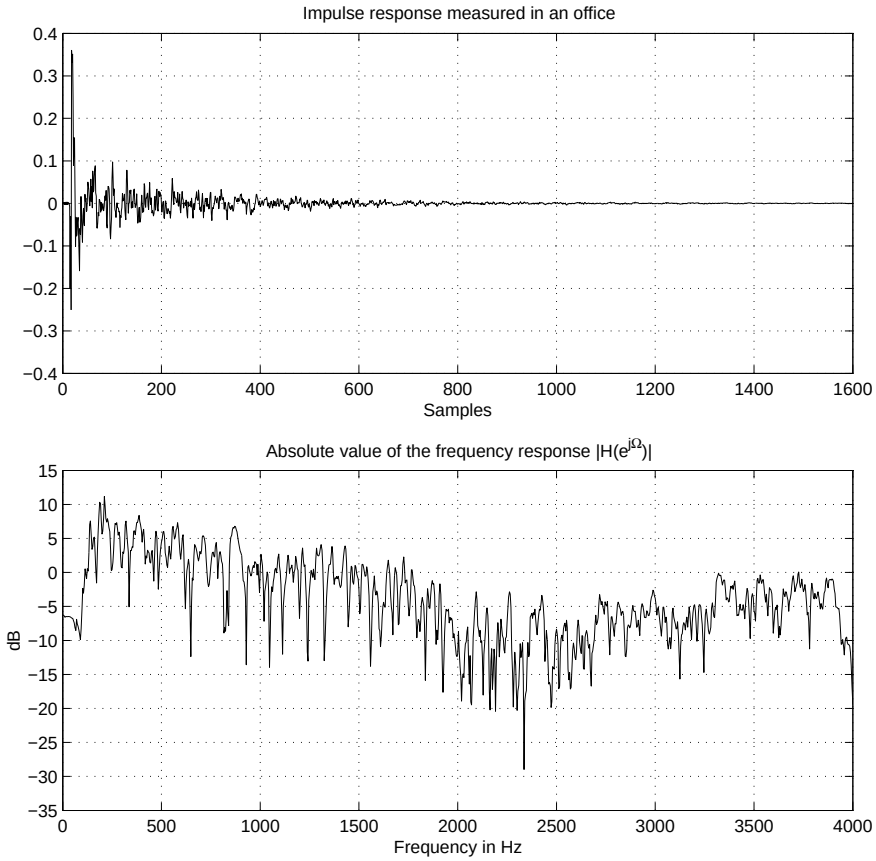

Fig. 3.15 Impulse response and absolute value of the frequency response of an office (sampling frequency $f_\mathrm{s} = 8000$ Hz).

where the echo $d(n)$ is equal to the microphone output signal $y(n)$ in case the loudspeaker is the only signal source within the LEM system; thus the local speech signal $s(n)$ and the local noise $b(n)$ are zero. Assuming, for simplicity, a stationary white input signal $x(n)$, the ERLE can be expressed as

$$ERLE(n) = \frac{\mathrm{E}\left\{x^2(n)\right\} \sum\limits_{i=0}^{\infty} h_i^2(n)}{\mathrm{E}\left\{x^2(n)\right\} \left(\sum\limits_{i=0}^{\infty} h_i^2(n) - 2 \sum\limits_{i=0}^{N-1} h_i(n)\, \hat{h}_i(n) + \sum\limits_{i=0}^{N-1} \hat{h}_i^2(n) \right)}. \tag{3.35}$$

An upper bound for the efficiency of an echo cancellation filter of length N can be calculated by assuming a perfect match of the first N coefficients of the adaptive

filter with the LEM system,

$$\hat{h}_i(n) = h_i(n) \text{ for } 0 \leq i < N. \tag{3.36}$$

In this case Eq. 3.35 reduces to

$$ERLE_{\max}(n, N) = \frac{\sum_{i=0}^{\infty} h_i^2(n)}{\sum_{i=N}^{\infty} h_i^2(n)}. \tag{3.37}$$

Figure 3.16 shows the impulse responses of LEM systems measured in a car (top), in an office (second diagram), and in a lecture room (third diagram). The microphone signals have been sampled at 8 kHz. It becomes obvious that the impulse responses of an office and of the lecture room exhibit amplitudes noticeably different from zero even after 1000 samples, that is to say after 125 ms. In comparison, the impulse response of the interior of a car decays faster due to the smaller volume of this enclosure.

The bottom diagram in Fig. 3.16 shows the upper bounds of the ERLE achievable with transversal echo cancellation filters of length N. An attenuation of only 30 dB needs filter lengths of about 1900 for the lecture room, 800 for the office and about 250 for the car.

3.3 STANDARDS

The operation of a worldwide network is only feasible on the basis of commonly accepted standards. For the telephone network regulations are formulated by the *International Telecommunication Union* (ITU) as well as by the *European Telecommunication Standards Institute* (ETSI). In addition, industry has set up rules and requirements for specific applications.

3.3.1 Standards by ITU and ETSI

ITU and ETSI set up requirements for for the maximally tolerable front-end delay and the minimal acceptable echo suppression of echo and noise control systems connected to the public telephone network.

3.3.1.1 Delay and Attenuation

For ordinary telephones the additional delay introduced by echo and noise control must not exceed 2 ms [120]. For mobile telephons up to 39 ms additional delay are allowed [64]. It is obvious that these are severe restrictions for the types of algorithms usable for any front-end processing. Especially in case of ordinary telephones, computationally efficient frequency-domain procedures cannot be applied.

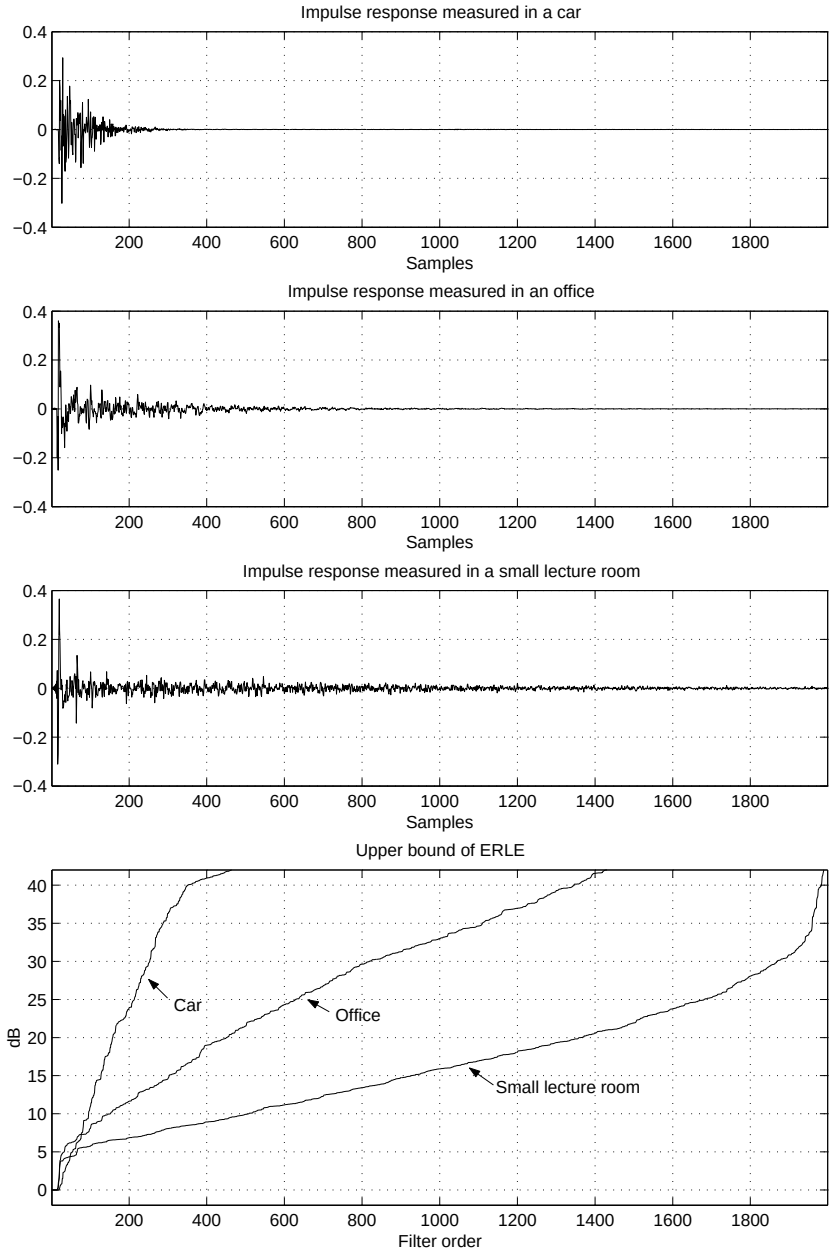

Fig. 3.16 Impulse responses measured in a car, in an office, and in a small lecture room (sampling frequency = 8 kHz). The bottom diagram shows the maximal achievable echo attenuation in dependence of the filter order of an adaptive filter placed in parallel to the LEM system (with permission from [100]).

As the level of the echo is concerned standards require an attenuation of at least 45 dB in case of single talk. During double talk (or during strong background noise) this attenuation can be lowered to 30 dB. In these situations the residual echo is masked at least partially by the double talk and/or the noise. Figure 3.17 shows the result of a study [208] that explains the dependence between echo delay and the acceptable level of the echo attenuation. The three curves mark the attenuation at which 50%, 70%, or 90% of a test audience were not annoyed by residual echoes.

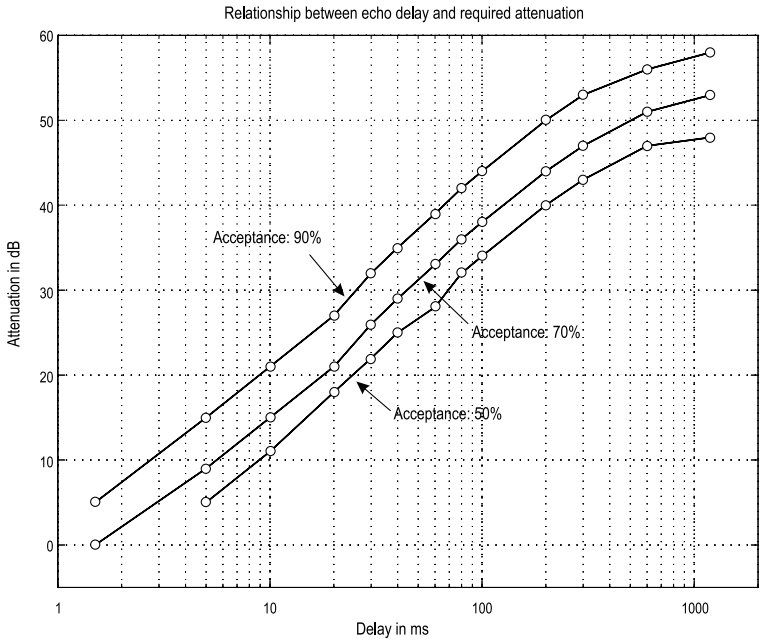

Fig. 3.17 Relationship between echo delay and required attenuation. (Based on a study of AT&T, published in [208].)

3.3.1.2 Test Signals

For measurement purposes in telecommunication systems an artificial voice signal is recommended by ITU [122]. It is mathematically defined such, that it models human speech. The long- and short-term spectra, the probability density function of speech signals, the voiced/unvoiced structure, and the syllabic envelope for male and female speech are emulated.

For communication devices that contain speech activated circuits a “composite source signal” has been defined [79, 123]. It consists of three sections: a 50-ms-long voiced signal taken from artificial voice [122] intended to activate speech detectors in the system, a pseudonoise signal of about 200 ms duration during which measurements can be taken, and a pause that is long enough to set the system back into its

quiescent state (see Fig. 3.18). The composite source signal can be repeated several times with alternating polarities.

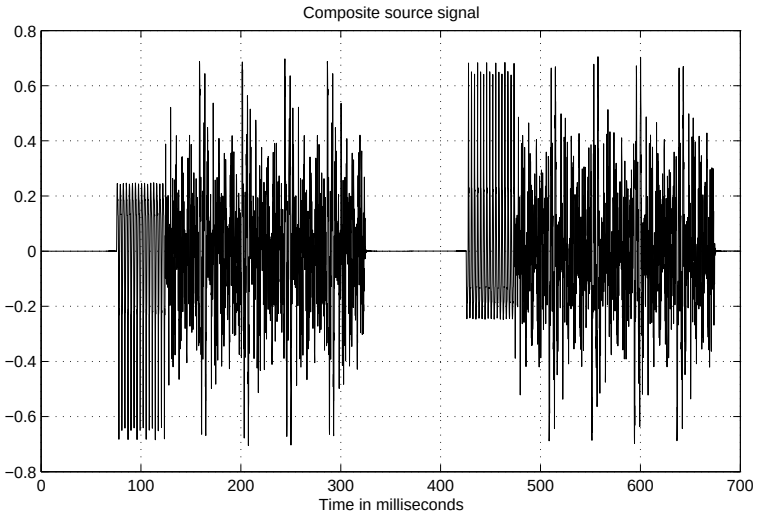

Fig. 3.18 Composite source signal: original and repetition with opposite polarity.