8

The Approach and Methodology of Psychoacoustics

Auditory psychophysics, more often called *psychoacoustics*, is important in understanding the systemic and information processing properties of auditory functions (see Section 1.3). It is, in principle, independent of physiological research and knowledge, but it is always most fortunate if physiological and psychoacoustic facts and models support each other. As will become obvious below, many (but not all) psychoacoustic phenomena find a correlate in the physiology of hearing. Modern psychoacoustics, based on systematic experimentation, has been carried out for roughly a century. Its development has been greatly influenced by engineering sciences, especially by the challenges of communication technology.

Psychoacoustics has the advantage that experimentation in its basic form is easy and non-invasive, and thus the subject under study is not in danger of physical injury. However, this does not mean that making such behavioural experiments or interpreting their results is easy. Another advantage of the psychoacoustic approach is that higher level functions of the auditory system can be studied where physiological knowledge is missing or too weak to support our understanding.

Studies on auditory sensation and perception can be compared to measurements of a very complex physical system that is inherently non-linear, time-varying with both short-term and long-term effects, and it also shows minor or major variation due to innumerable other factors than those specifically being studied. It is like having an unreliable measurement device for measuring an unpredictably behaving system. With proper methodology and for properly specified problems, the task is, however, manageable and leads to useful theories and models.

8.1 Sound Events versus Auditory Events

Understanding psychophysical experimentation can be based on the system diagram shown in Figure 8.1. The outer box encloses the subject under study. Sound stimuli consist of *sound events* (or *sound objects*) s that enter the auditory system of the subject. They may be anything from simple tones to combinations of complex sound sources, deterministic sounds, or noises.

The equivalent of an external sound event, internal to the subject in Figure 8.1, is an *auditory event* (or *auditory object*) h_i. The subject has more direct access to this internal event than any

Communication Acoustics: An Introduction to Speech, Audio, and Psychoacoustics, First Edition.
Ville Pulkki and Matti Karjalainen.

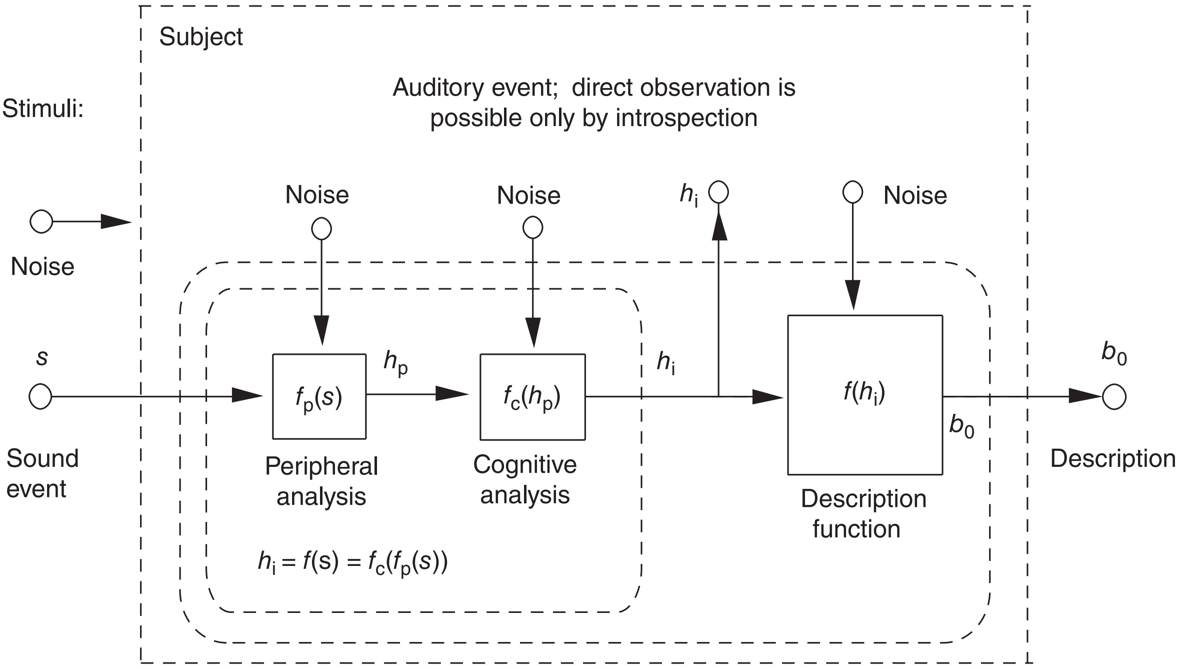

Figure 8.1 Psychophysical experimentation as a process where a subject is exposed to a sound event s (physical sound stimuli) and the external observer has only indirect access to internal auditory events h_i through an externalized description or reaction b_0.

external observer. This ability to observe one's own internal events is called *introspection*. This may be useful to a researcher in order to acquire a general picture of the auditory perception, but, on the other hand, introspective observation is seldom accepted as a scientific method as such. This is because of the subjective nature of such observations that may easily have a strong bias due to many disturbing factors and the fact that a single subject may not be representative of general behaviour. Therefore, in general, auditory events should be studied by an external observer, statistical methods should be applied (or at least one must be aware of the statistical nature of the process), and several subjects should be involved, unless the goal is to know the behaviour of a specific subject.

Figure 8.1 indicates that the peripheral analysis f_p produces result $h_p = f_p(s)$, which is available neither for the subject nor for the external observer. The output of the periphery is directly mapped to auditory event h_i with analysis f_c, as $h_i = f_c(h_p)$. h_i is available only to the subject through introspection. Both peripheral and cognitive transforms are disturbed by noise, which naturally has a different nature. The noise for peripheral analysis consists of such components as Brownian noise at the eardrum and tinnitus. The cognitive transformation may, in turn, be distracted by changes in concentration of the listener for various reasons.

The goal of an investigation may be to seek a relation $h_i = f(s)$, where $f(s) = f_c(f_p(s))$, which describes how the attribute(s) of the auditory event s is mapped onto the attributes of an auditory event. The external observer encounters another transform, $b_o = f(h_i)$, a mapping from the internal representation to the external reaction that can ultimately be registered objectively. This relation may be called the reaction function or the description function, depending on how the external observations are carried out. In a sense, it corresponds to the registration or display function of a measurement device used in physical or physiological experiments. Since the study of the mapping $h_i = f(s)$ is the focus of the investigation, the influence of $b_o = f(h_i)$

and noise should be eliminated or minimized. Statistical analysis is a powerful tool for this if the observed data are 'noisy'. If the effect of the reaction function $b_\mathrm{o} = f(h_\mathrm{i})$ is not known well enough, different ways to study the same internal events may help to improve the reliability of interpreting $h_\mathrm{i} = f(s)$. It is beneficial if there are physiological facts and knowledge that support the interpretation.

8.2 Psychophysical Functions

A *psychophysical function*, characterized by $h_\mathrm{i} = f(s)$ in Figure 8.1, represents the relation between one or several properties of a sound event s and one or several properties of an auditory event h_i. Psychophysical functions may be mappings from one continuous scale to another continuous scale (such as sound pressure level → perceived loudness), from a continuous scale to a discrete scale (say, sound pressure level → audible or inaudible), and so on. A specific type of psychophysical function called the *psychometric function* refers to the mapping from a continuous scale to a yes/no scale expressed as a probability function of the detection of a signal.

In psychophysical functions from a continuous physical attribute to a continuous sensation variable, the auditory analysis does not typically make a linear mapping. The first studies on psychophysics were conducted in the early 1800s, when the Weber–Fechner law was derived. It was assumed that these mappings followed logarithmic characteristics

$$h = \mathrm{a}\ \log(s), \tag{8.1}$$

where h is, for example, the subjective loudness of a tone; s is a physical attribute, such as sound pressure; and a is a constant. In more careful studies, it turned out that psychophysical functions can have different forms. As will be discussed below, with SPL above about 40 dB subjective loudness follows the power law (rather than a logarithmic law)

$$h = \mathrm{c}\, s^{\mathrm{k}}, \tag{8.2}$$

where c and k are constants ($\mathrm{k} \approx 0.6$; see Section 10.2.3). On the other hand, the pitch (height) of a tone is almost a logarithmic function of the frequency of the tone, as will be shown in Section 10.1.

Each attribute of an auditory event depends typically on many properties of the sound event. For example, the loudness of an event depends not only on the sound pressure level of the sound event, but also on, for example, the frequency content and the temporal duration of sound. Often, one of the physical properties of a sound event is dominant, like the fundamental frequency of a tone complex that mostly defines the pitch.

8.3 Generation of Sound Events

In psychoacoustic tests, the sounds should be designed in such a way that the subject can report the characteristics of the auditory event reliably, which will eventually reveal properties of the peripheral or cognitive functions in hearing. This section describes the methods most commonly used to generate the sound events for different listening conditions.

8.3.1 Synthesis of Sound Signals

Relatively simple stimuli are often used in fundamental research on psychoacoustics. The stimuli are presented in more detail in Section 3.1.2 on page 45. These include:

- *Pure tone*
- *Amplitude- or frequency-modulated tone*
- *Tone burst*
- *Sine-wave sweep*
- *Chirp signal*
- Single *pulses*
- *White noise*
- *Pink noise*
- *Uniform masking noise*
- *Modulated noise*

The first five stimuli can also be realized using other simple signal waveforms such as a square wave, a sawtooth wave, an impulse train, and their filtered (low-pass, high-pass, band-pass) forms.

Since the auditory system is highly developed to receive complex sounds from the environment and other communicating subjects, simple stimuli are insufficient for psychoacoustic research. It is increasingly important to study the perception of complex sounds such as:

- *Harmonic tone complexes*;
- *Complex combination sounds* including inharmonic sounds;
- *Combinations of sinusoids, noises, and pulses*;
- *Speech sounds*: real speech and synthetic speech;
- *Musical sounds*: acoustic and electronic music;
- *Sounds from nature*: from animals and inanimate nature;
- *Noise*: harmful, loud, or annoying sounds.

Alternatively, the sound signals to be tested may originate from an engineering task, where the effect of processing an input signal with a system having different parameters is of interest. For example, in audio coding applications, the effect of the data rate on perceived quality of sound can be studied by processing a sound sample with codecs using different settings.

The non-linear processing in the ear causes sound with different sound pressure levels to be perceived differently. The loudness differences between the samples may cause undesired effects if the loudness itself is not studied. Thus, when conducting a listening test, the effect of the level of presentation of the signals should be taken into account. For example, if the audio quality provided by different loudspeakers is tested, and if one loudspeaker delivers slightly but noticeably higher SPL to the listener, it quite probably will be rated to provide the best quality of the tested items.

The effect of loudness should typically be avoided in the tests, and equal loudness should be produced by the listening test signals. Depending on the case, the task may be simple or complex. In some cases, the equalization of the signal energies may be sufficient, while in others no computational metric available is sufficient, and separate listening tests have to be organized to set the perceived loudness levels to be equal. Interested readers are referred to

Bech and Zacharov (2006) for a detailed discussion on *level calibration*, which is a process that aims to equalize the levels of the test items.

8.3.2 Listening Set-up and Conditions

Psychoacoustic experiments require that attention is paid not only to the sound signals but also to the acoustic environment and to how stimuli are presented. The sound source can, in principle, be any source that generates a desired and well-controlled sound. In practice, the stimuli are most conveniently generated by computers and played by electroacoustic reproduction means:

- *Loudspeakers* (Section 4.1.1), one or many, controlled by a single or several audio signals. The best control over the sound field is achieved in non-reverberant, free-field conditions in an *anechoic chamber*. If a loudspeaker is close to the listener (≤ 1 m), the sound field can be approximated by a spherical field. A plane wave can be approximated by placing a loudspeaker far enough away (≥ 2 m) in an anechoic chamber. Multiple loudspeakers are often needed when special effects of spatial hearing are being studied.
- *Headphones* (Section 4.1.1), which are in some cases an ideal source of sound, since the reverberant environment can be eliminated, and some headphones also attenuate external noise. Headphones are the only choice if very different sound stimuli are needed in each ear. On the other hand, spatial attributes may be difficult to reproduce using headphones unless very careful binaural reproduction techniques are applied (see Chapter 12).

It is important to pay enough attention to the acoustic environment of psychoacoustic experiments. If a very carefully controlled free field (anechoic chamber) is not needed, conducting experiments in a specially designed *listening test room* that resembles a living room with good acoustics may be better. As an example, conducting listening tests in an anechoic space when studying audio quality would give misleading results, since the reverberant field in normal rooms immensely affects the listening experience. Loudspeakers are meant to be used in normal rooms, and the response obtained in an anechoic chamber can be very different from the response in a normal room (Bech and Zacharov, 2006).

Sometimes a special room is necessary, for example a reverberation chamber, if real reverberation is being studied. Background noise should be minimized, unless it is an integral part of the study. Eliminating visual and other undesirable cues is also an important issue, since they may easily bias results or draw attention away from the focus of the study. (Of course there are also cases where just these effects are studied and therefore they have to be included.) Hearing is particularly influenced by vision, especially in the processing of information where vision is more reliable, such as in localization, object identification, and size. In these cases we ‘often hear what we see’ rather than what our ears receive.

Computers and digital signal processing have made psychoacoustic experiments easier and more precise. Computers with high-quality audio interfaces are ideal for generating practically any sound with high quality and repeatability. An inherent lack of ideality in loudspeakers and headphones may be compensated for by DSP, as was discussed in Section 4.3.

8.3.3 Steering Attention to Certain Details of An Auditory Event

An auditory event is often a comprehensive percept, where it is difficult or impossible to concentrate on specific parts or characteristics. For example, it is often impossible to concentrate on a single harmonic of a harmonic complex tone. Depending on the case, different methods can be used to route the attention to a specific aspect of sound. For example, a partial modulated by amplitude or frequency is easier to perceive. Presenting the sound first without the harmonic and then with it is another way of focusing attention on the harmonic. A third approach is to first play only the harmonic tone and then the whole complex with the harmonic.

8.4 Selection of Subjects for Listening Tests

When conducting the tests, it is also necessary to control the level of listening test experience of the subjects. The panel of test subjects may be trained to be as sensitive as possible to the researched attributes of sound, unexperienced listeners, or something in between. A trained panel is needed if one wants to conduct reproducible tests at the finest resolution of a specific property of hearing. Training makes subjects more ‘analytic’: they learn how to analyse the auditory input and to describe it in an objective manner. On the other hand, the opinion of trained listeners may not represent the opinion of the larger audience, as, for example, in product sound quality questions. In such cases, a relatively large set of listeners from the population segment of interest must be used. However, here, too, the ability of the subjects to report the properties of interest in the sound under study has to be taken into account (Bech and Zacharov, 2006).

8.5 What are We Measuring?

In psychoacoustic experiments, sound events are presented, and a question is asked of the subjects, which they should answer based on the properties of the auditory event created by the sound event. Ideally, they should respond to the question such that the attribute being investigated is revealed in the results of the tests. Typical properties which are measured are different thresholds. The relation between attributes of sound events and attributes of auditory events, that is, psychophysical functions in general, is also of interest. Auditory scales quantify these relations. Thresholds and scales used are briefly introduced below.

8.5.1 Thresholds

A basic question in the research on hearing is if any kind of auditory event is formed with a sound event, either in silence or in the presence of noise. There are two main types of threshold values:

- *Absolute threshold*: For example, the threshold of the sound pressure level for detecting an auditory event; in other words, the hearing threshold measured in audiometry. It quantifies the value of an attribute of the sound event and the respective psychoacoustic quantity above which the auditory event emerges. The corresponding task is called the ‘detection task’.
- *Difference threshold*: The smallest change in one of the attributes of the sound event that is audible. A synonym is the term ‘just noticeable difference’ (JND). A special case of difference thresholds are different *modulation thresholds*, such as just noticeable amplitude or frequency modulation of a tone. The corresponding task is called the ‘discrimination task’.

Note that absolute and difference thresholds can be quantified on both acoustic and auditory scales.

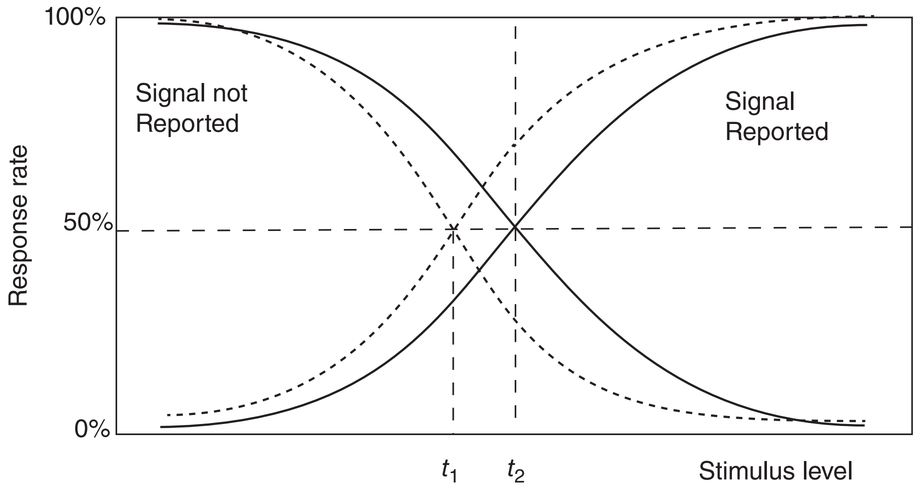

Figure 8.2 Psychophysical functions for the measurement of the absolute threshold of a sound event, where a subject is presented a signal with variable level and the rate of answers is measured. The dashed line shows the curve of the 'optimistic' type of subjects with threshold of t_1 and the solid line the curve for 'neutral' subjects with the threshold t_2.

The mapping of a physical quantity onto an auditory quantity is thus binary, either the value of h is above or below the threshold, as discussed in relation to Figure 8.1. However, the mapping does not change abruptly near the threshold, there is always a region where the auditory event may or may not be formed. This depends on some internal state of the subject, such as the level of inherent noise in sensory systems. When a listening test is organized and subjects are asked whether they perceive the auditory event, the description function $b_o = f(h_i)$ also comes in to play. For example, the personal character of the listener may have a large influence on the description, as typically some subjects tend to perceive auditory events although there are no sound events, and other subjects are more insecure and report an auditory event only when it is very clearly audible. This is illustrated conceptually in Figure 8.2. The rate of reporting that the signal is present is shown as a function of the signal level for two types of subjects: 'optimistic' and 'neutral'. Clearly, the test would produce very different threshold values if the stimulus level producing the rate of 'signal present' with a value of 50% was taken as the value of the threshold, as shown in the figure with vertical dashed lines.

Signal detection theory

This bias can be avoided if the subject is presented with an interval where the signal is or is not present. The interval is thus a period of time which can be indicated, for example, with a visual indicator on a computer screen. The length of the period is typically shorter than 10 seconds. The subject is asked whether the interval contains the signal, with the possibility of answering only 'yes' or 'no'. This gives information about the tendency of the subject to favour either 'signal present' or 'signal not present' cases. The analysis of the results from such testing is formalized in *signal detection theory*, which is a theory of perceptual mechanisms when measuring thresholds (Gescheider, 1997). It was developed during World War II to correctly detect planes in noisy radar measurements.

Signal detection theory can be applied to different cases, and we present it in its basic form. In the basic version, the subject is presented with a single interval containing either noise with the signal at a known low level near the threshold or just plain noise and asked to determine the presence or absence of the signal. This task is presented to the subject many times, and the rates for both ‘signal present’ and ‘signal not present’ are measured as percentages. The answers are then categorized as ‘hit’, the signal is present; ‘miss’, the signal is present but not perceived; ‘correct rejection’, the signal is not present and not reported; and ‘false alarm’, the signal is not present but reported. According to the theory, presenting the signal at a fixed level sets a value to the internal variable $h_{\mathrm{s}}(t)$, which is not constant but varies due to the presence of external and internal noise. The mean value of $h_{\mathrm{s}}(t)$ in this case is denoted as μ_{s} and the standard deviation as σ_{s}. When the signal is not present, the value of the internal variable $h_{\mathrm{n}}(t)$ is assumed to have a lower mean value μ_{n} than μ_{s}, and the external and internal noise then causes the standard deviation μ_{n}.

The sensitivity d' is a statistic that shows the separation between the probability density functions of $h_{\mathrm{s}}(t)$ and $h_{\mathrm{n}}(t)$, which represent the distributions of the internal variable h in the presence of signal+noise and only noise, respectively. The sensitivity d' of the subject perceiving the signal at a given level can then be written as

$$d' = \frac{\mu_{\mathrm{s}} - \mu_{\mathrm{n}}}{\sqrt{(\sigma_{\mathrm{s}}^2 + \sigma_{\mathrm{n}}^2)/2}}. \tag{8.3}$$

In some cases we cannot directly measure the variables and their standard deviations, since h_{s} and h_{n} exist only as internal variables. In such cases, d' can still be estimated as

$$d' = Z(\text{hit rate}) - Z(\text{false alarm rate}), \tag{8.4}$$

where Z is the inverse of the cumulative Gaussian distribution. For example, if the hit and false-alarm rates are both 50%, $d' = 0$. For corresponding rate pairs (hit rate, false-alarm rate), (0.7, 0.4)$\rightarrow$ $d' = 0.78$, (0.9, 0.1)$\rightarrow$ $d' = 2.1$ and (0.9, 0.3)$\rightarrow$ $d' = 1.8$. A more thorough description of the theory and further applications of it can be found in Gescheider (1997).

8.5.2 Scales and Categorization of Percepts

The absolute or difference thresholds can only be used to measure psychophysical functions that relate attributes of sound events to simple functions having only two values. Often, a psychophysical function which represents an auditory percept with a continuous scale is desired, for example the mapping of sound pressure levels between 0 dB and 120 dB to the perceived loudness of an auditory event. There are many methods to estimate such psychophysical functions.

The task is, naturally, very complicated, as the subjects cannot access the absolute measure of the auditory attribute. For example, the subject’s ability to report, say, the loudness of a sound on an absolute scale or in relation to a sound heard more than a few seconds earlier is limited. As will be shown later, some clever systems have been found to define complex psychophysical functions, such as the relation of the SPL and loudness, or the sound spectrum and the loudness spectral density.

When estimating the magnitude of sensation on a scale, subjects are asked to describe a characteristic of an auditory event on a response scale either by itself or in comparison to other auditory events. In such tests the whole psychophysical function spanned by the attributes of the sound event may be researched, although a number of problems are inherent in such measurements, as already discussed at the beginning of this chapter. Different scales used in the tests are described below:

- The simplest response scale is the *nominal* scale, where the response (a number or symbol) implies that the auditory event belongs to a certain class. The classes may be, for example, *rough*, *reverberant*, *bright*, and so on. In nominal classification, the auditory events or their characteristics are not compared on any quantitative scale.
- When the auditory events, or some of their audible characteristics, can be ordered in an array, such a scale is called the *ordinal* scale. The position of an auditory event in the array is denoted with a positive integer, which does not mean that the differences between the ordinal numbers can be used as a measure of dissimilarity. Arithmetic operations between values are not applicable.
- The *interval scale*, in turn, is a numerical scale, which defines the differences of classes quantitatively. The zero point of the scale is not meaningful, because only the differences of subsequent values on the scale are defined. The valid arithmetical operations are thus based only on the differences.
- The *ratio scale* is defined similarly to the interval scale, but with the zero point defined. The valid analysis methods also include such operations where the position of zero on the scale has meaning (such as the geometric mean).

8.5.3 Numbering Scales in Listening Tests

The user interface normally contains a numerical scale, which is assumed to help the subject to describe the auditory attribute being studied. The numerical scaling of an auditory event is often performed on a scale with easily conceivable numbers, such as $p_i \in [1, 5]$, $p_i \in [1, 10]$ or $p_i \in [0, 100]$. A special case is the Mean Opinion Score (MOS) scale, MOS $\in [1, 5]$, which is commonly used to evaluate audio quality. A verbal description may correspond to different positions on the scale, for example, the value 5 may correspond to ‘excellent’ quality. MOS scales are discussed in more detail in Section 17.4.1.

A scale can also be set up using two concepts exhibiting opposite values, the *semantic differential*. Such pairs are, for example,

- soft ↔ hard
- low ↔ high
- distorted ↔ clean

In the case of clearly opposite values, defining the scale symmetrically around zero can be meaningful, as in $p_i \in [-5, 5]$. The neutral case corresponds, in this case, to the value zero.

8.6 Tasks for Subjects

In formal listening tests, a task is given, sound events (or silence) are presented, and the subject responds to the auditory event according to the task. The subject typically has access to a human–computer user interface, such as a computer keyboard, a touch screen, or a specific response device. In some cases, speech- or movement-based reporting may also be used.

In any case, it is important to eliminate all sources of error and bias in psychoacoustic tests. The ideal situation is a blind test, where subjects have as little information as possible on the sounds they are hearing. The wording of the questions or tasks is very important in this respect.

Typical tasks used in listening tests are described below.

- *Detection*: The subject is presented with a single interval, which may contain a signal, noise or both, and a relevant question is asked, such as, ‘did you hear a sound?’
- *Discrimination*: The subject is presented with a single interval containing two sound events with a small difference in an attribute of the sound event. The question may be, for example, ‘do you hear a difference?’
- *Forced choice*: Subjects are presented with a number of temporal intervals from which they have to choose one based on the question asked. One of these intervals contains the signal while others are silent or contain some other sounds. The choice is thus based on the comparison of auditory events with a predefined criterion. Depending on the sounds presented to the subject, the task can be used to measure thresholds or can be scored. Such tests have different nomenclatures, such as *two-interval forced choice* (TIFC or 2IFC) or *two-alternative forced choice* (TAFC or 2AFC). The number of intervals can also be higher. Forced-choice methods are not sensitive to bias produced by subjects’ tendencies, as discussed in Section 8.5.1.
- *Direct scaling*: Subjects are presented with the sound event being studied, which may be presented only once or it may be accessible many times using a user interface, after which they must report the magnitude of the sensation on a given auditory scale. This is also called *magnitude estimation*, or *grading*. The question asked may be, for example, ‘how loud is the sound on a scale from zero to ten?’
- *Adjustment*: The task of the subject is to adjust the value of an attribute of a sound event until a desired attribute is obtained. This is more commonly called the *method of adjustment*, and it will be discussed more in detail in the next section.
- *Chronometric tasks*: Here, subjects are given a task where they must react to a specified auditory event as quickly as possible. An example task is ‘press button A as quickly as possible when you hear a voice.’ Different auditory events are then presented, and conclusions are drawn from the subjects’ reaction times.
- *Verbal description*: Subjects are required to describe verbally the sounds they perceive. The description can be done using questionnaires, with free-form textual or oral descriptions, or by other means. The subject may also be asked to answer a formal question or to perform a task, and the verbal description complements the results. There are also methods that apply statistical tools to analyse verbal descriptions. Often, the target of such tests is to seek the perceptual dimensions in the background of a complex sound event, such as in product quality. In some cases, informal descriptions can also be reported, although they are seldom found sufficient for drawing conclusions in psychoacoustic experimenting. However, they may be a useful addition to other results.
- *Other tasks*: Different types of tasks can be utilized depending on the topic being studied and on the subjects. For example, a psychoacoustic test can be implemented as an application similar to a computer game, where the scoring of the subject defines the result. A typical use of such a task is to test hearing aids with children, who typically cannot concentrate on mechanical tasks for long periods. The use of such applications can produce less biased results than the simple use of formal tasks.

8.7 Basic Psychoacoustic Test Methods

So far, we have discussed the generation of sound events and auditory events, the scales (or psychophysical functions) to be researched with the tests, and also the tasks for the subjects. Another dimension in designing psychoacoustic tests is the procedures – how the tasks discussed in the previous section are presented to the subject in succession to ensure that meaningful results will be produced about the phenomenon being studied. The test method is here defined to be the logic underpinning how the attributes of sound events are chosen into the subsequent tasks presented to the subject. Psychoacoustic tests can be conducted using many different methods. The most important methods are described below.

8.7.1 Method of Constant Stimuli

The *method of constant stimuli* is used to quantify thresholds. To this end, the experimenter chooses a relatively large number of sound event attribute values around the assumed value of the absolute or difference threshold. The listening test is conducted for each attribute value with a detection or discrimination task, or preferably with a multiple-interval forced-choice task to avoid bias. The task must be repeated a considerable number of times, and a psychophysical function such as that shown in Figure 8.2 is obtained. The actual value for the threshold can then be selected to be, for example, the value of the abscissa where the function has the value 50%.

In some cases the value of the psychophysical function does not approach 0% at the lower end of the scale. This happens, for example, if the difference threshold is measured with the 2AFC method, where the chance of guessing correctly is 50% when the difference between the signal and the reference is below the threshold. In this case, the value of the threshold may be selected to be 71% or 75% of the maximum of the function.

In principle, the method of constant stimuli is the best method to measure the value of a threshold, as it avoids many subject-related sources of error. Additionally, the shape of a psychophysical function is also revealed with the method, whereas other methods reveal only the value of a certain threshold. Unfortunately, it is a relatively slow method to conduct. The number of presented stimuli required to obtain reliable data is, in many cases, relatively high, and often some adaptive methods are used instead.

8.7.2 Method of Limits

In the *method of limits*, an attribute of a sound event is changed automatically or by the experimenter, and the sound is presented to the subject during an interval. The subject can be given a *detection task*, reporting whether the stimulus was present in an interval. The gathered data are then used to measure absolute thresholds.

In an ascending series, the stimulus attribute value is first set well below the threshold and is then increased until the response changes. The attribute value where the response changes is called the ‘limit’. In a descending series, the opposite is performed: the stimulus attribute is decreased from a level well above the threshold until the response changes. The experiment consists of many runs in both directions, possibly distributed randomly. The average of all the obtained limits is taken as the threshold.

Alternatively, in measurements of difference thresholds, two sound events are presented with a small difference in their attributes, and subjects perform a *discrimination task*, reporting if they perceive a difference in the auditory events. The method of limits, in general, is prone to bias, since the tendency to report the stimulus one way or another affects the results.

8.7.3 Method of Adjustment

The *method of adjustment* is actually a task, as already mentioned in the previous section. Here, the subject changes an acoustic attribute of the sound event until the auditory event corresponds to a reference value. For example, the level of a tone is adjusted to a level where it is just noticeable, or the frequency of a tone is adjusted to match the perceived pitch of another sound. The adjustments are conducted many times, and the results are averaged. The subject must be instructed to adjust the attribute value to one higher and lower than the reference before picking the final value. If possible, the adjustment is made in steps that are of the order of the JND to avoid the subject overestimating the change in the auditory event due to a minimal adjustment, the effect of which is actually not perceivable (Cardozo, 1965).

Besides finding thresholds, this method can be used in other tasks. For example, in *magnitude production*, the subject is asked to adjust a certain attribute of a sound event until the desired magnitude is reached. Similarly, in *ratio production*, the adjustment is made to obtain a ratio between the auditory attribute of each of two percepts. In early psychoacoustic experimenting, many of the tests were conducted by asking the subject to adjust an acoustic attribute of a test sound so as to produce the attribute that corresponded to, say, twice that of a reference value. By repeating this procedure for a new reference value every time, a relative psychophysical scale is derived, which can further be made into an absolute scale by choosing a single anchor point with a specified *anchor sound*. The subjective attribute value of the anchor sound corresponds to a certain value of the objective attribute value. This is used, for instance, for loudness and pitch scales defined later in this book. This technique has many variations.

8.7.4 Method of Tracking

In the *method of tracking*, the subject influences the direction of change of the studied attribute of sound. The historical example of this method is *Békésy audiometry*, where the task of the subject is to press a button whenever hearing a sound (von Békésy, 1960). The level of the tone whose frequency is swept gradually decreases as long as the button is pressed and increases when it is released. The local average of the level function with frequency can be taken as an absolute hearing threshold of the subject. The method is prone to bias, since the tendency to produce false positive and negative answers is not taken into account.

8.7.5 Direct Scaling Methods

Different methods are used to measure psychophysical functions with a direct scaling task. Measurements using a single auditory event only are prone to noise and bias, since the human senses have limited accuracy when evaluating the strength of a stimulus on an absolute scale. It is often beneficial to conduct the test so that the sound event being studied is compared to one or more known reference sound events, which are sometimes called *anchors*. The subject may or may not be aware which of the sound events are really the references. In the context of Figure 8.1, this means that the effect of the reaction function $b_{\mathrm{o}} = f(h_{\mathrm{i}})$ is minimized when the subject compares two auditory events with a minimal difference.

The task may also be to scale multiple sound events, which are compared with each other and potentially with some reference sound events. In audio quality measurements, such tests are often referred to as the multiple-stimulus-hidden-reference-with-anchors (MUSHRA) type of tests, which will be discussed in Section 17.4.2 in more detail.

8.7.6 Adaptive Staircase Methods

Adaptive staircase methods are similar to the method of tracking, except that, typically, forced-choice tasks are used. The value of the tested attribute is altered based on whether the subject's answer is correct or not. A wrong answer changes the attribute to make the task easier. Correct

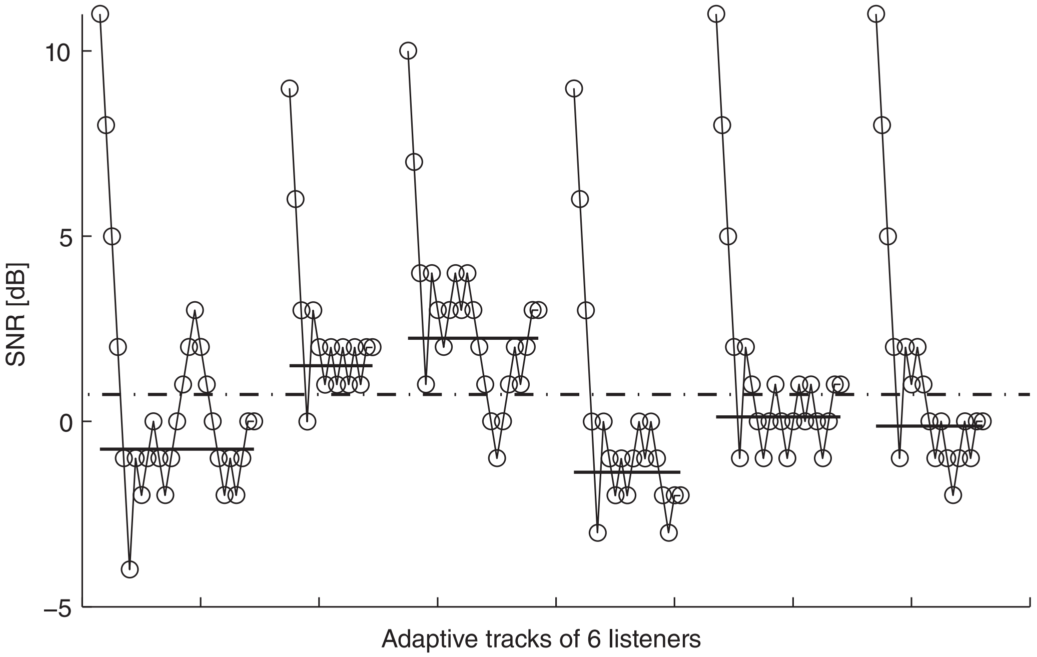

Figure 8.3 The detection threshold of a signal in noise measured with the adaptive procedure. The subject is presented with a signal+noise interval and two noise-alone intervals in random order with the forced choice approach. The SNR is reduced for correct answers and increased for wrong answers. The SNR step size is 3 dB first and is decreased to 1 dB after two reversals in the procedure. The results for six subjects are shown: the bold solid lines are the individual averages computed over the eight last turning points of the procedure. The bold dash-dotted line shows the average for 20 subjects.

answers, on the other hand, make the task harder. After a sufficient number of trials, the attribute ideally converges to the value of the threshold, and the average of last reversals in the tracking curve can be used as an estimate of the threshold. The value of the attribute can be plotted as a function of the number of trials, which often resembles a staircase, hence the name for the method. Staircases from such an experiment are shown in Figure 8.3 as an illustration.

The method is often designed with decreasing step size, starting with relatively large steps that are made smaller when convergence is thought to occur. There are variants of this procedure, where the level of convergence in the psychophysical function of a threshold is changed. The variants either apply different step sizes for the up and down movement or they require a different number of correct or wrong answers before changing the level of the attribute (García-Pérez, 2011; Levitt, 1971). Adaptive procedures are quite common, since they avoid the subject-related bias effects, and since the threshold value can be found with a smaller number of tests than with the method of constant stimuli. However, the down side of this method is that it does not guarantee convergence, and the experimenter must carefully select the parameters used and verify that the results obtained are meaningful.

8.8 Descriptive Sensory Analysis

The previous sections discussed listening tests, where it is often implicitly assumed that the difference between auditory events is in a single auditory attribute, such as in loudness or pitch. As will be discussed later in the context of sound quality, the auditory events studied may differ from each other in multiple dimensions, for example in both loudness and pitch. The situation becomes even more challenging if the attributes are not known *a priori*. For example, when perceptually motivated audio codecs are developed, and different versions of the codecs are tested, the listeners may perceive changes in 'crispness', 'noisiness', and 'loudness'. *Descriptive sensory analysis* is a family of methods which targets revealing the palette of attributes of a given set of stimuli, and in some cases also scaling (or grading) the stimuli in the attribute dimensions.

In the tests, a set of sounds is defined by the experimenter, and the perceptual properties that differentiate them are to be measured. Descriptive sensory analysis aims to identify, describe, and quantify the sensory attributes of stimuli using naive or trained human subjects (Piggott *et al.*, 1998), and it is often described as the most sophisticated tool in sensory science (Lawless and Heymann, 1998). An overview of the techniques in the context of audio is given by (Lorho, 2010), and is summarized here.

A number of techniques have been specifically designed for this purpose, mainly in food science but also in speech and audio. The term *elicitation* is often used in this context, which means 'the process of getting information from someone'. In this case, the elicited information is how the auditory events differ from each other, specifying all the attributes in which the differences are found, and also how much they differ in each of the dimensions specified by the attributes. This means that the researched sounds are projected to an N-dimensional space spanned by the elicited attribute dimensions.

The set or palette of attributes is often thought of as a *vocabulary*, which means a set of meaningful words that can be associated with the attributes. As shown in Figure 8.4, the techniques can be divided as follows:

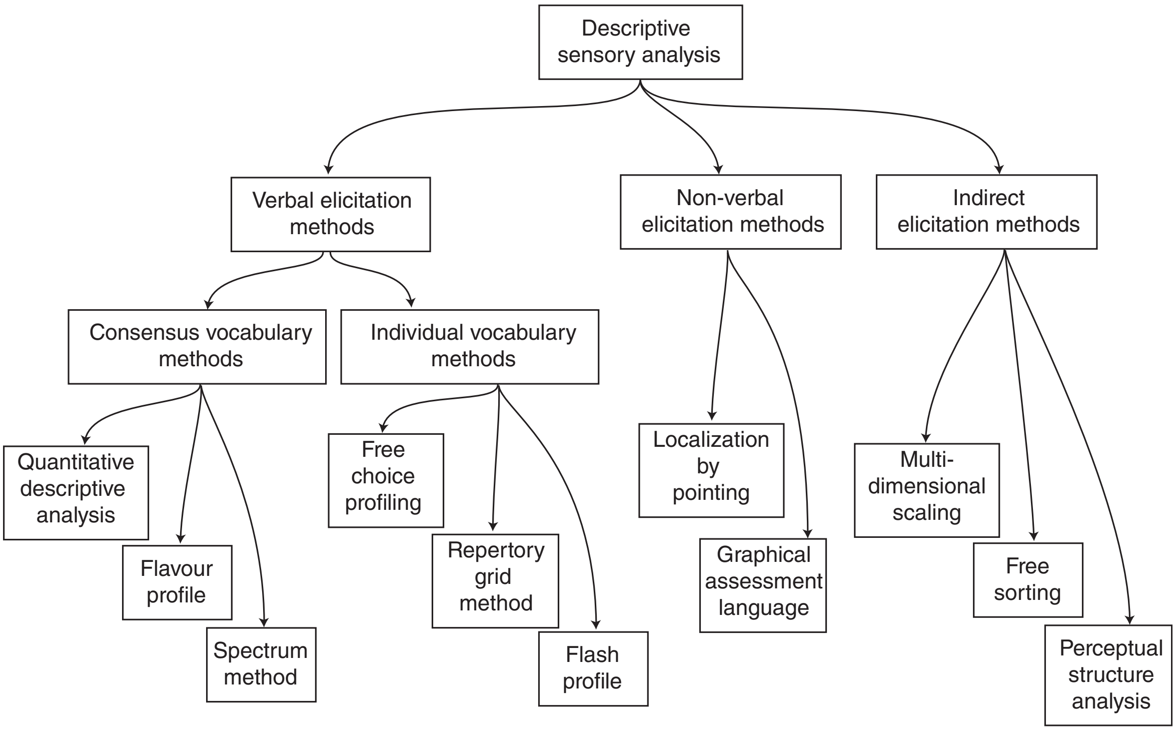
Descriptive sensory analysis
Verbal elicitation methods
Non-verbal elicitation methods
Indirect elicitation methods
Consensus vocabulary methods
Individual vocabulary methods
Quantitative descriptive analysis
Flavour profile
Spectrum method
Free choice profiling
Repertory grid method
Flash profile
Localization by pointing
Graphical assessment language
Multi-dimensional scaling
Free sorting
Perceptual structure analysis

Figure 8.4 Descriptive sensory analysis methods commonly encountered in the field of sensory science. Adapted from Lorho (2010).

- *Verbal elicitation*: Methods relying on a verbal description of perceived sensations, for instance, techniques employing a vocabulary development process with a group of subjects.
- *Non-verbal elicitation methods*: These techniques are based on bodily gestures.
- *Indirect elicitation*: Covers those methods working without direct sensation labelling.

8.8.1 Verbal Elicitation

Techniques based on verbal elicitation are extensively utilized in the field of sensory science, and they form the largest group of descriptive analysis methods. Two distinct categories of techniques exist for establishing the sensory descriptors, consensus vocabulary (CV) methods and individual vocabulary methods. CV methods use a panel of assessors to develop a common set of descriptors, or dimensions, characterizing the sensory properties of the stimuli being investigated. Examples of CV methods are:

- *Flavour profile method* (Cairncross and Sjöström, 1950): A major component of the flavour profile method is a highly trained panel of four to six members, who individually evaluate the stimuli and then work in discussion as a group to determine a consensus profile. This consensus leads to data that act as a representative value; this it is not an average of the panellists; scores, it is a single score agreed upon by all panel members. This component of the profile method was criticized in the 1960s and 1970s as offering too much potential for the panel leader or an opinionated panellist to introduce bias. It is also claimed that the appropriate selection of panellists, extensive training, and the blind nature of the testing can protect against bias.
- *Quantitative descriptive analysis* (Stone *et al.*, 2004): During training, a representative set of stimuli is used for the consensus language development. The panel leader facilitates communication without involvement and interference in panel discussions. Known reference stimuli can be used to generate sensory terminologies, especially when panellists disagree with each other on some sensory attributes. The subjects then conduct the actual analysis of the stimuli separately using the descriptors found in the training period.
- *Spectrum method* (Meilgaard *et al.*, 2006): In spectrum descriptive analysis, the panel consists of 10 to 15 screened subjects who develop technical expertise through a comprehensive training procedure. A descriptive terminology is built covering all the perceptual attributes using a set of absolute category scales calibrated to have equal intensity across the attributes. For example, grade 5 on a sweetness scale is defined to have equal intensity with grade 5 on a saltiness scale. The scales are based on the systematic use of reference points with corresponding reference samples, but the magnitude estimation of the attributes is made individually by the assessors.

Individual vocabulary methods let subjects in turn develop their own, individual set of sensory descriptors. Examples of such techniques are:

- *Free-choice profiling* (Williams and Langron, 1984): In free-choice profiling, subjects are assumed to differ mainly in the way they describe sensory characteristics and not so much in the way they perceive them. This allows assessors to first elicit the dimensions in the stimulus set and then to quantify the stimuli with the attributes following their own vocabulary. The effort used in panel training is thus considerably reduced because the difficult and time-consuming step of agreeing on the descriptors is side-stepped. The output of the analysis is thus a set of grades on individual scales. For example, when analysing a specific sample, subject 1 may give the value 6 to the dimension 'reverberance' and value 2 to 'bass', while subject 2 may give the value 7 to 'hall-sound' and 1 to 'warmth'. A data analysis procedure known as generalized procrustes analysis (Gower and Hand, 1995) is then used to project the results to a common set of dimensions representing the sensory attributes.

- *Repertory grid technique* (Kelly, 1955): The basic idea of this technique is to get subjects to define their own constructs by asking them to describe the ways in which elements and their associated meanings vary. This is done, for example, by presenting three samples, where each assessor states the characteristic for which two of the samples are similar to each other and different from the third. After a number of trials, an individual set of descriptors of the dimensions is obtained and can then be applied to evaluate all the stimuli. Different types of data analysis, such as principal component analysis, can be exploited to study individual and multiple aspects of the experiments.
- *Flash profile* (Delarue and Sieffermann, 2004): The individual elicitation approach of free-choice profiling and the pair-wise comparative evaluation technique are combined in flash profile. During the descriptive analysis process, all stimuli are compared in pairs, which apparently removes the need for a phase of familiarization and a phase of individual training with the attributes. In addition, flash profiling assumes that assessors are familiar with descriptive analysis, which ensures that discriminant attributes can be generated in a short time. Generalized procrustes analysis or a similar method has to be used to reveal the main dimensions in the data. As a result, a relative sensory grading of the stimuli on the scales found in the test can be obtained in just one to three sessions with this technique.

8.8.2 Non-Verbal Elicitation

In *non-verbal elicitation techniques*, which form the second group of descriptive analysis methods, the aim is to achieve a direct elicitation of perceived sensations, but without using a formal set of verbal descriptors (Mason *et al.*, 2001). Several techniques based on bodily gestures, such as *localization by pointing* in the direction of the tested or reference object, have been used (Choisel and Zimmer, 2003). The rationale is that verbal elicitation is not always appropriate to describe the complexity of an auditory space. Drawing techniques have also been exploited in the *graphical assessment language* to quantify the auditory perception created by spatial sound reproduction systems (Ford *et al.*, 2002).

8.8.3 Indirect Elicitation

The third group of descriptive sensory analysis methods comprises techniques based on *indirect elicitation*, as shown in Figure 8.4. The test methods included in this group are significantly different from the verbal and non-verbal elicitation methods discussed earlier, since the subjects do not elicit directly the perceived sensory characteristics of the stimuli. *Multidimensional scaling* (Carroll, 1972) is an example of an indirect elicitation method commonly utilized. A number of samples are produced, and the target is to find the main auditory attributes responsible for the dissimilarities between the samples. The listener rates the perceived dissimilarity pairwise between all combinations of the samples, thus forming distance matrices between the samples. The matrices are scaled to a lower-dimensional space for easier interpretation. The perceptual attributes are assumed to be present in the resulting space. However, the distance matrices alone do not offer a way to interpret the perceptual dimensions associated with the spatial map, because no labelling of the sensation is asked of the subjects.

Free sorting requires subjects to create groups containing stimuli that are perceived similar, based on their own criteria. In addition, they can be asked after the sorting task to describe each group of stimuli with verbal descriptors. This *a posteriori* labelling is assumed to facilitate the interpretation of perceptual dimensions during the analysis (Cartier *et al.*, 2006). The interview data may be analysed by means of the *grounded theory* (Corbin and Strauss, 2008), where a theory is systematically developed beginning from the collected data. The key points in the interview notes are labelled with codes, which are further organized into categories to form the basis of the theory, explaining, for example, sound quality.

Perceptual structure analysis is an example of an indirect elicitation technique recently used in the field of audio by Wickelmaier and Ellermeier (2007). This approach is based on Heller's theory of semantic structures (Heller, 2000), where the processes of identification and labelling of perceived characteristics are separated. In the test, the subjects are presented with three stimuli and are asked to indicate if the first two stimuli share a common feature with the third stimulus or not. After verifying that the subjects really use consistently the features that the data indicate, a representation of the individual perceptual structure can be derived indirectly. The method has been applied by Choisel and Wickelmaier (2007) to develop a set of auditory attributes that describe the differences in perception of multi-channel sound reproduction.

8.9 Psychoacoustic Tests from the Point of View of Statistics

An important part of the research on psychoacoustics is the statistical analysis of the results from listening tests. Actually, in some cases, the statistical considerations should be taken into account in the design of the experiments. Tests should be conducted with the proper number of subjects and a meaningful selection of stimuli and tasks, listening conditions, and repetitions. In many cases, such designs should be conducted carefully so that the results prove or disprove the existence of the phenomenon that is hypothesized based on informal listening before the test.

The attributes of sound events are frequently called independent variables in experiment design and statistical analysis. In subjective tests, the independent variables are manipulated to produce different stimuli for testing, and when the selected test method is applied to the subjects, their responses (possibly after some post-processing) then yield the 'dependent' variable(s). The results should then reveal the relation of the 'independent' variable(s) to the 'dependent' variable(s), or in general the psychophysical function $h_\mathrm{i} = f(s)$, as discussed in Section 8.1.

Testing typically produces a large data set to which proper statistical methods must be applied. In simple cases, some basic descriptors, such as means, variances, and 95% confidence intervals can be used. However, often the influence of attributes in the tests on the data obtained should be examined using appropriate parametric or non-parametric statistical tests. Quite commonly *analysis of variance* (ANOVA) is used, which answers the question, 'do any of the independent variables have an effect on the dependent variable?' If an independent variable is found to have an effect, *posthoc tests* can then be used to measure how the independent variable affects the result.

Summary

This chapter discussed various methods used to study the functionality of hearing mechanisms by psychoacoustic means; that is, by presenting sound events to subjects and asking them to perform some tasks in a formal listening test method. The field is quite mature: if a test is designed carefully, the results indeed provide valid information on the attributes of an auditory event, generated by acoustic attributes of a sound event. In other words, psychoacoustic test methods can be used to measure the psychophysical functions that transfer acoustic attributes into auditory attributes. Descriptive sensory analysis involves methods of finding, in a formal way, the attributes of auditory events perceivable by subjects.

Further Reading

An introduction to psychophysical research methods of perception from all senses is found in Goldstein (2013). In Bech and Zacharov (2006), various considerations of listening-test design are made in the context of audio quality, and Gelfand (2004) discusses the psychoacoustical methods in general in more detail. A good source regarding early investigations into the quantitative formulation of auditory sensation and perception is Fletcher (1995). The use of descriptive sensory analysis techniques in the field of audio is reviewed by Bech and Zacharov (2006); Neher *et al.* (2006). The statistical analysis of quantitative attributes resulting from descriptive sensory analysis is reviewed by Næs and Risvik (1996).

References

Bech, S. and Zacharov, N. (2006) *Perceptual Audio Evaluation – Theory, Method and Application*. John Wiley & Sons.

Cairncross, S. and Sjöström, L. (1950) Flavor profiles – a new approach to flavor problems. *Food Technology*, **54**(4), 308–311.

Cardozo, B. (1965) Adjusting the method of adjustment: SD vs DL. *J. Acoust. Soc. Am.*, **37**, 786–792.

Carroll, J.D. (1972) Individual differences and multidimensional scaling. *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences*, **1**, 105–155.

Cartier, R., Rytz, A., Lecomte, A., Poblete, F., Krystlik, J., Belin, E., and Martin, N. (2006) Sorting procedure as an alternative to quantitative descriptive analysis to obtain a product sensory map. *Food Quality and Preference*, **17**(7), 562–571.

Choisel, S. and Wickelmaier, F. (2007) Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference. *J. Acoust. Soc. Am.*, **121**(1), 388–400.

Choisel, S. and Zimmer, K. (2003) A pointing technique with visual feedback for sound source localization experiments *Audio Eng. Soc. Convention 115* AES.

Corbin, J. and Strauss, A. (2008) *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage.

Delarue, J. and Sieffermann, J.M. (2004) Sensory mapping using Flash profile. Comparison with a conventional descriptive method for the evaluation of the flavour of fruit dairy products. *Food Quality and Preference*, **15**(4), 383–392.

Fletcher, H. (ed.) (1995) *Speech and Hearing in Communication*. Acoustical Society of America.

Ford, N., Rumsey, F.J., and Nind, T. (2002) Subjective evaluation of perceived spatial differences in car audio systems using a graphical assessment language *Audio Eng. Soc. Convention 112* AES.

García-Pérez, M.A. (2011) A cautionary note on the use of the adaptive up–down method. *J. Acoust. Soc. Am.*, **130**, 2098–2107.

Gelfand, S.A. (2004) *Hearing: An introduction to psychological and physiological acoustics*. Marcel Dekker.

Gescheider, G.A. (1997) *Psychophysics: The Fundamentals*. Psychology Press.

Goldstein, E.B. (2013) *Sensation and Perception*, 9th edn. Cengage Learning.

Gower, J.C. and Hand, D.J. (1995) *Biplots* volume 54. CRC Press.

Heller, J. (2000) Representation and assessment of individual semantic knowledge. *Methods of Psychological Research*, **5**(2), 1–37.

Kelly, G. (1955) The Psychology of Personal Constructs. Norton.

Lawless, H.T. and Heymann, H. (1998) *Sensory evaluation of food. Principles and practices*, Chapmann & Hall.

Levitt, H. (1971) Transformed up–down methods in psychoacoustics. *J. Acoust. Soc. Am.*, **49**(2B), 467–477.

Lorho, G. (2010) *Perceived quality evaluation: an application to sound reproduction over headphones*. Ph.D thesis, Aalto University.

Mason, R., Ford, N., Rumsey, F., and De Bruyn, B. (2001) Verbal and nonverbal elicitation techniques in the subjective assessment of spatial sound reproduction. *J. Audio Eng. Soc.*, **49**(5), 366–384.

Meilgaard, M.C., Carr, B.T., and Civille, G.V. (2006) *Sensory Evaluation Techniques*. CRC Press.

Næs, T. and Risvik, E. (1996) *Multivariate Analysis of Data In Sensory Science* volume 16. Elsevier.

Neher, T., Brookes, T., and Rumsey, F. (2006) A hybrid technique for validating unidimensionality of perceived variation in a spatial auditory stimulus set. *J. Audio Eng. Soc.*, **4**, 259–275.

Piggott, J.R., Simpson, S.J., and Williams, S.A. (1998) Sensory analysis. *Int. J. Food Sci. & Technol.*, **33**(1), 7–12.

Stone, H., Sidel, J., Oliver, S., Woolsey, A., and Singleton, R.C. (2004) Sensory evaluation by quantitative descriptive analysis. In Gacula, M.C. (ed.) *Descriptive Sensory Analysis in Practice*. John Wiley & Sons, pp. 23–34.

von Békésy, G. (1960) *Experiments in hearing*. McGraw-Hill and Acoustical Society of America.

Wickelmaier, F. and Ellermeier, W. (2007) Deriving auditory features from triadic comparisons. *Perception & Psychophysics*, **69**(2), 287–297.

Williams, A.A. and Langron, S.P. (1984) The use of free-choice profiling for the evaluation of commercial ports. *J. Sci. Food Agri.*, **35**(5), 558–568.