Chapter 2
A Dimension-Based Approach to Mouth-to-Ear Speech Transmission Quality

2.1 General Research Scenario

A very basic schematic of the research scenario dealt with in this book is illustrated in Fig. 2.1.

The schematic reflects the classical two-party *speech communication* situation where two interlocutors exchange information via speech (O'Shaughnessy 2000). "Communication", in general, is the "intentional transmission of information" (Lyons 1977, p. 32). Speech is the acoustic part of language common to the conversation partners as an "established signaling-system" (Lyons 1977, p. 32), and can be considered as the most natural way in human communication.

Speech communication usually is bidirectional and requires the *transmission* of speech in both directions. In Fig. 2.1, the *transmission path* is represented by a "black box" starting from the mouth of the speaker to the ear of the listener. Each of the interlocutors represent a source and sink of information, thus both are taking the role of speaker and listener usually by turns, respectively.

In speech communication, the transmission of speech from one human being, acting as the *talker*, to another, acting as the *listener*, is often referred to as the *speech chain*, a term coined by Denes and Pinson (1993), see also Vary et al. (1998, pp. 5–8), for example. It suggests that, ultimately, the brains of the interlocutors represent the source and sink of the information being conveyed in end-to-end speech communication. The brains, in turn, are linked to the speech production and perception organs via motor and sensory nerves, respectively. Thus, the speech chain comprises three important processes:

- Speech production of the speaker (source of information),
- the physical or technical process of speech transmission, and
- speech perception of the listener (sink of information).

With their physiological and neuro-physiological capabilities, humans are able to encode information arising in the brain. The central nervous system is connected to the human speech-production organs by motor nerves and muscles. Employing the

M. Wältermann, *Dimension-based Quality Modeling of Transmitted Speech*,
T-Labs Series in Telecommunication Services, DOI: 10.1007/978-3-642-35019-1_2,

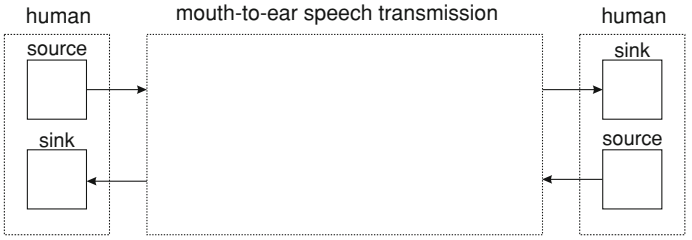

Fig. 2.1 Two-party speech communication

lungs for airflow provision, the larynx with its glottis is used primarily for airflow "control", whereas the vocal tract serves for modulation of the airflow through its articulators (O'Shaughnessy 2000, pp. 35–48). This way, speech *sounds* are produced and emitted primarily from the mouth of the speaker. Speech sounds, as all sounds, physically emerge in form of mechanic vibration or (longitudinal) mechanic waves, see Blauert and Xiang (2008). Speech is physically characterized by an average sound intensity level between 55 and 75 dB rel. $10^{-16}\,\mathrm{W/cm^2}$ for normal speaking. The long-term average speech spectrum shows lowpass behavior and ranges approximately from 50 to 10000 Hz. Between 100 and 600 Hz, that is, including the fundamental frequency and the first formant, the energy is highest (Denes and Pinson 1993, pp. 139–145).

For the moment, the speech transmission system is represented by a "black box" comprising all transmission elements from the mouth of the speaker to the ear of the listener. In direct face-to-face communication, the speech transmission "system" is a purely acoustic sound field. Under free-field conditions, it is the direct air path "mouth to ear". The acoustic waves, emitted by the mouth of the speaker, represent the vehicle carrying the information from the speaker (information source) to the listener (information sink). The information is encoded in the sound pressure variations as a function of time and space as a purely physical phenomenon that is studied in the field of "pure" acoustics (Blauert and Xiang 2008). The waves eventually arrive at the listeners ear(s) as an attenuated version of the original acoustic signal (assuming an otherwise noise-free and free-field environment).[1]

Telecommunication via speech is enabled by *speech-communication technology*, partially replacing the natural air-path and aiming at "supporting natural communicative possibilities [...] by means of technical devices and facilities" (Jekosch 2005b, p. 23). This comprises elements such as user terminals (electro-acoustic transducers), codecs, transmission facilities (e.g., circuit-switched or packet-switched networks), and other signal processing units. Most generally, according to Shannon (1948), transmission of information in technical terms consists of a *transmitter*, a (noise-prone) *channel*, and a *receiver* in order to enable telecommunication.

[1] Gestures as a part of human face-to-face communication are disregarded in this consideration, although they become more important, for example, in video telephony. The present work is focused solely on acoustic speech communication.

The transmitted speech signal finally arrives at the listener's outer ear(s) as a modified version of the source signal, whereas the modification depends on the transmission channel the signal traversed. Much is known about the *sound reception* involving the human hearing system on both a physiological and neuro-physiological level (Denes and Pinson 1993, p. 80). The pinna of the outer ear supports localization of acoustic events, for example. The acoustic waves travel through the ear canal and excite the ear drum, which in turn is connected to the chain of the ossicular bones hammer (malleus), anvil (incus), and stirrup (stapes) of the middle ear. Here, the acoustic vibrations are converted to mechanical vibrations, where the mechanical impedance of the entrance point of the inner ear, the membrane of the oval window, is matched through impedance transformation between air and the inner ear liquid (endolymph in the scala media, perilymph in the scala vestibuli and the scala tympani; the Reissner membrane and the basilar membrane form the boundaries to the scala media, respectively). The cochlea of the inner ear is basically a tube filled with these liquids. Its function is the transformation of the arriving vibrations in form of mechanical energy into electrical excitation (electrical energy) on the auditory nerve fibres (so-called "firings"). The neurons are mainly afferent, leading to the central nervous system. The basilar membrane with the organ of Corti resting on it contains sensory hair cells connected with the auditory nerve and performs a spectral decomposition through frequency-to-place transformation, which is of non-linear nature (see, e.g., Denes and Pinson 1993, pp. 131–137). Details on the anatomy and physiological as well as neuro-physiological mechanisms can be found, for example, in O'Shaughnessy (2000, pp. 109–119) or Purves (2008). Finally, the transformation of the neuronal representation into a linguistic message interpretable by the listener completes the brain-to-brain information transmission.

Perception takes place on a psychological level. The field of study concerned therewith is *psycho-acoustics*, that is, the study relating acoustic events in the physical realm to auditory events in the perceptual space (Blauert and Xiang 2008). For instance, much research has been done in psycho-acoustics of fundamental sounds such as tones and noise (e.g., Zwicker and Fastl 1999), including perception thresholds, pitch perception, masking, and localization.

Speech, as language in general, can be considered as a system of signs, where "a sign is a mental unit, which is processed as standing for something other than itself" (Jekosch 2005a). In terms of semiotics, the science of signs, a sign can be modeled, for example, by the triad of (a) sign carrier, (b) the object the sign stands for, and (c) its meaning. These constituents of a sign cannot be regarded in an isolated way, as for example the content has an influence on speech perception as well as the sign carrier. The present work focuses on the perception of *transmitted* speech in terms of its perceived *quality* and, in particular, the underlying *perceptual dimensions*. Thereby, the research presented in this work mainly is concerned with the acoustic speech signal as the sign carrier, that is, the form or the surface structure. Quality and its determining features can in this context also be seen as "additional factors governing speech perception". In telephony, "the user's appreciation of a

particular connection" (Raake 2006, pp. 13–14) can be influenced, for example, by unwanted distortions of the signal, that is, the sign carrier. Speech quality can thus be understood as "extending the comprehension-oriented view on speech": "Quality might not be perfect although the comprehension is reliable" (Raake 2006, pp. 13–14), whereas depending on the severity of the distortions, also comprehension might suffer.[2]

The central assumption of the present work is the notion that speech quality can be explained by its underlying perceptual features (Jekosch 2005b, pp. 11–21), both influenced by elements of the transmission path and perception processes of the recipient. Given that these features are measurable, they allow to diagnose and model integral speech quality. Moreover, given that the quality-relevant elements can be quantified and brought into relation with the perceptual features, both these features and integral quality can be *instrumentally* estimated. The relations between integral quality and its dimensions, as well as between technical parameters and the dimensions, are assumed to be describable by empirical functions. They can be illustrated on the basis of a "black box" schematic of the listener of transmitted speech. Therefore, the schematic depicted in Fig. 2.1 will be further extended by the building blocks of speech transmission technology and speech quality perception with a level of detail appropriate for the further investigations presented in this work.

In the next two sections it is shown how today's speech transmission technology might affect a speech signal in the physical domain (Sect. 2.2), and how this translates into the perceptual domain as *perceptual dimensions* and eventually overall *quality* (Sect. 2.3). Moreover, the terminology necessary for the precise formulation of the research topics that are dealt with in this book is introduced. In Sect. 2.4, selected auditory measurement methods and measurement implications with human listeners as measuring organs are presented. The dimension-based quality-modeling approach is introduced in Sect. 2.5, whereas important instrumental quality measures and measurement approaches are presented in Sect. 2.6. The resulting research topics aiming at an instrumental dimension-based quality model are summarized in Sect. 2.7 and constitute the basis for the following chapters of this book.

[2] Speech comprehension is related to several other terms. A message to be comprehended by the recipient, besides her/his willingness to do so, depends on a series of factors. Given that the percentage of correctly identified word fragments, syllables, phonemes, and meaningless words of a transmission path is measured by *articulation*, articulation is the prerequisite for *comprehensibility*. Comprehensibility describes how well the speech signal, that is, the sign carrier, is capable to convey information. Comprehensibility, in turn, constitutes the prerequisite for *intelligibility*, itself describing the percentage of correctly identified meaningful words, phrases, or sentences. In parallel, *communicability* refers to how well a message can serve to communicate. Context as well as recipient's knowledge factors influence the process of comprehension (whereas the definition of context depends on the level of comprehension). For more information, see Raake (2006, pp. 9–11), Jekosch (2005b, pp. 97–102), and Möller (2000, pp. 26–27).

2.2 Speech Transmission in Telecommunication

2.2.1 Introduction

This section provides an overview of the physical and speech—technological elements of the transmission path between the mouth of the speaker and the ear of the listener, that is, the middle block in Fig. 2.1. As it will be seen, these elements might have various impacts on the transmitted source speech signal, which eventually affect the perception of the received signal by the listener in terms of quality and its features, as discussed in Sect. 2.3.

Figure 2.2 provides a more detailed though still simplified view on the mouth-to-ear speech transmission system.

A speech communication system is a special case of the schematic of a general communication system (cf. Shannon 1948 and Richards 1973, pp. 14–19), consisting of a transmitter (sender), a transmission channel, and a receiver, connected in cascade. Information source and sink are the human beings communicating with each other (see Fig. 2.1, omitted in Fig. 2.2).[3] In mouth-to-ear speech transmission, all elements can be affected by noise or other phenomena causing distortions (e.g., room reflections in the sound field, frequency characteristics of the transducer, radio interference on the channel). The system in Fig. 2.2 is assumed to be symmetric in both directions. In real-world scenarios, however, this assumption does not necessarily hold, for example, due to different terminal equipment used by the conversation partners or different channel paths.

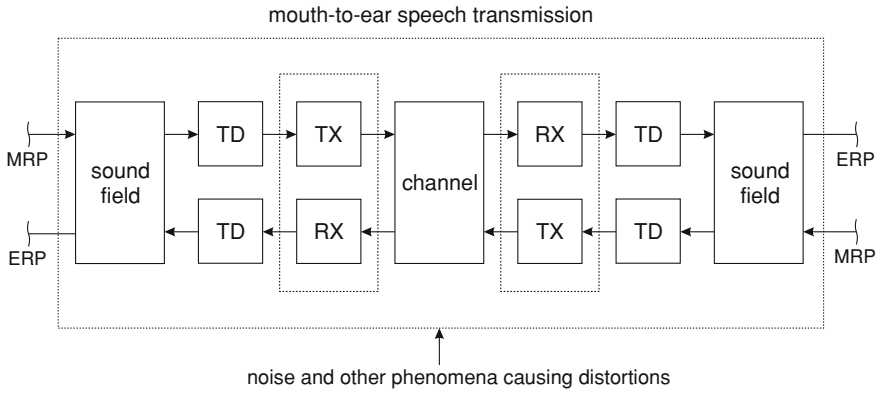

Fig. 2.2 General transmission system (Shannon 1948) and extensions based on Blauert and Xiang (2008). *MRP* Mouth reference point, *ERP* Ear reference point, *TD* Transducer, *TX* Transmitter, *RX* Receiver

[3] The terms transmitter and receiver are ambiguous as they may also refer to the human beings. In order to avoid confusion, the technical entities for sending and receiving signals may also be called sending and receiving apparatus, respectively (Richards 1973, p. 14).

In the present work, speech transmission designates the path from the mouth of the speaker to the ear (or the ears) of the listener. More precisely, the point in space right in front of the mouth of the speaker corresponds to the *Mouth Reference Point (MRP)* as defined in ITU-T Rec. P.64 (2007, Annex A), which is located at a distance of 25 mm in front of the lips. The speech transmission ends at the entrance of the ear canals, the *Ear Reference Points (ERPs)* as defined in ITU-T Rec. P.64 (2007, Annex A). As a consequence of this definition and according to Blauert and Xiang (2008), the communication "system" also comprises "pure" acoustics, that is, the soundfields existent at send side between the MRP and transmitter and at receive side between receiver and the ERPs.

The elements involved in telephony are described in the following sections. A fundamental concept for planning and describing systems in classical telephonometry are so-called *loudness ratings*. A loudness rating is a scalar value in decibels and denotes the insertion loss of a reference system that yields a perceived loudness that is equivalent to the loudness of the unknown system under investigation.[4] Loudness ratings can be measured in auditory experiments (ITU-T Rec. P.78 1996) or they can be calculated based on weighted amplitude spectra measured in the ISO-preferred 1/3-octave bands (ITU-T Rec. P.79 2007). The attenuation between the MRP and the transducer (electrical interface in the transmission path) is denoted by *send loudness rating (SLR)*, whereas the attenuation between the transducer and the ERP is denoted by *receive loudness rating (RLR)*. Loudness ratings are additive, that is, the sum of loudness ratings of the subsystems of a transmission system is approximately equal to the *overall loudness rating (OLR)* of the system, for example, $OLR = SLR + RLR$. See Möller (2000, pp. 19–26) for more details.

2.2.2 Mouth to Channel

Figure 2.3 depicts in an illustrative way the situation of a user located in a soundfield using a hands-free terminal.

If no free-field conditions can be assumed, the room in which the communicating person resides has an influence on the speech signal y to be transmitted and on the speech signal x that is received.[5] These effects are particularly prominent if hands-free terminals (HFTs) or hands-free capabilities of standard user terminals are employed. In contrast to handsets or headsets with a close-talking microphone located directly at the MRP and the loudspeakers (more or less well coupled) located

[4] In traditional narrowband telephony, the (modified) intermediate reference system (IRS) is used as a reference (see ITU-T Rec. P.48 1988 and ITU-T Rec. P.830 1996, Annex D). For wideband telephony (50–7000 Hz), frequency weights are given in ITU-T Rec. P.79 (2007, Annex G) for a reference system that yields a loudness rating of 0 dB when compared to an IRS.

[5] For simplicity reasons, no difference is being made between variables denoting signals in the time-continuous domain and the time-discrete domain. That is, x might represent the continuous signal $x(t), t \in \mathbb{R}$, or the discrete version $x(k)$, where $t = kT$ and $T = 1/f_s$. f_s denotes the sampling rate and $k \in \mathbb{Z}$. Moreover, the amplitude of time-discrete signals is assumed to be quantized.

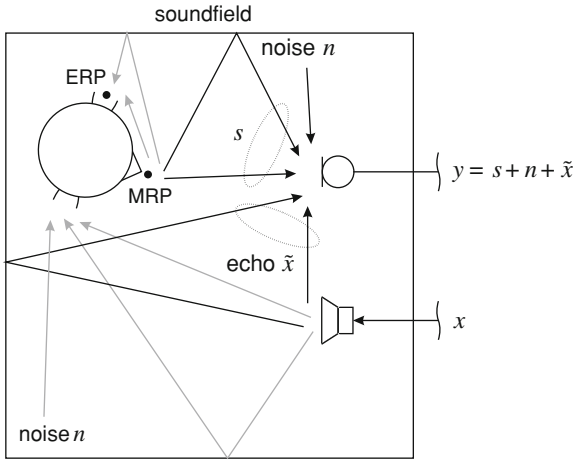

Fig. 2.3 Human being located in a soundfield using a hands-free terminal (modified from Vary et al. 1998, Fig. 13.1). *Black arrows*: acoustic waves captured by the transducer. *Gray arrows*: acoustic waves captured by the human being

directly at the ERP, loudspeaker and microphone(s) of HFTs are most often located with a significant distance to the ERP(s) and MRP, respectively. In this section, it is focused on the signal y to be transmitted and its components according to the acoustic situation in the room at send side, that is, distortions that might provoke a degradation of perceived quality for the *far-end listener*. These components are represented by the black arrows in Fig. 2.3.[6]

In particular, depending on the room and microphone characteristics, early and late reflections might be captured by the microphone of the transmitter and superimpose the speech signal at the MRP (the speech signal and possible reflections are subsumed by s). As there is always a certain distance between mouth and microphone, this can easily also occur for non-hands-free terminals. Another prominent effect, especially in hands-free telephony, stems from the acoustic coupling between loudspeaker and microphone. The received signal x of the far-end speaker emitted by the near-end loudspeaker is transmitted through the room (including reflections) and is fed as a modified version $\tilde{x}$ to the microphone as an interfering signal. Without countermeasures, the far-end interlocutor hears her/his own voice delayed while speaking (*talker echo*; a *listener echo* might occur in rare cases, see Möller 2000, p. 32, if the signal is reflected twice, that is, at both the near end and the far end). Moreover, environmental (or background) noise n of different kind might superimpose the sending signal. From the near-end person's perspective, these effects sum up through superposition to a sending signal $y = s + n + \tilde{x}$ (see Vary et al. 1998, p. 430 and Vary and Martin 2006, p. 506).

The (distorted) acoustic signal y is transformed into a proportional electrical signal by the electro-acoustic *transducer* (microphone), see Fig. 2.2. For auditory

[6] The acoustic properties of the room also have an influence on the receiving signal x, which, as a result, affects perception. These sound field components are represented by the gray arrows in Fig. 2.3 and are briefly discussed in Sect. 2.2.4.

tests as well as in narrowband telephonometry, the transducer's transfer functions in send and receive direction are often generalized to a model of common handsets, the (modified) Intermediate Reference System (IRS), see ITU-T Rec. P.830 (1996, Annex D). Apart from that, the transducers consist of input and output entities, which are often interconnected by a sidetone path. Considerations of the influence of terminal equipment are given in Sect. 2.2.4.

For speech transmission, as for media transmission in general, there is a clear trend towards the utilization of a unified, "All-IP" network. Motivated by this prospect, it is focused on packet-based transmission in the following.[7] Figure 2.4 depicts a schematic of an example realization of a VoIP transmitter/receiver-combination. Practical realizations might deviate from this simple schematic, however, it comprises the major elements of the technology encountered today.

Common to practically all transmitters in today's terminal equipment is the analog-to-digital conversion (including the obligatory anti-alias low-pass filter) of the input signal y. More precisely, the time- and amplitude-continuous signal $y(t)$ is low-pass filtered with a cut-off frequency of at most $1/2 \cdot f_s$, where f_s is the sampling rate, sampled and quantized, resulting in the time-discrete and amplitude-discrete signal $y(k)$. As a next step, several signal processing routines such as noise reduction and acoustic echo cancelation are usually invoked as "counter-measures" for different

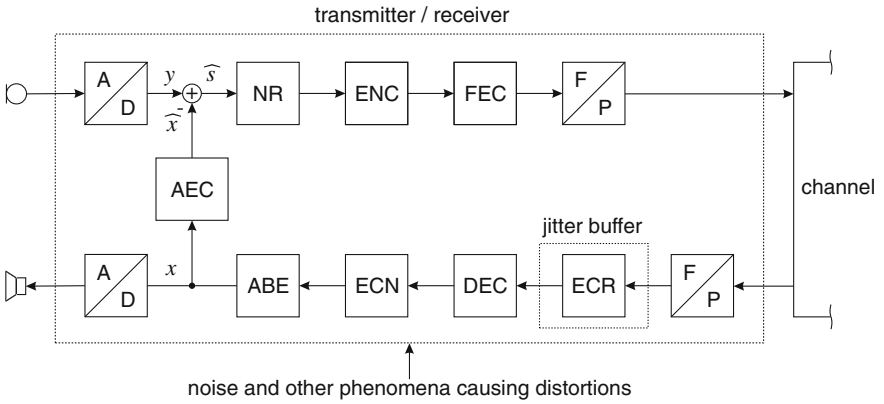

Fig. 2.4 Schematic of a possible realization of a single-input single-output VoIP transmitter/receiver, based on Raake (2006, Fig. 3.3), Vary and Martin (2006, Fig. 1.1), and Perkins (2003, Figs. 1.2 and 1.3). *A/D* Analog to digital conversion, *AEC* Acoustic echo cancelation, *NR* Noise reduction, *ENC* Speech encoding, *FEC* Forward error correction, *F/P* Packetization (frame to packet), *ECR* Error correction, *DEC* Speech decoding, *ECN* Error concealment, *ABE* Artificial bandwidth extension. See text for details

[7] Complementary information on Voice over IP (VoIP), in particular with regard to the time-varying behavior of packet-based networks and the resulting effects, can be found in Raake (2006).

kinds of distortions potentially occurring up to this point in the transmission path.[8] Speech signal processing routines, including most speech codecs, are usually realized in a block-oriented way and process speech frames of limited duration (typically 20 ms, e.g., 160 samples with $f_s = 8\,\text{kHz}$) due to the quasi-stationary character of speech in this time interval.

The aim of *acoustic echo cancelation (AEC)* is to prevent the signal $\tilde{x}$ from being transmitted to the far-end interlocutor, see Vary et al. (1998, pp. 429–464) and Vary and Martin (2006, pp. 505–568). The signal $\tilde{x}$ appears in the transmitter as a component of the sending signal y and it is a modified and delayed version of the received signal x appearing in the receiver. Thus, AEC requires an interconnection between transmitter and receiver at each side of the communication system. The cancelation of $\tilde{x}$ can be achieved by a voice-controlled echo suppressor: In sending and receiving direction, attenuators are adaptively applied depending on whether the near-end or far-end interlocutor is active. This often prevents the interlocutors from double talk, which, however, belongs to natural human communication and is thus desirable. More sophisticated echo cancelation approaches (where double talk is possible) are based on an estimation of the impulse response of the linear time-variant loudspeaker-room-microphone system. The compensation is done by an adaptive filter estimating $\tilde{x}$ and subtracting the estimate from y. Although the impulse response is in principle infinite, usually, finite-impulse-response (FIR) filters are used for this task as an approximation. The filter length (and thus the complexity) depends on the sampling rate f_s of the microphone signal y, the distance between loudspeaker and microphone, and the room acoustics. In practice, the parameters *talker echo loudness rating (TELR)* in decibels, which describes the level difference between the far-end speech signal and the echo, together with the mean one-way delay T of the echo path are often used for specifying the requirements of AECs, evaluating AECs, and in overall network planning (see Appendix B). Less common are listener echoes, quantified with the *weighted echo path loss (WEPL)* in decibels in conjunction with the round-trip delay T_r. More details on the large field of AEC can be found, for example, in Vary et al. (1998, pp. 429–464) and Vary and Martin (2006, pp. 505–568).

Noise reduction (NR) aims at eliminating the noise component n from the signal y. This can either be done in the time-domain by Wiener filtering or in the spectral domain (although both approaches are related, see Vary et al. 1998, pp. 377–428 and Vary and Martin 2006, pp. 389–504). For the latter, a practical approach to noise reduction is *spectral subtraction* (Boll 1979) of a noise floor estimate from an estimate of the spectrum of the noisy signal (Vary et al. 1998, pp. 387–397, Vary and Martin 2006, pp. 400–402). Due to the non-stationary character of the speech signal, this is done frame-wise for short-time spectra and can also be done in conjunction with the Wiener filter. The plain subtraction of the spectra potentially leads to (a) speech signal distortions and (b) a randomly fluctuating residual noise due to estimation

[8] Speech distortions that originate from early or late room reflections at send side are not compensated. In Brüggen (2001), however, methods are proposed for "sound decolorization" applicable in HFT-telephony.

errors in the frequency domain. The time-variant noise component is commonly referred to as "musical noise" (see Vary et al. 1998, pp. 393–395 and Vary and Martin 2006, p. 401). Several counter-measures exist in order to reduce the musical noise phenomenon, for example through noise "over-subtraction" (Vary and Martin 2006, pp. 403–408). On the other hand, this might lead to signal distortions in turn. Apart from these approaches, spectral subtraction can generally be done in a more robust way. Approaching the problem in terms of the a-priori SNR, that is, the SNR of the input speech and noise signals, with the estimation technique introduced by Ephraim and Malah (1984) (see also Vary and Martin 2006, pp. 402–403 and Vary et al. 1998, pp. 396–397), the quality of the enhanced speech signal is improved as compared to plain spectral subtraction. Furthermore, "musical noise" is reduced (Cappé 1994).

The performance of noise reduction algorithms can principally be improved by employing more than one microphone. Here, the spatial characteristics of the sound field (cf. Fig. 2.3) and statistical properties of the signals can be exploited, which is particularly beneficial for non-stationary noise (Vary et al. 1998, pp. 409–428 and Vary and Martin 2006, pp. 443–504). Using multi-channel systems with microphone arrays, it becomes possible to separate acoustic sources. Adaptive beamforming techniques can be applied aiming at spatial selectivity by forming a "beam" of high gain in the direction of the desired sound source by a combination of the different microphone signals (see Vary et al. 1998, pp. 422–428 and Vary and Martin 2006, pp. 477–481).

Prior to transmission of the enhanced signal over the transmission channel, the signal is encoded in order to reduce its bitrate (*source* coding). Several speech codecs of different complexity and quality are available. An ideal codec would exhibit high speech quality, low bitrate, low complexity, and low delay, see Vary et al. (1998, pp. 233–234) and Vary and Martin (2006, pp. 240–243). However, these criteria cannot be optimized at the same time in practice. In fact, some of the criteria are mutually exclusive (Vary et al. 1998, p. 234). For example, high quality codecs aiming at a low bitrate are usually more complex and thus introduce a noticeable delay. Hence, there are different codecs with different parameters optimized, depending on the application.

Speech codecs can be classified according to their underlying coding principle. *Waveform codecs* achieve a reduction of the bitrate by direct manipulation of the speech signal waveform. A simple procedure implemented by the G.711 codec (ITU-T Rec. G.711 1988) is a logarithmic *companding* technique, resulting in a bitrate of 64 kbit/s. *Com*pressing the amplitude of the input signal in a logarithmic fashion yields a signal-to-noise ratio acceptable also for small signal amplitudes predominant in speech.[9] At the receiver side, the compression is reversed by ex*panding* the signal's amplitude accordingly. Since the logarithmic curve is computationally problematic for values asymptotically reaching zero, a properly displaced

[9] Amplitudes of speech are not uniformly distributed. In contrast, Laplace or Gamma density distributions can be observed, indicating that low energy levels are most frequent in speech and should thus be quantified with a higher resolution than higher energy levels.

logarithmic curve is used instead (μ-law), or a combination of linear and logarithmic curve segments (A-law).

In *differential waveform coding*, the redundancy inherent to human speech is exploited. Linear predictive (LP) filters are employed for generating *residual* signals to be transmitted, which show a reduced dynamic range. The bitrate is reduced due to the shorter word length required for quantization. This technique is known as *differential pulse code modulation (DPCM)* (realized with a weighted first-order LP filter in the simplest case). An adaptive realization leads to even better results in terms of bitrate reduction and speech quality (ADPCM coding, see Vary and Martin 2006, pp. 260–261). A prominent example for ADPCM codecs is the G.726 codec family (ITU-T Rec. G.726 1990) used in DECT (Digital Enhanced Cordless Telecommunications) phones. Depending on the specific quantization technique, acceptable signal-to-noise ratios can be achieved. With noise shaping in the spectral domain, the quality can be further increased by exploiting spectral masking phenomena of the human ear.

Parametric codecs (vocoders) work more efficiently in terms of bitrate. Here, mainly the source-filter model of human speech production is employed to parameterize the given speech signal. Only few parameters are necessary for representing the source (such as the amplitude, fundamental frequency of the excitation signal, together with a voiced/unvoiced information). The vocal tract is modeled by an adaptive LP filter (synthesis filter) at the receiver, where the glottis signal stems from a signal generator. The filter coefficients and information on the source signal are determined by a respective inverse filter and residual signal analysis in the encoder (analysis filter). In parametric codecs, only the above-mentioned parameters are quantized and encoded. With this approach, very low bitrates of the encoded signal can be achieved (typically 2.4 kbit/s), see Vary et al. (1998, pp. 290–301) and Vary and Martin (2006, pp. 262–272), however, with less naturalness of the reconstructed speech signal.

For almost all codecs used in public telephony, the techniques used in waveform codecs and parametric codecs are combined to so-called *hybrid codecs*, leading to medium bitrates of 4–12 kbit/s (for NB speech, see Vary et al. 1998, pp. 301–331 and Vary and Martin 2006, pp. 273–304). In hybrid codecs, the source-filter model is still employed. The adaptive LP analysis filter of the encoder is commonly extended by an adaptive long-term predictor (LTP) allowing to parameterize the source signal. The resulting residual signal is quantized and transmitted, along with the parameters for the LTP and LP synthesis filter as side information. For hybrid codecs, there are some variants and refinements of these general coding principles. For instance, in baseband-RELP (residual-excited linear prediction) codecs such as the GSM full-rate (ETSI GSM 06.10 1988) used in mobile networks, a scalar quantizer is applied to the perceptually important low-frequency part of the residual signal and the resulting signal is transmitted in addition to the LP and LTP parameters, see Vary et al. (1998, pp. 311–319) and Vary and Martin (2006, pp. 282–289). Another variant of hybrid codecs are code-excited linear prediction (CELP) codecs. Here, the scalar quantization is replaced by vector quantization, see Vary and Martin (2006, pp. 290–301) and Vary et al. (1998, pp. 319–327). This technique is most often

encountered in today's employed codecs. Examples include the GSM half-rate codec (ETSI GSM 06.20 1996), the GSM enhanced full-rate codec (ETSI GSM 06.60 1996), and the adaptive multi-rate (AMR) codec, see ETSI GSM 06.90 (2000). The latter was designed for GSM and UMTS. In general, it is particularly suitable for potentially adverse or time-varying speech channels such as packet-switched networks since the bitrate can be adapted to the resource needs of the channel coding in order to increase error robustness.

Except for waveform codecs employing solely a speech-optimized quantization curve like the G.711, codecs of all three classes are based on LP filtering, being the reason why the term *linear predictive coding (LPC)* refers to all of the presented codec technologies.

Speech codecs can further be classified according to the audio bandwidth they are able to transmit. In traditional circuit-switched telephony, the audio bandwidth is restricted to approximately 300–3400 Hz. Due to the channel bandpass (ITU-T Rec. G.712 2001), speech codecs are restricted to that bandwidth as well (f_s = 8 kHz), cf. Sect. 2.2.3. This frequency range is commonly referred to as *narrowband (NB)*. Although not widely employed, in ISDN also *wideband (WB)* transmission is possible with the G.722 codec, a sub-band ADPCM codec (ITU-T Rec. G.722 1988; 50–7000 Hz, f_s = 16 kHz). In digital cellular networks of the more recent generation, WB transmission is in principle enabled by the WB version of the AMR (AMR-WB, see ITU-T Rec. G.722.2 2003).[10] There is no principle restriction of audio bandwidth in packet-switched transmission. Hence, *super-wideband (SWB)* and *fullband (FB)* speech transmission with audio bandwidths of 50–14000 Hz and 20–20000 Hz, respectively, are possible here.

In the present work, apart from the codecs mentioned so far, the following codecs are employed in auditory experiments: The G.723.1 codec (dual rate codec operating at 5.3 or 6.3 kbit/s; ITU-T Rec. G.723.1 2006) and the G.729A codec (conjugate-structure algebraic CELP, CS-ACELP; ITU-T Rec. G.729 Annex A, 1996), both often used in packet-based systems, and the G.722.1 codec (ITU-T Rec. G.722.1 2005). Other codecs can be found, for example, in the G-series of ITU-T recommendations.

Independent of the codec technology, further temporal bitrate savings can be achieved with a discontinuous transmission (DTX) in such a way that only segments of active speech are transmitted. Voice-activity detection (VAD), applied prior to the encoding process, is used to identify segments with speech activity. False classifications might lead to front/end or mid-speech clipping. In speech pauses, comfort noise is generated in the receiver in order to suggest an active connection to the user, see Sect. 2.2.4.

The coding and quantization techniques mentioned are just a coarse description of the complex processing steps actually involved in today's codecs. A more detailed description also of further techniques can be found, for example, in Vary et al. (1998, pp. 233–376) or Vary and Martin (2006, pp. 201–314).

[10] Note that the bandwidth for the highest bitrate mode is approximately 50–6600 Hz; the upper cutoff frequency decreases slightly with decreasing bitrate (Raake 2006, p. 59).

Further processing steps in order to prepare the encoded speech for transmission over fixed or mobile packet-based networks are introduced in the following.

Packet-based networks such as the Internet can be described with the OSI (Open Systems Interconnection) reference model (Day and Zimmermann 1983), see also, for example, Perkins (2003, pp. 16–23). The OSI model hides the physical interconnection between network nodes (*physical layer*) from the application (*application layer*), for example consisting of the user interface and audio stream of a software VoIP phone. In the OSI model, between the physical and the application layer there are several intermediate layers, each building on the services provided by the next lower layer. Higher layers thus provide higher abstraction. Each layer uses a defined protocol for communication. At the sender, the layers are traversed from the application layer towards lower layers, whereas the data subject to transmission is encapsulated by adding a *header* in each layer, containing layer-specific information about the *payload*, such as its size, a time stamp, or a checksum. This process is referred to as *packetization*. The headers comply with the protocol of the respective layer.

In VoIP, the encoded speech frames are embedded into packets compliant to the Real-time Transport Protocol (RTP, IETF RFC 3550 2003),[11] where one packet usually contains one or several frames (for a detailed description of packetization principles, see Perkins 2003, pp. 152–157). The packets are supplemented by an RTP header, containing information on the packet's content, and are passed to the lower layers. In addition, the RTP Control Protocol (RTCP) is employed for the reception of quality feedback of the receiver and other control functions. Further details on RTP and RTCP can be found in Perkins (2003, pp. 51–144).

RTP makes use of the service the *transport layer* provides. Here, two protocols are common: The connection-oriented and thus reliable Transmission Control Protocol (TCP), and the connectionless User Datagram Protocol (UDP). UDP is commonly used in VoIP speech and other media transmission. Although UDP is less reliable, it provides "timely transmission", which is desirable in "real-time" services (Perkins 2003, p. 45). The increase of loss probability (also detected in this layer) is concealed or corrected by techniques described below. TCP, in turn, is commonly used for VoIP call signaling and control using the Session Initiation Protocol (SIP; IETF RFC 3261 2002) in conjunction with the Session Description Protocol (SDP; IETF RFC 2327 1998).

Since speech data might get lost or corrupted during speech transmission, redundancy can proactively be created by sending a separate duplicate stream of encoded data. This is referred to as *Forward Error Correction (FEC)* (IETF RFC 2733 1999). Instead of FEC, it is also possible to create a second, low-bitrate encoded version of the stream (*Low Bitrate Redundancy, LBR*; LBR; IETF RFC 2198 1997). Another more passive possibility is packet retransmission (e.g., based on RTCP requests, see Perkins 2003, pp. 276–285). Both FEC and the technique of packet retransmission have potential problems with delay (in particular, retransmission requires at least one additional round-trip delay).

[11] The assignment of RTP to the OSI framework is ambiguous. RTP is related to the transport layer, the session layer, as well as the presentation layer (Perkins 2003, p. 57).

2.2.3 Channel

An important role for VoIP plays the Internet Protocol (IP), residing below the transport layer on the *network layer*. This *IP layer* provides higher layers with a unified addressing system and, apart from other functions, masquerades the next lower layers (*link layer* and *physical layer*), that is, the potentially heterogeneous underlying network technology with their rudimentary link protocols: The transmission physically can take place wired or wireless, over different gateways, network switches, and cables, over Digital Subscriber Line (DSL) or Ethernet. On these lower layers, several "distortions" may occur, which finally influence the IP layer.

Bit errors might occur on the physical layer, for example due to signal interference or noise, especially in mobile radio transmission channels. Such errors translate to corrupted packets in packet-based transmission. Corrupted packets can be identified by integrity checks on the IP layer (detection of faulty headers) or the transport layer (detection of faulty payload). Depending on the source coding, partial checksums, that is, checksums of groups of relevant bits are also possible (e.g., for the AMR codec). If such techniques are used in conjunction with IP-based networks, transport protocols need to be deployed that allow for partial checksums of the payload (such as UDP-Lite, IETF RFC 3828 2004).

However, corrupted packets are usually discarded (Perkins 2003, p. 41). Thus, looking top-down from the application layer, such packets appear to be *lost* (Perkins 2003, p. 32). Thus, in the present work, it is assumed that frame-corrupting distortions on the physical layer always translate into packet loss.

Apart from that, the IP layer transmission is done in a "best-effort" fashion, that is, datagram packets are delivered without a reliable or timely delivery (Perkins 2003, p. 19). The transmission "behavior" and performance in IP-based networks might be different in every link. Apart from distortions in lower layers as described above, this might depend, for example, on the route the packets take (number and geographic location of network links traversed) and on the type and amount of parallel traffic (potential network congestion). The underlying network characteristics might also change during the life-time of an established link. For example, congestion might occur or the network route might change. As a result, the IP packets may get lost (*packet loss*). Since the IP packets might take different routes through the Internet, and thus arrive at different points in time in a possibly different rank order, packet *jitter* is likely to occur, which adds time-varying delay (jitter refers to the variation in inter-packet delay, see Perkins 2003, p. 184). In turn, this may lead to an unordered packet arrival at the receiver.

Overall, the IP layer can be considered as the "primary factor determining the performance of a system communicating across the Internet" (Perkins 2003, p. 22). Due to the layer structure of the OSI model, any irregularities occurring in the network layer (resulting from distortions in even lower layers) are directly passed to the upper layers. For example, if an IP packet is lost, the respective content is missing in the session layer and eventually in the application layer as well, if no counter-measures are taken.

In order to quantify distortional effects on IP-level, there exists a number of metrics. Assuming *independent* packet loss, the average packet loss rate (random packet loss probability) P_{pl} can be used to get a broad picture of a given network setup and might reflect, for example, the network congestion. However, this coarse measure does not reflect the fact that packet loss is usually not an independent process, that is, the losses are not uniformly distributed over time. On the contrary, packet loss most often occurs in *bursts*, that is, there is some probability that the loss of one packet is *dependent* on the loss of the previous one, resulting in packet loss patterns. These patterns can often be modeled by n-state Markov chains or Gilbert/Gilbert-Elliot models (see Raake 2006, pp. 63–69 for an overview). One can further subdivide packet loss or, more general, time-varying channel behavior into *microscopic* and *macroscopic* loss (Raake 2006, pp. 71–76), where both terms refer to user perception: Microscopic time-related degradations are perceived as such, without leading to time-varying quality perception, that is, the magnitude of the percept is stationary. Macroscopic time-related degradations, in turn, consist of multiple microscopic segments of sufficient duration and with different amount of degradation, thus leading to instantaneously changing features and quality perception. The focus of the present work is on microscopic loss behavior. For more details on the measurement of IP-network performance, see Perkins (2003, pp. 25–42).

Despite of the distortions packet-based networks potentially introduce, many of the drawbacks of circuit-switched networks can be overcome with packet-based transmission. Electrical echoes due to signal reflection at 2-wire-4-wire hybrids or circuit noise in analog networks become less important. In analog circuit-switched networks, analog frequency division multiplexing with a frequency grid of 4 kHz and sub-audio signaling was foreseen, limiting the transmission bandwidth to the range 300–3400 Hz in NB telephony (ITU-T Rec. G.712 2001), which was widely maintained in digital circuit-switched transmission for compatibility reasons. For further information on traditional telephone networks, refer to Richards (1973) and Möller (2000), where the latter includes more recent technologies such as mobile telephony. In packet-based transmission, in the contrary, no direct reference to the physical channel is being made. Due to the more abstract transport architecture, bandwidth constrains are rendered obsolete. WB transmission with a bandwith of 50–7000 Hz (ITU-T Rec. G.712 2001) or beyond is possible, cf. Sect. 2.2.2.

2.2.4 Channel to Ear

Once the transmission physically took place, either over a wired or a wireless channel, the OSI layers are traversed in reverse order at the receiver, starting from the physical layer. The collected packets are unwrapped (depacketized) layer for layer, until reaching the transport layer and higher layers where the RTP resides. The bit frames (i.e., the encoded frames) are stored in a playout buffer and sorted according to the RTP time-stamps. In addition to handling the media RTP packets, the receiver is responsible for sending reception “quality” statistics back to the sender

using RTCP. Counter-measures are taken in the receiver in order to *repair* or *conceal* degradations occurred during transmission. The fact that packets may arrive late or in an unordered fashion is accounted for by the *jitter buffer* in the receiver. The size of the jitter buffer directly corresponds to its delay. The size is chosen as a trade-off between the additionally introduced delay, and an inevitable *discard* of packets due to non-timely arrival. Thus, jitter can eventually be translated into packet loss, similar to the effect of corrupted frames as described earlier. Jitter buffers are usually realized in an adaptive manner in order to account for varying network conditions (evaluated, e.g., by RTCP statistics). Adaptation of the buffer size optimally takes place in silence periods between talk spurts in order to avoid audio glitches. There exists a number of possible realizations (e.g., Perkins 2003, pp. 166–192).

The decoder generally performs the inverse procedure that was done in the encoding process. For hybrid codecs, this comprises, for example, LP synthesis filters, as already discussed in Sect. 2.2.2. The decoded speech frames are stored in a playout buffer. The decoder also takes care of the comfort-noise insertion at speech pauses in the DTX mode.

Exploiting the forward error-correction techniques of the sender, transmission errors can be corrected to a certain extent (depending, e.g., on the amount of redundancy information). From the perspective of the decoder, jitter that led to packet discard in the jitter buffer and packet loss that could not be repaired are translated to *effective* packet loss (Raake 2006, pp. 72–74). Here, packet loss concealment algorithms attempt to (perceptually) mask the lost information. That is, errors that cannot be corrected (e.g., if the redundant packets do not arrive in time) can be *concealed* at the receiver (decoder) by *packet loss concealment (PLC)* techniques. The amount of error correction and error concealment often is a trade-off: Since both FEC and retransmission result in increased bandwidth usage, in congestion-prone network conditions, concealment of errors might be more appropriate. Again, the behavior of the sender-receiver interaction can be adaptively and seamlessly scaled through RTCP.

In the trivial case of packet loss concealment, missing packets are substituted by a vector of "zeros" or white or shaped noise of appropriate length (*silence* or *noise substitution*). Other basic error concealment schemes substitute the parameters of bad frames by repetition of the decoded model parameters of the last good frame. If groups of bits are classified a-priori according to their perceptual importance, depending, e.g., on which model parameters they represent, frames might be marked as bad only if the most relevant bits are corrupted (see Vary and Martin 2006, pp. 317–321). Besides these basic PLC strategies, more advanced concealment strategies were developed with varying complexity and quality, such as timescale modification or parameter interpolation based on the codec state. For an overview, see Perkins (2003, pp. 227–241).

After decoding, speech enhancement algorithms might be deployed. One of those also employed in this work's experimental studies is described in the following: Despite the advent of WB technology in mobile radio networks and VoIP telephony, the majority of terminal equipment used today is still only capable of NB transmission. In a connection involving a traditional terminal with NB coding and a

WB-capable terminal, the common NB-codec must be used due to compatibility reasons. In the WB terminal, however, a mechanism for *artificial bandwidth extension (ABE)* might be incorporated to artificially widen the incoming NB signal towards WB by adding the missing spectral components.[12] The technique used today is based on exploiting redundancies of the speech production mechanism modeled by the source-filter model of human speech production as introduced by Carl (1994) and Carl and Heute (1994): The bandwidth extension towards lower frequencies (below 300 Hz) and higher frequencies (beyond 3400 Hz) can be treated separately for the excitation (glottis) signal and the vocal tract filter (spectral envelope). The latter has been shown to be perceptually more crucial and thus the more challenging task in terms of accuracy, in particular if no auxiliary information is transmitted (which is the case for the described application scenario). The missing high-frequency components of the excitation signal can be produced by modulation of the bandlimited excitation signal (see Carl 1994; Carl and Heute 1994 and Vary and Martin 2006, pp. 375–377). For the WB ABE of the spectral envelope, different proposals can be found in the literature. In the present work, the approach proposed by Carl (1994) and Carl and Heute (1994) is employed. The coefficients for the WB LP filter that is used to model the vocal tract are obtained by "code book mapping" between the coefficient set of the given NB speech and a pre-trained WB code book. For an overview of alternative approaches and a theoretical framework for ABE, see Vary and Martin (2006, pp. 361–388). Variants of ABE are also employed in certain codecs. For example, the "spectral folding" employed in the GSM full-rate codec (ETSI GSM 06.10, 1988) can be considered as ABE with auxiliary information. In the AMR-WB codec scheme (ITU-T Rec. G.722.2 2003), the 6.4–7 kHz frequency band is synthesized by ABE, see Vary and Martin (2006, pp. 298–300).

The analog electrical signal is transduced to a proportional acoustic signal x via the receiver's loudspeaker. Various kinds of terminal equipment can be used at the terminating end, such as wirebound or mobile handsets, monaural or binaural headsets, or hands-free terminals (HFTs). The transduction results in a further (most often) linear distortion, also dependent on the device geometries and properties as well as the D/A converters. Particular attention should thus be spent to the design of these devices (cf. Raake 2006, pp. 90–91).

Moreover, there exist several *talker sidetone* paths describing through which routes a handset-telephone user hears her/his own voice. Besides the acoustic direct air path due to the leakage between the engaged ear and device and the bone conduction path to the engaged ear, there is a mechanical feedback path through the handset. Moreover, an electrical sidetone path might be employed for a feedback of the talker's own voice introduced from the microphone to the loudspeaker in order to compensate for the acoustic shielding on one hand, and feedback on whether the device is working on the other hand (this electrical path is not shown in Fig. 2.4). Furthermore, the *listener sidetone* comprises all paths through which ambient noise is transmitted to the engaged ear.

[12] With a preceding upsampling from $f_s = 8$ to 16 kHz.

The "average insertion loss", that is, the integral measure of the frequency-dependent sensitivity of the sidetone path, is denoted by the *sidetone masking rating (STMR)* in decibels for talker sidetone. The coupling of background noise to the engaged ear is measured by the *listener sidetone rating (LSTR)* in decibels. *STMR* and *LSTR*, both reflecting loudness ratings, can be calculated similarly to the procedure described in Sect. 2.2.2, see ITU-T Rec. P.64 (2007). The so-called *D*-factor characterizes the difference between the sensitivity for the listener sidetone related to the diffuse sound and the transmitting sensitivity for the talker sidetone related to the direct sound, in a frequency weighted way. Thus, the *D*-factor depends, e.g., on the shape of the handset (Möller 2000, pp. 35–36). More details on sidetone can be found in Möller (2000, pp. 28–30).

Note that a talker's own voice, environmental noise, as well as room reflections also arrive at the listener's free ear(s), or they are superimposed to the received signal x when using a HFT (see gray arrow in Fig. 2.3).

2.3 Perception of Transmitted Speech

2.3.1 Introduction

The transmission of speech signals through communication systems obviously leads to a potentially detrimental modification of the original physical signal, the speech sound, emitted from the mouth of the speaker. The signal can be affected by a myriad of entities (or combinations thereof) involved in the transmission process, as described in Sect. 2.2. In simple terms, "components the function of which can be seen as a contribution to the quality forming process" (Jekosch 2005b, p. 16) are referred to as *quality elements* (see Sect. 2.3.4 for a more formal definition).

Möller (2000, pp. 18–19) categorizes "perceptive factors" of (traditional) analog and digital transmission systems as follows:

- Loudness,
- articulation,
- perception of the effects of bandwidth and linear frequency distortion,
- perception of sidetone,
- perception of echo,
- perception of circuit noise,
- effects of environmental noise and binaural hearing, and
- effects of delay.

This list is extended in Möller (2000, pp. 201–203) by perceptive effects caused by more recent speech transmission technology:

- Effects of interruptions and time variations (e.g., caused by speech clipping in imperfect VAD),

- effects of transmission errors (e.g., caused by bit errors),
- perception of speech degradation (e.g., caused by low-bitrate codecs),
- effects of additional delay and residual talker echo (e.g., caused by packetization and imperfect AEC, respectively).

Much is known about the influence of transmission components with respect to quality perception in a broader sense, see, e.g., Möller (2000, pp. 17–35 and pp. 199–203) and Raake (2006, pp. 51–91) for an overview (the latter focussing on VoIP transmission).

However, a categorization of perceptive factors caused by transmission components can also be conceived to be literally based on "pure perception". This approach has several advantages: Transmission components causing similar effects in the perceptual domain can be grouped together and related to quality according to their perceptual importance, thus providing diagnostic information of overall perception. Provided that the purely perceptual factors are valid also for future transmission components, the perceptual factors appear to be more universal and stable over time, while transmission components are subject to constant change.[13] Furthermore, insight into the way these perceptive factors are combined by the human user is gained. As a prerequisite, however, a definition of criteria in the perceptual domain is needed. As it will be seen in this section, *perceptual features*, and—given that these features are independent from each other—*perceptual dimensions* spanning the *perceptual space* of the user provide a means for this. A framework for feature-based or dimension-based quality modeling is provided in the following sections.

2.3.2 Schematic of a Listener

As it was defined in Sect. 2.2.1, the ERPs are regarded as the entrance points to the listener. Depending on the terminal equipment, at least at one of the ERPs, a modified version of the acoustic signal emitted by the speaker is present.[14]

The received speech sound appears at a certain time with certain physical features: The sound is distinct in time, space, and other characteristics (Blauert 1997). Hence, it is referred to as *sound event*. As a sound event can generally be determined by means of more than one physical characteristic (e.g., bandwidth or level), it is generally of multidimensional nature, and thus can geometrically be represented by a *point* in a multidimensional *sound space*. For its description, the (position) vector $\boldsymbol{s}_0 = [s_{0,1}, s_{0,2}, \ldots, s_{0,N}]^T$ is used in the following. The nth element of this

[13] Although this cannot ultimately be guaranteed, as new perceptual factors might emerge or perceptual factors that are important today might become irrelevant in the future.

[14] In general, it is differentiated between monaural and binaural listening. In monaural listening, it is assumed that only one ear is engaged, typically with the usage of a handset terminal in a more or less quiet environment. In binaural listening with headphones, for example, it is further differentiated between monotic, diotic, and dichotic listening.

vector is denoted by $s_{0,n}$, where $n = 1, 2, \ldots, N$, N being the number of physical characteristics.[15]

From this point onwards, the physical universe is left. It is of interest how the received sound is actually perceived—in terms of perceptual features and finally in terms of quality: These perceptual characteristics are considered as the most important metrics of a communication system from a user's point of view. To this end, human listeners are required to attend auditory experiments.[16] Such experiments are usually of formal nature and require, for example, a precisely controlled laboratory environment, listeners with sufficient hearing capabilities, and a defined stimulus inventory.

The perception processes of (speech) sounds are described following the system-theoretic analysis of the auditory experiment depicted in Fig. 2.5. It describes the relevant processes of human perception and judgment by "black boxes" and empirical functional relations between their respective input and output ports. The schematic thus serves as a simplified and symbolic representation of the vast number of peripheral and central physiological and psychological mechanisms involved in perception. It was initially proposed by Blauert (1997, pp. 5–12) and extended in Raake (2006, pp. 16–18), see also Jekosch (2005b, pp. 71–74), for the case of quality assessment.[17]

As it can be seen from Fig. 2.5, several sub-systems are assumed to be part of the overarching black-box "listener", that is, only the sound event and descriptions thereof are accessible to the outside world.

The auditory event $\boldsymbol{w}_0$, its elements (perceptual features or dimensions), and its descriptions $\boldsymbol{\beta}_0$ are explained in Sect. 2.3.3. The formation of the quality event q_0 (and its description b_0) by a comparison between the expectation $\boldsymbol{r}_0$ (whose elements are the desired features) and the auditory event $\boldsymbol{w}_0$ is described in Sect. 2.3.4. The other elements depicted in Fig. 2.5 ($\boldsymbol{w}_0(t_1)$, $\boldsymbol{w}_0(t_2)$, and δ_0) are explained in later sections.

Note that although important definitions (e.g., quality) are provided in the context of the *listener* schematic, the definitions are very general and apply also, for example, for the conversation situation.

---

[15] All events that are of multidimensional nature can be represented by *points* in a multidimensional space. For a description of these points, vector notation is used from now on. In particular, multidimensional events are represented by *position vectors*, indicated by boldface variables. By such vectors, points in a space are represented in relation to the origin of this space. Thus, a vector's elements are equal to the coordinates of a point in space representing the event.

[16] Experiments for investigating speech perception are regarded to be a special case in psychoacoustics (Jekosch 2005b, p. 60).

[17] Although this schematic is here employed for the auditory modality (i.e., for psycho-acoustic measurements), it can analogously be used for other psycho-physical measurements analyzing visual, tactile, olfactory, or gustatory perception. Instead of the terms sound event and auditory event, it can more generally be referred to physical event and perceptual event, respectively (Möller 2010, pp. 23–25).

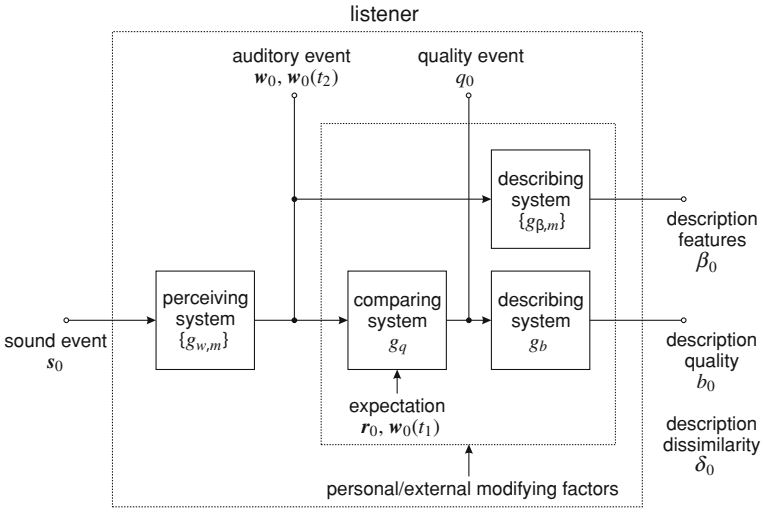

Fig. 2.5 Schematic of a listener in an auditory test according to Raake (2006, Fig. 1.7) (based on Blauert 1997, Fig. 1.1 and extensions by Jekosch 2005b)

2.3.3 Perceptual Features and Dimensions

Assuming that the properties of the sound event $\boldsymbol{s}_0$ match the hearing capabilities of the listener, an *auditory* (or, more generally, *perceptual*) *event* is caused by $\boldsymbol{s}_0$ and transformed into the perceptual domain via the "perceiving system". The perceptual event is generally of multidimensional nature and might be composed of a multitude of perceptual *features*. According to Jekosch (2005b, p. 14), a *feature* is a

> recognizable and nameable characteristic of an entity,

where the "entity" in the context considered here is an auditory event $\boldsymbol{w}_0$. Example features might be describable with attributes like loudness or timbre. The perceived features arise inside the listener in a reflection process (cf. Raake 2006, Fig. 1.6)[18] by decomposing the *perceived composition*, that is, the

> totality of features of an entity [...] (Jekosch 2005b, p. 16),

see also Raake (2006, Fig. 1.6).

The *perceptual event* and the *perceived composition* are interchangeably used in the following. The perceptual event can geometrically be thought of as a point in a multidimensional *perceptual space* and is described by the position vector $\boldsymbol{w}_0 = [w_{0,1}, w_{0,2}, \ldots, w_{0,M}]^T$ in the following, where its mth element is denoted by $w_{0,m}$ ($m = 1, 2, \ldots, M$ and M being the number of perceptual features). If the coordinate system of the perceptual space is Cartesian, that is, its basis is orthogonal, and

[18] "By *reflection*, the real experience of perception is interpreted and thus given intellectual properties" (Jekosch 2005b, p. 58).

each of the auditory features has the property of lying along one of these orthogonal axes and thus are themselves orthogonal, these features are in the remainder referred to as *perceptual dimensions*. The dimensionality of this perceptual space is equal to M.

In the following, functions mapping one event onto a different type of event or event component are called *psycho-physical functions*[19] (more specifically *psycho-acoustic functions*), see Blauert (1997, pp. 8–9), and are denoted by g, where the subscript indicates the dependent (target) value (see Fig. 2.5). Thus, the functions g_q and g_b represent the "comparing" and one of the two "describing systems", see Fig. 2.5. For the "perceiving system", more than one psycho-acoustic function is in general needed that describes the relation between the sound event and one particular perceptual feature of the perceptual event. The functions $\{g_{w,m}\}$, with $m = 1, 2, \ldots, M$, can be considered as components of the "perceiving system". Likewise, the functions $\{g_{\beta,m}\}$ represent components of the second "describing system".

The functional psycho-acoustic relation between the sound event $\boldsymbol{s}_0$ and the element $w_{0,m}$ of the perceptual event $\boldsymbol{w}_0$ is represented by the psycho-acoustic function $g_{w,m}$ according to

$$w_{0,m} = g_{w,m}\left(\boldsymbol{s}_0\right) \ . \tag{2.1}$$

As mentioned in Sect. 2.3.2, apart from introspection or physiological measurement, the auditory events are not accessible from outside the listener. Thus, a *description* $\boldsymbol{\beta}_0$ of the auditory event via the "describing system" is necessary to be obtained from the listener. Here, the elements $\{\beta_{0,m}\}$ correspond to the elements $\{w_{0,m}\}$, modified by the "describing system", itself being dependent on personal and external modifying factors (such as experience of the listener or the environmental context). Implications of this modification are discussed in the following sections.

2.3.4 Integral Quality, Quality Features and Dimensions, and Quality Elements

Given that the perceptual event $\boldsymbol{w}_0$ is present, the *quality event* q_0 arises following the definition provided in Jekosch (2005b, p. 15): Quality is the

> result of [the] judgment of the perceived composition of an entity with respect to its desired composition.

Quality thus is a result of an internal comparison between the *perceived composition* and the *desired composition*, and thus quality corresponds to the output of the "comparing system" in Fig. 2.5. The properties of these two building blocks of quality are detailed in the following.

[19] "Psycho-physics measures the relationship between physical phenomena and phenomena of perception" (Jekosch 2005b, p. 61).

Those *features* of the perceived composition $\boldsymbol{w}_0$ that are *relevant* for quality are referred to as *quality features*, the definition of it being an extension of the feature definition provided earlier. Jekosch (2005b, p. 17) formally states that

> a quality feature is a recognized and designated characteristic of an entity that is relevant to the entity's quality.

Correspondingly, if the features are orthogonal in addition, they are referred to as *quality dimensions* in the following.

The

> totality of features of individual expectations and/or relevant demands and/or social requirements

is referred to as the *desired composition* (Jekosch 2005b, p. 16). The desired composition, in the following denoted by the position vector $\boldsymbol{r}_0 = [r_{0,1}, r_{0,2}, \ldots, r_{0,M}]^T$, can geometrically be represented by a point in the same perceptual space in which also the perceived composition $\boldsymbol{w}_0$ is located. Hence, the desired composition $\boldsymbol{r}_0$ and the perceptual event $\boldsymbol{w}_0$ share the same dimensionality M. The desired features again are the result of a reflection process inside the listener (cf. Raake 2006, Fig. 1.6). The mth desired feature of $\boldsymbol{r}_0$ is denoted by $r_{0,m}$, with $m = 1, 2, \ldots, M$.

By the definition given above, it is implied that for every perceptual feature value $w_{0,m}$, there exists a desired feature value $r_{0,m}$. Hence, the desired composition constitutes a multidimensional *internal reference* of the listener storing reference values for the characteristics of the auditory event $\boldsymbol{w}_0$. This process can be regarded as an anticipation of the percept $\boldsymbol{w}_0$ by the listener (Raake 2006, p. 15 and Fig. 1.6), and thus her/his *expectation*. The expectation $\boldsymbol{r}_0$ can, for example, be conceived as an *ideal point* in the perceptual space to which all perceptual events are *compared* in the process of formulating the integral quality (cf. Sect. 2.5).

In contrast to Eq. (2.1), a direct relation between the sound event $\boldsymbol{s}_0$ and the quality event q_0 cannot be established, since the quality formation process depends on the variable $\boldsymbol{r}_0$, the expectation of the user. According to the definition of quality given in this section, quality is the result of a comparison between what is perceived, that is, $\boldsymbol{w}_0$, with what is expected, that is, $\boldsymbol{r}_0$. In mathematical terms, quality events result from the relation

$$q_0 = g_q\left(\boldsymbol{w}_0, \boldsymbol{r}_0\right) \ . \tag{2.2}$$

Stepping back to the technological domain of speech transmission as introduced in Sect. 2.2, it can now be formalized by which elements "quality perception" can actually be caused. Components in the speech transmission chain can constitute *quality elements*, for which a general definition is given in Jekosch (2005b, p. 16):

> [A quality element is the] contribution to the quality
>
> - of a material or immaterial product as the result of an action/activity or a process in one of the planning, execution, or usage phases
> - of an action or of a process as the result of an element in the course of this action or process.

Quality elements are the physical counterpart to quality features: "While [a quality element] is the building block for designing an entity, a quality feature is the analyzed result of the perceived, designed entity and is therefore the basis of any description of its quality" (Jekosch 2005b, p. 17).

In the present work, the quality of transmitted speech (and also its quality features) is considered from mouth to ear (the elements involved in this transmission path were described in Sect. 2.2) according to the definition given in ITU-T Rec. P.10/G.100 (2006): Mouth-to-ear quality is the

> speech quality as experienced by the user of a voice communication system. [It] includes the whole transmission path from the mouth of the talker to the ear of the listener.

In order to stress that quality is a scalar value obtained by an integration of different perceptual features or dimensions, the term *integral quality* is used in the following (cf. Möller 2000, p. 11).[20]

Similar to the perceptual event, see Sect. 2.3.3, the quality event q_0 is not accessible from outside the listener. Thus, only a *description* b_0 of the quality event can be obtained via the "describing system", resulting in a modification of q_0 depending on personal and external modifying factors. In addition, the expectation $\boldsymbol{r}_0$ is not invariant over time and is highly dependent on personal and external factors. Considerations on factors influencing the quality description are given in the following sections.

2.3.5 QoS and QoE Terminology

Quality, in general, can be regarded from the perspective of service provision (referred to as *Quality of Service, QoS*) and from the perspective of the user of this service (referred to as *Quality of Experience, QoE*).

The term QoS is typically employed for describing all aspects related to the acceptability [21] of a service. Accordingly, ITU-T Rec. E.800 (2008) gives the following definition:

> [QoS is the] totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service.

The term QoE, in turn, strongly emphasizes the "perception of QoS" and is defined in ITU-T Rec. P.10/G.100 (2006) as follows:

> The overall acceptability of an application or service, as perceived subjectively by the end-user.

---

[20] Note that quality is said to be of multidimensional nature in everyday language. However, according to the definitions provided here, integral quality is a one-dimensional (scalar) value, whereas both the perceptual event and the internal reference are of multidimensional nature.

[21] Acceptability is typically measured as the ratio between the number of potential users and the number of actual users of a service, cf. Möller (2000, p. 14).

> NOTE 1 Quality of experience includes the complete end-to-end system effects (client, terminal, network, services infrastructure, etc.).
>
> NOTE 2 Overall acceptability may be influenced by user expectations and context.

The notes indicate that the complete mouth-to-ear system is considered (see Sect. 2.3.4), and that modifying factors such as the user's expectation play a role. Thus, QoE is in line with the notion of *quality* as defined by Jekosch (2005b) and which is used throughout this book.

Möller (2000) developed a comprehensive taxonomy for QoS that was slightly updated in Möller (2005), revealing the relevant aspects of QoS and their interactions and how these aspects are related to the acceptability of a service. According to this taxonomy, QoS builds on three groups of factors: *speech communication factors*, *service factors*, and *contextual factors*. Thus, according to the above definition, QoE is implicitly part of this framework. Besides *ease of communication* and *conversation effectiveness*, *(one-way) voice transmission quality* is one of the building blocks of *speech communication*. Together, they describe the *communication efficiency*.[22] According to this taxonomy, as it is mainly focused on the listening-only situation, the work presented in the following chapters focuses on *(one-way) voice transmission quality*. For further details, for example, on how factors such as economical benefit integrate into QoS, see Möller (2000, pp. 11–15 and pp. 43–44) and Möller (2005). Considerations on measures to guarantee a certain QoS in VoIP such as packet prioritization can be found in Raake (2006, pp. 55–56).

2.4 Auditory Quality Measurement

2.4.1 Listener as a Measuring Organ

The aim of psycho-acoustic measurement in auditory tests is to arrive at *quantitative* descriptions $\boldsymbol{\beta}_0$ and b_0 of the perceptual and quality events $\boldsymbol{w}_0$ and q_0, respectively, in order to obtain utilizable and communicable results. According to Jekosch (2005b, p. 91), an auditory (speech) test is

> a routine procedure for examining one or more empirically restrictive quality features of perceived speech with the aim of making a quantitative statement on the features.[23]
>
> A formalization of the description process is achieved by *measuring*, which is the entirety of all the activities in the measurement chain [...],

for example, the choice and practical application of the measuring technique (Jekosch 2005b, p. 63). In psycho-physical measuring, these techniques are referred to as *psychometric methods*. The *measurand* is the

---

[22] *Efficiency* is defined as "the resources expended in relation to the accuracy and completeness with which users achieve specified goals" (ETSI Guide EG 201 013 1997).

[23] Note that the term "features" is here meant to include "quality" as defined in Sect. 2.3.4 as well.

feature of an object to be measured which can numerically be described in the course of the measuring process (Jekosch 2005b, p. 61).

Scaling, a part of the measuring process, is the concrete assignment of numbers (*scale values*) to measuring objects according to consistent rules, cf. Stevens (1951, p. 22) and Jekosch (2005b, p. 63). The relation between the attributes under test, that is, the measurands, and the assigned numbers must be isomorphic, and the numbers must be chosen in such a way that each element under consideration can be assigned to a number. The set of numbers is called a *scale* (Blauert 1997, p. 7).

The (physical) measuring object is the sound event $\boldsymbol{s}_0$, where the measurand is a characteristic n of a sound event $\boldsymbol{s}_0$. *Instrumentally*, scaling is achieved by means of a physical *measuring instrument*. In analogy, in the perceptual universe, the measurement *subjectively* takes place, where the listener acts as a *measuring organ* (Jekosch 2005b, p. 65). More precisely, the "describing system" is considered as a "psycho-physical measurement instrument" in the listener schematic in Fig. 2.5 (cf. Blauert 1997, Fig. 1.3). The measurand is either a feature m of the auditory event $\boldsymbol{w}_0$ or quality itself.

2.4.2 Scaling Functions

Similar to the proposal in Blauert (1997, pp. 8–9), however, following the vector notation introduced so far, multidimensional *scale vectors* are denoted by $\boldsymbol{s}$, $\boldsymbol{w}$, and $\boldsymbol{\beta}$ for the corresponding sound event $\boldsymbol{s}_0$, the perceptual event $\boldsymbol{w}_0$, and the description $\boldsymbol{\beta}_0$, respectively. The *scale values* are denoted by the elements $s_1, s_2, \ldots, s_N$ of $\boldsymbol{s}$, the elements $w_1, w_2, \ldots, w_M$ of $\boldsymbol{w}$, and the elements $\beta_1, \beta_2, \ldots, \beta_M$ of $\boldsymbol{\beta}$.

As a convention, scaling functions mapping the single elements of events to scale values are denoted by f, where the subscript indicates the dependent (target) value. The scaling functions are denoted with $\{f_{s,n}\}$, $\{f_{w,m}\}$, and $\{f_{\beta,m}\}$. The scaling function $f_{s,n}$ with $s_n = f_{s,n}(s_{0,n})$ represents the physical measurement by means of a physical measuring instrument for a given physical characteristic n. The scaling function $f_{\beta,m}$ transforms the feature description $\beta_{0,m}$ for the perceptual feature m to a scale value β_m according to $\beta_m = f_{\beta,m}(\beta_{0,m})$. However, as the auditory event $\boldsymbol{w}_0$ is hidden and not accessible from outside the listener, the scaling function $w_m = f_{w,m}(w_{0,m})$ does not exist (Jekosch 2005b, pp. 71–72).

As the description $\beta_{0,m}$ represents a measurement of the mth perceptual feature value $w_{0,m}$, that is, $\beta_{0,m} = g_{\beta,m}(w_{0,m})$, it follows that

$$\beta_m = f_{\beta,m}(\beta_{0,m}) = f_{\beta,m}(g_{\beta,m}(w_{0,m})) \,. \tag{2.3}$$

Hence, the scaled description β_m equals the mth auditory feature $w_{0,m}$, modified both by the conversion through the "describing system" $g_{\beta,m}$ and the scaling function $f_{\beta,m}$. As can be seen from Fig. 2.5, these functions are both affected by personal and modifying factors. Thus, β_m can be considered as a *biased* version of the sought

perceptual feature $w_{0,m}$ (as further described in Sect. 2.4.4). In Sect. 2.4.3, further details are provided on how β_m can be measured.

Analogous to the measurement and scaling considerations above, a scale value q reflecting the (scalar) *quality event* q_0 is defined by means of a scaling function $q = f_q(q_0)$. Values b representing scaled versions of the quality description b_0 are related via the scaling function $b = f_b(b_0)$.

Analogous to the auditory event $\boldsymbol{w}_0$, the quality event q_0 is not accessible outside the listener, and the scaling function $q = f_q(q_0)$ is practically not existing (Jekosch 2005b, p. 73). Thus, the listener describes what is perceived through the respective "describing system", see Fig. 2.5, resulting in the scalar quality *description* b_0 that represents a measurement of q_0, that is, $b_0 = g_b(q_0)$. It follows that

$$b = f_b\left(b_0\right) = f_b\left(g_b\left(q_0\right)\right) \ . \tag{2.4}$$

Hence, the scaled description b equals the quality event q_0 modified by both the conversion through the "describing system" g_b and the scaling function f_b. Again, both functions can be attributed to the scaling process affected by personal and modifying factors. The scaled description b can be considered as a biased version of the sought quality event q_0, see Sect. 2.4.4. In Sect. 2.4.3, further details are provided on how b can be measured.

2.4.3 Psychometric Methods

2.4.3.1 General Categorization

The psycho-acoustic measuring principles, that is, the practical methods aiming at obtaining scaled descriptions b and $\boldsymbol{\beta}$, can be different and their usage depends on the psycho-acoustic measurement task. According to Blauert (1994), see also, for example, Möller (2000, pp. 48–49), Möller (2010, pp. 32–35), and classical literature such as Guilford (1954), psychometric methods can be categorized according to the following criteria:

- The *scaling method* and the resulting *scale level*,
- the *presentation method*,
- the *test "modality"*, and
- the *mediacy*.

These criteria are further described in this section.

Depending on the *scaling method*, the numbers assigned to the stimuli in the scaling process can have different properties according to which measurement scales can be classified. Four different scale levels are distinguished (Stevens 1946), see also Blauert (1994, pp. 7–8), Möller (2000, pp. 66–67), and Möller (2010, pp. 41–42):

- *Nominal scales* are based on the identity property: Each number is identical with itself and is used as a label without any meaningful relations between the numbers.
- In addition to the property of nominal scales, *ordinal scales* are based on the rank order property: The numbers are arranged in a specified order. The steps of the scale, however, are not necessarily equidistant.
- In addition to the properties of ordinal scales, *interval scales* are based on the additivity property of intervals: The steps of the scale, therefore, are equidistant.
- In addition to the properties of interval scales, *ratio scales* contain an absolute "true" zero point (i.e., not only an arbitrary assignment of the number zero, which can be done on nominal, ordinal, or interval scales), as well as ratio relations.

The scale level determines the permissible mathematical (and thus also statistical) operations applicable to the scale numbers.

Ratio scales are obtained by the majority of physical measurements. In psychophysics, this scale level can result from, for example, magnitude estimation (Stevens 1957): The task of the test participants is to assign numbers to test objects, where the numbers are required to reflect the relation between perceptual intensities or *apparent magnitude*. Since ratio scales provide the highest level of information, this scale level is desirable in a psycho-physical scaling task. However, since only the ratios between pairs of values on this scale are meaningful, magnitude estimation does not provide absolute scale values across test participants, which is a major drawback for use in telephony (Möller 2000, p. 68). Category scaling, which is often used in telecommunication-related subjective testing (see Sect. 2.4.3.2), mostly results in ordinal to interval scale levels (Möller 2000, p. 68). *Paired comparisons* such as *similarity judgments* of pairs of stimuli reach interval level at the most due to their relative nature. Möller (2000, pp. 72–74) proposes the use of a scale for combined category-ratio scaling, originally developed by Borg (1982), for use in telephony. With this scale, drawbacks of category scales and ratio scales can be potentially overcome, resulting in "absolute" ratio scale values. In classical psychophysics, most scale levels are nominal or ordinal and stem from tasks like detecting absolute thresholds, differential thresholds (just noticeable differences), or detecting equivalents (point of subjective equality), see Jekosch (2005b, pp. 75–76) and Möller (2010, pp. 34–35). Judgments from identification tasks or from the determination of just noticeable differences are examples for nominal scale level.

Common *presentation methods* are the method of adjustment (a stimulus is adjusted by the test participant or the experimenter until a certain condition is met, e.g., equal loudness of two stimuli) and constant methods (the test participant is asked to scale a constant stimulus).

In speech-based telephony, the following *test "modalities"* are conceivable: Conversation test, listening-only test, talking and listening test.

The *mediacy* of measurements differentiates *direct* from *indirect* measurements. In its classical sense (see, e.g., Guilford 1954, pp. 259–260 and Bech and Zacharov 2006, pp. 67–80), the measurement mediacy is linked to the scaling process: Indirect scaling occurs whenever the test participant does not establish a one-to-one assignment of measuring objects to scale values, but rather gives nominal

(e.g., identification task) or ordinal (e.g., rank ordering task) responses according to a discrimination between the stimuli. The target scale is obtained by a separate transformation of these ratings by the investigator (e.g., through the law of comparative judgment, Thurstone 1927a, see also Guilford 1954). In direct scaling, in turn, the predefined scale is the target scale and the test participant employs this target scale to formulate her/his rating.

In this book, the definition of *indirect* measurement is used in a more general fashion: Rating scales that do not provide the target values the investigation actually aims at, that is, subjective judgments that require a substantial, non-trivial mathematical operation in order to arrive at the target scale(s), are referred to as an *indirect* measure of the target scale through an intermediate scale.

As an example, the similarity between pairs of stimuli can be rated *directly* by means of a similarity scale.[24] However, the similarity scale is regarded as an intermediate scale, because the obtained similarity data is input to a mathematical algorithm performing multidimensional scaling (see Sect. 2.4.3.2 and 3.2), transforming the dissimilarities into distances. The conditions can be represented as points in a multidimensional space. The coordinate axes of this space represent the sought scales, which reflect the perceptual dimensions, that is, orthogonal components of a feature vector $\boldsymbol{\beta}$. According to the generalized definition used in this book, these new scales are obtained *indirectly* via similarity judgments. In Chap. 4, a method is developed to scale the perceptual dimensions in a *direct* way.

Instead of extending the classical meaning of the measurement mediacy as it is followed in the present work, Bech and Zacharov (2006, pp. 44–65) denote methods "based on the assumption that there is a close relationship between a given sensation and the verbal descriptors used by the subject to describe the sensation" as *direct elicitation*, whereas *indirect elicitation* methods "try to separate sensation and verbalization", such as multidimensional scaling (as a method without actually labeling the sensation) or body gestures.

Jekosch (2000, p. 82) differentiates between mediate and immediate measurements based on the measurement-process perspective instead of the scaling perspective (cf. Sect. 2.4.1). According to the author's remarks, mediacy depends on the degree to which intermediate subjective processes are necessary for the psychophysical measuring task. Hence, every measurement involving quantitative scaling and thus implying a cognitive *encoding* process (or intermediate process) according to the predefined scale is a mediate measurement, whereas nominal and ordinal ratings are of immediate nature.

In other contexts, further definitions of measurement mediacy are possible. Instead of scaled descriptions $\boldsymbol{\beta}$ or b, task performance measures or physiological measures like eye movement, skin conductance, pulse rate, blood pressure, or electroencephalography data might be used as quality or quality feature indicators in the future (Möller 2010, p. 27 and p. 34). For first studies in these directions, the interested reader is referred to Durin and Gros (2008) and Antons et al. (2010), for example.

[24] Alternative, indirect methods for measuring similarity are given, for example, in Tsogo et al. (2000).

An additional dichotomy can be applied to subjective tests that discriminates *analytical* type of tests from *utilitarian* type of test, see Quackenbush et al. (1988, pp. 15–16) and Raake (2006, p. 16). Subjective tests aiming at the measurement of perceived features described in Sect. 2.3.3 and reflected by $\{\beta_m\}$, especially in the context of speech-transmission quality testing, are referred to as analytical type of auditory tests, in this book also referred to as *multidimensional analysis*. The results are sometimes referred to as *sensory* judgments (Bech and Zacharov 2006). In turn, *utilitarian* type of auditory tests aim at descriptions of the perceived quality as illustrated in Sect. 2.3.4, reflected by b. The results are sometimes referred to as *affective* judgments (Bech and Zacharov 2006). In utilitarian tests, the transmission "system performance [is measured] with respect to one or more criteria [...]" (Jekosch 2005b, p. 107). This measurement process is in the following referred to as *quality assessment*.

2.4.3.2 Common Test Methods

Speech telecommunication systems are usually designed for bidirectional communication, cf. Sect. 2.2. Thus, conversation-type of subjective testing reflects this situation in an ecologically valid way. Here, test participants interact as actual interlocutors exchanging information (Möller 2000, pp. 59–60). In practical tests, conversation scenarios are given to the test participants, where each of the interlocutors takes a specific role to play in order to simulate a natural telephone conversation. Möller (2000, pp. 75–81) developed and experimentally tested economical *short* conversation scenarios that allow for balanced dialogs, short conversation duration, and thus up to 15 conditions to be tested within a single test session. Alternatively, conversation scenarios can be simulated involving only one test participant (ETSI Technical Report TR 102 506 2011), see Weiss et al. (2009) for an application example.

Listening-only is the most common "modality" in subjective tests in telecommunications (Möller 2000, p. 52), probably because such tests are relatively easy to implement while the results are (within limits) directly related to results from the more natural conversation tests. Only one test participant is required at a time (i.e., not two conversation partners), the scaling is faster due to short samples, and no online testing equipment is required (listening-only presentation methods allow to process the sentence material offline in prior to the actual test). Hence, listening-only tests are less expensive in terms of time and money. Listening-only tests rely on more or less purposeless sentence material transmitted through a speech telecommunication system in one direction. Most signal-based instrumental quality measures aim at estimating results from listening-only tests, see Sect. 2.6.2.

A systematic investigation comparing conversation test results and listening-only test results is presented in Möller (2000, pp. 129–133). Although no general relationship between both test modalities could be found, test ratings obtained in conversation tests turned out to be higher, indicating that, amongst other things, in listening-only tests the "surface structure" of the speech signal, cf. Sect. 2.1, is more important to test participants and thus the judgments tend to be more critical in this respect. With

this presentation method, subsystems of a transmission path that require a bidirectional signal flow such as AECs, see Fig. 2.4, or the effects of echo, sidetone, and delay cannot be tested.

Talking-only tests reflect a rather unrealistic communication situation leading to different results as compared to conversation or listening-only tests, at least for untrained test participants (Möller 2000, p. 51 and p. 58).

In utilitarian-type tests aiming at scaling descriptions b of the quality events, *absolute category rating (ACR)* has been proven over the past decades to provide reliable results for telephone applications, both in conversation and listening-only tests (ITU-T Rec. P.800 1996). In integral quality assessment, the scale consists of five discrete categories with the attributes "excellent", "good", "fair", "poor", and "bad", see Fig. 2.6.

The labels suggest an absolute meaning, however, quality is not an absolute measurand and depends on personal and external modifying factors, cf. Sect. 2.4.4.[25] In listening-only tests, the labels are pre-annotated with numbers from 5 to 1, suggesting an interval scale level to the listener (this scale is referred to as *listening-only scale* in ITU-T Rec. P.800 1996). The labels are post-annotated in conversation tests, being the reason that this scale is said to be of ordinal level (*conversation quality scale*). In both cases, however, the per-condition mean, denoted by the *Mean Opinion Score (MOS)* $\in [1; 5]$, is calculated.[26]

Alternative category rating methods given in ITU-T Rec. P.800 (1996) focus on different quality aspects, both absolute and relative in the form of paired comparisons. They include the *listening-effort scale*, the *loudness preference scale*, the *degradation category rating scale*, and the *comparison category rating scale* for listening-only tests (see Möller 2000, pp. 52–55, for details and other, non-standardized scales), and the *difficulty scale* for conversation tests (see Möller 2000, pp. 59–60, for details). Note that paired comparisons are said to offer a higher sensitivity because the "differential sensitivity of the ear is much higher than the absolute sensitivity" (Möller 2000, p. 49).

Quality of the speech:

excellent	good	fair	poor	bad
5	4	3	2	1

Fig. 2.6 Listening-quality scale according to ITU-T Rec. P.800 (1996), taken from Möller (2000, Fig. 4.1)

[25] If the personal and external modifying factors can be assumed to be fixed, for example for a particular auditory experiment, the expectation $\boldsymbol{r}_0$ can be assumed to be fixed as well. For this specific setting, quality can be regarded as absolute.

[26] In ITU-T Rec. P.800.1 (2006), a terminology is presented in order to avoid ambiguities between *MOS* values obtained in different types of tests, as well as from different instrumental models. It is refrained from using this terminology due to simplicity reasons: Listening-only tests are exclusively considered in this book.

In ITU-T Rec. P.835 (2003), a subjective testing method is described for the particular case of conditions including noise and noise reduction (NR) algorithms, cf. Sect. 2.2.2. Depending on the level of NR, either the noise can be reduced by only a small amount, leaving the speech signal unaffected, or the noise is suppressed completely, however, leading to an adverse impact of the speech signal due to imperfect NR. This, in turn, might lead to confusion of the listeners and thus inter-individual (or even intra-individual) differences in the ratings because it might be unclear whether they should base their quality judgment upon the (potentially degraded) foreground speech signal, or the fact that background noise is absent. ITU-T Rec. P.835 (2003) was developed to steer the attention of the test participant in order to rate three aspects on separate scales: The speech signal alone, the background noise alone, and the overall effect, that is, speech and noise on three separate ACR scales.

Analytic-type tests, that is, multidimensional analysis, aim at scaling one or more feature descriptions $\{\beta_m\}$, the components of the vector $\boldsymbol{\beta}$ of the perceptual event description. Depending on the mediacy of the judgments with respect to the target scale, this can be achieved in different ways.

If information on the features of the auditory event are desired, these features can be represented by meaningfully labeled *attribute scales* and be scaled in a *direct* way, where the attributes verbally describe the features to be judged. That is, the scale values $\{\beta_m\}$ represent the target scores of the test. A prominent example of attribute scaling is the Semantic Differential (SD) technique developed by Osgood et al. (1957) using a set of continuous bipolar scales with antonym labels at each end. The technique was deployed in many fields of psychological research.

The Diagnostic Acceptability Measure (DAM), see Voiers (1977) and Quackenbush et al. (1988, pp. 67–82), is an example for attribute scaling intended for application to speech communication systems. In DAM, 19 unipolar scales with different attributes are presented to the test participants, where each scale is labeled with numbers 0–100, synonym attributes, and "negligible" and "extreme" at both ends (Quackenbush et al. 1988, Fig. 3.1). Moreover, separate scales are used for assessing the speech signal and the background. The scales are condensed into 10 "parametric" scales (ignoring the total quality scales), see Quackenbush et al. (1988, Table 3.1).[27] In addition, four "metametric" and "isometric" scales closely related to integral quality are presented (Quackenbush et al. 1988, p. 77). Highly trained subjects are necessary for the application of DAM (Quackenbush et al. 1988, pp. 78–79).

The components of $\boldsymbol{\beta}$ represent the *perceptual dimensions* of the listener's perceptual space if the components are orthogonal to each other, see Sect. 2.3.3. Orthogonal components, that is, dimensions, are mostly *indirectly* obtained by two independent paradigms:

- Principal component analysis (PCA) or other types of factor analysis of attribute scales, or

[27] The term "parametric" here is used as an indicator that the 10 scales are parameters of quality, "each [measuring] one aspect [...] of composite acceptability". Thus, these parametric scales, except for the two scales directly related to quality, are here understood in a very similar way as the *quality features*, see Sect. 2.3.4.

- multidimensional scaling (MDS) of pairwise similarity (PS), for example, dissimilarity δ or other proximity data.[28]

With PCA, the correlating attribute scales used, for example in an SD experiment, can be subsumed to principal components, reflecting the perceptual dimensions of the perceptual space of the listener. With MDS, on the other hand, the perceptual space is derived by converting the PS of pairs of stimuli into distances between points represented in that space. The two contrary paradigms of attribute scaling and pairwise similarity scaling are to some degree complementary in the sense that specific drawbacks can be counterbalanced by advantages of the respective other method. For example, PCA of SD data allows an easy interpretation of the resulting space due to the pre-defined attributes, whereas it cannot be guaranteed that all possible features are covered by this type of test. The PS paradigm, on the other hand, does not require a pre-definition of attributes. However, the interpretation of the results provided by MDS might be more difficult. The procedures of PCA and MDS are described in more detail in Sect. 3.2.

Note that the listener scheme in Fig. 2.5 does not exactly represent the paradigm of PS, however, the processes involved in judging the PS are akin to the processes aiming at judging the quality. Instead of the measurement of quality, the listener is asked to provide a scalar description δ_0 in terms of the similarity of two stimuli (sound events) presented subsequently. The stimulus (sound event) presented first causes an auditory event $\boldsymbol{w}_0(t_1)$ at time t_1. This auditory event can be thought as being stored inside the listener. That is, for the given task of comparing two stimuli in terms of their similarity, the long-term internal reference $\boldsymbol{r}_0$ is replaced by a short-term reference $\boldsymbol{w}_0(t_1)$ containing the auditory nature of the first stimulus. Correspondingly, the stimulus presented second causes a second auditory event $\boldsymbol{w}_0(t_2)$ at time t_2. Just as for the quality assessment, both events are represented by points in the auditory space and are fed to the "comparing system", which eventually leads to the dissimilarity description event δ_0, see Fig. 2.5.

It is one of the research goals of the present work to develop a *direct* scaling method for the subjective measurement of perceptual dimensions, see Sect. 2.7 and Chap. 4. The main benefits of such a method are discussed in detail in Chap. 4.

The choice of a suitable measurement method (including, e.g., the measurement scale) is to a high degree dependent on the measurement task. Hence, there is no optimal method that serves all requirements. An overview, in-depth discussion, and new results for speech assessment, including more rarely used methods such as performance tests, user surveys, usability evaluation, and the assessment of cost-related factors can be found in Möller (2000, pp. 47–88).

[28] Note that the proximity data itself can be obtained *indirectly* by subjective tests, see Tsogo et al. (2000), for example.

2.4.4 Personal and External Modifying Factors and Some Countermeasures

Every measurement, be it physical or psycho-physical, is only of limited accuracy and reproducibility (Blauert 1997, p. 9). Due to the difficulty in handling the interdependencies between varying physical events and their scaled versions, several simplifying assumptions are made in the following. Clearly, it can be assumed that a sound event $\boldsymbol{s}_0$ and the corresponding scaled version $\boldsymbol{s}$ can be both accurately produced and reproduced (Blauert 1997, p. 10). Thus, any invariants in the psycho-physical measurement process are to a great extent inherent to the listener, the *measuring organ*.

Blauert (1997) further assumes that the "describing system", that is, the "psycho-physical measurement instrument" (cf. Blauert 1997, Fig. 1.3), are invariant between and within subjects (achieved, e.g., by proper instructions), while the "perceiving system" and thus the resulting auditory event $\boldsymbol{w}_0$ is assumed to be variable and responsible for measurement errors. In contrast, Möller et al. (2010) argue that it can reasonably be assumed that the "perceiving system" rather than the "describing system" is stable and time-invariant, although perception might be influenced by attention, for example, and might be different across listeners (cf. Sect. 2.4.6).

The simplified assumption that perception (rather than description) is a more or less invariant process is followed here as well, since it can be argued that the description process is dependent on the underlying scale and other experimental circumstances as described in this section. Measurement errors can thus be attributed to the place where the measurement of the auditory event $\boldsymbol{w}_0$ or the quality event q_0 actually takes place, namely inside the "psycho-physical measuring instrument", that is, the "describing system" of the listener who acts as the measuring organ. That is, the scaling functions $\{f_{\beta,m}\}$ and f_b as well as the components $\{g_{\beta,m}\}$ and g_b of the "describing systems" characterize the bias, see Sect. 2.4.2. Moreover, since q_0 depends on the expectation $\boldsymbol{r}_0$ of the listener, see Sect. 2.3.4, b depends on $\boldsymbol{r}_0$ as well.

Psycho-acoustic measurement errors can be manifold and both of systematic (bias) and random nature. Random errors are mostly caused by human measuring inconsistencies, both within and between listeners. Given that a central tendency of the individual score distribution exists, random errors can easily be ruled out by averaging, which is assumed to be possible in the following. Bias, that is, systematic errors, subsumes all consistent and repeatable deviations of the scaled values $\{\beta_m\}$ and b from theoretical "psychological" scale values $\{w_m\}$ and q, respectively, see Sect. 2.4.2. Due to the fact that the true scale values $\{w_m\}$ and q are hidden, however, it is often difficult to identify bias in scale values. Moreover, different kinds of bias might occur simultaneously (Zieliński et al. 1998).

Personal as well as *external* modifying factors have an influence on the final scale values, as depicted in Fig. 2.5.

The *momentary state* of several *personal factors* of the listener (also referred to as *user factors*, see Möller 2005) is defined by influences such as the general experience (with telephony) or affinity to audio, the motivation of the call, the attitude (towards

the communication system), or the emotional state. Personal factors are usually ruled out also by taking a number of listeners into consideration that are naïve, that is, they are non-experts, representative (sample of the telephone-user population), and have normal hearing capabilities. It is often argued that experts or highly trained listeners can produce more analytical results (Möller 2000, p. 50), however, these listeners are not representative in terms of the population and more difficult to obtain. Therefore, the results are biased as well.[29]

Moreover, the subjective measurement task is influenced by *external modifying factors*. Such factors can be acoustic or non-acoustic. In VoIP, a trade-off might be chosen between the service quality, negotiable per call, and the price of a connection, such that the price is directly related to the user expectation (Raake 2006, p. 99–102). An important group of external modifying factors depends on the *context*: The mobility advantage in mobile telephony, or more general the "advantage of access", leads to a reduced expectation in terms of quality for experienced users compared to wireline telephony (e.g., Möller 2000, pp. 137–141). Due to the "advantage of access", in contrast to the VoIP scenario, users are often willing to pay more than for a standard telephone call (Möller 2000, pp. 141–145), despite of the worse quality. Both acoustic as well as non-acoustic properties of the terminal equipment play a role for expectation,[30] or the environment (e.g., whether or not it is noisy).

Moreover, in laboratory experiments, in contrast to user surveys (cf. Möller 2000, pp. 61–63), each test participant is rating a larger number of conditions due to efficiency reasons. That is, the *test context* is a further (external) modifying factor in this case.

As Allnatt (1983, p. 11) states, "the most important factor affecting opinion rating is undoubtedly the choice of the rating scale in terms of which the subject is to form his opinion". Test context effects arise since a context-dependent relationship between the stimuli and the scale has to be established. The participant's rule how to "assign numbers to objects" (cf. Sect. 2.4.1) usually is created with the first stimulus and consistently applied for the rest of the experiment. This has a direct influence on the *distribution* of the scale values (e.g., bias regarding the spacing, the centering, or the contraction of the scaled values, see Zieliński et al. 1998). Category rating shows some specific bias, in particular absolute category rating (Möller 2000, pp. 68–72), due to factors such as language-dependent interpretation of the meaning of the scale labels, personal use, intervals between scale labels not necessarily being perceptually equidistant (bias due to perceptually nonlinear scale, see Zieliński et al. 1998), low sensitivity (due to a low number of categories), avoidance of extreme categories (contraction bias, Zieliński et al. 1998), or saturation of extreme categories. Some of the disadvantages of ACR scales can be avoided, for example, by means of a continuous scale (higher sensitivity) with extended extreme positions (avoidance of

[29] It is a part of this book to develop an analytic test method for non-expert listeners (see Sect. 2.7 and Chap. 4).

[30] A systematic investigation on the "psychological role" of different user interfaces for speech transmission quality assessment as a function of the NB and WB channel bandwidth is presented Raake 2006, pp. 192–197.

contraction bias and saturation effects), see Fig. 2.7 (cf. ITU-T Contribution COM 12-39 2009).

In order to avoid the participants' sole orientation by means of the scale, a preceding training session with some stimuli reflecting the range of conditions to be expected in the test can be used for familiarization. Hence, the finding of the "scaling rule" is facilitated for the participant and leads to a more balanced usage of the scale and less unwanted variability in the ratings. These effects are referred to as "anchoring" (Möller 2000, p. 117). The range of conditions should be sufficiently large and the stimuli should reflect realistic conditions (Möller 2000, p. 52).

Several reference conditions can be presented reflecting the overall range to be expected. In past telephony experiments, these references often consisted of modulated-noise reference unit (MNRU) conditions (ITU-T Rec. P.810 1996), with different signal-to-correlated-noise-ratios Q. However, this type of distortion does not reflect the variety of distortions encountered in modern telephony such as the perceptual effects modern low-bitrate codecs provoke, and thus might itself introduce scaling difficulties for the test participants. In other words, not only the kind of features of the auditory events is different for MNRU distortions, but also the number of features (the dimensionality) is lower than the number of features of the auditory events caused by more recent speech technology. In fact, Möller (2000, pp. 121–129) found in direct and indirect comparison tests that allegedly equivalent MNRU settings remarkably differ from other distortion types in subjective terms, concluding that similar-sounding reference conditions should be used for comparison and transformation.

Including a diverse set of reference conditions in a training sessions is particularly useful to ensure that the complete perceptual space is stimulated, that is, that all relevant perceptual features are covered. A perceptually balanced stimulus inventory in general leads to a balanced scale usage and ensures that no perceptual effect is artificially emphasized (such as MNRU distortions, see above) or even disregarded. In analytical tests where it is the measurement aim to *identify* relevant perceptual dimensions (as it is one of the research aims of the present study, see Chap. 3), the inclusion of all practically relevant stimuli is therefore inevitable.

Another important reason for using the same or a similar set of reference conditions in different tests is that the test results can be averaged across these tests. This way, test-specific context factors can be "averaged out", for example, if the tests were conducted with a different group of participants, in different laboratory environments, or in different languages. From such average values, in turn, reference values might be derived that to some extent are freed from personal and external

Fig. 2.7 Continuous rating scale according to Bodden and Jekosch (1996) and ITU-T Rec. P. 851 (2003), German version, taken from Möller (2000, Fig. 4.10). English translations: "extremely bad", "bad", "poor", "fair", "good", "excellent", "ideal"

modifying factors. An example constitute the equipment impairment factors listed in ITU-T Rec. G.113 (2007), reflecting average quality indices for different codecs. The procedure to derive such values is described in ITU-T Rec. P.833 (2001).[31] If no reference conditions are available, linear transformations on the *MOS* scale are common if the subjective ratings obtained in one test do not cover the whole scale range. Then, it can be decided to transform the values such that the "best" condition match the upper scale boundary, for example (Raake 2006, pp. 252–254).

Borg (1982) presents a "category ratio scale", the so-called CR-10 scale, that aims at achieving absolute quality ratings combined with the advantage of direct scaling provided by category scales, cf. Sect. 2.4.3.1. In Möller (2000, pp. 147–155), an attempt was undertaken to rate quality directly on ratio level. In fact, it could be shown that the ratings obtained on the CR-10 scale are linearly related to the transmission rating scale of the E-model, assuming a NB context (cf. Sect. 2.6.3). In Sect. 2.4.6, considerations are presented how a universal continuum can be obtained that represents quality *without* the assumption of a specific context.

In modern telephony, one important context factor is the transmission bandwidth. The decades-long exposure to PSTN-type NB telephony led to a NB expectation by the users, which will probably change with the advent of WB (and even SWB and FB) speech transmission enabled by VoIP. If beyond-NB speech is expected by the listener, NB speech is perceived with a lower quality than traditionally (see, e.g., Raake 2006, pp. 175–203 or Wältermann et al. 2010d). In subjective experiments, the *immediate* expectation can be adjusted by the maximum available transmission bandwidth used for the stimuli, and thus can be regarded as a test context factor. In general, the transmission bandwidth leads to equalizing or contraction bias and an overall shift of the scores (e.g., scale values for NB conditions are compressed and shifted towards the lower scale range in a WB context), cf. Zieliński et al. (1998).

The bandwidth context can be conceived to be directly attributed to the expectation $\boldsymbol{r}_0$ in utilitarian tests. However, the bandwidth context influences the scale values $\boldsymbol{\beta}$ in analytical tests as well (see Chaps. 4 and 5), which is attributed to the "describing system". Hence, the separation of factors influencing the internal reference and the "describing systems" is to some extent hypothetical, since only the scaled values are accessible. A clear separation between the adjustment of expectation and the scaling process itself, however, is hardly possible (since both are not directly accessible). Zieliński et al. (1998), for example, remark that "it might be possible that the listeners undertake [the judgment and mapping task] together as one task" and that "there might be a significant overlap between mapping and judgments".

Other context effects specific to auditory tests can be partly ruled out and are partly inherent to the experiment. The terminal equipment and presentation method (e.g., monaural vs. binaural) are usually fixed (as long as they are not subject to

[31] In the past, test results from different experiments were recommended to be transformed according to the "equivalent Q-method", using MNRU stimuli as reference conditions and a normalization procedure based on a fixed relation between *MOS* and Q of these references, see Möller (2000, pp. 123–129) for details. This method is not recommended today due to the perceptual inappropriateness of MNRU distortions compared to the distortions introduced by low-bitrate codecs, see discussion above.

assessment themselves) or the test "modality" (a listening-test is more artificial than, e.g., a conversation test; listeners focus more on the "surface structure" of the speech and not on the content, see Möller 2000, p. 52). Other factors can be ruled out more conveniently, at least partly. Examples include dependencies on speech material (such as speakers and content; usually, several different female and male speakers and different sentences are used). *Order effects* might occur due to interdependencies of judgments made for consecutive stimuli (Möller 2000, pp. 116–117). In order to reduce the impact of such effects, the stimuli are presented in randomized order per listener and the corresponding ratings are averaged over the participants.

In the ITU-T Handbook on Telephonometry (1992) and ITU-T Rec. P.800 (1996), recommendations for common experimental conditions are provided that should be satisfied in order to achieve a high accuracy and reproducibility, and thus highly comparable results across laboratories. Further suggestions with respect to standardized test procedures are given in Bech and Zacharov (2006, pp. 97–105) and Zieliński et al. (1998), for example. Möller (2000, pp. 47–88), provides a thorough discussion of test context factors. See Bech and Zacharov (2006, pp. 105–141) and Möller (2000, pp. 50–51) for considerations on the group of test participants. For a thorough overview and further studies of response-modifying factors specifically with respect to quality expectation refer to Möller (2000, pp. 133–144) (focus on differences in expectations and thus quality depending on the user group and trade-offs between costs and quality) and Raake (2006, pp. 99–102 and pp. 190–203) (especially focussing on VoIP-enabled technology such as WB speech transmission and new types of user interfaces). More information on measuring bias, in particular with regard to the scaling process, can be found in Bech and Zacharov (2006, pp. 84–96) and Zieliński et al. (1998).

As long as the measurement fulfills the criteria of *validity*, *reliability*, and *objectivity*, the measurement can be regarded as meaningful (Möller 2010, p. 21): The validity describes "how well [the measurement process] measures what it should measure" (Möller 2000, p. 153). Reliability is closely related to accuracy and reproducibility, that is, the stability of a measurement when the measurement is repeated. A distinction can be made between different types of reliability. For example, the inter-test reliability compares the results of different tests, whereas the test-retest reliability compares the results of repeated measurements with the same sample of subjects. Finally, a measurement can be regarded as objective if it is independent of the observer. More in-depth considerations on measurement criteria can be found, for example, in Jekosch (2005b).

2.4.5 Scale Transformation

One particular scale bias is discussed in this section in more detail, because its compensation will be important in subsequent parts and leads to beneficial properties, see Sect. 2.5.3.

The rating scales most often employed in subjective tests aiming at the assessment of the quality of a service, see Sect. 2.4.3.2, can be considered as being finite at

both ends. Along with the finiteness of these rating scales, an empirically recurring phenomenon can be observed: As the peak of a distribution of ratings approaches either of the scale ends, the distribution becomes skewed and narrow due to the "piling up" of ratings against a boundary (saturation of extreme categories, see Sect. 2.4.4 and Zieliński et al. 1998). However, Thurstone (1927b) (see also, e.g., Torgerson 1958, pp. 156–158) postulated that scores on a psychological continuum should form normal distributions. This is linked to the idea that *just noticeable differences* in the magnitude of a stimulus yield equal intervals on the scale (whereas it was criticized in related literature that equal intervals along an (artificial) scale do not necessarily correspond to equal subjective intervals, e.g., Allnatt 1983, p. 24). This requires, however, the continuum to be infinite in both directions, which can be achieved by transforming the values from the finite scale to an infinite continuum by means of a sigmoidal relation. Allnatt proposes the *logistic on logistic transform (LOLT)* (Allnatt 1983, pp. 26–32), yielding normal per-condition distributions (see also Weaver 1959). The core function transforming the subjective rating scale t onto the new continuum T is a simple logistic function, cf. Eq. (A.1)[32]:

$$t = \frac{1}{1 + \mathrm{e}^{-T}}. \tag{2.5}$$

It is $0 < t < 1$ and $-\infty < T < \infty$, see discussion in Appendix A.[33]

However, the transformation problem can also regarded from a different perspective: It is argued in Allnatt (1983, p. 23) that the *apparent magnitude* continuum, that is the scale arising from ratio scaling (cf. Sect. 2.4.3.2), is bounded by zero at one end, and is infinite at the other, taking account of the fact that the participant "can envisage an unlimited range of possible apparent magnitudes, extending from zero effectively to infinity". Allnatt (1975) and Allnatt (1983, pp. 32–40) argue that these scales can be related by a sigmoidal, more specifically, *log*-logistic function.

Allnatt indirectly provides evidence by linking (a) the known empirical relation between (physical) stimulus magnitude d and qualitative rating t, and (b) the relation between stimulus magnitude d and the apparent magnitude N. Allnatt states that, at least for television impairments, the psychometric relation between stimulus magnitude d and subjective rating t empirically follows a log-logistic function, cf. Eq. (A.2)[34]:

---

[32] General properties of the logistic as well as the log-logistic curves are described in Appendix A and, e.g, in Allnatt (1983, pp. 6–8).

[33] Note that the variable t in Eq. (2.5) corresponds to the scaled quality b in the context of the present work. In the remainder of this section, the nomenclature used in Allnatt (1975) and Allnatt (1983) is used.

[34] Note that Stevens and Galanter (1957) compared several data sets obtained both by ratio scaling and by category scaling. For the "prothetic" type of concepts (i.e., stimuli that change in perceptual intensity such as loudness), they showed that the relationship between the two scales is non-linear, and sometimes the category ratings are linearly related to the *logarithm* of the values obtained on the ratio scales, that is, apparent magnitude. As Allnatt (1975) argues, however, it is likely that such a relationship only applies over a limited range since conceptually there is no definite limit to magnitude, both with regard to physics and perception.

$$t = \frac{1}{1 + \left(\frac{d}{d_M}\right)^{-G}} \,. \tag{2.6}$$

Here, d_M is the mid-opinion value d for which $t = 1/2$. Note that with $G > 0$, the slope of the curve is positive, which is assumed here (see Allnatt 1975). Thus, t has the meaning of "negative quality" (impairment) instead of "quality", as t increases with increasing stimulus magnitude. It is $0 \leq t < 1$ and $0 \leq d < \infty$. Compared to the logistic function in Eq. (2.5), the log-logistic function in Eq. (2.6) is determined for $d = 0$ where it takes the value $t = 0$. See Appendix A for a discussion of other properties.

According to Stevens and Galanter (1957), the apparent magnitude on a ratio scale N is a power function of the stimulus magnitude d (Steven's Power Law):

$$N = a \cdot d^b \,, \tag{2.7}$$

where a is a normalizing constant for the stimulus magnitude, and the exponent b characterizes the particular kind of stimulus.

Solving Eq. (2.7) for d and substituting d in Eq. (2.6), it follows

$$t = \frac{1}{1 + \left(\frac{N}{a \cdot d_M^b}\right)^{-G/b}} \,. \tag{2.8}$$

This equation can be simplified by choosing the constant $a \cdot d_M$ in such a way that $N = 1$ when $t = 1/2$:

$$t = \frac{1}{1 + N^{-k}} \,, \tag{2.9}$$

with $k = G/b$.

As can be seen, the log-logistic form is preserved, and Eq. (2.9) represents the relation between apparent magnitudes N on ratio scales and finite rating scales, denoted by t. Experiments show that k usually takes values $k \in [0.7; 1.5]$, that is, values around unity (see Allnatt 1975 and Allnatt 1983).

Regardless from which perspective the scale transformation is approached, apparently some kind of sigmoidal transformation curve is necessary to counterbalance the bias introduced by a finite rating scale. This way, it is assumed that a scale rating b or β_m better approximates the true underlying quality event q_0 or the perceptual feature value $w_{0,m}$. The transformation according to Eq. (2.5) leads to an infinite continuum, whereas the transformation according to Eq. (2.9) leads to a continuum bounded at one end, namely at an apparent magnitude of zero and thus zero impairment (optimal quality). It is argued in Allnatt (1983, pp. 23–26) that the underlying continuum of a finite rating scale, however, is finite as well, because if the assumption of a direct

physical analog pertains, the rating scale would otherwise be infinite as well, which practically cannot be possible. Thus, the log-logistic curve seems to be preferable due to its definite zero for zero impairment, see Appendix A. However, in practice (e.g., for a specific auditory experiment), there certainly also exists a maximum impairment.

Note that the particular bias presented here can be partly ruled out by the scale design depicted in Fig. 2.7 as well. In ITU-T Contribution COM 12-39 (2009), this extended and continuous scale was experimentally compared to the standard ACR scale (ITU-T Rec. P.800 1996). It was found that an empirical S-shaped transformation rule exists for a mapping between the scales, indicating that with the extended scale range, the saturation at the ends is less extreme.

2.4.6 Towards a Universal Continuum for Perceptual Value

As discussed in Sect. 2.4.4, even psychometric methods aiming at absolute ratings are never really absolute in a universal sense due to personal and external modifying factors such as context effects. Thus, "[s]peech quality measurement results do not generally have any absolute value, but are always to a certain extent relative and specific" (Jekosch 2005b, p. 89).

A truly absolute scale, which is inter-individually valid and independent of the measurement context (i.e., where personal and external modifying factors are excluded), is highly desirable because results from different sources such as different laboratories can directly be compared on the common continuum. This is beneficial for communicating test results to non-experts, for example, to plausibly explain why one system is better than another by means of comparing two one-dimensional values. Context-independency is also advantageous for instrumental measures as the quality estimations should be as general as possible. In fact, the theoretical considerations in this section are applied in the E-model, a tool for network planning, see Sect. 2.6.3.

As discussed in Sect. 2.4.4, it is assumed that the perceptual event $\boldsymbol{w}_0$ is stable across time, that is, independent of the context (Möller et al. 2010). Hence, this "perceptual level" is the appropriate stage in the listener schematic depicted in Fig. 2.5 to assume such an absolute scale. The quality event q_0 results from a comparison to the internal reference $\boldsymbol{r}_0$, and thus necessarily depends on personal and external response modifying factors. As context-free quality does not exist per definition, see Sect. 2.3.4, the one-dimensional scale value "can be considered as an index reflecting the *perceptual value* of a particular characteristic of the perceptual event [$\boldsymbol{w}_0$] with respect to the quality event [q_0]" (Möller et al. 2010). As the perceptual event is multidimensional, the one-dimensional index is a context-independent aggregation of the different perceptual features or dimension values $\{w_{0,m}\}$ according to a function h_w mapping the features from $\boldsymbol{w}_0$ onto the one-dimensional continuum w:

$$w = h_w\left(\boldsymbol{w}_0\right) = h_w\left(w_{0,1}, w_{0,2}, \ldots w_{0,M}\right) \ . \tag{2.10}$$

How the continuum of perceptual value w can practically be obtained is addressed later in Sect. 2.6.3. It can, however, be noted that the scale level of w should at least be interval level, cf. Sect. 2.4.3.1, rather than solely ordinal level (information on the rank order): Equidistant steps on this scale hence should correspond to equidistant variation in perceptual value. The assignment of numbers, however, is arbitrary. It is also noteworthy that since the scale is reference-free, neither the minimal nor the maximal value is known. If this was the case, the statement would be equal to knowing what the absolutely best quality is ("maximal expectation"), regardless of personal and external factors, or what the absolutely worst quality is. Ultimately, if w was bounded, it would not be reference-free (which, however, is required). The continuum w is therefore probably open at both ends; its range is undefined. A scale where no absolute zero is defined cannot be a ratio scale, see Sect. 2.4.3.1. Therefore, the universal absolute scale for "perceptual value" might be an interval scale with undefined, that is, open boundaries. Note that as soon as expectation comes into play, a maximal quality (or zero impairment) indeed can be defined, namely if the perceptual composition $\boldsymbol{w}_0$ matches the desired composition $\boldsymbol{r}_0$. As discussed in Sect. 2.4.5, the practical rating scale as well as the underlying continuum is assumed to be finite.

2.5 Dimension-Based Quality Models

2.5.1 Principle

Considerations on how transmitted speech is perceived, how quality arises (Sect. 2.3), and how auditory measurement is practically done (Sect. 2.4) were presented with reference to the listener schematic depicted in Fig. 2.5.

According to its definition, see Sect. 2.3.4, quality q_0 is the result of an internal comparison between the perceptual event $\boldsymbol{w}_0$ and the expectation $\boldsymbol{r}_0$, and thus the output of the "comparing system". According to this definition, Eq. (2.2) was formulated. This equation constitutes the foundation of a new approach for predicting the mouth-to-ear speech transmission quality pursued in this book: Speech quality can be predicted on the basis of relevant quality dimensions. To this aim, the function g_q reflecting the "comparing system" of the listener schematic needs to be determined.

However, Eq. (2.2) implies that, in order to determine the unknown psychoacoustic function g_q, both the independent variables, $\boldsymbol{w}_0$ and $\boldsymbol{r}_0$, and the dependent variable, q_0, have to be accessible. However, neither of these variables is directly measurable. Nevertheless, there exist potentially biased versions $\boldsymbol{\beta}$ and b of $\boldsymbol{w}_0$ and q_0, respectively, as obtained by auditory measurement. The emergence of measurement bias by personal and external modifying factors was explained in Sect. 2.4.4. Since the "perceiving system" is assumed to be independent of such factors, Eq. (2.3) and (2.4) state that the occurrence of bias can be attributed to the scaling functions $\{f_{\beta,m}\}$ and f_b, as well as to the "describing systems".

In Sects. 2.4.4 and 2.4.5, strategies were presented in order to reduce known bias effects (e.g., due to the employed rating scale). By such measures, the scaling functions and the influence of the "describing systems" can be assumed—at least partly—to be inverted.

In the following, b' and $\boldsymbol{\beta}'$ denote the values obtained from auditory experiments, where some of the bias effects were partly counterbalanced, for example, by applying normalization procedures and scale transformation of the raw values b and $\boldsymbol{\beta}$. Thus, it is *assumed* that $b' \approx q_0$ and $\boldsymbol{\beta}' \approx \boldsymbol{w}_0$, and it can be written:

$$b' = g_q\left(\boldsymbol{\beta}', \boldsymbol{r}_0\right) \ . \tag{2.11}$$

That is, the sought function g_q can principally be determined based on ratings from subjective tests. As it cannot be guaranteed that b' and $\boldsymbol{\beta}'$ are completely free of bias, the function g_q might be biased in practice.

Given that the features $\{\beta'_m\}$, that is, the elements of $\boldsymbol{\beta}'$, are orthogonal and all relevant for integral quality, they can be regarded as *quality dimensions*, see Sect. 2.3.4. A realization of the function g_q then is a *dimension-based quality model* as it will be derived in Chap. 5.

In the following two sections, modeling approaches from literature are discussed that can be considered as potential candidates for the model function g_q. Note that the original nomenclatures are used in the following.

2.5.2 Vector Model and Ideal-Point Model

The structure of the mapping function g_q, though being a functional relation, can be chosen on the basis of conceptual assumptions regarding the relation between the perceptual dimensions and integral quality. Carroll (1972) provides a useful framework for this purpose, a *linear-quadratic hierarchy* of four models for g_q, the mapping process itself referred to as *external preference mapping*.[35] The emphasis of Carroll's considerations is to account for individual differences in the models. Although the modeling of individual differences is not the aim of the present work as personal factors are assumed to be eliminated by averaging over participants, see Sect. 2.4.4, the models are equally well applicable to an "average listener".

In the following, a given set of points arranged in a multidimensional space is assumed. The points represent stimuli presented to listeners in a test. The configuration can be the result of an MDS experiment, for example (see Sect. 2.4.3.2). The simplest realization of the mapping function g_q towards integral quality is a linear one. Geometrically, this relationship can be conceived as a *vector model*: A quality vector resides in the space, pointing towards optimum quality. Integral quality is

[35] The term "external" means that analysis of preference takes place in relation to a given set of a-priori determined dimensions. In contrast, *internal* preference mapping is entirely based on a set of preference data (see, e.g., Mattila 2001 for an application example).

then monotonically related to the *projection* of a point onto this vector. Hence, the cosines of the angles between the vector and the coordinate axes measure the (relative) importance of a dimension with regard to integral quality. In algebraic terms, these "importances" are represented by the coefficients of a linear combination of the dimensions. The model can be interpreted as relating the dimensions towards quality in a "the more the better–the less the worse" fashion (if integral quality is positively related to the vector): The higher the dimensions' values, the better the quality (and vice-versa). Thus, the vector model concept fulfills the assumptions made for a continuum for perceptual value, see Sect. 2.4.6. However, also the vector model is of theoretical nature. At least an increase of quality *ad infinitum* cannot be conceived in practice (Carroll 1972) (cf. the considerations on the finite nature of rating scales in Sect. 2.4.5). In the vector model, it is possible to determine separate vectors for each individual in order to visualize individual differences in terms of preference, assuming a common *group* space that all individuals share (Carroll 1972). Alternatively, a single vector can be used to reflect the *average* individual.

The function g_q can alternatively be assumed to follow an *unfolding model* (Carroll 1972; in the present work, this model variant is referred to as *ideal-point model*): Integral quality here is *inversely* monotonically related to the *distance* between a stimulus point and the ideal point, assuming a metric on the space, which can, for example, be Euclidean. The ideal point directly corresponds to the expectation $\boldsymbol{r}_0$, see Sect. 2.3.4. *Iso-quality contours* can be described by concentric circles, spheres, or hyper-spheres, depending on the dimensionality. In a more general model version, the dimensions are allowed to be weighted in distance calculation according to their importance (see above). Hence, iso-quality contours can more generally be described by ellipses, ellipsoids, or hyper-ellipsoids. The model can further be generalized to allow for a rotation (in combination with the weighting) of the whole point configuration. Several ideal points might reflect the *points of view* of individual subjects (Carroll 1972). In addition, the two generalized model variants then allow for individual weights (assuming that distinct individuals weight the dimensions differently) or individual rotations (assuming that distinct individuals base their preference mapping on different sets of dimensions). As for the vector model, it is also possible to reflect an average individual by a single ideal point.

The vector model is a special case of the ideal-point model, with an ideal point lying far from the stimulus points in the direction of the quality vector. It is intuitively obvious that in this case, the rank order of distances from the ideal point is asymptotically identical to that of the projections of stimulus points onto a vector. Iso-quality contours in the region of the stimulus points can be described by straight lines perpendicular to the vector (Carroll 1972).

The vector model, the ideal-point model, and the two generalizations of the ideal-point model are included in the common hierarchical framework (in inverse rank order, starting with the most general Model I and ending with the most specialized Model IV, the vector model). Detailed mathematical derivations can be found in Carroll (1972). Note that the original nomenclature is used to describe these models.

Let $\boldsymbol{X}$ be an $n \times r$ matrix storing the coordinates of the n stimuli in r (perceptual) dimensions (the number of dimensions is denoted by M in this work). $\boldsymbol{X}$ can, for

example, contain the point configuration of a group space derived from MDS. A single point is represented by the row vector $\boldsymbol{X}_j$, with $j = 1, \ldots, n$ (in this work, such a row vector corresponds to the scaled version of the perceptual event, that is, $\boldsymbol{\beta}$ resulting from the jth sound event). Moreover, let $\boldsymbol{S}$ be an $m \times n$ matrix containing preference scale values for m individuals. The elements of $\boldsymbol{S}$ are s_{ij} with $i = 1, \ldots, m$ and $j = 1, \ldots, n$, as above (according to the nomenclature used in the present work, s_{ij} corresponds to a scaled quality event b of an individual test participant i and the jth sound event). Assuming an ideal-point model, the distance d_{ij} between $\boldsymbol{X}_j$ and the ideal point $\boldsymbol{Y}_i$ (in the present work reflected by the expectation $\boldsymbol{r}_0$) is monotonically related to s_{ij}. In Carroll (1972), the stronger assumption is made that the square of the weighted Euclidean distance is linearly related to s_{ij}, that is, $s_{ij} \approx a_i d_{ij}^2 + b_i$, where a_i and b_i are arbitrary constants and $a_i \geq 0$.

The squared distance can be expressed in summation or matrix notation. Note that $\boldsymbol{X}_j^* = \boldsymbol{X}_j \boldsymbol{T}_i$ and $\boldsymbol{Y}_i^* = \boldsymbol{Y}_i \boldsymbol{T}_i$ denote transformed values of $\boldsymbol{X}_j$ and $\boldsymbol{Y}_i$ according to an orthogonal transformation matrix $\boldsymbol{T}_i$, respectively:

$$d_{ij}^2 = \sum_{t=1}^{r} w_{it} \left(x_{jt}^* - y_{it}^* \right)^2 = \left(\boldsymbol{X}_j^* - \boldsymbol{Y}_i^* \right) \boldsymbol{W}_i \left(\boldsymbol{X}_j^* - \boldsymbol{Y}_i^* \right)^T . \tag{2.12}$$

Here, $t = 1, 2, \ldots, r$ denotes the dimension (in the present work's nomenclature, a perceptual dimension is denoted by m). The matrix $\boldsymbol{W}_i$ is a diagonal matrix, where the diagonal elements are the weights w_{it}.

In matrix notation, Model I results in

$$s_{ij} \approx \boldsymbol{X}_j \boldsymbol{R}_i \boldsymbol{X}_j^T - 2 \boldsymbol{Y}_i \boldsymbol{R}_i \boldsymbol{X}_j^T + c_i , \tag{2.13}$$

where $\boldsymbol{R}_i = a_i \boldsymbol{T}_i \boldsymbol{W}_i \boldsymbol{T}_i^T$ represents an individual rotation matrix containing both information on the orthogonal transformation as well as the weights. The parameter c_i is an arbitrary individual constant. With $\boldsymbol{B}_i = -2 \boldsymbol{Y}_i \boldsymbol{R}_i$, Eq. (2.13) can be rewritten in summation notation as

$$s_{ij} \approx \sum_{t} \sum_{t'} r_{tt'}^i (x_{jt} x_{jt'}) + \sum_{t} b_{it} x_{jt} + c_i \tag{2.14}$$

where $r_{tt'}$ denotes the elements of $\boldsymbol{R}_i$, b_{it} the elements of $\boldsymbol{B}_i$, and x_{jt} the elements of $\boldsymbol{X}$. From this notation, it can be seen that the model contains quadratic and linear terms, and, in particular, interaction terms.

In Model II, the more constrained ideal-point model and a special case of Model I, the transformation matrix R_i in Eq. (2.13) is replaced by the diagonal matrix $\boldsymbol{W}_i$ containing the weights w_{it}, that is, $\boldsymbol{T}_i$ is the identity transformation $\forall i$. The abandonment of rotation let the interaction terms present in Eq. (2.14) vanish and thus the number of free parameters decrease:

$$s_{ij} \approx \sum_t w_{it} x_{jt}^2 + \sum_t b_{it} x_{jt} + c_i \ . \tag{2.15}$$

Model III is the most constrained ideal-point model. The model is very similar to Eq. (2.15), however, only weights $w_{it} \in \{-1, 1\}$ are allowed.

Finally, Model IV corresponds to the vector model and reflects only the linear part of Eq. (2.15), in summation notation:

$$s_{ij} \approx \sum_t b_{it} x_{jt} + c_i \ . \tag{2.16}$$

The model coefficients in Eqs. (2.14), (2.15), and (2.16) can be obtained by curve fitting. Models I–IV can all be seen as realizations of the "comparing system" g_q in Fig. 2.5 for an individual i and a stimulus j, with $s_{ij} = b'$ and $x_{ij} = \beta'_m$.

2.5.3 Combination of Impairments

Based on early findings of Lewis (Lewis and Allnatt 1965), Allnatt (1975) and Allnatt (1983, pp. 133–144) develop a simple formula for combining "different kinds" of impairments for video pictures. It turns out that the term $\left(\frac{d}{d_M}\right)^{-G}$ in Eq. (2.6) can be regarded as a total impairment J composed of "summable" impairment units J_r,[36] such that

$$J^\nu = \left(\frac{t}{1-t}\right)^\nu = \sum_r \left(\frac{d_r}{d_{Mr}}\right)^{-\nu G_r} = \sum_r J_r \ , \tag{2.17}$$

with J_r being the rth summable impairment and $J_r \geq 0$. Note that t again describes "negative quality", cf. Sect. 2.4.5. Allnatt found for his experiments that $\nu \equiv 1$ can be assumed, resulting in the *law of additivity of impairments*. The J-scale, hence, has at least interval scale level, see Sect. 2.4.3.1. Comparing Eq. (2.17) with Eq. (2.9), it can be shown that for $k\nu = 1$, it follows that summable impairment J_r and apparent magnitude N_r are one and the same thing (Allnatt 1975). Hence, the scale transformation presented in Sect. 2.4.5 does not only counteract the scaling bias of saturation of extreme scale values, but apparently also leads to the practical property of impairment additivity.

As it will be seen in Sect. 2.6.3, the simple law of additive impairments also holds in the E-model, a modern tool for quality prediction. There, impairments are denoted by I, and the psychophysical functional relations do not necessarily follow Steven's Power Law anymore.

[36] Allnatt distinguishes between I and J, depending on whether the variables represent subjective values gained from a discrete category scale, or from a continuous scale.

An important prerequisite for the law of additivity of impairments to hold is that the impairments need to be "unrelated" (Allnatt 1983, pp. 137–138), meaning that the impairment caused by two distortions of the same type (e.g., a cascade of two equal bandpass filters restricting the transmission bandwidth) is in general not equal to the sum of impairments each single distortion provokes (for the case of the cascade of two equal bandpass filters, the impairment is probably even equal to the impairment caused by only one of the bandpass filters).

However, no precise definition of "unrelatedness" could be provided by Allnatt. As mentioned in Allnatt (1975) (see also Möller 2000, p. 98), one possible solution is to define unrelated as being *orthogonal* and to think of each of the unrelated impairments being represented along a single orthogonal axis of a multidimensional space, as obtained by MDS, for example (see Sect. 2.4.3.2). In fact, the work presented in this book allows to verify this assumption. Moreover, Eq. (2.17) can be regarded as a realisation of the "comparing system" g_q in Fig. 2.5, where the total impairment J is a measure for q_0 and the impairments J_r measures for the perceptual dimension values $\{w_{0,m}\}$. Later in this book, an equation similar to Eq. (2.17) will be reanalyzed with a variable equivalent to ν being a free parameter.

2.6 Instrumental Quality Measurement

2.6.1 Introduction

The quality or quality features of transmitted speech can a-priori only be assessed in auditory tests, since only human listeners, acting as measuring organs, can make the quality event q_0 or the perceptual event $\boldsymbol{w}_0$ "visible" in form of scale values b and $\boldsymbol{\beta}$, respectively, see Sects. 2.3.3 and 2.3.4. However, such tests are expensive both with regard to time and money. Therefore, tremendous effort has been put into the development of tools in the past in order to replace auditory tests. These tools aim at measuring quality or quality features instrumentally. More precisely, quality scale values b or feature scale values $\{\beta_m\}$ are estimated, resulting in quality estimates $\widehat{b}$ or feature estimates $\{\widehat{\beta}_m\}$, respectively.

Measuring physical phenomena by instrumental means, for example, the values $\{s_n\}$ of a sound event $\boldsymbol{s}_0$, can usually be done with high accuracy and repeatability, see Sect. 2.4.4. In contrast, instrumental quality or quality feature predictors aim at estimating magnitudes determined by human listeners, the measuring organs, in auditory tests (Sect. 2.4) *by means* of measured physical feature values $\{s_n\}$. Hence, instrumental measurement methods mimic in one way or the other the response-forming process of listeners in a subjective test. In order to reduce any bias of the estimation, for example, due to test context effects, see Sect. 2.4.4, adjusted estimations $\widehat{b}'$ and $\widehat{\boldsymbol{\beta}}'$ are ultimately targeted and thus employed for training such models.

In any case, however, quality estimations are valid only for a defined context, as quality is always dependent on a reference, see Sect. 2.3.4. For example, a model

developed for the NB context does not provide valid results for a WB context. It will be shown in Sect. 2.6.3, however, that the concept of "perceptual value" introduced in Sect. 2.4.6 can practically be implemented by instrumental methods and thus leads to a universal context-independent scale.

Instrumental methods differ with respect to the mapping between physical features and the estimated quality (or quality feature) ratings, ranging from purely empirical mapping functions to complex models of the human auditory hearing system. In fact, the psycho-acoustic relationships often have turned out not to be a trivial one-to-one relation between quality and, for example, a basic signal-to-noise ratio. The development of instrumental models, thus, involves a certain amount of engineering effort and is a continuous process, due to the inherent context-dependencies and due to new kinds of degradations potentially evolving by introducing new kinds of speech technology (e.g., new codecs). Therefore, similar to the situation in psychometry, there does not exist a best model, but different models for different applications.

A particular paradigm for instrumental quality and quality feature measurement is based on the dimension-model introduced in Sect. 2.5, which itself is a result of the basic definition of quality provided by Jekosch (2005b), see Sect. 2.3 and Fig. 2.5 for visualization: If quality can be conceived to be based on a set of dimensions, also an *estimation* of quality should be obtained based on dimension *estimates*, given that a model g_q and the expectation $\boldsymbol{r}_0$ are known.[37] Correspondingly, Eq. (2.11) can be rewritten as:

$$\widehat{b'} = g_q\left(\widehat{\boldsymbol{\beta}'}, \boldsymbol{r}_0\right) . \tag{2.18}$$

In Heute et al. (2005), this principle is referred to as *diagnostic speech-quality measurement*. In their article, the authors revisit this relatively new approach by describing the state of the art, current problems, and proposals for improvement (see also Sect. 2.6.2).

Assuming g_q to be known from Sect. 2.5 and $\boldsymbol{r}_0$ known and fixed for a given context, the dimension-based instrumental models rely on the realization of the components $\{g_{w,m}\}$ of the "perceiving system" mapping physical features $\{s_n\}$ of the sound event $\boldsymbol{s}_0$ onto single dimensions $\{\beta'_m\}$ according to Eq. (2.1).[38] Note that according to the ideas explained so far, common instrumental models do not make explicit assumption of the listener schematic outlined in Fig. 2.5 and implement the relation between the sound event $\boldsymbol{s}_0$ and the quality estimation $\widehat{b}$ without the "detour" of quality feature estimates. However, two major advantages are expected from *diagnostic* models, see Heute et al. (2005) and Raake (2006, p. 42):

- It is assumed that the dimension-based approach rests on a more or less complete set of perceptual features of modern telephone connections that are described

[37] In the following, it is assumed that the expectation $\boldsymbol{r}_0$ is known. In a further generalization of the following equation, it can in principal be replaced by an estimate $\widehat{\boldsymbol{r}}_0$ of the expectation.

[38] Similar to Sect. 2.5, the perceptual event $\boldsymbol{w}_0$ is not directly available for practical modeling. Thus, the context-effects-reduced version $\boldsymbol{\beta}'$ is used instead.

by orthogonal dimensions. As argued in Sects. 2.4.4 and 2.4.6, the "perceiving system" and thus the perceptual features can be assumed to be context-independent. Hence, it can be expected that feature-based models can more easily be adapted to future degradations than approaches relying on integral quality alone (given that the features remain the same) by adjusting the mapping function g_q (e.g., in terms of the weights of a vector or ideal-point model) and/or the expectation $\boldsymbol{r}_0$, see Sect. 2.5. Therefore, it is assumed that such a measure will even be able to reliably judge unknown kinds of degradations current (signal-based) models have problems with (Heute et al. 2005).

- Besides predicting integral quality, the single dimension estimators provide perceptually adequate *diagnostic information* on the composition of the quality. This allows for identifying the potential source of a quality degradation for system developers or network providers, which a scalar estimate of integral quality does not provide.

According to the above considerations, instrumental quality or quality feature models can be classified according to the *output* they provide, either being conventionally an integral quality estimate $\widehat{b'}$ or being (in addition) diagnostic information $\{\widehat{\beta'}_m\}$. Moreover, instrumental quality or quality feature models can broadly be classified as *parametric* and *signal-based* models, according to the required *input* the estimation is based upon.[39] Signal-based models rely on concrete speech signals. Thus, it is possible to estimate quality or quality features for an individual sample (*per-sample* or *per-file* estimation). The estimation results can be averaged across samples emerged under the same conditions of the transmission path to *per-condition* estimates. In parametric models, parameters are used describing a specific condition (for example, a network setup), and thus the quality or quality feature estimations are only possible *per-condition*. Whereas per-file estimations can be more precise for individual samples, the per-condition estimations are of more general nature. *Hybrid* models combine signal-based and parametric approaches, often also including protocol information of packet-switched networks. Both signal-based and parametric models can be useful for monitoring purposes. See Raake (2006, pp. 46–48) for a more detailed discussion.

For an overview of recent standardized and non-standardized instrumental methods for speech quality estimation, and also for more details on protocol-information-based models and hybrid approaches that combine signal-based and parametric approaches, see Möller et al. (2011b). Selected signal-based and parametric instrumental models are presented in the following.

[39] It is noteworthy that signal-based models usually transform the input signal(s) to an internal representation, which itself can result in parameters. Thus, the term "parametric" model, though being established, is not very precise.

2.6.2 Signal-Based Instrumental Models

Signal-based instrumental models can use either the input and the output signal of a transmission system (full-reference models) or only the output signal (reference-free models) to determine a quality estimate. A further classification can be based on the fact whether or not the models are capable of taking not only electric but also acoustic effects into account in their estimation. Models capable of dealing with electric signals expect the measurement probes being at the electrical interface of the transducers, cf. Fig. 2.2, whereas models taking acoustic effects into account, are better suited for the mouth-to-ear situation by considering also, for example, the acoustic interface. Representative full-reference models recommended by the ITU-T are PESQ (Perceptual Evaluation of Speech Quality, ITU-T Rec. P.862 2001) for NB speech and WB-PESQ (ITU-T Rec. P.862.2 2007) (both electric), see also Rix et al. (2001), and its successor POLQA (Perceptual Objective Listening Quality Assessment, ITU-T Rec. P.863 2011) for up to SWB speech and covering most recent advances of speech technology (see Côté 2011, pp. 134–135 for details). For WB-PESQ, optimizations have been proposed in Côté (2011, pp. 87–104). An alternative model is TOSQA (Telecommunication Objective Speech-Quality Assessment), see ITU-T Contribution COM 12-34 (1997) and Berger (1998). ITU-T Rec. P.563 (2004) is an example for a reference-free model.

The mentioned models solely estimate integral quality according to the ACR listening-quality scale, see Sect. 2.4.3.2. To put it simply, the assumption of these models is that integral quality is proportional to a perceptually weighed distance between the reference (corresponding to a transformed version of the input signal in full-reference models or a synthesized reference in reference-free models) and a transformed version of the output signal.[40] However, this model principle does not correctly reflect the situation in the underlying listening test paradigm. Rather than presenting explicit references against which the transmitted signal should be judged in a paired comparison fashion (as it is assumed in full-reference models), the listeners compare single transmitted signals with an internal reference, see Sect. 2.3.4. Thus, the comparison unit of these signal-based models is not a direct realization of the "comparing system" g_q, see Fig. 2.5. In turn, the "comparing system" is more closely mimicked by reference-free models. The "internal reference" of reference-free models is based on various models, for example, speech production models. Using the speech-production model, a reference signal is generated from the transmitted signal via LPC and additional parameters. In general, full-reference models are more accurate as they have more information for the estimation at disposal.

On the other hand, there are only few and not yet standardized models for diagnostic quality prediction as proposed in Heute et al. (2005) that are capable of predicting quality features, that is, $\{\beta'_m\}$. Examples include Quackenbush et al. (1988) as well as Halka and Heute (1992), see also Halka (1993).

[40] In full-reference models, this transformation usually includes signal preprocessing such as level- and time-alignment, a perceptual transformation modeling part of the peripheral human auditory system, and a comparison unit.

In Quackenbush et al. (1988), more or less basic signal measures like the global or segmental SNR or spectral distances showed only weak correlations with DAM auditory features (Voiers 1977), cf. Sect. 2.4.3.2. A combination of the fundamental measures and thus more complex measure performed better. A more recent study estimating a subset of the DAM attributes can be found in Sen (2004).

Halka and Heute (1992) (see also Halka 1993) found correlates of the dimensions identified in Bappert and Blauert (1994) on the basis of a decomposition of a system into a linear and a non-linear component according to Schüssler (1987). The peculiarity of their approach is to use a speech-like random process as an input signal rather than a reference speech signal as it is commonly used. Thus, this model is useful as a single-ended measure. The signal features that led to high correlations between $\{\widehat{\beta}_m\}$ and $\{\beta_m\}$ also served as independent variables for modeling overall quality, however, with several unexplained outliers (Heute et al. 2005).

Existing dimension-based or feature-based approaches were qualitatively analyzed in Heute et al. (2005) with the conclusion that the disadvantages of current dimension-based models are not seen in the quality modeling by means of the "detour" via attributes, but rather in the nature of the underlying attributes. In Halka and Heute (1992), for example, attributes like "clearness" and "naturalness" were used, which are more related to quality than to quality features. This principal issue is further discussed in Sect. 3.3 of Chap. 3.

Signal-based models that are founded on results of the work presented in this book are briefly described in Sect. 5.5 of Chap. 5. In the following, some related though not strictly dimension-based approaches are sketched.

Beerends et al. (2007) followed a technically driven approach for quality diagnostics that is not strictly based on perceptual dimensions. The authors derive three indicators based on PESQ (ITU-T Rec. P.862 2001) in order to display the causes of quality degradations in terms of "specialized" *MOS* values. These *MOS* indicators reflect the presence of additive noise, linear time-invariant frequency response degradations, and time-varying behavior such as packet loss and "clicks". Thus, they provide physical rather than perceptual diagnostics. Integral quality could be successfully estimated on the basis of the degradation indicators.

As mentioned in Sect. 2.4.3.2, quality assessment can be difficult for conditions affected by noise and different levels of noise reduction. A full-reference model that mimics the ITU-T Rec. P.835 (2003) paradigm is recommended by the ETSI in ETSI Guide EG 202 396-3 (2008). It is based on the "relative approach" proposed in Genuit (1996).

For a comprehensive overview of recent signal-based models, see Côté (2011, pp. 64–84), for example. A detailed overview of the above-mentioned and other diagnostic instrumental signal-based measures is provided in Scholz (2008, pp. 52–55) and Côté (2011, pp. 77–82).

2.6.3 The E-Model, a Parametric Instrumental Model

Parametric instrumental models are purely computational models. They base the quality estimation on a set of transmission parameters. As parametric models do not rely on the speech signals, their output holds true per condition rather than per sample. Due to their signal-independency, parametric models can be used for network planning in order to ensure user satisfaction and avoid network over-engineering. Nevertheless, the transmission parameters can in principle be also derived from signals (payload in VoIP) or from protocol information (e.g., RTCP, see Sect. 2.2.2), which makes parametric models attractive for monitoring. However, due to the per-condition estimation character, the individual per-call quality estimation accuracy is often not sufficient for practical use. A prominent example of a parametric model is the E-model, a network planning tool for the prediction of conversational and listening speech quality as gained from experiments according to ITU-T Rec. P.800 (1996).

The E-model has a relatively long history and is in fact an enhanced version of several different network-planning models for mouth-to-ear speech transmission. In particular, four models contributed to the core of the E-model. They were described in ITU-T Suppl. 3 to P-Series Rec. (1993): The Bellcore Transmission Rating model (BcTR, see also Cavanaugh et al. 1976), allowing to predict the quality in case of distortions due to circuit noise, overall loudness rating, talker echo, listener echo, attenuation distortion (including bandwidth), the SUBMOD/CATNAP model by British Telecom (cf. Richards 1974, see also Möller 2000, pp. 91–94), the Information Index from France Télécom, and the OPINE model developed by NTT (see also Osaka and Kakehi 1986 and Osaka et al. 1992). A first version of the E-model (the capital E being the first letter of ETSI, the European Telecommunications Standard Institute) was compiled in ETSI Technical Report ETR 250 (1996), see also Johannesson (1997). Major contributions to the E-model came from the BcTR model (ETSI Technical Report ETR 250 1996, p. 89), see ETSI Technical Report ETR 250 (1996, p. 107) and ITU-T Contribution COM 12-37 (1997) for a full list of sources. The BcTR model is based on purely empirical modeling, whereas the other models take also processes of human perception into account.

The E-model does not only consist of this core, which is based on well established models, but also new features that allow for covering "modern" distortions such as low-bitrate codecs or large absolute delays. In the past decade, several updates have been incorporated into the E-model in order to better take into account the effects of room noise at send side, quantizing distortion, and low talker sidetone levels. Moreover, the effects caused by random and (short-term) bursty packet loss for different codecs were included. These improvements are to a great extent based on research presented in Möller (2000) and Raake (2006) and several contributions directed to Question 8 of the ITU-T Study Group 12. The E-model is now recommended by the ITU-T for network planning, see ITU-T Rec. G.107 (2011).

The E-model has recently been adapted to also provide speech quality estimations for WB speech, including several WB speech codecs, packet loss, circuit and sent-side

noise, talker echo, and delay. The E-model version for WB is based on findings presented in Raake (2006), Möller et al. (2006), Raake et al. (2010), and ITU-T Contribution COM 12-278 (2011). These extensions to the traditional NB model are now incorporated in the new recommendation ITU-T Rec. G.107.1 (2011).

Both the E-model and the WB E-model are under continuous development in Question 8 of Study Group 12 of ITU-T. For example, in Wältermann et al. (2010d) and ITU-T Contribution COM 12-119 (2010), a first experiment is presented that enables the E-model to be extended towards SWB speech.

In the E-model, distortion types are categorized into so-called *Impairment Factors*, describing the amount of impairment from end-to-end of the transmission chain due to the basic signal-to-noise ratio (R_o), the signal-simultaneous distortions (I_s), and the delayed impairments (I_d), such as transmission delay or echo. Distortions originating from codecs are subsumed under the *Equipment Impairment Factor* I_e, which is extended to the *Effective Equipment Impairment Factor* $I_\mathrm{e,eff}$, taking packet loss effects into account. The impairment factors are calculated from parameters describing the end-to-end transmission system, such as loudness ratings, delay, noise levels, or packet loss rate, which were introduced in Sect. 2.2. An overview of these input parameters and computational details for impairment factor calculation are given in Appendix B. Due to the lack of a parametric description of the perceptual effects of the various types of codecs, values for I_e are given in tabulated form in ITU-T Rec. G.113 (2007).

The impairment factors are assumed to be additive on a psychological scale, the so-called *Transmission Rating Scale*, or R-scale, reflecting a scalar overall-quality estimate:

$$R = R_\mathrm{o} - I_\mathrm{s} - I_\mathrm{d} - I_\mathrm{e,eff} + A, \text{ with } R \in [0;\, R_\mathrm{o,max} + A]\,. \tag{2.19}$$

The factor A denotes the quality-advantage related to a given technology as perceived by the user and mainly serves as an explicit adjustment factor if the expectation of the user is shifted due to context effects other than the maximum bandwidth, see below. Both the factors I_e and A were introduced in ETSI Technical Report ETR 250 (1996) (see also Johannesson 1997).[41]

As mentioned in Sect. 2.4.4, one particular context effect in telephony is that of the transmission bandwidth. This effect is modeled by R_o in Eq. (2.19). It requires the R-scale to be extended beyond the traditional maximum of 100 if the bandwidth is extended beyond NB, corresponding to a shift of the reference toward "minus impairment". In other words, WB speech *exceeds* the NB expectation. This fact supports the existence of a "psychological continuum" that is theoretically open towards *both* ends,[42] see the more theoretical discussion in Sect. 2.4.6. As long as a transparent transmission can be assumed, $R_\mathrm{o} = R_\mathrm{o,max}$, where $R_\mathrm{o,max}$ depends on the

[41] Note that although R is defined for the range $R \in [0;\, R_{o,max} + A]$ in Eq. (2.19), the zero is arbitrary and not absolute in the sense of a ratio scale, cf. Sect. 2.4.3.1.

[42] Although for bandwidth distortions, the scale is practically bounded according to human hearing capabilities.

(test) context in terms of the maximum available bandwidth. In the E-model (ITU-T Rec. G.107 2011), $R_{\mathrm{o,max}}$ is set to $R_{\mathrm{o,max,NB}} \approx 95$ for NB conditions[43] (theoretically, $R_{\mathrm{o,max,NB}} = 100$ for a direct NB channel). For wideband transmission, it is set to $R_{\mathrm{o,max,WB}} = 129$ in the WB E-model (ITU-T Rec. G.107.1 2011). The scale extension for the WB case is described in Raake (2006, pp. 175–181) and Möller et al. (2006).

Linear distortions are not covered by the E-model, although they were part of the BcTR model (cf. ETSI Technical Report ETR 250 1996, p. 90). Also, the SUBMOD/CATNAP model does not provide a successful modeling of linear distortions (Möller 2000, pp. 176–183). Raake (2006, pp. 181–189) introduced the *Bandwidth Impairment Factor* I_{bw}, a novel description of linear distortions that fits both the E-model and the WB E-model, see Wältermann et al. (2008b),Wältermann and Raake (2008), and Raake et al. (2010). However, it is so far not included in either of the recommendations ITU-T Rec. G.107 (2011) and ITU-T Rec. G.107.1 (2011).

The "additivity property" of the E-model is borrowed from the OPINE model, see Johannesson (1997) and ETSI Technical Report ETR 250 (1996, p. 106), whereas the actual transmission rating scale (R-scale) was taken from Cavanaugh et al. (1976). There, it is stated that "psychological factors on the psychological scale are additive", which is based on early findings by Allnatt (cf. Sect. 2.5.3). Interestingly, the additivity property found in OPINE is based on Thurstone's assumption of normal rating distribution on the psychological scale (cf. the discussion in Sect. 2.4.5). Allnatt's additivity assumption, however, was based on a relation between finite rating scales with apparent magnitude ranging from zero to infinity, see Sects. 2.4.5, 2.5.3 and Allnatt (1975). Theoretically, Thurstone's and Allnatt's assumptions have different implications with regard to the transformation curve between the R-scale and the finite rating scale, cf. discussion in Sect. 2.4.5. In either case, the relation between subjective scores and a "psychological continuum" seems to follow some sort of sigmoidal curve (which can be logistic or log-logistic function, for example). In the E-model, the transformation function is a monotonic third-order polynomial (ITU-T Rec. G.107 2011, Annex B), see Eq. (B.31), relating the R-values to P.800 *MOS* values.

Besides the desirable additivity assumption of the transmission rating scale, another advantage of the R-scale is stressed in ITU-T Suppl. 3 to P-Series Rec. (1993, p. 2) and Cavanaugh et al. (1976): In order to provide *unique* ratings, the transmission rating scale should be *independent* of subjective factors modifying ratings in auditory tests, such as those discussed in Sect. 2.4.4. It is believed to better achieve such unique ratings by a *separation* of the problem into (a) a *unique* relation between factors describing the transmission characteristics and transmission rating and (b) the relation between transmission ratings and subjective scores.

In other words, the R-scale should be universal and of absolute character as defined in Sect. 2.4.6, thus free of any modifying or personal factors, see Sect. 2.4.4. This

[43] If all parameters are set to their default values and all impairment factors are taken into account, an R-value of $R = 93.2$ is obtained, corresponding to a standard ISDN connection (ITU-T Rec. G.107 2011).

is partly achieved by the sigmoidal scale transformation, which can be regarded to counterbalance some of the context effects stemming from the finite rating scale (saturation effect), see Sect. 2.4.5. Other context effects were already eliminated in the developing phase of the E-model: The E-model formulae and impairment factors were derived based on large amounts of data gathered from different laboratories over many years in order to achieve estimates averaged over samples (per-condition, see above), participants, and tests. Extensions or modifications of the E-model require a prior normalization of the new data in order to assure that existent formulae remain valid, see ITU-T Rec. P.833 (2001) and ITU-T Rec. P.833.1 (2009). Some context effects are reflected by the advantage factor A, for example, modeling the fact that higher quality ratings are commonly obtained in mobile telephony for a given network setup (cf. Sect. 2.4.4).

A closer look at Eq. (2.19) allows to bring the E-model variables into relation with the listener schematic in Fig. 2.5, the quality definition provided in Sect. 2.3.4, as well as the ideas presented in Möller et al. (2010) and Sect. 2.4.6.

Let I_{tot} be the *total impairment* with

$$I_{\text{tot}} = R_{\text{o,max}} - R \,. \tag{2.20}$$

$A \equiv 0$ is assumed in the following for simplicity. With Eq. (2.19), Eq. (2.20) corresponds directly to Allnatt's *law of additivity of impairments* given in Eq. (2.17), Sect. 2.5.3. It can be considered as a realization of the "comparing system" g_q in Fig. 2.5. Therefore, I_{tot} can be conceived to be inversely proportional to the quality event q_0 and is thus the output of the "comparing system". As such, I_{tot} is dependent on the internal reference $\boldsymbol{r}_0$, as quality per definition depends on the internal reference, cf. Sect. 2.3.4. As an example, for $R = R_{\text{o,max,NB}} = 95$, $I_{\text{tot}} = 0$ holds in case of a NB transmission bandwidth context, but $I_{\text{tot}} = 129 - 95 = 34$ in a WB transmission bandwidth context, due to the different definitions of $R_{\text{o,max}}$ in both cases.

According to Fig. 2.5, one of the two inputs to the "comparing system" are the elements of the perceptual event $\boldsymbol{w}_0$, which are, according to Möller et al. (2010), assumed to be independent of the expectation $\boldsymbol{r}_0$ and any other personal or external modifying factors, see Sect. 2.4.4. In Eq. (2.20), R is a context-independent variable, which can be considered as a model input, see the above example. Considering the fact that R, as opposed to the perceptual event $\boldsymbol{w}_0$, is a scalar value, R is an appropriate measure for the perceptual value w, see Sect. 2.4.6, as suggested in Möller et al. (2010).

As it was argued for the w-scale in Sect. 2.4.6, the R-scale is an interval scale with undefined boundaries. As R is independent on the expectation and any other context-factors, a definition of absolute boundaries for R is impossible as well. The expectation, however, comes into play for the impairment-scale: The term $R_{\text{o,max}}$ can be considered as a one-dimensional version of the expectation $\boldsymbol{r}_0$. Similar to w, it can be considered as an aggregation of different *desired* features $\{r_{0,m}\}$, cf. Sect. 2.4.6.

Due to Eq. (2.20), the total-impairment scale can be interpreted as an *inverted and rigidly shifted* version of the R-scale, dependent on the expectation $R_{\text{o,max}}$, whereas

the interval scale character remains unaffected (the scale steps are equal on both the R-scale and the impairment scale, cf. Sect. 2.4.3.1).

Further interpretations are possible, for example, with regard to Carroll's modeling terminology presented in Sect. 2.5.2, which will be elaborated in Chap. 5 (Sect. 5.2.3).

2.7 Research Topics Covered in this Book

As mentioned in Chap. 1, one of the aims of the present book is to provide the basis for instrumental quality prediction based on perceptual dimensions. As discussed in Sect. 2.6, the following advantages over current approaches are seen:

- Given that the set of dimensions underlying quality is completely known and valid also for future degradations, quality prediction can be adapted to different contexts by adjusting the mapping function and/or the model of the internal reference.
- Along with integral quality prediction, diagnostic information on the perceptual causes of the quality estimate is delivered.

This aim is based on the hypothesis that quality can be explained and modeled by means of perceptual dimensions according to the definitions given in Sect. 2.3.3 and Sect. 2.3.4, and the practical model presented in Sect. 2.5. The notion that estimates of these dimensions can be employed in order to estimate integral quality was explained in Sect. 2.6.1.

In order to reach the aim based on the hypothesis, the following *research topics* are addressed in the present book:

1. *Exploration*: Identification of the relevant quality dimensions; the exploration allows to determine the number M and the kind of components of the perceptual event $\boldsymbol{w}_0$ and the expectation $\boldsymbol{r}_0$; Chap. 3.
2. *Elicitation*: For the development of the models, training data is required. Therefore, based on the knowledge of the dimensions, a new efficient test method is developed that allows the identified dimensions to be quantified by $\{\beta_m\} = \{\beta_{dim}\}$ in a direct way; Chap. 4.
3. *Modeling*: Based on the new data, models for (a) the "comparing system" g_q and (b) the relevant components $\{g_{w,m}\} = \{g_{w,dim}\}$ of the "perceiving system" are derived, leading to a new parametric instrumental model that is based on perceptual dimensions; Chap. 5.

The present work systematically addresses each of these prerequisites and thus lays the foundation for dimension-based instrumental quality estimation. In particular with the exploration (Chap. 3) and elicitation (Chap. 4) parts of this book, a fundamental basis is set for developing both signal-based and parametric diagnostic instrumental measures that estimate quality.

The line of investigation represented by the research topics is supported by the thoughts mentioned in Jekosch (2005b, pp. 82–86): "finding out the measurand (the

perceptual dimension) [research topic 1] and the process to measure the value of the dimension [research topic 2] constitutes the *elements of quality of the measurements* [...]", and in addition, "[...] taking into consideration the known elements of quality of the speech signal [...]", that is, the physical phenomena (research topic 3).

The present book mainly focusses on the *listening* aspect of a telephone conversation, that is, the perception of *transmitted* speech, which is one of the three building blocks of *communication* efficiency, see Sect. 2.3.5.

Existing knowledge on dimension-based quality assessment and modeling is extended in multiple directions. The following novelties have so far not been addressed in related literature:

- Perceptual dimensions that are valid for new kinds of speech transmission technology (such as ABE, state of the art noise suppressors, WB codecs), but at the same time valid for traditional networks. For the identification of the dimensions, the underlying separation of features and quality according to the listener schematic in Fig. 2.5 is clearly followed.
- A new auditory test method for directly assessing orthogonal dimensions that is efficient and feasible for naïve listeners.
- A thorough investigation of and new insights into the "comparing system" g_q of the listeners where the perceptual dimensions are integrated to integral quality.
- A parametric model giving diagnostic information on perceptual dimensions.

http://www.springer.com/978-3-642-35018-4

Dimension-based Quality Modeling of Transmitted Speech

Wältermann, M.

2013, XII, 204 p., Hardcover
ISBN: 978-3-642-35018-4