

Chapter 2

Fundamentals

This book is about the quality assessment, diagnosis, and estimation of transmitted speech in a telephone conversation. This chapter will give insights into the main concepts, definitions, and approaches to introduce important fundamental knowledge of speech that is transmitted over a telephone system. After a brief description of the general research scenario in Sect. 2.1, the characteristics of speech production, transmission, and perception are given in Sect. 2.2. This detailed introduction leads to the definition of perceived quality of transmitted speech, its assessment methods, as well as analyzing and diagnosing concepts that are described in Sect. 2.3.

2.1 Research Scenario

A simple and very basic schema of the general research scenario that this book is about can be seen in Fig. 2.1. The illustration shows the classical concept of two-party human-to-human *speech communication*. Here, the term *communication* is defined as the “intentional transmission of information” [4]. Thus, two humans participating in the communication intend to exchange information. The information exchange in this research scenario is based on speech, while also written language or sign language could be considered at this point.

Speech communication between two participants requires the *transmission of speech* in both directions, thus speech communication is bidirectional. The transmission of speech usually starts at one participant, the *speaker* that produces speech, that is the source of information. The speech is then transmitted through a *transmission system*. Finally, the transmitted speech reaches the sink of an information, the second participant named the *listener* who perceives speech. This introduced speech communication with two participants that intentionally exchange informa-

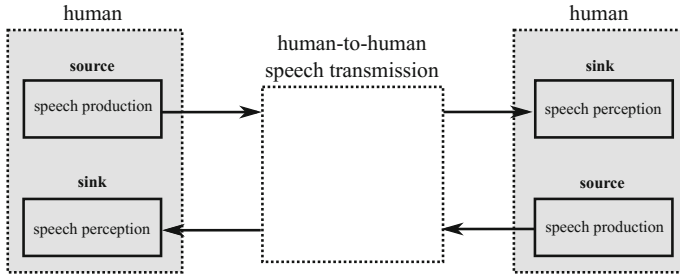


Fig. 2.1 Two-party speech communication based on [2]

tion and thus take turns in the role of listening and speaking is from now on called a *conversation* and the two participants are called *interlocutors*.

This schematic transmission of speech is often referred to as the so-called *speech chain* [5]. Starting from the brain of the speaker as the source of information and ending at the brain of the listener as the sink of an information, it describes the causal concatenation of physiological and physical processes of speech communication. Thus, the speech chain covers three fundamental processes:

- *Speech production* as the source of the information
- *Speech transmission* to exchange information
- *Speech perception* as the sink of information.

In the present book, the focus is on the perception of transmitted speech in a conversation in terms of its perceived *quality* (see Sect. 2.3). As a traditional transmission service, this book is targeting at *telephone communication* services. The only definite methods for testing and evaluating bidirectional communication services in terms of quality are conversational methods [6]. This is because only in a conversation all three processes of a the speech chain can be respected. In turn, the three processes are all important for the perceived quality in a conversation (see Chap. 3 and [7]). A telephone communication is composed of specific *elements* that might introduce *degradations* to the transmitted speech. For example, a corrupt microphone (element) might lead to an attenuated transmitted speech (see Sect. 2.2.2). The focus of this book, that is the quality perception of transmitted speech in a conversation, is explained by analyzing and optimizing a given transmission system by diagnosing (identifying elements that are responsible for degradations) the perceived speech quality (see Sect. 2.3.3).

In sum, the research scenario of this book is the bidirectional speech communication in a conversation over a telephone transmission system. To give more information about the aforementioned fundamentals, the three processes of the speech chain will be described in Sect. 2.2. In addition, Sect. 2.3 introduces and defines the quality of transmitted speech, its assessment methods, as well as approaches of analyzing the quality of transmitted speech in terms of diagnosis and optimization. Section 2.4 concludes the fundamentals and summarizes their relation to the present book.

2.2 Human Speech

Speech production and speech perception are two complex research fields that have been studied in different scientific area such as acoustics, physiology, and linguistics. This section will provide fundamental knowledge about the two processes in terms of basic concepts of acoustics, perception, and cognition. Following the time line of the speech chain, the first subsection is about the production of speech (see Sect. 2.2.1). Afterwards, the produced speech is transmitted through a transmission system. Section 2.2.2 will provide substantial information about telecommunication systems, as this book will concentrate on telephone speech transmission systems. Finally, the transmitted speech reaches the listener. The hearing process, the analysis of speech in the listener's brain, and consequently the speech perception are briefly introduced in Sect. 2.2.3.

2.2.1 Production

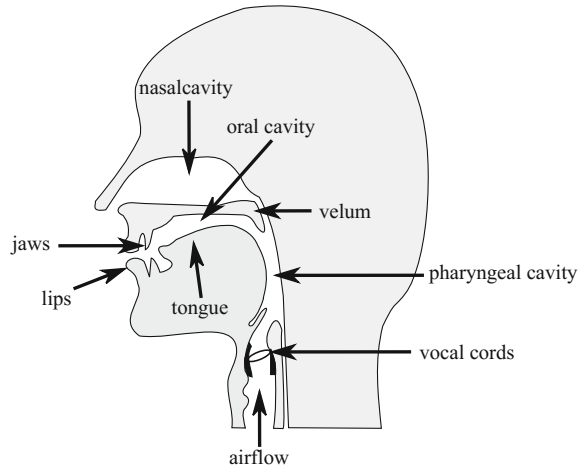
As described in Sect. 2.1, the first process of human-to-human communication is the speech production. In other words, no production of intelligible speech sounds, no conversation. The basic concept of generating a *speech sound* is built up in two consecutive events, the *excitation* and the *sound shaping* [8].

2.2.1.1 Excitation

During the excitation, from the lungs (serving as an “energy reservoir”) the chest musculature (serving as a “pump”) produces an airflow. This airflow is then led through the *glottis*. The *vocal cords* can either be opened (the airflow will pass without limitation), or they are closed at first and become opened by the airflow. This leads to two different “excitation types”:

- The first excitation type is called *periodic* excitation. As said before, first the vocal cords are closed. The airflow lets the air pressure at the vocal cords rise and at a certain pressure level, the vocal cords jump open. In turn, this leads to a fast drop in pressure, and the vocal cords close again. The repetition of these steps results in a quasi-periodic opening and closing of the vocal cords. The period of the opening and closing is defined with $T_0 = 1/f_0$. f_0 is called the *fundamental frequency* (also referred to as *pitch frequency*). The fundamental frequency depends on the length of the vocal cords. Hence, the fundamental frequencies between males (approximately 132 Hz), females (approximately 223 Hz), and children (300 Hz or higher) vary [9]. The fundamental frequency can change rapidly, however it is constant for about 20 ms. Therefore, periodically excited speech can also be seen as constant for 20 ms [10].

Fig. 2.2 Schematic illustration of the human vocal apparatus based on [8]



- The *aperiodic* excitation can be separated into two different modes. (I) If the vocal cords are already open, the airflow can pass through. This leads to turbulences behind the vocal cords (in the *vocal tract* (see Sect. 2.2.1.2)) that build a *noise-shaped* excitation signal. (II) The vocal cords are open, but the airflow is hold by a barrier (e.g. the lips) at a different point in the vocal tract. With raising air pressure this barrier is suddenly opened, it “explodes” or “bursts”. This results in a so-called *explosive* (or *plosive*) excitation signal [8].

The introduced extraction types can also be combined or adapted, for example when humans whisper (see [10]). At this point, whatever excitation signal is produced, no information are so far “printed” on the signal. For this, the signal at hand has to be shaped to carry information. This happens during the so-called *sound shaping*.

2.2.1.2 Sound Shaping

Responsible for the sound shaping is the *vocal tract*. The vocal tract consists of pharyngeal, oral, and nasal cavities. It is limited at the lower end by the vocal cords, and at the upper end by the mouth. The vocal tract serves as a resonator that introduces different energy levels in the frequency spectrum of the pressure wave. The maximum energy levels are called *formants* and are approximately equal over different speakers. The positions of the so-called *articulators* (lips, jaws, tongue, or velum) have a strong effect on the sound wave traveling through the vocal tract and therefore shape the speech sound. Together with the excitation (source), the vocal tract (filter) forms a so-called *source-filter-model* that produces and shapes a specific speech sound. A schematic illustration of the vocal apparatus can be seen in Fig. 2.2.

2.2.1.3 Speech Sound

As described before, the sound shaping, together with the excitation, characterize a *speech sound* for a short period of time. The basic components of speech production are referred to as *phones*. They are grouped in the abstract class called *phonemes*. Phonemes are the smallest distinctive unit of a language that do not carry a meaning in itself (e.g. /b/ in bit and /p/ in pit) [8]. They can be divided into two groups: the vowels and the consonants [11]. Vowel sounds (e.g. [i], [e], or [u]) are characterized by a periodic excitation (constant airflow through the vocal tract) and they can be defined by their formant frequencies (see, e.g. the formant maps for German vowels in [12]). Consonant sounds are mostly created with an aperiodic excitation and can for example be divided into plosive ([p] or [t]) or fricative ([s] or [f]) consonants. Formants are also important for consonants, but for these phonemes they often vary in time. Different phonemes are concatenated into phoneme sequences by a speaker in order to form the words to be expressed, in which the exchange of one phoneme leads to a change in the meaning (hat—fat).

The produced speech sound leaves the speaker at the output of the mouth. The spectrum of the produced human speech typically ranges between 100 and 7000 Hz [13]. The bandwidth is dependent on the phoneme that is produced. Vowels concentrate most energy between 300 and 3000 Hz, while fricatives have little energy below 7000 Hz. However, between 100 and 600 Hz (here the first formants as well as the fundamental frequency are included) the energy of human speech is usually the highest. For more detailed information and a broader overview about speech production, see for example [8] or [10].

2.2.2 Transmission

As described in Sect. 2.1, the second process of human-to-human communication is the speech transmission. Again, in other words, no speech transmission, no conversation. In Fig. 2.1, the speech transmission is represented as a “black-box”. This combines all possible “systems” that transmit speech from the mouth of the speaker to the ear of the listener. In a face-to-face conversation the transmission “system” is the pure acoustic sound field, the direct air-path from mouth to ear. However, in this book the focus is on telephone communication systems. In this context, speech telecommunication is realized by applying a *speech transmission system* that is replacing the natural air-path in a face-to-face conversation. A speech transmission system “follows the aim of supporting natural communicative possibilities involving people by means of technical devices and facilities” [14]. This section will give a brief overview of the basic concepts of speech transmission and its possible implementations.

2.2.2.1 Speech Transmission in a Telephone Communication System

In Sect. 2.1, Fig. 2.1 shows a very simple schema of speech communication. An (again very basic) example of a one-way (source to sink or speaker's mouth to listener's ear) speech transmission in a telephone communication system is displayed in Fig. 2.3. After the speech sound is produced (see Sect. 2.2.1), the acoustic signal $x(t)$ is received by the microphone of the speaker's handset. However, the acoustic signal might be disturbed by different acoustic signals $n(t)$ that are produced by sources surrounding the speaker. Thus, the microphone does not only receive the pure speech signal, but rather a "signal mixture" consisting of the speech signal $x(t)$ and the background noise $n(t)$.

Subsequently, the acoustic signal is converted into an electrical signal. This signal is digitized, that means that it is *sampled* and *quantized* in a form of $x(k)$ (k representing the sample index). In addition, the "signal mixture" is pre-processed to remove the undesired signals like background noise.

Finally, the digitized and pre-processed signal is compressed and sent through the *telephone network* (see Sect. 2.2.2.2) where it passes several gateways and nodes. After the signal has passed the network it arrives at the listener's side. Here, the digitized electric signal is synthesized by several post-processing steps, resulting in a digital signal $y(k)$. Then, the signal is again converted into an acoustic signal $y(t)$ that is played back at the speaker(s) of the listener's handset where it is presented to the listener (see Sect. 2.2.3).

In a conversation, speaker and listener change roles and this mouth-to-ear schema (Fig. 2.3) could be flipped. As it is described, it follows that multiple elements of the speech transmission could effect and possibly degrade the transmitted speech signal. For now, these elements are illustrated as "black-boxes" in Fig. 2.3 (Element A, B, and C). An overview of these elements is given in Chap. 3.

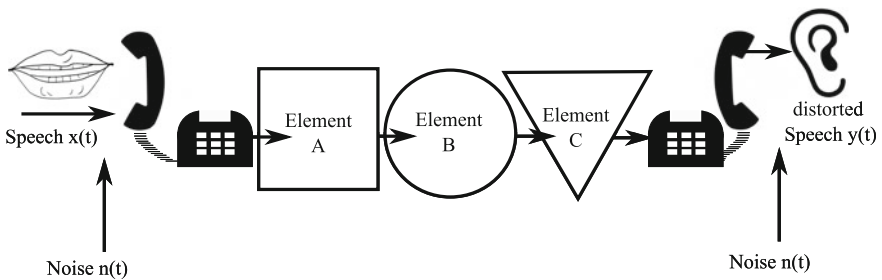


Fig. 2.3 Schema of one-way speech transmission in a telephone communication system based on [15]

2.2.2.2 Telephone Networks

At the very beginning of the evolution of analog telephone networks, two telephones were directly connected to each other. It was not possible to reach other telephones. Later, it was possible to manually connect two telephones. If one person wanted to reach another one, the operator was called and he manually set up the connection. This process was then automated which led to the public fixed-line telephone networks, also called *Public Switched Telephone Network* (PSTN).

In PSTN the bandwidth of transmitted speech corresponds to the transmission of frequencies between 300 and 3400 Hz. This bandwidth is referred to as *Narrowband* (NB). In PSTN the two interlocutors are connected by a physical circuit that is held for the conversation and duration of the call.

The next step of evolution was the introduction of digital transmission. The *Integrated Services Digital Network* (ISDN) was introduced and standardized in the mid-nineties [16]. ISDN replaced analog transmission, however, the digitally encoded information are still transmitted by an analog electrical signal.

Besides PSTN, also mobile phone networks were established over the world in the last decades. Other than in fixed-line telephony networks, the two interlocutors can move during their conversation. This, however, is dependent on the radio channel between the mobile phone and the antenna. Also, during movement the interlocutors may switch from one transmission channel to another one (referred to as *handover* see e.g. [17]). This and for example interferences produce errors and degradations (e.g. bit errors or frame drops) that are new in comparison to fixed-line networks. The global expansion of mobile phone networks began with the *Global System for Mobile Communications* (GSM) network [18]. The GSM network was first replaced by the third generation network (“3G”) called *Universal Mobile Telecommunications System* (UMTS) [19] and later by the fourth generation network (“4G”) called *Long Term Evolution* (LTE) [20].¹ The new networks provide a higher transmission bit-rate and thus enable new services as well as usage possibilities.

Apart from PSTN and mobile phone networks, computer networks like the *Internet* have been established. They are also known as *packet-switched* networks as they transmit packets of data. In the context of speech telephony, the system providers adapted the networks to transmit speech as data in packets of equal size (*frames*). Since speech is assumed to be stationary for a short period (see Sect. 2.2.1.1), for the digital speech signal, the source signal is chopped into frames of equal length. At this point, each sample in each frame is represented by *bits*. The *bitrate* specifies how many bits per second the network can transmit, it is also called the *network rate*. The packet transmission is handled by protocols like the *User Datagram Protocol* (UDP) or the *Real-time Transport Protocol* (RTP). Thus, in packet-switched networks the transmission of speech (or voice) is called *Voice over Internet Protocol* (VoIP).

VoIP services have been spread widely in the last decade and are nowadays one of the most widely used transmission paths. This is because of their flexibility and the higher audio bandwidths that can be transmitted. While PSTN is limited to NB, it

¹Note that LTE is a pure packet-switched network.

is possible to transmit speech with a bandwidth of 50–7000 Hz (referred to as *Wideband* (WB)), 50–14000 Hz (referred to as *Super-Wideband* (S-WB)), or 20–20000 Hz (referred to as *Fullband* (FB)). However, also packet-switched networks introduce new kinds of degradations in comparison to fixed-line networks. For example, the process of creating the packets could lead to a higher overall transmission delay that may introduce talker echoes or affect the interactivity of a conversation. In addition, packets could be lost in the transmission path, referred to as *packet-loss*, which leads to time-varying degradations and discontinuities. An overview of degradations introduced in packet-switched networks can be found in [21].

2.2.3 Perception

As described in Sect. 2.1, the third process of human-to-human communication is the speech perception. And yet again, in other words, no speech perception, no conversation. So, as described in Sect. 2.1, a speech signal transmitted through a transmission system finally arrives at the sink of an information, the speakers ear. The *sound reception* and the human hearing system have been researched in numerous studies. Detailed information can for example be found in [5, 8], or [10]. Basically, the human ear consists of three parts (see Fig. 2.4):

The *outer Ear* is composed of the *pinna*, the *ear canal*, and the *eardrum*. The sound wave is received by the pinna. The pinna supports the localization of the sound source and its funnel-shaped form serves to send the sound wave through the ear canal. At the end of the ear canal the eardrum is excited.

The *middle Ear* is a leverage that transmits the pressure from the eardrum on to the *oval window* of the *cochlea*. It is composed of three connected small bones that form the chain of ossicular bones, the *malleus*, *incus*, and *stapes*. Thus, these ossicles are an impedance-matching system that converts the acoustic sound wave into a mechanical wave.

The main component of the *inner Ear* is the cochlea. It is formed by three fluid-filled parallel and snail-shaped (2 and a half turns) chambers, where two of them are connected at their ends. The membrane separating these two chambers is called *basilar membrane*. The basilar membrane contains sensory hair cells connected with the *auditory nerve* system. Thus, the oval window excites the fluid in the cochlea and the hair cells convert the mechanical wave into electrical pulses. At this point, a spectral decomposition through a frequency-to-place transformation, which is of non-linear nature, is performed. Finally, the brain transforms the electrical pulses into a message that is interpretable by the listener.

The actual *speech perception* is taking place in the human brain on a psychological level. The study of relating physical acoustic events to auditory perceptual events is called *psycho-acoustics* [23]. An overview of psycho-acoustic research for noise and tones (including masking, thresholds, or localization) can be found in [24] or [25].

While much is known about how speech signals are converted into patterns of auditory nerve signals, much less is known about how the brain translates these

Fig. 2.4 Anatomy of the human outer, middle, and inner ear based on [3, 22]

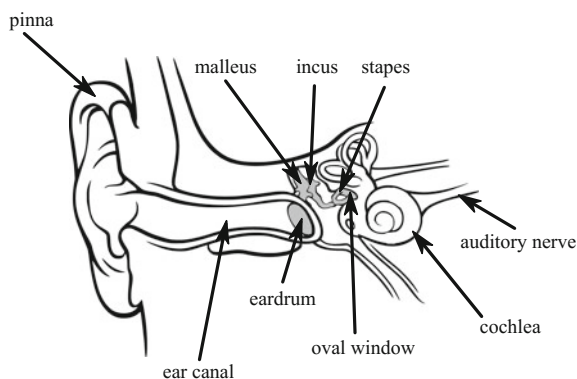
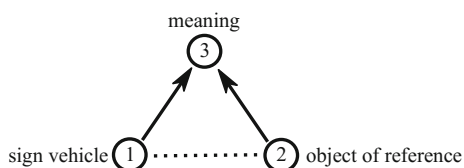


Fig. 2.5 Semiotic triangle based on [27, 28]



signals into an interpretable message [26]. For speech, it has to be considered that a speech signal does not “mean” anything by itself [14]. Thus, speech can be seen as a system of signs (something which can stand for something other than itself) that can be represented with a triangle (see Fig. 2.5) of three correlates in a semiotic (the science of signs) approach (following [27]). A sign can be modeled by a (I) sign vehicle, (II) its meaning, and (III) an object of reference [28]. Thus, the listener extracts the meaning from the sign vehicle (the form of a speech signal, e.g. sound wave) and the object the sign stands for.

In addition, an important part in speech communication is the *auditory memory* that gives the base for different theories about speech perception [10]. One of these theories implies that in the auditory process acoustic features are stored in the auditory memory. The interpretable message is then based on these acoustic features and stored again. Thus, different types of storage at different stages of perception are defined (see for example [29, 30], or [31]):

In the *Echoic Memory* only the important acoustic features for durations of 150–300ms are stored. Following, the features are transferred to the *Short-Term Memory* (STM) where the phonetic process appears. Here, the information is stored for 2–20s. If sounds are not directly processed in the STM they could be stored in the echoic memory for up to 3–4s (for example no speech sounds). The third component of the memory is the *Long-Term Memory* (LTM). Here, information is stored from a few days up to decades. The LTM allows people to recognize speaker and instruments.

Finally, in speech communication the target of the two interlocutors is that their messages are understood. This process of speech recognition is called *comprehension* and describes the perceptual analysis of the speech signal [14]. According to [6],

comprehension is the last stage in the speech perception process that is composed of four successive steps:

The *Comprehensibility* describes the capability of the speech signal to transport phonemic information (sign vehicle). A high comprehensibility indicates a perfect recognition of each phoneme from the speech signal.

The *Intelligibility* describes the capability to extract the content of the speech signal (object of reference).

The *Communicability* describes the capability to understand the speech signal (meaning).

The *Comprehension* describes the result of the speech perception process in terms of (a) achieved communication efficiency, and (b) understood messages.

2.3 Quality of Transmitted Speech

The previous section briefly introduced the processes of speech production, transmission, and perception. Obviously, a transmitted speech signal in a telephone conversation contains numerous information: for example linguistic (the actual message) or paralinguistic (identity or emotion of the speaker) information. Thus, for a speech transmission system, it is important to provide intelligibility and comprehension. In modern telephone transmission systems, the intelligibility and comprehension of a speech message is almost always given. However, the system user's perception in a telephone situation is different from a natural face-to-face conversation. Other factors, like annoyance due to an undesired sound additional to the desired speech sound for example, influence the user's perception of a particular telephone connection. Hence, speech intelligibility and comprehension are important, but not enough to fully quantify the user's perception of a speech transmission system [32]. In fact, besides intelligibility and comprehension a fundamental part of the speech transmission system is the user's perception of the *quality* of the transmitted speech signal. The quality of transmitted speech reflects the undesired degradations introduced by, e.g. the transmission system and perceived by the system users. In this section, the terms related to speech quality are introduced, providing the fundamental understanding and the scope of research this book is about.

2.3.1 Definition of Perceived Quality

According to [14],

perceived **quality** is “[the] result of [the] judgment of the perceived composition of an entity with respect to its desired composition.”

Here,

the **perceived composition** is “[the] totality of features of an entity.”

and

the **desired composition** is “[the] totality of features of individual expectations and/or relevant demands and/or social requirements.”

while

a **feature** is “[a] recognizable and nameable characteristic of an entity.”

In the context of *speech quality* the mentioned entity is the acoustic speech signal experienced by a human. According to [33],

experiencing “is the individual stream of perceptions (of feelings, sensory percepts and concepts) that occurs in a particular situation of reference.”

Thus, quality is a subjective value dependent on the human experiencing an entity. However, the quality formation process inside a human (or here a listener) is not a comparison of a desired and a perceived composition alone. In fact, according to [21] and based on the work in [14], the *speech quality formation process* can be divided into five successive steps (see Fig. 2.6) resulting in the *perceived quality* (see also [33]):

In the *perception* step, the speech signal is perceived by the listener, resulting in the perceived composition. In terms of speech quality, this is the perceived *auditory* composition composed of all perceptual aspects like the duration, loudness, or the phonetic information. The modifying factors are formed by the *personal* and the *context characteristics* [21]. Personal characteristics are for example the motivation, knowledge, experiences, or memory of the listener. The context characteristics describe the listener’s environment and the particular communication situation. Thus, the modifying factors form the desired *auditory* composition (also called *internal reference*) for a listening situation. Each human has his or her own desired composition and therefore different quality formation results are produced when the same speech sample is presented to two different listeners. The desired components are stored in the LTM (see Sect. 2.2.3) and are expected to change with training (see for example [34]).

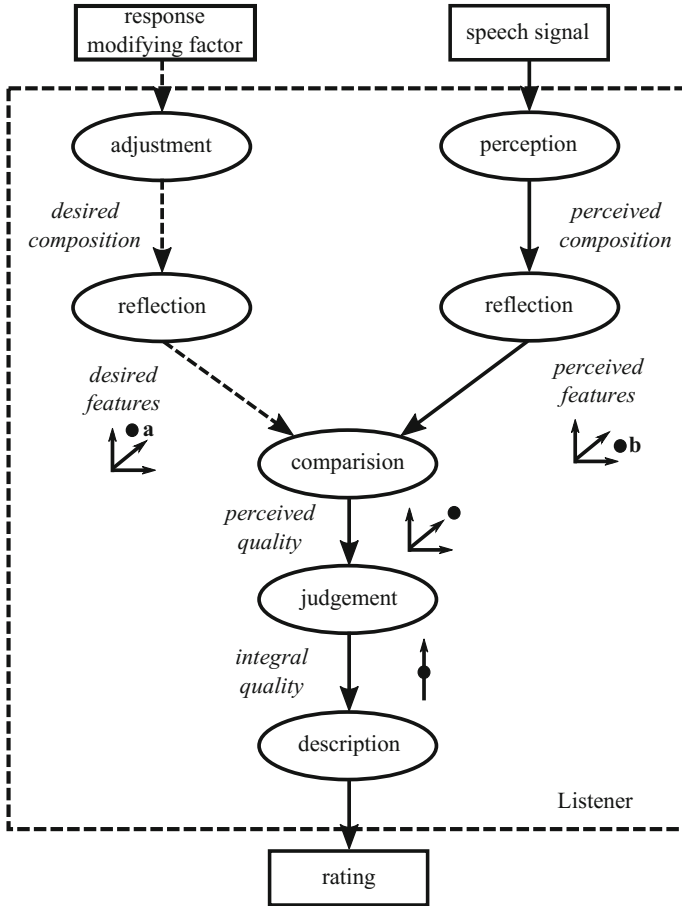


Fig. 2.6 Speech quality formation process based on [3] according to [14, 21]. *Ellipses* represent processes, *italicized names* storages and *rectangles* the inputs and outputs of the human/listener

During the *reflection*, the listener reflects the perceived signal characteristics and transforms them into a set of perceived features. These quality-related features define the perceived composition and form a specific position in the multidimensional space (**b** in Fig. 2.6). Accordingly, the desired characteristics are decomposed and the desired features are identified to form a specific position in the desired composition space (**a** in Fig. 2.6). Thus, for each perceived feature, there exists a desired feature.

The *comparision* step covers the required comparison between the desired and the perceived features **a** and **b**.

In the *judgment* step, the listener uses the comparison step to actually judge the perceived quality. Here, the features are associated into a single quality value by weighting each feature with its influence on the quality. At this point the single

quality value is called *integral quality* or *overall quality*² as it is composed of the different features (see also Sect. 2.3.2). Consequently, high quality values can only be formed if the desired and the perceived composition are as similar as possible.

In the final *description* step the listener has to describe the perceived quality. Typically, listeners are asked to do that on a rating scale. One famous rating scale is the 5-point rating scale defined in [35], where listeners choose between *bad*, *poor*, *fair*, *good*, and *excellent*. However, more information about subjective quality ratings can be found in Sect. 2.3.4 and Chap. 3.

In addition to the term perceived quality, the term *Quality of Experience* (QoE) was introduced in [36]. Based on [33], the definition of QoE was extended to:

Quality of Experience “is the degree of delight or annoyance of a person whose experiencing involves an application, service, or system. It results from the persons evaluation of the fulfillment of his or her expectations and needs with respect to the utility and/or enjoyment in the light of the persons context, personality and current state.”

Here, following [36]

an **application** is “[a] software and/or hardware that enables usage and interaction by a user for a given purpose. Such purpose may include entertainment or information retrieval, or other.”

Here, an application refers to an *Information and Communication Technology* (ICT) like a gaming, video, television, or communication service. In this book, the focus is on a telecommunication system and in this case QoE includes the effects of the complete *end-to-end* (speaker-to-listener and back—conversational situation) system (terminal, all system elements, and the network itself). Taking these definitions into account, the human experiencing an entity is not only passively describing the perceived quality (see step V of the quality formation process) but rather actively using an application. Thus, the person who is proactively using and interacting with a particular technology or an ICT product is “introjected” in the role of a *user* [37]. Regarding the topic of this book, an interlocutor in a conversation is therefore rating the QoE as an end-user of a telecommunication system.

This implies that a certain degree of *interaction* of a user is required and respected when talking about QoE. The introduced schema of the quality formation process is mainly targeting at a passive situation (*listening*, see Chap. 3). An extended version for the QoE formation process regarding ICTs in general can be found in [33]. A detailed introduction to interactive situations (conversations) is given in Sect. 3.5.

²From now on the term *overall quality* will be used.

As QoE is regarding quality from the perspective of the user, quality can also be regarded from the network perspective, referred to as *Quality of Service* (QoS). According to [38] QoS is defined as:

Quality of Service is “[the] totality of characteristics of a telecommunication service that bear on its ability to satisfy stated and imply needs of the user of the service.”

However, QoS covers technical parameters and the performance of a physical system. Thus, QoS is targeting at the technical quality that is different from the perceived quality as the context and human factors are not taken into account. An overview about the QoS taxonomy can be found in [6, 39].

2.3.2 *Quality Features and Quality Elements*

As described in Sect. 2.3.1, during the speech-quality formation process in the comparison and judgment steps the features of the desired and the perceived composition are compared and associated to form the single overall-quality value. From this it follows that the overall speech quality is a *multidimensional* value—it is composed of the perceived features that are the perceived characteristics of the speech signal such as the loudness or timbre. These features of the perceived composition that are obviously relevant for the overall quality are called *quality features*. Along the definition in Sect. 2.3.1 and according to [14] is a

quality feature “[a] recognizable and nameable characteristic of an entity that is relevant to the entity’s quality.”

In contrast, looking at the technical domain of speech transmission systems (see Sect. 2.2.2), specific elements of the transmission system can actually cause an impact on the quality perception. These elements of the speech transmission system (see Sect. 3.2) are referred to as *quality elements*. Following [14]

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.

Hence, quality elements are the physical counterparts to quality features: “While an element of quality 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” [14]. Quality elements and quality features serve to give a deeper insight into the single overall speech-quality value. Their relation and the concept behind this are given in the next section.

2.3.3 *Perceptual Quality Space and the Concept of Diagnosing Speech Quality*

As mentioned in the section before, the perceived quality is a multidimensional value that is formed by the comparison between the desired and the perceived quality features. Geometrically, it is illustrated in Fig. 2.6 that the perceived composition is a specific point (**b**) in a multidimensional space defined by the quality features. This multidimensional space is called *perceptual quality space* and is defined as follows:

The **perceptual quality space** is the geometrically multidimensional space formed by the quality features in which the perceived composition is located.

If the perceptual quality space is a Euclidean space with an orthogonal basis and each perceptual feature is lying along one of the orthogonal axes, and thus are themselves orthogonal (independent and not correlated), the quality features are called *perceptual quality dimensions*. Consequential,

perceptual quality dimensions are the orthogonal, and thus independent, dimensions of the perceptual quality space.

Looking again at Fig. 2.6, the desired composition (internal reference) (**a**) and the perceived composition (experienced speech signal) (**b**) are positioned in the perceptual quality space spanned by the perceptual quality dimensions. This implies that the overall quality Q can be modeled with a mapping function m_q on the basis of perceptual quality dimensions, such that:

$$Q = m_q(\mathbf{a}, \mathbf{b}). \quad (2.1)$$

If a perceptual quality space is defined (see Sects. 2.3.4 and 4) this equation determines that the overall speech quality can be assessed on the basis of relevant perceptual quality dimensions [2]. In that case, the dimensions are (quality) *underlying* perceptual dimensions. The mapping function that defines the dependency between

the overall quality and its underlying perceptual dimensions introduces the term *quality profile* that is defined as

the **quality profile** is the relation between the overall quality and its underlying perceptual dimensions.

Concerning the structure of the mapping function m_q , in [40] two possible approaches are introduced:

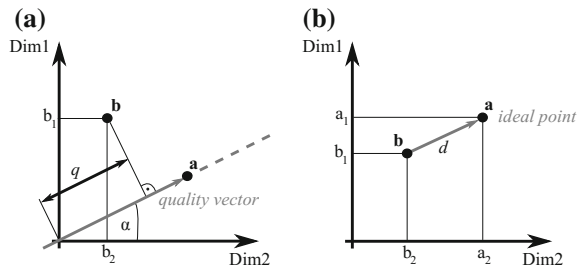
The first approach follows the realization of a mapping function towards the overall quality with a linear relation. In a geometrical sense, this approach can be seen as a *vector model*: A quality vector in the perceptual quality space is pointing at an optimum quality point (desired composition or internal reference). Now, the perceived composition forms a point in the same perceptual quality spaces and its projection on the optimal quality vector is considered as the overall quality. An example in a two-dimensional perceptual quality space can be seen in Fig. 2.7a: If $\mathbf{b} = (b_1, b_2)$ is the multidimensional representation of a speech signal, and the quality vector of a user has an orientation of α , the resulting quality Q is proportional to the projection q :

$$Q \propto q = b_2 \cos(\alpha) + b_1 \sin(\alpha) = \gamma_1 b_1 + \gamma_2 b_2. \quad (2.2)$$

Thus, in algebraic terms, the influence of each perceptual dimension with regard to the overall quality Q is represented by the weighting coefficients $(\gamma_{1,2})$ of a linear combination of the dimensions. In turn, this implies that the overall quality can be assessed from the linear combination of the different dimensions. Regarding the relation between dimensions and overall quality, the vector model can be seen in a “the higher the better—the lower the worse” sense—the better the dimensions, the higher the quality and vice versa. For each user of a system, a single quality vector can be determined; to cover an average user, a single vector that reflects the average individuals should be used [40].

The second approach is called the *ideal-point model*: A specific point in the perceptual quality space corresponds to the desired composition $\mathbf{a}$ (the internal reference

Fig. 2.7 Illustration of the vector model (a) and the ideal-point model (b) after [21, 40]



or here the *ideal point*). The perceived speech signal forms the perceived composition that is also located as a specific point in the same perceptual quality space $\mathbf{b}$. Both points can be described as vectors, such as:

$$\mathbf{a} = (a_1, \dots, a_{N_{Dim}}), \quad \mathbf{b} = (b_1, \dots, b_{N_{Dim}}). \quad (2.3)$$

Here, N_{Dim} gives the number of dimensions. Now, the overall quality Q of the speech signal is anti-proportional to the squared Euclidean distance d between the two vectors $\mathbf{a}$ and $\mathbf{b}$ (see Fig. 2.7b for a two-dimensional example):

$$Q \propto q = \sqrt{\sum_{i=1}^{N_{dim}} \gamma_i (a_i - b_i)^2}, \quad (2.4)$$

where a_i and b_i are the values of one of the N_{Dim} perceptual dimensions. The γ_i variable is again a weighting coefficient that corresponds to the influence of each dimension on the overall quality ($\gamma_i > 0$). Again, (2.4) implies that the overall quality can be assessed from the different dimensions. In difference to the vector model, the ideal point model cannot be seen in a “the higher the better” sense. In fact, degradations are introduced if the ideal point is not met in an ellipse form [40].

The introduced concepts allow to analyze the overall quality in terms of its underlying perceptual dimensions and thus the perceived quality features. As described in Sect. 2.3.2, the quality elements are the counterparts of the quality features. This means that, if a perceptual quality space is defined (see Sect. 2.3.4 and Chap. 4), the assessment of the perceptual dimensions allows to determine the quality elements that are potentially responsible for possibly low perceptual quality-dimension judgments. For providers of telecommunication systems, this process allows to identify weak points in their systems (that can be improved or eliminated) which introduces the concept of *diagnosing* speech quality. Traditionally (in medical science) and according to [41], diagnosis is defined as

[the] identification of the nature of an illness or other problem by examination of the symptoms.

Applied to the quality of transmitted speech, this definition modifies to the following:

Diagnosis is the identification of the quality elements of a telecommunication system that are potentially responsible for a drop of the overall speech quality (nature of an illness) by examination of the defined perceptual quality space (the symptoms).

Again, having diagnostic information at hand gives the telecommunication-system providers the possibility to analyze and improve their systems. To give an example: two different impaired speech signals (one degraded by, say, a level attenuation and the other one by background noise) may be judged with an equal overall speech-quality value. Now, having only the overall quality value at hand gives the telecommunication system providers no information about why the signals are impaired and how to improve the transmission system. Applying the introduced concept of diagnosing speech quality on the basis of perceptual dimensions, a defined perceptual quality space gives diagnostic information for the system providers. The signal degraded by a level attenuation gets lower ratings on a perceptual dimension that is connected to the “loudness” of a speech signal. This dimension in turn, is directly connected to a quality element that is responsible for the “loudness” (for example the gain control (see Sect. 3.2)). The same with the second speech signal, the background noise causes a lower rating of a perceptual dimension that is connected to the “noise” of a speech signal and that perceptual dimension is connected to a responsible quality element, here for example the noise reduction element (again, see Sect. 3.2).

The introduced concept of diagnosing the quality of transmitted speech is based on the assumption, or expectation, that perceptual features are one to one connected to specific quality elements. In theory, the concept promises to provide meaningful diagnostic information. However, in practice, the one to one connection between perceptual features and quality elements, or technical causes, is not always given. It is possible that one quality element is connected to more than one quality feature, and in turn, that one quality feature is connected to more than one quality element. Hence, diagnosing the quality of transmitted speech on the basis of perceptual dimensions might lead to ambiguous results. Thus, diagnosing the quality of transmitted speech on the basis of perceptual quality dimensions is only one possible approach to provide insights into the reason for a low quality rating. A second approach could be the identification of the technical causes of sub-optimum quality, in terms of characteristics of the signal or the transmission system that cause the lower quality rating. For this, the *International Telecommunication Union* (ITU) developed and proposes a specific test method for performing expert annotations after listening to transmitted speech files. The method is planned to be standardized in a future recommendation called *Technical Causes Analysis* (P.TCA) [42], and aims at identifying signal characteristics such as sub-optimum speech level, speech spectrum, noise level, echo, or alike. A detailed description of the proposed test method, first results, and possible

improvements can for example be found in [1, 43, 44]. However, in this book, the focus will exclusively be on diagnosing the quality of transmitted speech following the approach of perceptual quality dimensions.

Thus, to sum up, the concept of diagnosing speech quality with the perceptual quality space exhibits inherent advantages for telecommunication system providers:

- Identifying of quality elements responsible for a drop of the overall speech quality,
- Differentiating between equally rated speech signals,
- Analyzing and improving the telecommunication system on the basis of the perceptual quality space,
- Understanding the user's perception of speech quality,
- Robustness against the quick evolution of signal processing elements in the transmission system [45],
- Modeling the overall quality on the basis of its underlying perceptual dimensions.

However, before using the introduced diagnostic concept, the perceptual quality space has to be defined. The perceptual quality space can be defined in four successive steps: (i) Generating configurations of the transmission system that span the whole perceptual space under study, (ii) conducting an auditory test using *analytic* assessment methods, (iii) incorporating a *multidimensional analysis* of the test results (for (ii) and (iii) see Sect. 2.3.4.1), and (iv) identifying and interpreting the resulting perceptual dimensions.³ The next section gives an overview about the principles of quality assessment methods for the overall quality and also for analytic methods. Nevertheless, examples and standards for quality assessment methods are also introduced in Chap. 3.

2.3.4 Assessment Methods

In Sect. 2.3.1 and in Fig. 2.6, the speech quality formation process with its five successive steps was introduced. In the last step (the description) the listeners describe their quality perception on provided scales which leads to the output of the quality formation process, the quality rating. This “encoding” process [21] of assigning numbers to objects [46] is referred to as *assessing* the quality of a particular speech sample or a transmission system. In other words, speech quality assessment is the quantitative description of the perceived quality. Speech quality assessment methods can be divided into two groups: auditory (or subjective) methods and instrumental (or objective) methods. Both assessment methods are introduced in the following. Note that the following introductions are of general meaning, a differentiation for different contexts of a telephone conversation is given in Chap. 3.

³The identification of the perceptual space in a telephone conversation is described in detail in Chap. 4.

2.3.4.1 Auditory Methods

According to [14],

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 these features”.

Here, the term features includes quality as well. The quantitative statement, that is the description step, is implemented by

measuring, that is “the entirety of all the activities in the measurement chain up to determining the value of a dimension” [14].

The term measurement chain corresponds to the actual practical application of the measuring technique. In addition,

the **measurand**, is the “feature of an object to be measured which can numerically be described in the course of the measuring process” [14].

In an auditory speech test, the measuring object is the acoustic speech signal. In a speech quality test, this acoustic speech signal is the output signal of a speech transmission system (as introduced in Sect. 2.2.2). The transmission system affects the perception of the speech signal in terms of its different characteristics regarding the quality elements. Due to the dependencies on the speech material (usually female and male speakers as well as different sentences are used [6]) often more than one speech signal is processed for one test. In this case, the composition of one transmission system configuration is referred to as a *test condition* composed of multiple acoustic speech signals called *stimuli* (different content and/or speakers but always the same system configuration).

The measurand is either a feature of the speech signal or the quality itself. A fundamental part of the measuring process is the *scaling*, that is the assignment of numbers to measuring objects (the speech signals) according to consistent rules [14, 46]. The numbers must be provided in a way that all aspects of the measuring object under consideration can be assigned to a number [2]. The set of the provided numbers is called a *scale* [47]. In such a test, the listener acts as the *measuring organ* and thus on a perceptual base, the measurement subjectively takes place [14]. Therefore, the listener is also called *subject* and auditory speech tests are also referred to as *subjective tests*. To sum up, in a subjective speech test, a test subject is asked to judge the perceived quality of a speech signal on consistent scales.

To guarantee that a subjective speech test is really quantifying the perception of a user, specific characteristics of a test have to be met. In [48] (and summarized in [3]), six characteristics were introduced to ensure a satisfying test method:

- *Objectivity*: results are reproducible over different listeners.
- *Reliability*: results are not widespread if the test is repeated by a listener.
- *Validity*: the test measures what is intended to measure.
- *Sensitivity*: differentiations and distinctions should be as fine as those made by the listener.
- *Comparability*: enables comparisons between different transmission systems (conditions and stimuli) and its groups.
- *Utility*: usefulness of the provided information.

Speech Quality Test Classification

According to [49], subjective speech quality tests can be divided into four categories, classified into two dichotomies: *analytic* versus *utilitarian*, and *subject-oriented* versus *object-oriented*. This classification is also illustrated in Table 2.1. In this book the focus will be on object-oriented tests.

The first dichotomy describes whether the focus of the test subject is on the overall quality or on its quality features. In utilitarian test methods, the test subjects are asked to judge the overall speech quality as a single quality value on a one-dimensional rating scale. A very famous example for these tests is described and standardized by the ITU in ITU-T Recommendation P. 800 [38]. However, more examples will be given in Chap. 3.

In analytic test methods, the subjects are asked to judge specific quality features that compose the overall quality. Either the subject judges one specific quality feature in a one-dimensional scale, or the subject judges on several scales, one per quality feature. The latter forms the basis for a multidimensional analysis that is necessary to identify orthogonal perceptual quality dimensions of the perceptual quality space. These tests serve for diagnostic quality assessment. Three famous examples for these test method are (i) the *attribute scaling* method like the *Semantic Differential* (SD) with a subsequent *Principal Component Analysis* (PCA) [50, 51], (ii) the *Diagnostic Acceptability Measure* (DAM) method according to [52], or (iii) the *Multidimensional Scaling* (MDS) method in which perceptual differences of pairwise presented stimuli are scaled, and then the perceptual distances/dissimilarities are mapped to a multidimensional space [53]. More information about these methods and their results can be found in Chap. 3 and Sect. 4.2.

Table 2.1 Speech quality test classification based on [3] according to [49]

	Subject-oriented tests	Object-oriented tests
Utilitarian judgments	Psycho acoustic research	Overall speech quality assessment
Analytic judgments	Audiological evaluation	Diagnostic quality assessment

The second dichotomy describes the two possible types of analysis that speech quality tests could end in. On the one hand, the analysis is focused on the quality of the transmission system conditions (object-oriented), and on the other hand, the analysis is focused on the subject in the quality formation process (subject-oriented).

Speech Quality Test Characteristics

The introduced definitions and classifications show that planning and implementing a subjective speech quality test is a complex task in which numerous aspects have to be considered. In [6, 21], five characteristics of subjective quality tests are introduced that define exact test results. Based on the number of transmission system conditions under test and the measurand, the test supervisor chooses the appropriate characteristics:

- The *presentation method*: defines whether a comparison (relative) or an absolute assessment will be conducted (examples for both will be given in Sect. 3.3.2).
- The *scale level*: defines whether an interval-, ordinal-, ratio-, or nominal-scale is used (see [54] for more information).
- The *scaling method*: defines whether the rating process will be conducted on a single or on multiple scales (see Sect. 3.3.2 for examples).
- The *test modality*: defines whether a listening, a speaking, or a conversation test is conducted (more information about different test contexts are given and explained in Sect. 3.1).
- The *analysis method*: defines whether a simple average analysis or a multidimensional analysis is performed (the two analysis methods will also be introduced in Sects. 3.3.2 and 4.2).

Schematic of a Listener

In Sect. 2.3.1, Fig. 2.6 illustrates the quality formation process of a listener. This concept can also be adapted to a listener in a subjective test. In Fig. 2.8, the schematic of a listener in an auditory subjective test according to [2] (and based on [47]) is shown: A speech signal arrives at the listener. As the receiving speech sound is distinct in time, space, and other characteristics [47], it is called a *sound event*. At the first step, the sound event causes an *auditory event* (more generally a *perceptual event*) that is transformed from the physical domain to the perceptual domain by the *perceiving system* (see the perception step in Sect. 2.3.1). If the listener participates in a utilitarian test (the listener is asked to judge the overall quality), the perceptual event is compared to the internal reference (comparing system) and transformed into a *quality event*. This quality event is not accessible from the outside of the listener. Hence, the listener is asked to describe the quality event (describing system) by judging it on a scale.

If the listener participates in an analytic test (the listener is asked to judge perceptual features), the perceptual event (that is of multidimensional nature and composed of multiple quality features) is transformed into feature descriptions by the second describing system. This way dissimilarities between specific features can

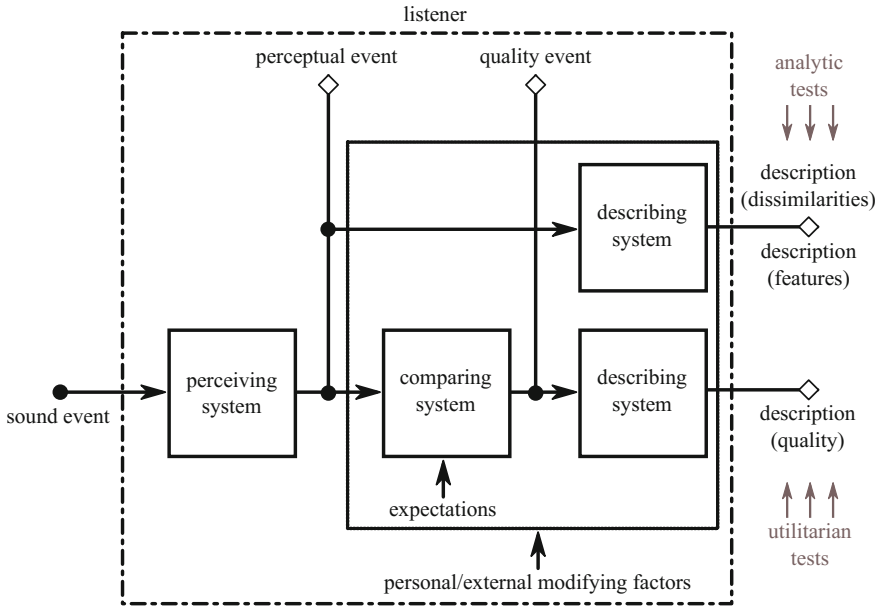


Fig. 2.8 Schematic of a listener in an auditory subjective test according to [2] (based on [47] and extensions by [14, 21])

be described (see Sect. 4.2.1 for more details on dissimilarity judgments). Note that other than for the quality formation process, this concept of the schematic of a listener is very general and also applies for *speaking* and *conversation* tests (see Chap. 3 for more details about different test contexts). For more details about the schematic of a listener in a auditory test see [2, 47].

Test Subjects

Depending on the purpose of the test, care should be taken when selecting test subjects. In principle, test subjects are classified according to their knowledge about subjective quality tests and the speech transmission systems under test. In general, test subjects are classified into two groups, *expert* subjects and *naïve* subjects. Utilitarian tests usually targeting at providing quality ratings of an “average” system user. Thus utilitarian tests are commonly carried out with naïve subjects. According to [38], naïve test subjects are defined as:

Naïve test subjects “taking part in listening tests are chosen at random from the normal telephone-using population, with the provisos that:

- they have not been directly involved in work connected with assessment of the performance of telephone circuits, or related work such as speech coding,

- they have not participated in any subjective test whatever for at least the previous six months, and not in any listening-opinion test for at least one year,
- they have never heard the same sentence lists before”.

Unfortunately, to the best of the author’s knowledge, no similar definition of expert subjects exists. However, experts could be defined by the contrary of a naïve test subjects. Experts should be connected with assessment of the performance of telephone circuits, or related work such as speech coding, they can have participated in any subjective test in the last six month, and may know the sentence list. Thus, experts are for example researchers that plan and conduct subjective tests, developers of new coding techniques, or audio engineers. Nevertheless, experts are more costly and difficult to recruit than naïve test subjects (students, colleagues, or friends).

Both, expert and naïve test subjects, have to fulfill two further requirements to participate in a speech quality test. First, it must be guaranteed that the test subjects are free of any hearing (or speaking in case of conversation tests) impairments. The hearing ability is usually evaluated using an audiometer that determines the hearing threshold. Second, the test subject should be a native speaker of the language that is used in the experiment. In addition, if a test is supposed to be representative for a particular user group, the test subjects should be selected according to the corresponding requirements (age, gender, experience in telecommunication, and the like).

Subjective Test Effects

In a subjective quality test, many different factors can influence and bias the way the test subject perceives the speech signal that is under test. This subsection will give a very brief overview about the aspects of a subjective quality test that may affect the judgment of the test subject. A detailed review can be found in [6, 55], or [56]. In [14], the aspects that may affect the judgment of the test subject can be classified into three groups:

The *Scaling Effect* describes all effects that can result from the scale itself. For example, the intervals between the categories on a category scale may be unequal. This leads to a non-linear scale where simple mean values should not be calculated. To overcome this effect, each category is connected to a number (see Sect. 3.3.2). Also, the sensitivity of scales may be different or the saturation at the extreme categories may be neglected by naïve test subjects. Both effects are discussed and compared on two different scales in [57].

The *Subject Effect* describes the effect that can result from the different internal references of the subjects. Each test subject has his or her own opinion about the perceived quality because each subject has his or her own internal reference due to different prior experiences. To minimize this effect, the number of test subjects should be large enough to ensure an average rating of the quality. The ITU-T Handbook of Telephonometry [58] recommends around 30 subjects. Also, so-called anchor

conditions (for example the highest and the lowest quality of a system condition in the test) are presented before (i.e. training) and during the test to ensure high reliability. At last, also the fatigue of test subjects may have an effect on the quality judgments. It is recommended to insert short breaks in the test at regular intervals (for example every 30 min).

The *Context Effect* describes the effect that the assessment situation may have on the subject's judgments. The main bias introduced in subjective quality tests is the listening environment. Most tests are conducted in a laboratory context that is different from a real-life situation. In the laboratory, the quality judgment is restricted to one modality (sound), in real-life the perception is multimodal (vision, sound, temperature) [59]. Therefore, tests are conducted in standardized test rooms where the surrounding conditions are kept stable to minimize this effect [35]. In addition, the corpus and the order of the speech material influence the subject's judgment (*corpus effect* and *order effect*). Thus, a subject's judgment depends on the range and distribution of degradations within the test corpus. For example, in [6] it is found that a NB condition obtains a higher quality rating in a pure NB corpus than in a mixed NB-WB corpus. The order-effect corresponds to the influence of a preceding stimulus on the current stimulus. It is assumed that the most recent stimulus has the strongest influence. Each test subject is presented with a different order of the conditions under test to minimize the bias of the order-effect.

Regarding QoE, these effects are proposed to be called *Influence Factors* according to [60]. Three types of influence factors are proposed, *human*, *system*, and *context* influence factors. Each of the three influence factors are connected to one of the effects mentioned above. For more information see [60].

2.3.4.2 Instrumental Methods

The introduced auditory subjective quality tests depend on the test subjects and their quality ratings. Also, they are complex in terms of planning and conducting. Thus, even though these subjective test are valid, reliable, and sensitive means for evaluating a speech transmission system, they are very time and money (compensation for the test subjects) consuming. So for example, if a company develops or modifies a codec, it would not be economic to conduct a new subjective test to evaluate the enhanced transmission quality for every development phase. Hence, computer programs that *estimate* (often also the term *prediction* is used) the perceived speech quality were developed. These *quality models* are called *instrumental* speech quality models. However, these instrumental models must be based on the results of subjective tests. Thus, they are the “counterparts” of subjective test and are also often referred to as *objective* methods.

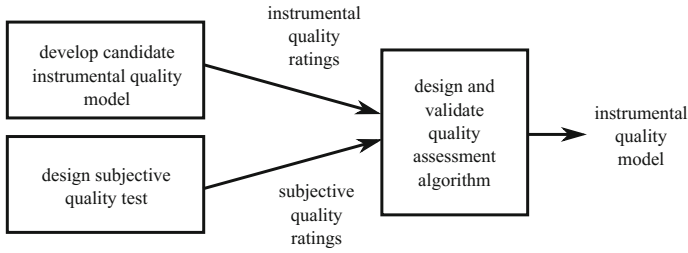


Fig. 2.9 Development of an instrumental quality model based on [61]

In [61], the development of an instrumental speech quality model is described (see Fig. 2.9). There should be at least three major steps involved in the development of an instrumental model:

First, a valid subjective quality test has to be designed and conducted. This test provides subjective quality ratings serving as a ground truth for the instrumental model. Second, a candidate instrumental quality model has to be developed and implemented. This model provides instrumental (objective) quality ratings. The third major step is the design and validation of the instrumental quality model. Here, the subjective and the instrumental quality values are compared in terms of correlation and error (see Chap. 7 for an introduction). Instrumental models that prove to be redundant or not significant may be discarded at this point. If no accurate model can be found, maybe step one (subjective quality test) and two (candidate instrumental quality model) have to be iteratively repeated. However, if the instrumental model proves to provide valid and reliable correlates between the instrumental and the subjective quality ratings, this model can be used to estimate subjective quality ratings. Note that for a robust estimation, the model has to be validated on more data provided by the subjective test design in step one, for example by other laboratories.

Instrumental models use variable input information and are used for different applications and targets (for example, an instrumental model designed for NB transmission systems does not provide valid estimations for WB transmission systems). According to these characteristics, instrumental models are classified in three different groups; *parameter-based models*, *signal-based models*, and *packet-layer models*. Note that in the literature the names for the different model classes vary, alternative names are given in parentheses (based on [21, 62, 63]).

Parameter-Based Models

Parameter-based models (or *Opinion models* or *Network planning models*) use parameters that characterize the different quality elements of a transmission system to plan future transmission systems. If telecommunication-system providers plan a new speech transmission system, they only have information about the elements they could use. So, based on these elements, the transmission system and its elements can be described by specific parameters that are mathematically combined to estimate the quality of a future transmission system. Examples for parameter-based instrumental models are given in Chap. 3.

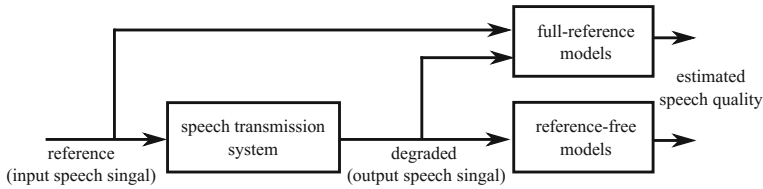


Fig. 2.10 Overview of signal-based instrumental models based on [3]

Signal-Based Models

Signal-based models (or *Speech-layer models*) use signals that are transmitted over a speech transmission system (or degraded by a speech processing chain) to estimate the perceived speech quality of these systems (or processing chains). If a transmission system is developed and implemented, it is possible to transmit speech through it. Signal-based models try to extract specific indicators of the speech signal to estimate the perceived quality. Based on what signal the model is using, signal-based models can be classified into two groups (see Fig. 2.10):

- (i) *Full-reference* (also known as *intrusive* or *double-ended*) models depend on a reference (system input) speech signal and a corresponding degraded (system output) speech signal. They time-align and compare both speech signals and use the difference of the indicators to estimate the speech quality.
- (ii) *Reference-free* (also known as *non-intrusive* or *single-ended*) models depend only on the degraded signal at the system output. They extract information from the degraded signal only and use these information to estimate the speech quality. Examples and further information about signal-based instrumental models (both types) are also given in Chap. 3.

Packet-Layer Models

Packet-layer models analyze parameters of the transmission network (for example the packet pattern in VoIP networks) to monitor the speech quality. They are used in packet-switched networks and measure in gateways or at the listener's side several network related parameters (for example the delay of packets, the packet-loss rate, or the burst ratio). The current ITU-T standard for a packet-layer model is described in [64]. Packet-layer models will not further be addressed in this book, for more information see [63].

2.4 Conclusion

This chapter is targeting at giving the definitions and explanations of the fundamental knowledge important for the research that is conducted in this book. At this point, the reader should be familiar with the following topics: First the basic research scenario,

a telephone conversation between two interlocutors, was presented. Second, an introduction of the three main parts of a conversation: speech production, transmission, and perception, was given. This introduction led to the definition of perceived quality of transmitted speech. In addition, the concepts of diagnosing speech quality on the basis of perceptual dimensions that form a perceptual quality space were shown. Finally, the end of this chapter explains the general methods to assess the perceived quality of transmitted speech. This covers the schema of subjective test and the different techniques for instrumental quality assessment. However, the given introductions are for now rather general than precise in case of a telephone conversation. The next chapter will give a deeper insight into the speech quality in a telephone conversation, including the different phases of a conversation and their corresponding quality elements, perceptual spaces, and subjective as well as instrumental quality assessment methods.

Multidimensional Analysis of Conversational Telephone
Speech

Köster, F.

2018, XIII, 184 p. 60 illus., Hardcover

ISBN: 978-981-10-5223-1