1

Speech Quality in Telephony

1.1 Speech

Language can be regarded as a communication system particular to humans. Its signs can be available in written or acoustic form[1]. Speech is a subsystem of language, that is, the '[. . .] communication by means of language in the acoustic channel' (Sebeok, 1996). Human interlocutors, for example, in a telephone conversation, communicate by exchanging speech signs. Thus, they are able to convey abstracted information in acoustic, that is, physical form (with little effort). The acoustic speech signal present at the recipient's ear, the *sound event*, causes an *auditory event* in the listener's brain (Blauert, 1997). On its way from sound event to auditory event, speech is processed at different levels in the brain, starting from the preprocessing provided by the auditory periphery. Ultimately, meaning is established on higher processing levels from the percept and additional contextual information, such as the emotional state of the listener.

The auditory event related to a particular sound event as well as the meaning the listener attaches to it are not accessible from the outside. Hence, knowledge on the perception related to specific sound events can only be gained through introspection (in case of own perception) or the description provided by another listener[2]. The science of formally studying auditory perception is referred to as *psychoacoustics*. Because of the absence of instrumental means for directly accessing perceptional events, human subjects are the only appropriate measurement instruments. A systems-analytic view of auditory experiments was proposed by Blauert (1997). A corresponding schematic representation of a listener in an auditory test scenario is depicted in Figure 1.1 (Blauert, 1997, pp. 5–12). Here, s_0 denotes a sound event and h_0 the related auditory event; b_0 refers to a description of the auditory event h_0.

As in all measurements, the description is typically given as an assignment of numbers on an appropriate measurement scale, for example, as a value b in certain scale units, to the measurement object (Stevens, 1951, p. 22). As in instrumental measurements, the measurement method has to be chosen in such a way that b (i.e. the rating or judgment) *validly* and

[1] Or in other more specific forms such as sign language.

[2] Alternatively, knowledge on perception can be indirectly gained from the reaction of subjects, or their performance on specific tasks.

Speech Quality of VoIP: Assessment and Prediction Alexander Raake

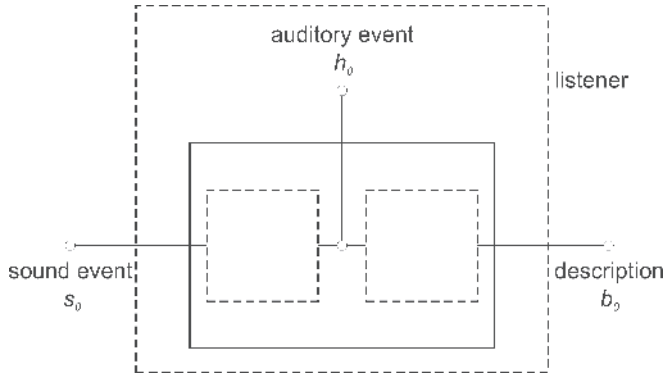

Figure 1.1: Schematic representation of a test subject in an auditory test, see Blauert (1997).

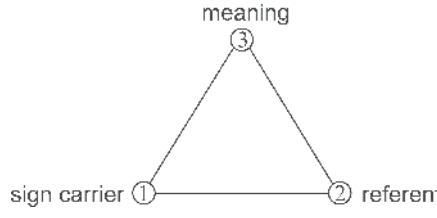

Figure 1.2: Semiotic triangle in general form according to Nöth (2000).

reliably quantifies the measurement object, the auditory event h_0. Here, validly means that the test method actually measures what it is intended to measure. Reliably means that the method is accurate and produces results without large scattering, and similar to the results obtained when the test is repeated (Guilford, 1954, pp. 349–357). The applied rating scale itself constitutes a sign system for conveying information on the perceptional event.

It has to be noted that signs are not signs *per se*: A sign becomes a sign when meaning can be associated with its carrier by an interpreting subject. Hence, '[. . .] a sign is a mental unit which is processed as standing for something other than itself, as pointing at something else' (Jekosch, 2005a). The science of signs is generally referred to as *semiotics*. Different diadic and triadic sign models have been proposed to capture the relation between the sign carrier and the associated meaning (Ogden and Richards, 1960; Peirce, 1986). Triadic sign models of semiotic theory differentiate three constituents (correlates) of a sign, which form the so-called semiotic triangle, shown here in a general representation (Figure 1.2, according to Nöth, 2000).

In case of language, the *sign carrier* or sign vehicle is the written word or the acoustic speech signal, thus the form, in which the sign is presented. The *referent* is the (possibly abstract) object the sign stands for. The *meaning* is the sense made of the sign by its interpreter. Hence, used as the correlate of a sign, *meaning* can be regarded as the role the sign plays, or the function it has for a sender or receiver. Here, the situation in which the sign 'happens' – for example, during a telephone conversation – is an important factor both for the reception and for the conveyance of the sign (Jekosch, 2000, 2005a,b).

Most of this book is concerned with the acoustic speech signal, i.e. the *form*, and the perception of its quality. However, many of the perceptual experiments were carried out as conversation tests, which include the actual application of speech: communication. Communication, between a person and other persons or objects in the outside world of the person, is the main subject of semiotics (Jekosch, 2005a). From the point of view of semiotics, it becomes clear that the three constituents of a speech sign cannot be separated, neither in terms of speech production nor in terms of speech perception. For example, only in case of small degradations of transmitted speech (affecting primarily the carrier), will it remain intelligible without additional listening effort, and the referent and sense may remain unaltered. The role of content (i.e. referent) and meaning (Figure 1.2) for speech quality perception is specifically addressed in Section 5.4.2 of Chapter 5.

1.1.1 Speech Acoustics

By speech, linguistic information is conveyed in the form of pressure waves traveling from the speaker's mouth to the listeners' ears. The longitudinal waves of atmospheric pressure variation are time-varying by nature. They reflect the variation pattern of the vocal tract articulators, such as the vocal folds (glottis), lips, tongue and jaw used for conveying the linguistic information. Hence, speech sounds can be viewed as a carrier signal filtered by the adjacent tubes the vocal tract is formed of, and amplitude-modulated by the articulatory movements modifying the physical properties of these tubes. The carrier signal produced by the glottis is a broadband signal with a strong tonal component, the pitch of which is due to the periodically opening and closing vocal folds. Hence, the glottis signal has a broadband line spectrum. It is characterized by the spacing of the lines, corresponding to its inverse pitch period, generally referred to as *fundamental frequency, F0*. For an overview of acoustic phonetics see, for example, Vary *et al*. (1998, pp. 5–28) or O'Shaughnessy (2000, pp. 35–107).

Speech can transmit information to one or more listeners as a result of the combination of both *temporal* as well as *spectral* properties, which are used by the recipients for decoding the information. The acoustic speech signal is characterized by short periods of a particular acoustic behavior, typically referred to as *phones*. They correspond to the acoustic realization of the smallest meaningful contrastive units in the phonology of a language, the *phonemes*. Phonemes can be coarsely divided into two classes, the vowel type and the consonant type. Vowel phonation is characterized by an unrestricted airflow through the vocal tract, while consonant phonation is due to a restriction of the airflow at some point in the vocal tract. During periods of 'stationary' acoustic behavior – related to a certain position of the articulators – characteristic acoustic properties of the different phones can be extracted from the spectrum of the speech sound. For example, the different vowel sounds of a particular language can be differentiated from the peaks they show in the amplitude spectrum. The first three peaks result from the three major resonances of the vocal tract, corresponding to the first, second and third *formants* F1, F2 and F3[3]. In the speech perception process, they are used by the auditory system as cues for vowel discrimination.

[3]The formants lie in different frequency ranges, depending on the type of vocalic sound and on the speaker: F1 is typically in the range between 0.2 and 1.2 kHz, F2 in the range 0.7 to 3.3 kHz, and F3 in the range 1.5 to 3.7 kHz.

Information is coded in the *temporal variation* of the smallest building units of speech. Different phones are concatenated into phone sequences by a speaker in order to form the words to be expressed. In practice, the resulting phones differ from the phonemes the speaker intended to articulate, both in articulation and in acoustics. This is due to the fact that subsequent articulatory gestures overlap and ultimately yield an articulation that depends strongly on the phonetic context of the phones (*coarticulation*). In phonology, words are traditionally divided into *syllables*, that is, phoneme sequences typically containing one vocalic sound. Linguistic information is not only coded in the particular composition of these sequences but is also carried by the segmental duration[4] of the phones in the syllables (Greenberg, 1999; Klatt, 1976). For example, certain words may be stressed more than others by the lengthening of particular phonemes. Phone durations vary broadly in the range between 20 and 240 ms, depending, for example, on the phone type, and on whether the phone is stressed (Umeda, 1975, 1977, as an example for read American English: 50 to 240 ms for vowels and 20 to 150 ms for consonants). Average syllable durations lie in the range from 100 to 300 ms (Arai and Greenberg, 1997) depending on the durations of the phones the syllable is composed of. The control of duration is typically combined with variations in fundamental frequency and speech level. This interplay is generally referred to as *prosody*. Although prosody may provide additional contextual information supporting word identification, its main functions are stressing information, indicating the purpose of a particular utterance (i.e. whether it is an exclamation or question), or providing information on the emotional state of the speaker (e.g. Murray and Arnott, 1993). Higher order formants (FN, $N > 3$), the fundamental frequency F0 and properties of the speaking style may also be used by listeners to identify the speaker (e.g. Mersdorf, 1996).

1.1.2 Speech Perception

From the above considerations it can be concluded that in order to process and understand speech signals, the auditory system has to (a) be able to resolve overlapping spectral cues and (b) decode the information ultimately associated with the dynamics of speech.

The stages involved in speech processing by the brain range from the auditory pre-processing over phonetic, phonological, lexical, and syntactic to semantic analysis, until meaning can finally be extracted (for an overview see O'Shaughnessy, 2000, pp. 141–172). In audition, these processing stages do not necessarily have to be sequential; some may act in parallel, and others may be skipped entirely (see e.g. McAdams, 1993, pp. 149–150). The following considerations depart from the incoming acoustic speech signal, and discuss aspects of lower- and higher-level speech perception considered relevant for this book.

1.1.2.1 Spectral Processing

To provide the functionality of spectral resolution, the ear is equipped with a mechanism for transforming spectral information into place information (see Zwicker and Fastl, 1999, pp. 23–60, for an overview on the information processing in the auditory system). After excitation of the eardrum by the incoming speech signal, the middle ear's ossicles supply the necessary impedance matching from the airborne sound outside the ear to the liquid-borne sound propagation in the inner ear (Figure 1.3). The cochlea in the human inner ear

[4]The durations of phones and syllables play a role in the impact of packet losses on the quality perception of packetized speech, as will be discussed in Chapters 3 and 4.

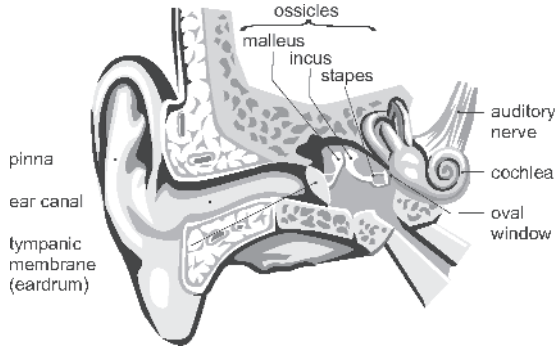

Figure 1.3: Schematic illustration of the outer, middle and inner ear. Copyright 2006 A. Raake, COREL® and their licensers. All rights reserved.

consists of three parallel tubes (or channels, referred to as *scalae*) filled with liquid, which are rolled up in the form of a snail. The basilar membrane separates two of the channels from the third (the scalae vestibuli and media from the scala tympani), and bears the organ of Corti. The organ of Corti hosts the sensory cells (*hair cells*) that convert the analog sound information into neural spike trains further processed at different levels in the brain.

The above-mentioned frequency–place transformation is achieved in the cochlea by exploiting the properties of traveling waves. The incoming sound waves are coupled via the eardrum and ossicles to the *oval window* connecting to the cochlea, where they excite the traveling waves. In case of a pure tone of a given frequency f, the traveling wave shows an excitation spanning over a certain area of the basilar membrane. The traveling wave yields a maximum excitation at a particular location, corresponding to the frequency of the tone transformed into place coordinates. A tone complex is thus decomposed into traveling waves showing main excitations at certain places on the basilar membrane, which correspond to the frequency components the tone complex is composed of. In case of tones of higher frequency, the maximum displacement of the basilar membrane occurs closer to the oval window, and tones of lower frequency lead to excitations closer to the end of the cochlea, the apex.

The relation between frequency and place is neither linear nor logarithmic. In pitch perception experiments, however, it was found that the pitch-ratio scale[5] shows a linear relationship to the place of main excitation on the (rolled-out) basilar membrane (Figure 1.4). For describing different phenomena of psychoacoustics, it has proven to be very useful to apply a transformation of frequency onto a scale showing a linear relation to the place of basilar membrane excitation. One of the most widely used examples of such a scale is the *Bark scale*, which relates frequency to a measure called *critical-band rate* (Zwicker, 1961; Zwicker and Fastl, 1999, pp. 149–164). It is measured in units of *Bark* and transforms the frequency range most relevant for human audition (0 to 16 kHz) to critical-band rates ranging from 0 to 24. Integer numbers of the critical-band rate correspond to the starting

[5]The scale on which the perceived pitch of tones can be displayed in such a way that ratios on the scale correspond to the ratios of tone perception, for example, answering the question how many times higher or lower a tone 'A' is perceived as a tone 'B'.

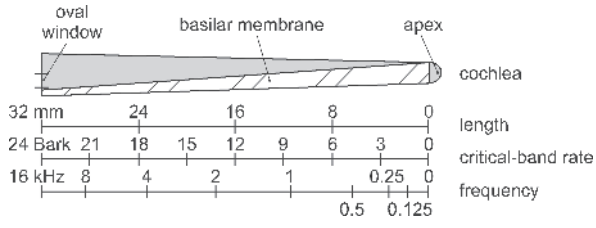

Figure 1.4: Schematic cross section of the rolled-out cochlea showing the basilar membrane (Zwicker and Fastl, 1999, pp. 149–164, mod. from Fig. 6.11). Below the figure, schematic scales are provided for location on the basilar membrane [mm] (top), critical-band rate [Bark] (middle) and frequency [kHz] (bottom).

points of segments that subdivide the basilar membrane into equally long *critical bands*, beginning at the apex (Figure 1.4).

The concept of critical bands is closely related to different psychoacoustic phenomena, such as spectral masking. The relation between critical bands and masking can be illustrated by the following experiment (Zwicker and Fastl, 1999, pp. 149–173): A test tone of a certain frequency f is presented centered between two band-pass noise signals, whose cutoff frequencies are separated by Δf. Then, for increasing width Δf of the noise gap, the masked threshold, i.e. the sound pressure level necessary for the test tone to be just audible, remains constant until a certain critical bandwidth is reached. For values of Δf above this *critical bandwidth* Δf_G, the masked threshold decreases. The critical bandwidth Δf_G determined this way is a constant associated with the center frequency f, the frequency of the test tone. Transformed from frequency to place, the different critical bandwidths correspond to constant distances on the basilar membrane, or in terms of critical-band rate, to constant increments of one Bark. Obviously, the processing performed by the cochlea can be viewed as a spectral decomposition of the incoming signal by an array of overlapping band-pass filters. Another psychoacoustic phenomenon related to the critical bands is associated with loudness perception: for equal sound pressure levels, the loudness of a band-limited noise signal is perceived as constant until the bandwidth is increased beyond the critical bandwidth associated with the center frequency of the noise. For larger bandwidths, the loudness increases: In loudness perception the excitation within the region of one critical band is grouped (Zwicker and Fastl, 1999, pp. 149–173).

A convenient approximation of the experimentally determined relation between critical-band rate and frequency is given by Equation (1.1) (Zwicker and Fastl, 1999, pp. 158–160):

$$z/\text{Bark} = 13\arctan(0.76\,f/\text{kHz}) + 3.5\arctan\left[(f/7.5\text{kHz})^2\right] \tag{1.1}$$

In some cases, an inverse relationship for deriving frequency from critical-band rate may be useful. A good fit of experimental data (Zwicker and Fastl, 1999, pp. 158–160) was found by the author of this book using Equation (1.2).

$$f/\text{Hz} = 1285.93\left(\frac{\mathrm{e}^{(z/\text{Bark})^{2.64}}}{1836.93} - 1\right) + 93.3\frac{z}{\text{Bark}} \tag{1.2}$$

Both formulae play a role in the model of speech quality in case of bandwidth restrictions, which is presented in Chapter 5. The spectral processing provided by the hearing system is of particular relevance for the processing of speech sounds. For example, the formants revealed in a spectral representation of a vowel type phoneme are replicated in the excitation pattern on the basilar membrane. By the hair cells, the excitation associated with the formants is translated into firing rates processed further on higher levels in the brain (Young and Sachs, 1979). Here, this information can be used as (spectral) cues supporting the identification of lexical units.

1.1.2.2 Temporal Processing

The processing of spectral cues like harmonics used to detect certain phonemes are paralleled by the temporal processing performed by the auditory system. The temporal speech cues that can be exploited by the auditory system can directly be deduced from the speech acoustics summarized in Section 1.1.1. Rosen (1992) suggested a framework for describing the temporal cues of speech:

- The coarse structure, that is, the envelope of the speech signal, is associated with cues on the syllable level, corresponding to time frames greater than 20 ms, with syllable durations typically in the range from 100 to 300 ms (Greenberg, 2004).
- The periodicity provides pitch information (2–20 ms $\Rightarrow$ $F0 = 50$–500 Hz).
- The fine structure contains information on the identity of the speaker (time frames below 2 ms).

After the preprocessing by the auditory periphery, the auditory nerve transmits the temporal information faithfully to the brain (Wang *et al.*, 2003). On higher processing levels, however, sparse acoustic events (like voice onsets) are marked with precise spike timing, while rapidly occurring acoustic events are transformed into firing rate–based representations (Lu *et al.*, 2001).

Obviously, a high relevance is given to the exact representation of envelope information and timely coding of sparse acoustic events like onsets. A syllable-centric view on speech perception underlining the importance of the envelope structure has been proposed by different authors (e.g. Greenberg, 2004; Mehler and Segui, 1987; Warren, 1999, pp. 73–76). A possible reason for less precise temporal coding of the periodicity and fine structure information, and instead representing it in a transformed way as firing rates, may be the necessity of integration with the information from other, slower senses, like the visual or tactile (for physiological details on temporal auditory processing, see e.g. Wang *et al.*, 2003; Young and Sachs, 1979).

1.1.2.3 Speech Perception and Auditory Memory

Different theories of speech perception have been reported in the literature, which are accompanied by different views on human auditory memory (for an overview see O'Shaughnessy, 2000). One theory assumes a dual perception process: bottom-up auditory processing and top-down phonetic processing (e.g. Mehler and Segui, 1987; Samuel, 1981).

According to this model, the auditory process analyzes acoustic features and stores them in auditory (echoic) memory (150–350 ms; see e.g. Baddeley, 1997; Massaro, 1975). The phonetic process is then assumed to yield syllable perception relying on features stored in echoic auditory memory. The phonetic process is driven by the expectation of the listener, mainly resulting from lexical contextual information.

This implies a storage-type memory, with different types of 'stores' at different stages of speech perception (see Crowder, 1993, for a more detailed description). These 'stores' can be distinguished as follows (e.g. Baddeley, 1997; Cowan, 1984):

Echoic memory: Peripheral storage of auditory features for durations of 150–350 ms. A classical experimental technique to determine the capacity of this type of storage was developed by Massaro. In his tests on backward recognition masking, he found that the second of two similar short sounds (e.g. pure tones of similar frequency, see Massaro, 1975) presented in fast succession prevented the identification of the first, when the delay between the second and the first sound was less than around 250 ms. Backward recognition masking was strongest for very short delays and decreased until the threshold of 250 ms was reached. The criticism of this paradigm is that the 250 ms may not represent the decay time or duration of the storage, but the duration needed to extract the information from the 'store' (Cowan, 1984).

Short-term auditory memory: Storage for longer durations (2–20 s), presumably applying some form of recoding of the auditory information. This 'store' is related, for example, to the *recency* and *suffix effects* (Crowder, 1993): In memory tests, it was shown that the last of acoustically presented verbal items like digits were remembered best (recency). The maximal number of items that can be stored was found to correspond to the 'magical' number seven reported as a limit for different types of human information processing tasks (Miller, 1956). The so-called suffix effect was observed when, after the presentation of the list of items, an additional, redundant speech item (suffix) was presented, which was considerably degrading the retention of the items presented before (Crowder and Morton, 1969). The suffix effect was found to be restricted to speech type suffixes similar to the list items in terms of pitch, voice quality and spatial location, without the meaning of the suffix being of any relevance for the effect to occur. Obviously, the suffix erases or inhibits the auditory storage of the last item(s). Neither suffix nor recency effects were found in serial recall experiments with visual verbal stimuli (modality effect). The modality and suffix effects are not fully explained: the suffix effect was found to also occur in case of lip-read interference of word lists presented acoustically, or of lists presented in a lip-read fashion (instead of both suffix and list being presented acoustically; for details see, for example, the literature reviews by Cowan, 1984; Crowder, 1993). It was thus hypothesized that the short-term auditory memory involved is related to mental mechanisms generally concerned with language perception and analysis (Crowder, 1993). However, the recency effect was also found in other tests on auditory perception, like tests on the loudness perception of nonstationary signals (Susini *et al.*, 2002) or tests on the perception of time-varying speech transmission quality (Gros, 2001; Gros and Chateau, 2001). In the latter tests, recent periods of bad quality were observed to have a larger influence on the final quality judgments than previous periods of bad quality (see Chapter 3). In Baddeley (1997),

some examples of the recency effect are also summarized for other than the auditory modality.

Long-term auditory memory: This memory spans over periods of time up to several years or a lifetime, and allows recognition and identification of, for example, musical instruments, melodies and voices. It also refers to a memory of acoustic-lexical items, for example, used for comparison with the auditory features extracted from recently processed speech units during speech comprehension[6]. Similar to the recency effect described above, a small effect of primacy can be observed in serial recall tasks, showing better recall for the first in a particular list of items. While recency is generally ascribed to short-term auditory memory, the primacy effect is thought to be related to long-term auditory memory: from the recall tests reported by Glanzer and Cunitz (1966) and Postman and Philips (1965) it can be concluded that the primacy effect results from a verbal representation of the first item(s) in memory, while the 'classical' recency effect is related to the storage of the item's features in short-term auditory memory.

Instead of this storage-based view on human auditory memory, Crowder (1993) advocates a procedural paradigm: particular perceptual events and possible additional cognitive processes (e.g. mood, attention) yield activation of different regions in the brain. Depending on the type of signal, different, possibly parallel, stages are involved in the perception process. During perception, auditory periphery information is partly recoded into other forms of representation at different processing levels (e.g. into verbal representations). The related activation patterns in the different areas of the brain reflect memory: the information is retained in those areas of the brain that were active during the initial processing (learning).

1.1.2.4 Comprehension

The main function of speech is communication. The sender of a speech message wants to make herself/himself understood, and the recipient intends to understand what the speaker wanted to convey. In the literature, the result of the speech recognition process is referred to as *comprehension* (see e.g. Jekosch, 2000, 2005b, p. 103). Miller (1962) has described the speech comprehension process as a combination of 'decision units' (which are certainly not consecutive) carrying out the phonetic, lexical, syntactic, semantic and pragmatic decisions taken in the course of everyday speech communication. In the following, a list of terms related to speech comprehension is provided for the purpose of this book. It is synthesized from the terms used in the literature and corresponding considerations by Möller (2000, pp. 26–27) and Jekosch (2000, pp. 100–105). The different terms and a simplified illustration of the interrelations are shown in Figure 1.5.

Comprehensibility addresses how well the speech signal allows content to be related to it. It concerns only the form of the speech sign. It reflects the ability of the carrier to actually convey information. With regard to transmission systems, the term *articulation* is sometimes used to quantify the capability of the speech link to faithfully transmit information (e.g. French and Steinberg, 1947; Möller, 2000, pp. 26–27).

[6] Obviously, the long-term memory beyond auditory memory is composed of different levels of encoding, like lexical, semantic and pragmatic stages.

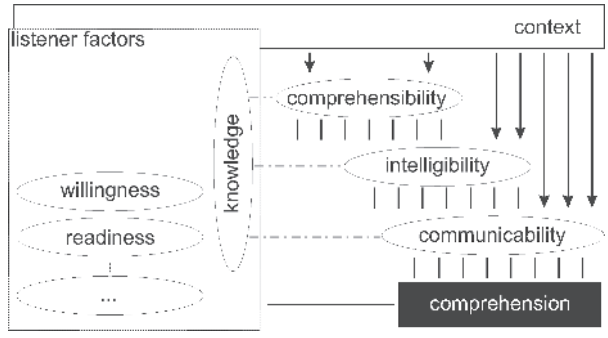

Figure 1.5: Simplified illustration of the terminology used to describe the factors involved in comprehension, and their interrelations. Note that the depicted process is not a linear one.

Comprehensibility relates to the identification of phonemes and phoneme clusters without a lexical activation. Often, the expressions *segmental intelligibility* or *syllable intelligibility* are used synonymously for comprehensibility as it is defined here. According to these definitions, articulation is a prerequisite for comprehensibility.

Intelligibility refers to how well the content of an utterance (i.e. the referent, see Figure 1.2) can be identified on the basis of the form (i.e. the carrier). Intelligibility, according to this definition, strongly depends on lexical, syntactic and semantic contexts. Considerations on the role of context for intelligibility are provided in Benoît (1990). A model for quantifying the role of context for intelligibility was proposed by Bronkhorst *et al.* (1993, 2002), relating intelligibility to the recognition of smaller speech units such as syllables or phonemes (i.e. to *articulation/comprehensibility*).

Communicability means that a speech message is such that it can serve to communicate, that is, can fully be understood by a recipient, ideally as it was intended by the sender (assuming that none of the interlocutors has the intention to deceive the other). It is related to functional aspects of speech and the entire communication process. Depending on the semantic and situational context (e.g. the topic of the conversation, the relationship between the interlocutors, etc.), communicability as it is used here requires a certain degree of intelligibility, which in turn requires a certain level of comprehensibility.

Communicability is closely related to perceived quality: speech quality can be considered as the perceived degree of communicability relative to the desired or expected degree of communicability (see also Section 1.2). From Figure 1.5 it becomes clear that speech communication links are conceivable where comprehensibility and intelligibility may be perfect, but communicability is not: For example, in case of a long transmission delay, the transmitted speech may be highly intelligible, but communicability may be affected severely. As timing plays an important role for the speech communication process, any impairment of the conversation time structure

will modify comprehension (see e.g. Brady, 1968, for an analysis of timing aspects in telephone conversations).

Note that in the speech quality-related literature, the term communicability is sometimes used in a slightly different way, namely to express the extent to which a communication *link* (and not the exchanged messages) is able to provide an efficient medium of communication (see e.g. Wijngaarden *et al*., 2001, 2002).

Comprehension is the result of the speech perception process. It presupposes that a perceived utterance is communicable, and that the recipient is ready and willing to comprehend.

The terms 'context' and 'knowledge' refer to different factors, depending on the level in the process of comprehension (see Figure 1.5): For *comprehensibility*, context refers to aspects such as coarticulation, and knowledge refers to phonological aspects like phoneme or syllable recognition, assisted by phonotactic knowledge, for example, on whether certain phone clusters are elements of the set of clusters characteristic of the respective language. For *intelligibility*, context refers to prosodic and syntactic, as well as to lexical and semantic aspects; knowledge, on the other hand, refers to the ability of the listener to extract the content from a speech message (i.e. lexical, syntactic and semantic knowledge). *Communicability* requires contextual information related also to the sender of the message. Here, context refers to the semantics of the utterance, and the entire situation the utterance was spoken and perceived in (i.e. related also to pragmatic aspects). Then, knowledge refers to the competence of the listener to fully understand what the speaker has said (including knowledge of the speaker, the situation, etc.).

When auditory tests are to be designed to measure either comprehensibility, intelligibility or communicability, context and knowledge associated with respective higher-level concepts should be excluded from the test. For example, in a comprehensibility test, the context information is reduced by using nonsense words, in order to prevent – as far as possible – any lexical access by the listener.

1.1.2.5 Restoration of Missing Sounds

Already on low levels, auditory perception is capable of employing context information to recover missing information, and to yield meaningful recognition results: in the perceptual illusion of *phonemic restoration*, listeners perceive speech samples with some phonemes replaced by noise or other sound segments as complete[7]. For example, in the classical study by Warren (1970), the first /s/ in 'legislatures' was replaced by sounds like a cough. The sound used for replacement is typically perceived as being an additional sound, which can only poorly be localized within the utterance. Dependencies of restoration on different factors were observed:

- Context: Better restoration was found for longer words than for shorter ones. Also, a small effect of word frequency was observed (Samuel, 1981). Moreover, restoration is better in meaningful sentence contexts than for independent speech units or words (Bashford and Warren, 1987; Bashford *et al*., 1988). The context also determines

[7]This effect plays a role in mechanisms applied to recover lost packets in packet-based speech transmission networks like VoIP (see Chapter 3).

whether intelligibility is increased by phonemic restoration: if the underlying speech material consists of complete linguistic units such as sentences, the insertion of noise into the silence gaps of periodically interrupted speech leads to an increase of intelligibility (Powers and Wilcox, 1977). In turn, if no context is provided, intelligibility is not improved (see Miller and Licklider, 1950, who used monosyllabic word lists).

- Confirmation: The higher the spectral similarity between the missing speech sound and the sound used for replacement, the better (or the more certain) is the restoration (Bashford and Warren, 1987).

- Duration: In case of periodic interruptions of read discourse, it was found that filling the interruptions with white noise leads to illusionary continuity up to maximal durations of around 300 ms (Bashford and Warren, 1987; Bashford *et al*., 1988). With speech material presented at different speaking rates, it was confirmed that the maximally restorable duration for periodically interrupted speech approximately equals the average word duration (Bashford *et al*., 1988). Speech interrupted by periodic silence gaps – in turn – was perceived as unnatural already for durations around 10 ms. Only for interruption durations longer than 100 ms was the perceptional effect clearly identified as being due to interruptions (Bashford and Warren, 1987; Bashford *et al*., 1988).

Several authors have explained phonemic restoration with the dual perception process described in Section 1.1.2.3 (top-down phonetic and bottom-up acoustic processing). As the phonemic restoration by white noise was found to be more efficient for consonants than for vowels, a great importance of the bottom-up process, i.e. of the delivery of auditory cues, was hypothesized (Samuel, 1981).

The motivation for this explanation is the fact that phonemic restoration is related to a more fundamental perceptual illusion, which is typically referred to as *continuity illusion* (e.g. Carlyon *et al*., 2002) or more generally *temporal induction* (see Warren, 1999, pp. 134–154 for an overview). The simplest form of temporal induction occurs, for example, when two spectrally identical short noise signals (e.g. of 200 ms duration), one louder than the other, are presented alternating with each other. In this case, the softer noise seems to continue behind the louder one. Under certain conditions, the continuity illusion also occurs when the two sounds are perceptually different (*heterophonic continuity*, see Warren, 1999, pp. 137–141). The first study mentioning this type of continuity illusion was reported by Miller and Licklider (1950). In their tests, they presented listeners with speech signals periodically interrupted at different interruption rates (50% duty cycle). When the interruptions were filled with white noise at interruption rates between 10 and 15 times per second, the speech signal seemed to continue behind the noise. Miller and Licklider referred to this phenomenon as 'picket fence effect', in analogy to the visual modality: when watching a landscape through a picket fence, the fence interrupts the view periodically. However, the landscape is seen to continue behind the fence.

The continuity illusion is closely related to the principle of closure described by the Gestalt psychologists (for an introduction to the laws of 'Gestalt' (German: *Gestaltgesetze*[8]) see e.g. Katz, 1969, pp. 33–39): For example, in case of a basic geometric form like a square, one edge of which is covered by another form, the visual system 'closes' the interruption, and the square is perceived as a continuous unit.

[8]Founded among others by Koffka, Köhler, Wertheimer and Von Ehrenfels.

The auditory processing aspects of phonemic restoration, where missing *linguistic* information is restored, can be summarized as follows (items (1)–(3) are the basis for the underlying continuity illusion; Bregman, 1990; Warren, 1999):

(1) The on- and off-transitions of the deleted sound are masked by the restoring sound.

(2) The restoring sound defines its own limits and not the edges of the sound it restores.

(3) By streaming processes, the neural activity caused by the restoring sound is partly associated to the interrupted sound, rendering the restoring sound softer than when it is presented in isolation (Bregman, 1990; Warren, 1999, pp. 134–154). The newly associated neural activity implies that the missing sound is still there.

(4) After recoding of the auditory feature information, pattern recognition processes associate a linguistic unit with the restored section, if sufficient context information is provided.

1.1.2.6 Human Adaptation to 'Noisy' Communication Channels

Apart from perceptual mechanisms, humans have additional means of enhancing the comprehensibility and communicability of speech, by adapting their communication behavior to the environment. If some information gets lost and cannot be restored by perception, interlocutors may recover the loss by question and reconfirmation, that is, by re-sending the respective message[9]. In noisy environments, interlocutors adopt a Lombard speaking style, that is, raising their voice and stressing syllables differently (e.g. Köster, 2003; Lane *et al*., 1961, 1970). Related strategies apply when speaking to hearing-impaired or foreigners: for example, using clear speech, i.e. overarticulated and slowed-down louder speech, a talker can enhance the comprehensibility of his utterances (e.g. Payton *et al*., 1994). During telephone conversations, impairments like long transmission delays, which have been recognized by the conversation partners, may be compensated for by adopting a walkie-talkie type conversation strategy: in order to avoid erroneous turn-taking, the listener may wait until the talker has finished, before he speaks himself; vice versa, the talker can try to code continuous information into one continuous message, instead of awaiting frequent back-channeling (i.e. confirmations) from his counterpart (Krauss and Bricker, 1966). To overcome severe loudness loss on a transmission system, a talker can raise his voice, and the listener can press the handset to his ear in order to improve the acoustic coupling (e.g. Krebber, 1995).

1.2 Speech Quality

Obviously, a multitude of information is contained in the speech signal typical of everyday communication, as it may happen, for example, during a telephone conversation. Although context and perceptual mechanisms as well as adaptation of the communication behavior yield reliable comprehension even in unfavorable conditions, deteriorations of spectral and/or temporal cues of the speech signal degrade the contained information: be it the linguistic information (the actual message), or paralinguistic information regarding the speaker

[9] Assuming they are aware of the loss!

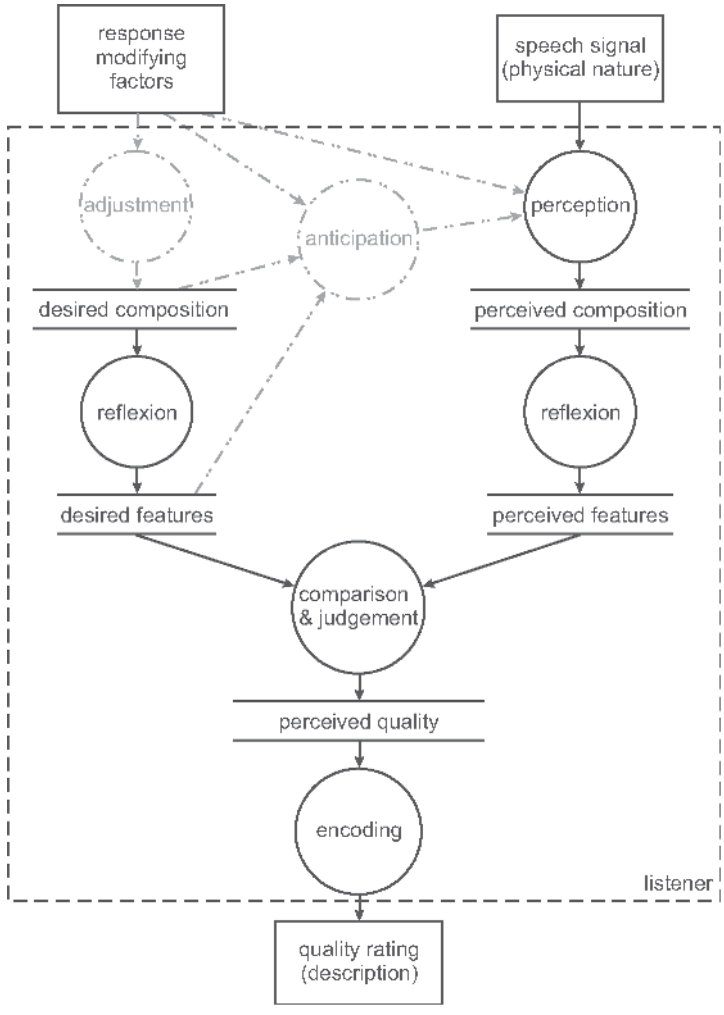

Figure 1.6: Quality: Perception and rating process as seen from the listener's perspective (based on ideas from Jekosch, 2000, 2004, 2005b)). In the figure, circles represent processes, two horizontal parallel lines storages and rectangles the inputs and outputs to the listener. Note that the comparison of features can be assumed to be carried out both on a feature-by-feature- and on an integrated 'Gestalt' level.

identity or his emotional state. Even if comprehension is not at stake, other factors such as annoyance, for example, due to time-varying unwanted sounds additional to the wanted signal, modify the user's appreciation of a particular connection. In this context, the term *speech quality* is typically used, extending the comprehension-oriented view on speech to additional factors governing speech perception. In this section, a working definition of the quality-related terms is presented, providing the scientific paradigm for the research described in this book.

1.2.1 Definition of Quality

Quality is the

> 'result of [the] judgment of the perceived composition of an entity with respect to its desired composition'.
> (Jekosch, 2000, 2005b, pp. 15).

Here, the ***perceived composition*** is the

> '[t]otality of features of an entity',
> (Jekosch, 2005b, pp. 16).

and a ***feature*** is

> '[a] recognizable and nameable characteristic of an entity.'
> (Jekosch, 2005b, p. 14).

If the entity under consideration is a sound, the 'perceived composition' refers to the perceived *auditory* composition, and the 'desired composition' to the desired *auditory* composition of the sound (see Figure 1.6; instead of perceived or desired 'composition', the terms 'perceived nature' and 'desired nature' are sometimes used, e.g. Jekosch, 2004). Two sounds can be distinguished on the basis of their loudness, pitch, duration, timber and spaciousness (see Letowski, 1989). The totality of these attributes or *features*[10] describes the *perceived composition*, *perceived nature* or *character* of a sound (Jekosch, 2004, 2005b; Letowski, 1989, respectively). In this book, the term 'sound' refers to speech sounds.

The listener is a part of a particular communication situation, communication either with one or several other persons, or, in a more abstract sense, with (parts of) his/her environment (e.g. in case of nonspeech sounds). Because of the framework of communication, the listener anticipates the percept to some extent (see Figure 1.6). On the basis of *modifying factors*[11] such as the context and situation in which the sound occurs, and on personal factors such as her/his mood or motivation, the listener relies on mental, *desired characteristics* of the percept[12]. If the listener has prior experience with the particular nature of the sound, the *desired* characteristics correspond to a stored schema, which is accessed on the basis of the particular context, and potential preceding auditory percepts related to the same act of communication. In a reflection process[13], the listener decomposes the desired characteristics of the sound and identifies the desired (or expected) features. Correspondingly, the perceived characteristics are reflected and transformed into a set of perceived features. Finally, the judgment on the comparison of the desired and the perceived features constitutes the perceived quality.

In this context, the ***desired composition*** is defined as

> '[the] totality of features of individual expectations and/or relevant demands and/or social requirements.'
> (Jekosch, 2005b, p. 16).

[10] For simplification of the notation, this book uses the term *feature* sometimes to refer to the feature itself (i.e. its identification), but sometimes also to additionally refer to its markedness or magnitude (owing of course to the multidimensional 'nature' of many features). In case of possible ambiguities, a differentiation will be made.

[11] The modifying factors lead to an *adjustment* of the desired nature, see Figure 1.6.

[12] The *desired characteristics* (i.e. the expectation of the user) will be addressed in more detail in sections 3.7 and 5.4

[13] The reflection process may be more or less 'conscious', for example, depending on whether the communication situation is a directed one, as in the case of an auditory quality test, or an undirected one, as in the case of a natural telephone conversation.

1.2.2 Speech Quality Assessment

In speech quality tests, the test subjects are asked to describe their quality perception, or in other words, to *assess* the quality of a particular speech sample or system used for its transmission: they are sought to assign numbers to objects (Blauert and Jekosch, 2003). In the same way as it was mentioned for auditory listening tests in Section 1.1, the subjects make their judgment on a rating scale provided by the experimenter. This process of 'encoding' ultimately leads to a quality rating, the output of the speech quality assessment process depicted in Figure 1.6.

With reference to the above quality definition, 'assessment'[14] can be considered as a subordinate term to the more generic term 'appraisal': assessment assumes a measurement, where the test subjects' perception process is to some extent directed by certain test design factors like the rating procedure or the test scenario being used. Consequently, the directives given to the test subjects are part of the modifying factors depicted in Figure 1.6. They exert an influence on the desired nature of the perceived speech sounds as well as on the perception process.

In contrast, during an everyday conversation, undirected, individual perception takes place. This undirected perception is subject to the particular attention listeners or interlocutors pay to certain of the different signals they are faced with; only those the individual person considers relevant for the communication situation are actually perceived ($\equiv$ *stimuli*; Jekosch, 2005a). In the case of directed perception (as in a speech quality test), different stimuli or features may appear more relevant to the subjects than others. For example, in an everyday conversation among friends the content, that is, what is said, may be more important than in a laboratory listening test focusing primarily on the form of certain speech signals. Hence, one of the main goals of speech quality tests – that is, of auditory tests in general, see Section 1.1, – is to appropriately choose the test design and directives, in order to achieve valid and reliable measurement results, as far as possible reflecting the quality perception of users during undirected communication.

There are further correspondences to the more general auditory test situation discussed earlier: Figure 1.7(a) shows a listener in an auditory quality test situation, in analogy with the auditory test on particular auditory features of a sound depicted in Figure 1.1. The bold lines in Figure 1.7(a) represent the configuration as it is displayed in Figure 1.1. In this case, the test subject delivers a description of all or of certain features of the perceived sound. Such a test not focusing on quality judgments but on the description of perceived (quality) features, is referred to as *analytical* type of speech quality test. If an actual description of the perceived quality is sought, the process of comparison and judgment comes into play (Figure 1.6). The corresponding type of speech quality tests is referred to as *utilitarian* type (Figure 1.7(b); Hecker and Guttman, 1967; Quackenbush *et al*., 1988, pp. 15–16). In this type of tests, quality is typically rated on a unidimensional quality or impairment scale. The outcome of such tests is particularly interesting for applications like network planning or efficient preference testing to determine the best of different implementations of a network component or system, or of other types of systems, such as hearing aids.

At this point, a second dichotomy for auditory tests in addition to the *analytical – utilitarian* one can be mentioned: test methods can be distinguished according to

[14] '*Assessment*: Measurement of system performance with respect to one or more criteria. Typically used to compare like with like, whether two alternative implementations of a technology, or successive generations of the same implementation. . .' (Jekosch, 2000, 2005b, p. 109).

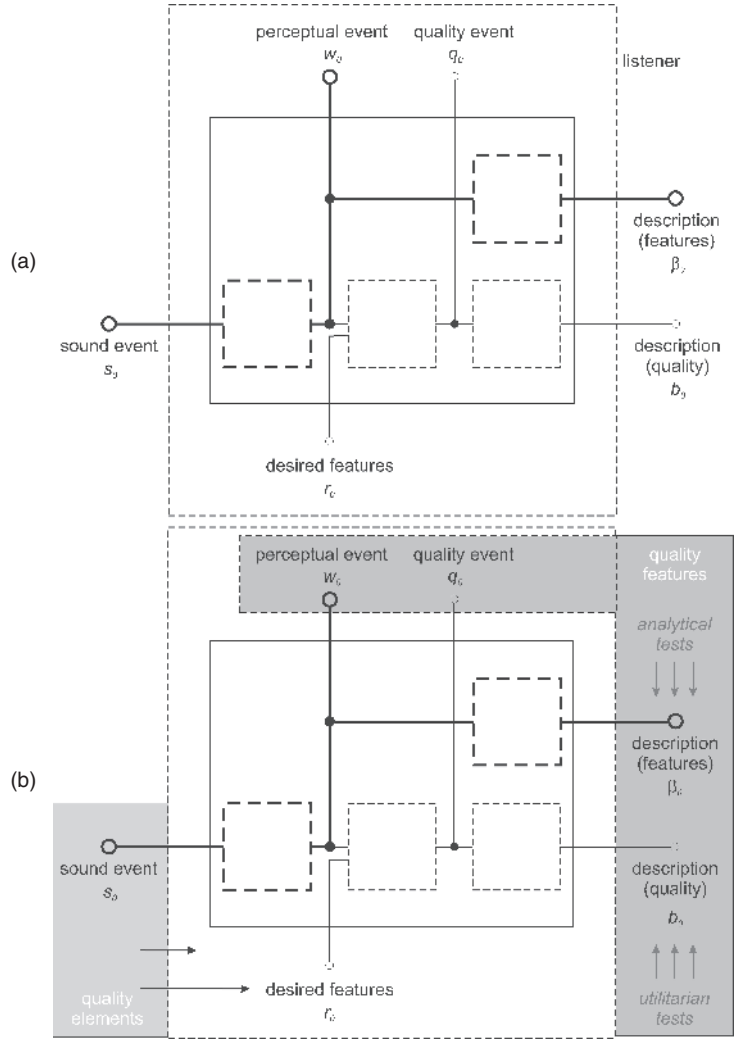

Figure 1.7: (a) Schematic representation of a test subject in a listening quality test (combining the concepts by Blauert, 1997; Jekosch, 2000, 2004, 2005b). (b) Additional illustration of the terms *quality elements* and *quality features* (see Jekosch, 2000, 2004, 2005b), and of the differentiation between *analytical* and *utilitarian* type speech quality tests (Hecker and Guttman, 1967; Quackenbush *et al*., 1988, pp. 15–16).

whether they are *subject-oriented*, that is, are carried out to gather information on human perception, or whether they are *object-oriented*, that is, investigate how the sound produced by or transmitted across certain systems is perceived (Letowski, 1989). In summary, the two dichotomies lead to four types of quality tests, as depicted in Table 1.1 (for an overview of some relevant analytical and utilitarian test methods see Chapter 2, Section 2.1).

Table 1.1: Quality tests: The four different applications resulting from the two dichotomies *analytical–utilitarian* and *subject-oriented–object-oriented* (combination of Quackenbush *et al.* (1988, pp. 15–16), and Letowski (1989). In the latter, the terms *heuristic* and *diagnostic* are used instead of *utilitarian* and *analytical*, with different definitions).

	Subject-Oriented	Object-Oriented
Utilitarian	Quality perception	Assessment of system quality
Analytical	Quality features and their perception	Quality features and acoustic or system correlates

The main emphasis of this book is on object-oriented tests, seeking to relate the quality perceived by users of certain speech communication networks to instrumentally measurable network parameters. Some of the research presented here can also be considered as subject-oriented, as aspects of quality perception are addressed[15].

1.2.3 Quality Elements

Up to this point, speech quality has been discussed mainly from a user's perspective. The remainder of this book discusses the implications of the design and planning of networks involving Voice over Internet Protocol (VoIP) transmission on speech quality as perceived by users. Although looking at quality from the perspective of the system, a network planner has to have the envisaged users in view if the network is to reach broad acceptance. If speech quality is regarded from the point of view of network planning or monitoring, the effect of certain network components or of certain parts of the network infrastructure on quality is of interest. For all the factors that have an impact on the quality perceived by the user and are in one way or another related to aspects of the design, technical realization or usage of the particular telecommunication system or service, Jekosch has proposed the following definition (see also Figure 1.7):

A ***quality element*** is the

> 'Contribution to the quality
>
> – of an immaterial or a material 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.'
>
> (according to Jekosch, 2005b, p. 22, modified from DIN 55350, Part 11).

From the point of view of perception, a complementary definition can be given for the quality-relevant perceptual features: a ***quality feature*** is

[15] In general, it has to be noted that tests that are *a priori* subject-oriented may also provide object-oriented information and vice versa, depending on the interpretation of the results.

> '[a] recognized and designated characteristic of an entity that is relevant to the entity's quality.' (Jekosch, 2000, 2004, 2005b, p. 17).

The dichotomy of *quality features* and *quality elements* is used as an important tool to structure the content of this book: In Chapter 3, the quality elements of speech communication networks involving VoIP are summarized, and an overview of the related quality and quality features is provided. The author's research on the impact of the most important of these quality elements on quality and quality features is described in Chapters 4 and 5. Therefore, we employ simulation tools to generate the quality elements in a laboratory situation, as described in Appendix B On the basis of the description of the quality elements and the knowledge of the related quality, modeling approaches will be discussed that quantify the relationship between the quality elements and the quality features.

1.2.4 Speech Quality and Quality of Service

Different terms are typically used in the literature to refer to the (speech) quality of speech communication systems (see Möller, 2000, p. 11): 'Mouth-to-ear' or 'end-to-end' quality refers to the quality of the entire system from the mouth of the talker to the ear(s) of the listener. The term 'integral quality' is used when the quality due to the totality of quality dimensions (or features) is considered. Another term frequently used in the literature is 'overall quality', which is used synonymously for mouth-to-ear by some authors, and for 'integral quality' by others. To avoid this ambiguity[16], the term *integral quality* will be used in this book in cases where the quality resulting from the totality of quality features is referred to. When referring to the (integral) quality instantaneously judged by subjects – for example, during listening to a speech signal degraded by time-varying distortions – the term *instantaneous quality*[17] will be employed (see Gros and Chateau, 2001; ITU–T Rec. P.880, 2004)[18]; the (mathematically obtained) time average of the instantaneous quality will be referred to as *average instantaneous quality*. Consequently, in this book, *integral quality* is the quality subjects relate to an entire conversation or speech passage, taking the history and evolution of the conversation or passage into consideration; it corresponds to the final quality judgment obtained at the end of the conversation or listening sample. More generally, in this book the term *speech quality* refers to the quality perceived in a conversational situation. If a listening-only situation is referred to, the term *speech transmission quality* is used.

Speech quality is only one of the different factors ultimately determining the acceptability of a telecommunication service. To refer to all aspects related to the acceptability of a service, the term *quality of service* (QoS) is typically employed following the definition provided in ITU–T Rec. E.800 (1994):

Quality of Service (ITU–T Rec. E.800, 1994):

> 'The collective effect of service performance which determines the degree of satisfaction of a user of the service.'

QoS, according to this definition, is composed of four aspects, namely, *service support*, *service operability*, *serveability* and *service security* (ITU–T Rec. E.800, 1994). *Service*

[16] Adopting the terminology and argumentation used by Möller (2000).

[17] Instead of 'instantaneous integral quality'

[18] Here, the terms *instantaneous judgment*, *time-varying speech quality* and *continuous speech quality evaluation* are used.

support is related to services like directory assistance or technical assistance; *service operability* refers to how easily a service can be operated by a user; *serveability* comprises aspects like the accessibility and retainability of a service (i.e. how faithfully a service can be obtained and be provided for a given period of time), as well as the level of speech quality it supplies; finally, *service security* is concerned with issues such as the protection against unwanted access to, or monitoring of, the transmitted data by third parties (e.g. 'spoofing' or 'sniffing' of data in packet-based systems[19]).

QoS can be looked at from two different perspectives: that of the service provider, and that of the user of the service. In order to reflect the user's perspective of QoS, the QoS framework has recently been complemented by the framework of *quality of experience* (QoE) (ITU–T Delayed Contribution D.197, 2004). In this context, an implicit distinction is made between the quality element side and the quality feature side. According to the proposed redefinitions, QoS now refers to the network, that is, the quality element side, and QoE refers to what the user actually perceives of QoS:

Quality of Service (revision proposed in ITU–T Delayed Contribution D.197, 2004):

> 'The collective effect of *objective* service performance which *ultimately* determines the degree of satisfaction of a user of the service.'

Quality of Experience (new definition proposed in ITU–T Delayed Contribution D.197, 2004):

> 'A measure of the overall acceptability of an application or service, as perceived subjectively by the end user.'

In order to reduce complexity, the term QoS will be used throughout this book, pointing out whether the user's or the service provider's perspective is considered by employing the concept of *quality elements* (service provider) and *quality features* (user).

A taxonomy integrating speech quality into the framework of QoS for telephone services was developed by Möller (2000). It was recently modified in order to better match a corresponding taxonomy of QoS developed for telephone-based services involving spoken dialogue systems (Möller, 2005a,b). According to this taxonomy (see Figure 1.8), QoS is composed of three factors, which constitute the quality elements of the service: *speech communication factors* concern the actual communication between the (two or more) interlocutors over the speech transmission system. According to the above QoS definition from ITU–T Rec. E.800 (1994), these factors contribute to the serveability of the system. *Service factors* cover service support and security, and parts of service operability and serveability. They summarize the impact of the service characteristics, however, excluding the speech communication factors. *Contextual factors* do not directly form a part of the QoS definition provided above. They relate to nonphysical aspects of the service, like the costs for the user (investment, monthly and per-call charges, etc.), and the contract conditions (e.g. the contract period or the period of notice).

Similar to the modifying factors depicted in Figure 1.6, the contextual factors have an impact on the user (i.e. on his attitude, emotions or motivation), and hence on the desired nature of the transmitted speech. Because of the user's role for all the intermediate levels of appraisal shown in Figure 1.8, and for the ultimate service acceptability, the user is

[19] Sniffing: Network-traffic monitoring, possibly with access to the actual data content. Spoofing: The other endpoint is deceived by his interlocutor who pretends an incorrect identity. Examples are email- or caller-ID-spoofing. On a lower level, network traffic could, for example, be rerouted to a malicious recipient.

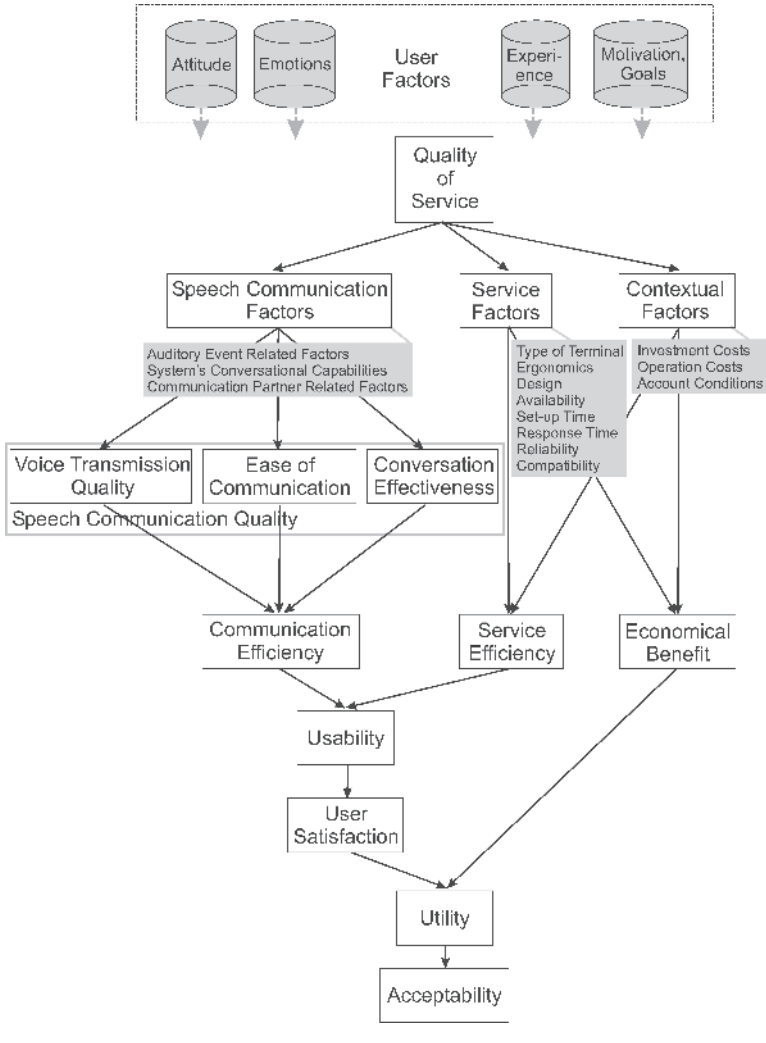

Figure 1.8: Quality of Service (QoS) taxonomy developed by Möller (2005a,b). The gray frame in the middle of the picture highlights the three constituents of speech communication quality (voice transmission quality, ease of communication and conversation effectiveness).

depicted on top of the figure. Different user factors are shown that contribute to the quality perception, like his attitude or goals.

The speech communication factors contribute to the quality features and to the corresponding speech quality[20] perceived by the user. They result from the quality elements, which in turn depend on the decisions taken during the network planning phase. The speech communication factors can be further subdivided into three constituents:

[20] With reference to Möller (2000), the term *speech communication quality* is used in Figure 1.8 instead of *speech quality* as it is used in this book.

Voice transmission quality considers all those quality elements that ultimately lead to an impact on quality effective already in a pure listening situation, that is, to factors like noise on the line, a loss of signal level introduced by the transmission, or transmission errors like packet loss. In this book, voice transmission quality is referred to as *speech transmission quality*.

Conversation effectiveness refers to the quality elements affecting the conversational capabilities of the system, that is, factors only relevant in a conversational situation. Examples are talker echo and transmission delay.

Ease of communication concerns conversation partner related factors, which also include a potential adaptation to an adverse acoustic environment (i.e. Lombard speech, etc., see Section 1.1.2.6).

As already stated in Section 1.1.2, all components of speech (communication) quality together determine the communication efficiency[21], that is, the resources a user has expended in relation to the accuracy and completeness with which he/she has performed a particular communication task (ETSI Guide EG 201 013, 1997), as well as his/her level of awareness of it. The service efficiency describes the resources expanded during the usage of the service, disregarding all issues related to the actual communication. Together, communication and service efficiency determine the usability of the service, that is, the aptitude of the system to be used for completing a specific task. Utility, the next step in the taxonomy, can be considered as the trade-off between the system's usability and the related costs. Ultimately, the acceptability measures the ability of the service to find acceptance. It is typically determined as the relation between the number of actual users and the number of potential users. It has to be emphasized again that it is the user who dictates every step in the chain, from the system- and service-related factors to the service's acceptance.

An example from the domain of early VoIP networks may serve for illustration: the first means to establish an internet telephone connection for a user not technically inclined were based on software tools quite complicated to handle. Consequently, the service[22] factors were counterproductive for a high usability. Moreover, communication efficiency was very low, as time-varying degradations like packet loss and delay jitter on the one hand, and more importantly severe amounts of mean transmission delay on the other, made effective communication difficult. The potential economical benefit of making long-distance calls free of charge was counterbalanced by the small amount of potential conversation partners, who had to be equipped with the same software client, and, in particular, be online at the same time: the serveability was very low. As a consequence, during the early tests carried out in the framework of this book, the number of subjects with prior experience with Internet telephony was rather low (around 4–8%, in 2001), slowly increasing to around 15–20% for later tests (in 2004), owing to both the improvement of the serveability and service operability of Internet telephony 'services', and their increasing recognition by potential users (see Chapter 4).

[21] In Section 1.1, the term *communicability* was used to describe the capability of a particular conversation to lead to a certain level of communication efficiency. Correspondingly, the *communicability of the system* summarizes both the voice transmission quality and the conversation effectiveness provided by the system.

[22] Early days internet telephony cannot be regarded as actual service, since the existing network structure and Internet service was 'reused' on the basis of software clients not stemming from dedicated service providers assuring aspects of QoS.