17

Sound Quality

The concept *quality* has two meanings. 'Quality' is used in this book as a synonym for 'excellence', to grade or rank objects on a subjective scale of preferability such as 'good–poor', based on some explicit or implicit criteria. The other common meaning is related to categorization by type or class of objects. When two observations or entities cannot be compared on the same (metric) scale they are said to be qualitatively different. Such category-related sound quality pertains to perceived features, attributes, factors, dimensions, or properties of auditory events, such as loudness or roughness. However, in this book, the term 'sound quality' is limited to the meaning involving preferability or acceptability.

The inherent topic of the discussion on quality after its definition is evaluation. We experience some objects or states of the world as more desirable, valuable, positive, appealing, useful, or what have you than others. Although often weakly formulated and structured, such conceptions and rankings help us to set goals of action and to find better solutions to problems at hand. A widely used term in this context is *quality of experience* (QoE) (Le Callet *et al.*, 2012), which denotes the overall acceptability of an application or service as perceived subjectively by the end user.

The theory of psychoacoustics, discussed in Chapter 8, uses the human as a simple metering device, where sound events evoke auditory events with attributes that can be measured using psychoacoustic techniques. In psychoacoustics, expectations, mood, and other cognitive factors of individual subjects are minimized when the values of attributes are measured. In the context of sound quality, cognitive factors can no longer be disregarded, since 'quality of sound' means the suitability of a sound to a specific situation, and such suitability cannot be judged without cognitive functions. The same sound may produce different sound quality in different contexts, depending on the mode of operation and the expectations of a subject. For example, higher *intelligibility of speech* improves the sound quality in mobile phones, but the ability of a worker to concentrate in an open-plan office is impaired by intelligible speech from neighbouring cubicles. The properties of sound can thus have either negative or positive effects on sound quality.

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

Although the interpretation of 'sound quality' varies widely in different domains of acoustics, audio, and speech, the concept has come into increasingly widespread use (Blauert and Jekosch, 1997), and in one form or another may be considered generally applicable to all sounds that humans encounter.

17.1 Historical Background of Sound Quality

The concept of sound quality has a relatively long history of emergence. Probably the oldest sounds associated with a quality rating have been human speech and singing, then theatre and music-making, including musical instruments. The first quality rating factors were subjective and implicit, based on emerging a esthetic factors and how the sounds had a desired effect in practice. The centuries-long evolution of present-day acoustic musical instruments is an early example of 'product sound quality' development by gradual experimentation. The acoustics of concert halls and other performing spaces is another similar case, where, until the beginning of the 20th century, sound quality evaluation had little, if any, scientific basis.

The development of physics and related mathematics started to enable a relationship between objective factors and subjective quality of sound. The sound spectrum (including sounds from musical instruments) and related hearing processes were studied in the late 1800s and early 1900s by Helmholtz (von Helmholtz, 1954) and the basics of concert hall acoustics by Sabine (Sabine, 1922). Inventions in electronic communications – the telephone, gramophone, and radio – had a strong impact on our understanding of sound quality. Particularly in telephone transmission, there was a practical need to know how the distortion caused by, and the limitations of, the early technology affected the intelligibility of speech and the recognition of individual speakers. Starting from the 1920s, Harvey Fletcher and the Bell Laboratories research group (Fletcher, 1995) made fundamental studies that laid the groundwork for engineering psychophysics (psychoacoustics) as a systematic experimental science by making quantitative formulations for articulation and intelligibility of speech. Subjects in listening tests were used as 'meters' to 'measure' desired factors in speech transmission. This was the basis, for example, for setting the standard of the telephone bandwidth that is still in use today.

The goal of high sound quality was clearly necessary in sound reproduction using microphones, tape recorders, amplifiers, record players, and loudspeakers, nowadays called audio techniques. To maximize the a esthetic experience of reproduced music, the HiFi (high fidelity) movement emerged. It was partly an engineering-oriented attempt to minimize distortion and colouration of sound in a reproduction channel and partly a highly subjective 'golden ear' and 'expensive gadget' hobby. Only the emergence of audio coding in digital audio at the end of the 1980s forced the modelling of auditory perception to become a central engineering challenge. In a similar manner, multi-channel and 3D sound reproduction have elevated studies and modelling in spatial sound perception to a higher scientific and engineering level.

Since the 1980s, the investigation of noise control techniques has been increasingly directed also towards qualitative aspects, that is, *noise quality* (Marquis-Favre *et al.*, 2005b), not only to simple quantitative measures, such as the A-weighted sound level for estimating the risk of noise-induced hearing loss. Earlier studies on the subjective effects of noise also exist, but the signal-analysis-based approach was introduced to understand such effects as annoyance caused by noise. This gradual shift of focus is natural, since, in many cases, hearing loss is not the primary problem anymore and quality-of-life aspects are found to be increasingly important.

The notion of sound quality, applicable to both positive effects (music and speech) and negative effects (noise), finds a generalization in the concept of *product sound quality*

(Blauert and Jekosch, 1997). In its most general sense, this concept also covers traditional sound quality aspects, since a concert hall, a musical instrument, a musical performance (even a music composition), audio equipment, a noisy working machine, or a car making noise are equally 'products' in the wide sense of the term. In all these cases, the goal is that the sound of a product meets the needs and requirements of the customer at hand and optimizes the sound quality factors against the cost of the product.

17.2 The Many Facets of Sound Quality

The discussion above brings up the question of whether a universal approach to assess or evaluate sound quality exists. Furthermore, if this is not possible, what are the different scientific methods and engineering techniques to evaluate sound quality? Different domains of sound quality, as discussed above, turn out to be truly different, so finding a simple general model of sound quality does not seem possible. Perhaps the only common factor is the listener: we must start by using subjects in listening experiments to get data on the factors affecting sound quality and, based on these data, build models and theories of sound quality.

The formalization and quantification of the concept of sound quality may raise conflicting opinions. On the one hand, a subjectivist believes that the experience of quality is highly individual and there are no grounds for generalization and formalization, while on the other hand, an objectivist opines that a coherent general theory for measuring sound quality can be developed. Both views are partly right and partly wrong. A further conceptual discussion can help to understand these issues more thoroughly.

How sound affects us can be categorized as follows:

- *Physical and physiological effects*. Only a very intense sound (above 120 dB) can have a considerable physical effect. Physiologically, the most important factor is the risk of hearing impairment, which typically occurs after long exposure to levels above 85 dB (A-weighted daily equivalent; see Section 19.3 for more details). From this point of view, a criterion for high-quality sound design is to keep the sound level low enough not to harm humans, animals, or nature. Compromises are needed when the cost of noise control becomes excessively high.
- *Information and knowledge.* That sound transmits information and knowledge is a desirable and valuable characteristic, although too much exposure to information can lead to negative effects. Information conveyed by environmental sounds is important for orientation in everyday life. Thus, it is desirable that, for example, appliances and vehicles make sounds that inform us about their existence and functioning, as long as the negative factors of this sound do not exceed the positive ones. The information aspect is most essential in speech communication, where the speech quality provided in transmission techniques is needed for undistorted transfer of information.
- A *esthetic and emotional effects.* These are the most demanding aspects of quantifying sound quality from a scientific and engineering point of view. Using listening experiments and statistical analysis, it is always possible to seek factors affecting the perceived quality of sound. Reactions to a sound are often strongly dependent on the sociocultural background of the subject, the context of presentation, and various other factors. For example, many objective factors from speech transmission as well as from high-quality sound reproduction that affect the a esthetic and emotional aspects of the percept have been identified as also

being properties that make up good musical instruments. In the context of noise, we may study the psychological and emotional factors that have a negative effect on our quality of life.

In the discussion below, sound quality is related primarily to the informational and a esthetic aspects of sound. Physiological (or physical) effects are considered only when needed.

The following sections discuss the concept of sound quality from a methodological point of view and in different problem domains, such as speech transmission, concert hall and auditorium acoustics, audio reproduction of sound, noise quality, and the general concept of product sound quality.

17.3 Systemic Framework for Sound Quality

Sound quality is primarily subjective. The most reliable method to study an informative feature or the a esthetic value of a sound is to conduct psychophysical experiments with a group of human subjects, or *assessors*. Often, this is an implicit activity, like in engineering prototyping or product development, where engineers apply their intuitive or introspective knowledge on what constitutes quality of sound. Sometimes the attributes of sound that make it high quality are obvious and sometimes not. In general, understanding the relations between attributes of physical sound and sound quality requires systematic experimentation and statistical analysis of the gathered data. When optimizing the sound quality of a product for a specific market, an extensive experimental basis is necessary.

The problem with subjective evaluation is that the required experiments are typically very laborious and time consuming. The experiments should be carried out in conditions that correspond to real usage of the sounds or related products. Thus, it would be much easier and less expensive to use objective criteria or models for sound quality. Ideally, a computational model provides an estimate of quality from signal analysis which correlates well with subjective data. The development of such a model (Figure 17.1) is typically based on proper subjective listening experiments from which data are collected and analysed statistically. Using various techniques, the factors and features of sounds that best explain the subjective behaviour are sought. The computational model can be a simple linear regression model or a more complex non-linear model, such as a neural network trained to map the feature parameters to quality indices. Figure 17.2 characterizes the general structure of computational sound quality models.

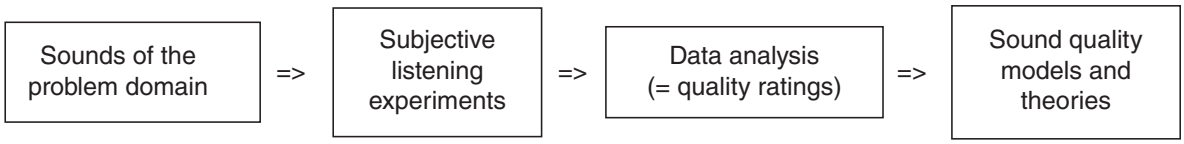

Figure 17.1 The development of sound-quality models and theories.

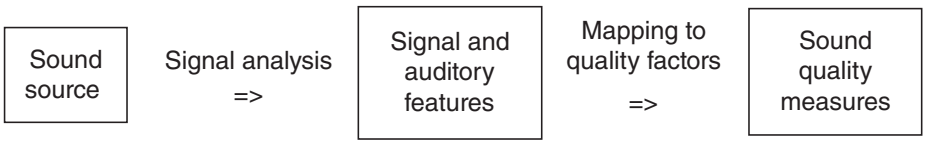

Figure 17.2 A general structure of a computational sound quality model.

The development of such models is also a tedious and demanding task, and objective models can never fully replace subjective evaluation. The advantages of objective models lie in efficiency, rapid evaluation, and repeatability of results. The development of such models may also yield a deeper understanding of the phenomena than subjective results. The disadvantage of objective models is that they never take into account all factors in full detail, and thus their domain of validity (meaning that they correlate well with subjective results) is limited.

17.4 Subjective Sound Quality Measurement

In speech and audio techniques, the final 'truth' about the sound quality achieved lies in the general opinion of the larger public. A number of techniques have been developed to measure the quality of sound associated with a set of sounds. The sounds in the set may be produced, for example, with a set of different audio systems, handheld devices, vacuum cleaners, or concert halls. In the technical development of systems, the choice of parameters for the system often changes the sound output of the system in quite an unpredictable manner. Typically, the only possible method to evaluate the differences between the sounds is to organize subjective tests.

Such tests can be conducted in various ways. A basic approach is to ask the subjects to sort sounds in order of preference, or to ask them directly to rate the quality of sound. The quality can be rated 'in general', or, alternatively, with associated a certain aspect of sound, such as 'rate the quality of speech in terms of intelligibility'. This is perfectly adequate for many applications, and many of the psychoacoustic techniques described in Chapter 8 can be used to measure the value either of the overall quality or of a specific attribute of sound.

In some cases it is not known in advance in which perceptual dimensions the sounds being studied differ from each other. In such cases, *descriptive sensory analysis* techniques can be used to characterize the dimensions, as described in Section 8.8. Furthermore, two concepts often employed in the context of sound quality are the mean opinion score (MOS) scale and the MUSHRA (multiple-stimulus hidden reference with anchors) method for scaling. These methods were not discussed in detail in the chapter on psychoacoustics, and thus they are briefly reviewed here.

17.4.1 Mean Opinion Score

The *mean opinion score* (MOS) value is often used to quantify the sound quality in general or in terms of a specific aspect of sound. Additionally, separate MOS scales have been defined for cases where the degradation of quality or the relative quality is measured. The MOS is a subjective measure obtained using psychoacoustic testing. The selection of the scale and the methodology to measure it depend on the task. There are a number of published and even standardized methods to measure MOS, for example ITU-T P.800 (1996), and this section merely provides a brief introduction to the topic. Detailed methods to measure the MOS in different cases and references to standards for MOS measurements are covered by Bech and Zacharov (2006).

The MOS scale consists of a numerical scale ranging from 1 to 5, either as whole numbers or with fractional intervals, and a descriptor associated with each number. Thus, the scale links a discrete numeric scale to verbal categories. It also enables the measurement of the quality on a

Table 17.1 An example of MOS and DMOS numeric and qualitative scales.

Value	Quality (MOS)	Impairment (DMOS)
5	Excellent	Imperceptible
4	Good	Perceptible, not annoying
3	Fair	Perceptible, slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

Table 17.2 An example of a CMOS numeric and qualitative scale.

Value	Categories (CMOS)
3	Much better
2	Better
1	Slightly better
0	About the same
−1	Slightly worse
−2	Worse
−3	Much worse

very broad scale without a reference sound. The MOS scale is also called an *absolute category rating* (ACR), as defined in ITU-T P.800 (1996) and ITU-T P.910 (2008).

There are many variants of the MOS. The direct measure of quality is simply called the MOS value, and the measure of impairment between reference and test cases is called the *degradation mean opinion score* (DMOS). Example MOS and DMOS scales are presented in Table 17.1.

The *comparative MOS* is the third basic MOS scale, and it is typically defined to have values between −3 and 3. It can thus be used where given test case can be ranked better than reference. This can occur, for example, in speech enhancement techniques. An example of a CMOS scale is give in Table 17.2.

Since MOS contains the word 'mean', it is clear that some kind of average of a large number of subjective ratings is taken. The statistical analysis of the results is an important part of the work. Without proper analysis of the results, the validity of the results cannot be shown. Some general concepts from statistical analysis were introduced in Section 8.9, but a detailed description of relevant methods is beyond the scope of this book. The reader is again referred to Bech and Zacharov (2006) for techniques to analyse MOS values.

17.4.2 MUSHRA

The *multiple-stimulus hidden reference with anchors* (MUSHRA) method is often used in sound quality measurements where multiple stimuli are scaled during the same task onto one MOS scale (ITU-R BS.1534-1, 2003) and was originally developed to test speech and audio codecs. MUSHRA provides reliable results if the sample differences are large, such that the

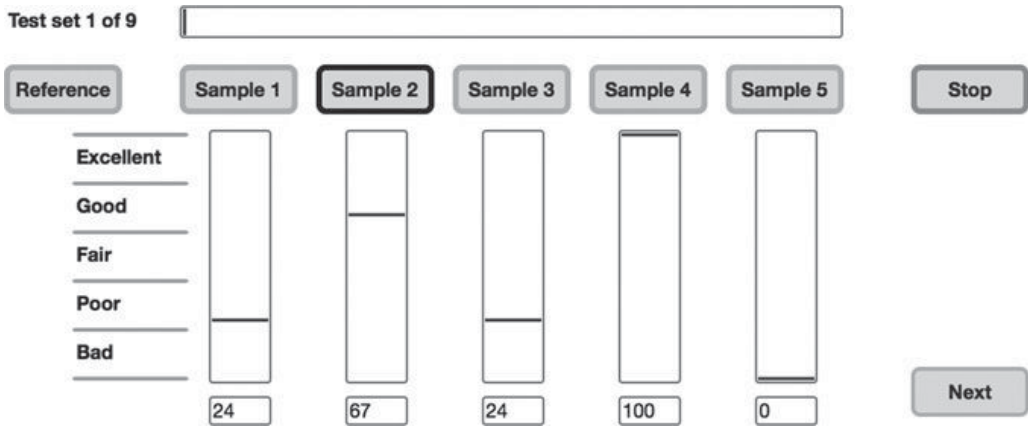

Figure 17.3 A graphical user interface for multiple stimulus hidden reference with anchors (MUSHRA) testing. Courtesy of Tapani Pihlajamäki.

listeners can perceive them without careful concentration. If the differences are small in the test cases, more accurate methods should be used, such as the method described in ITU-R BS.1116-1 (1997).

In a MUSHRA test, subjects listen to different sound samples: the same programme material that has been processed differently. The samples contain a reference known to the subjects, several test samples, a hidden reference, and anchors. The hidden reference is simply an unmodified copy of the original, and the anchors are modified versions of the original, processed by, say, low-pass filtering up to 3.5 kHz, the addition of noise, or the loss of packets during transmission. The subjects can freely listen to the samples, switching from one sample to another by pressing buttons on a user interface, as shown in Figure 17.3. On making a choice, the current sample quickly fades out and the chosen sample fades in. The new sample is played back, continuing from the temporal position where the playback of the current sample ended, so that the programme material continues playing without notable interruptions when the sample is changed. This helps to reveal differences between the samples. The test samples, the hidden reference, and anchors are positioned randomly on the user interface. The subjects rate the samples based on a given task, such as 'rate the quality of reproduction', using the sliders in the user interface.

The number of samples is recommended to be less than 15 (including anchors and references) and the perceptual differences between the samples should not be too small, since otherwise the comparison may be too challenging. The multiple-stimulus methodology has become popular, as relatively reliable results can be obtained faster than with pair-wise comparisons.

The method has also been criticized. For example, if the samples differ from each other in multiple dimensions, the listeners may be confused as to how to rate them. Let us imagine that the sample set consists of audio content with a reference case of 5.1 audio, stereophonic and monophonic down-mixes, and low-pass-filtered versions of the 5.1 audio content. The listeners then have to judge the degree of degradation of sound quality in two dimensions, since both spatial and timbral aspects vary. This may result in data that are difficult to interpret.

Interested readers are referred to Sporer *et al*. (2009) for details on running MUSHRA tests and corresponding data analysis. A revised version of MUSHRA is currently (2014) being standardized, and the final version will include elaborated methods for data analysis and test conduction.

17.5 Audio Quality

Section 14.2 discussed audio content production, and it was noted that the final modifications and final approval of a piece of audio content are conducted typically in the mastering studio. Thus, perfect authenticity in sound reproduction would require an identical listening set-up in a room identical to the mastering studio. Fortunately, this is not necessary, since sufficient quality of the audio experience can also be obtained in other listening conditions. The acoustic differences of listening rooms have an effect on the quality, but the ability of humans to adapt to different acoustic conditions mitigates the differences, as discussed earlier. Impairments in audio devices, on the other hand, often have a significant effect on the quality of the experience.

Historically, the bottleneck in audio quality has been the storing and transmission systems, such as gramophones, vinyl players, and cassette decks. The concept of *high fidelity* was developed to measure and minimize different deviations from perfect responses, such as linear and non-linear distortions. Nowadays, digital transmission has practically removed such problems, and currently the biggest bottlenecks are the microphones and loudspeakers. However, new challenges have emerged with perceptually based lossy audio codecs.

17.5.1 Monaural Quality

Let us first consider deviations that are already audible if only one channel is listened to. The traditional measures affecting sound quality are listed below, and they have already been introduced earlier in this book:

- *Magnitude response* and the perceptual effects of deviations from the ideal (see Sections 4.2.3 and 11.5.1).
- *Phase delay and group delay* and their perceptual consequences (see Sections 4.2.4 and 11.5.2).
- *Non-linear distortion* measures of signal differences between the output and the input with simple signals (see Section 4.2.5).
- *Signal-to-noise ratio* (see Sections 4.2.6 and 17.6.2).

Other traditional concepts concerning analogue audio techniques are, for example, fluttering and rumbling caused by mechanically rotating devices, *dropouts* in magnetic tapes, and pops and crackles in grooves of rotating discs. The advent of digital audio has brought new types of quality degradation that the traditional quality measures fail to detect. These degradation include, for example, quantization noise (see Section 3.3) and aliasing in the time–frequency domain processing of audio (see Chapter 15).

17.5.2 Perceptual Measures and Models for Monaural Audio Quality

The '13-dB miracle' example discussed in Section 15.3.1 shows that human capabilities in listening have to be taken into account when measuring audio quality. This leads directly to the idea of using auditory models (Chapter 13) for quality evaluation. If we can simulate the functioning of the auditory system in all of its relevant stages, theoretically we are able to implement a computational model of hearing that can explain auditory perception when listening to any sounds. Unfortunately, the current status of modelling is far from this, although

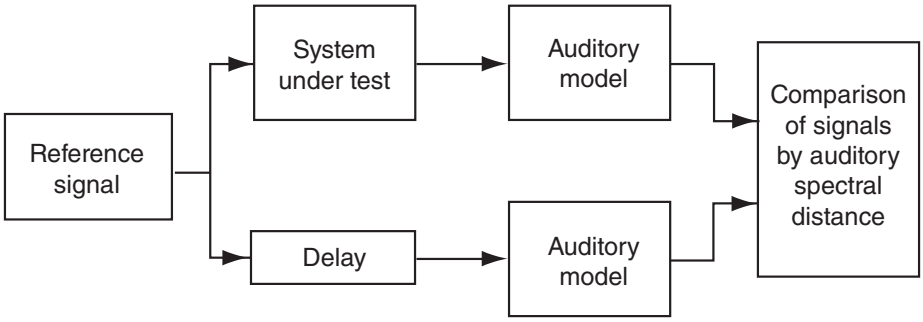

Figure 17.4 The principle of using auditory models to estimate quality degradation in an audio system.

the auditory modelling results explain much better the perceived audio quality of digital codecs than traditional distortion-based measures do.

A method of using auditory models in audio quality evaluation is shown in Figure 17.4; this was suggested by Karjalainen (1985), following pioneering work by Schroeder *et al*. (1979). A reference signal is fed to two identical auditory models, through the system under test to one and suitably delayed to the other so that the two auditory model outputs are aligned in time. The auditory models estimate auditory attributes for both signals, such as *auditory spectrum*, *pitch*, and *localization*. A distance measure is computed between the estimated attributes, from which the degradation of quality caused by the tested system is evaluated. For example, if the auditory spectra differ by more than 1 dB in any critical band, the degradation is audible.

The principle in Figure 17.4 has the advantage that any audio signal can be used as the reference signal. Thus, the audio quality produced by the system can be estimated with real signals, such as music and speech, and not only with simple test signals, such as sinusoids or impulses.

A relatively simple auditory model is presented in Figure 17.5; this implements the principle of audio quality evaluation shown in Figure 17.4 (Beerends and Stemerdink, 1992). The model is based on computing the specific loudness for each critical band using time-windowing, DFT-transfer, and frequency warping, in a similar manner to the Matlab script shown in Section 13.1. The specific loudness spectra are compared, and a time-dependent estimate of the audible difference is obtained, which is then averaged to obtain the final estimate of quality degradation. There are many parameters in the model that are selected to match the estimate with results obtained in MOS listening tests. The measure is called the *perceptual audio quality measure* (PAQM).

The *perceptual evaluation of audio quality* (PEAQ) method (Thiede *et al*., 2000) is an evolved version of the PAQM computation. PEAQ includes an auditory model implemented either with the computationally lighter DFT processing or with the temporally more accurate filter-bank processing. The computation to estimate the MOS has two stages. First, an auditory model is used to compute auditory features, such as excitation patterns, specific loudness patterns, and modulation patterns, for each auditory frequency band. The features are computed for both the reference signal and the signal reproduced using a device and the simulated or real network. The estimated features and the metrics describing differences in the features between the reference and test cases are provided as an input to a *cognitive model*, which then estimates the MOS value. The cognitive model is, in principle, a pattern recognition algorithm, such as an artificial neural network. The algorithm is trained using a large set of examples from real devices and acoustics conditions which have been analysed by a large panel of listeners.

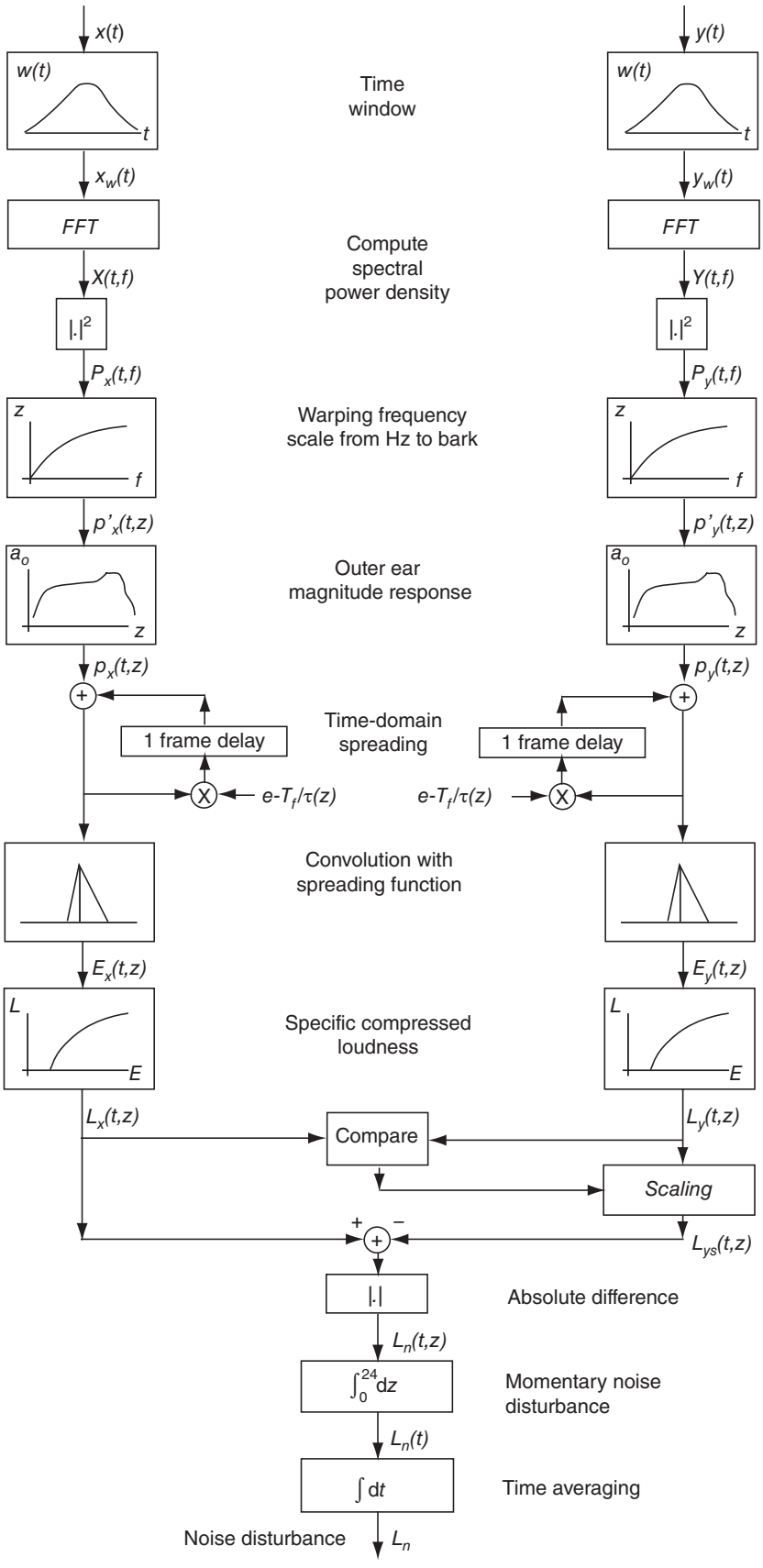

Figure 17.5 The computation of the perceptual audio quality measure (PAQM). Adapted from Beerends and Stemerdink (1992), and reproduced with permission from The Acoustical Society of America.

As the method is trained with a particular set of examples of reference and impaired audio samples, the method has a specific area of applicability. PEAQ has been used successfully to analyse audio quality with masking-based audio codecs without the effect of the listening room. For example, PEAQ estimates low MOS values for signals containing jitter or a slight deviation in the sampling frequency. However, PEAQ cannot operate with acoustically captured signals. It is probable that PEAQ would estimate a degraded MOS if the ear canal signals of a listener in a domestic environment were compared to the ear canal signals of the audio engineer in the mastering studio, although the listener would be satisfied with the audio quality.

Even though current auditory models are not yet able to evaluate all flavours of sound quality, it is clear that similar auditory models will eventually replace traditional audio quality measures.

17.5.3 Spatial Audio Quality

As discussed in Chapter 14, several methods have been developed to reproduce or to synthesize the spatial characteristics of sound as well. The overall advantage in the sound quality resulting when upgrading from monophonic reproduction to spatial audio with multi-channel loudspeaker set-ups has been researched by Rumsey *et al.* (2005). The quality of sound was measured with monophonic, stereophonic, and 5.1 surround reproduction with produced audio content. In addition, the quality was also measured with low-pass timbral degradations of the programme material. It was found that timbral quality corresponded to about 70% of the overall quality, whereas spatial reproduction corresponded to the remaining 30%. An interesting finding was that, when low-pass timbral artefacts were present, the spatial reproduction resulted in no advantage over monophonic reproduction.

This research was conducted with produced audio content, where the reference was 5.1 surround audio reproduction. In the most general case, the reference should be a real acoustic scenario, such as a case where sound sources are located around the listener in 3D, with natural reverberation arriving from all directions to the listener. Should the reproduction of such a case be targeted and the obtained authenticity measured, a major problem emerges. The perception of the original scenario must be compared to reproduction in a listening room, requiring that the subject is moved between the listening room and the original room to make the comparison. Unfortunately, the human auditory memory is too short to make accurate comparisons with delays of more than a few seconds between perceptions. Hence, such an approach is not feasible. In some cases, the reference scenario can be synthesized in an anechoic chamber using a large set of loudspeakers, and then the scenario can be recorded using microphones and reproduced using the same loudspeakers (Vilkamo *et al.*, 2009). This enables direct comparison, although the synthetic nature of the reference casts some doubts on the generalizability of the approach.

A considerably simpler case is the evaluation of spatial sound synthesis methods without aiming for reproduction of the recorded acoustic conditions. In synthesis methods, a specific spatial attribute is to be controlled, and the reference case can also often be created. For example, in virtual source positioning methods, the synthesized position of the virtual source can be measured relatively simply by using appropriate psychoacoustic test methods. The possible degradation of overall quality should also be measured, which is relatively simple to do, since the reference case can be generated by positioning a real source in the panning direction, thus enabling the comparison of the virtual and real sources in listening tests.

Auditory-modelling-based objective quality measurement systems could potentially solve the problem of the impossible comparison of the reproduced spatial sound to the original conditions. Binaural models, which were introduced in Section 13.5, have been suggested for use in measuring the quality of spatial sound reproduction. Indeed, in limited cases, auditory models are applicable for evaluating spatial audio quality (Blauert, 2013b). Although a large number of binaural models have been proposed, the explanation of human perception of spatially complex scenarios still seems challenging. The research in this field is thus incomplete. The need for models for the analysis of spatial audio in industry exists, since an effort to extend the PEAQ standard to the measurement of quality over stereophonic and multi-channel audio formats has been initiated. Unfortunately, the current results in the process are not promising (Liebetrau *et al*., 2010), and the process is on hold as of now (2014).

17.6 Quality of Speech Communication

This section covers some basic concepts and aspects related to speech communication over different channels which are key factors of sound quality in the context of speech. The discussion touches different layers of speech quality: we start from intelligibility and discuss some slightly higher-level concepts as well. The field is wide, as speech quality is needed in many applications with different needs, such as telephony, voice-over-internet, radio, and public address. The discussion merely scratches the surface of the topic.

Speech intelligibility (Blauert, 2005, Chapter 7; Quackenbush *et al*., 1988; Steeneken and Houtgast, 1985) is a property referring to how well the meaning of a spoken message is transmitted to a listener. As such, intelligibility depends on three factors:

- the ability of the *speaker* to produce a message with acoustically and linguistically clear contents;
- how well the *transmission channel* is able to transmit the message; and
- how well the *listener* is able to receive and analyse the message.

In speech communication without electrical devices, the transmission channel is the acoustic path from the lips of the speaker to the ears of the listener. When electrical devices are used, a microphone and a headphone and loudspeaker are needed together with an electrical transmission line, possibly with coding and transmission technologies.

The technical interpretation of *speech intelligibility* is related to the attributes of the transmission channel. To measure speech intelligibility subjectively, a set of speech signals is specified and delivered over the channel. The listener reports the message he or she perceived, and the proportion of correct identifications is taken as the subjective measure of speech intelligibility. There also exist objective measures, both instrumental and computational, that correlate well with subjective speech intelligibility.

When the quality of speech is at a level where the intelligibility is relatively good, certain other dimensions in sound quality are of interest. Such attributes are, for example, *speaker recognizability* and *speech naturalness*. For example, in telephony, the minimum requirement on quality is intelligibility of speech, but usually recognizability of the speaker is also required. The quality of synthetic speech also needs to be measured. For example, if speech synthesis is used in announcements in public spaces, the intelligibility of the messages has to be known. The same methods can be used with synthetic speech as with natural speech.

Different subjective and objective methods have been developed to measure the quality of speech, indicating the articulation, intelligibility, and quality of the reproduction of timbre (Quackenbush *et al*., 1988). We will list some relevant techniques and later present some of them in greater detail.

17.6.1 Subjective Methods and Measures

- *Articulation*. The term articulation here means the overall functioning of the speech transmission channel, not just the functioning of the speech organs, as discussed in Section 5.1.3. A measure for the quantity is obtained from a listening test, where the task of the subjects is to listen to nonsense phoneme sequences composed as a catenation of consonants (C) and vowels (V), such as /CV/ or /CVC/, and to report the sequences perceived. The percentage of correct answers gives the *articulation score*. The *articulation index* is the articulation score modified to obtain additivity, just as the values of loudness are additive but the values of loudness level are not (Fletcher, 1995).
- *Intelligibility* and *intelligibility score*. The articulation test, but this time conducted with real words or sentences measures the intelligibility of the communication channel. The percentage of correct answers is the intelligibility score.
- *Rhyme test*. The test uses rhyming words or one-syllable words where changing the first phoneme changes the meaning of the word, such as pay/may/day/say/way. The percentage of correct answers measured gives this measure of speech quality. Different variations of this test exist, differing in the application and realization.
- *Speech interference test*. Here, noise is added to interfere with a reference speech signal and the speech signal to be tested. The level of noise is first adjusted so that the articulation in the reference speech is 50%, after which that level of noise is sought that produces the same score with the test speech signal. The difference in the levels of noise in the test and reference cases is the quality factor Q.
- *Quality comparison methods*. The quality of multiple speech samples is compared, and the subject is asked to rank them in order of preference. The subject may also be asked to focus on a specific perceptual attribute of the sounds.
- *Isopreference method*. In this method, a set of recordings is first made where a signal at different levels is transmitted through the channel accompanied by additive background noise at different levels. The listener evaluates the different recordings and forms a map of preferences with coordinates defined by the level of speech signal and the level of noise. Numerous variations on this approach exist.
- *Mean opinion score* (MOS). This is a commonly used scale for sound quality in which numerical values from 1 to 5 are associated with verbal category ratings. See Section 17.4.1.
- *Indirect judgement tests* (Quackenbush *et al*., 1988). These tests aim to evaluate speech quality by measuring factors assumed to affect it. Such methods include, for example, the *paired acceptability rating method* (PARM), the *quality acceptance rating test* (QUART), and the *diagnostic acceptability measure* (DAM). The DAM method utilizes 20 different given scales with values between 0 and 100, and measures such as 'rasping', 'hissing', and 'acceptability' are evaluated. The measures are assumed to be related to the features of the speech signal itself, to features of background sound, or to the general impression, respectively.
- *Communicability tests*. Here, the task of a subject is to communicate with another subject through a channel, and to conduct a defined task together. For example, one of the subjects might instruct the other how to draw a picture. Immediately on completion of the task the

subjects are asked to rate the ease of communication, for example, on a scale from '1 = no meaning understood using reasonable effort' to '5 = completely relaxed communication; no effort required'.
- *Task recall tests*. In these tests, the subject has to remember as many words as possible that he or she hears through a channel. The test measures the ease of communication, since a flawed communication channel makes remembering words difficult.
- *Noise suppression tests*. Many mobile communication devices have algorithms to suppress background noise. Such processing affects 1) overall quality, 2) intrusiveness of background noise, and 3) the quality of the speech signal. A subjective test specified in ITU-T P.835 (2003) is often used in industry, which is discussed in slightly more detail in Section 17.8.2.

The speech material used in the tests is typically a subset of the vocabulary of a language. When the words for the material are chosen, they have to be phonetically balanced for the targeted purpose. For example, for mobile speech communication tests, the material has to contain the phonemes in the same probabilities as in the everyday language of the test subjects.

17.6.2 Objective Methods and Measures

- *Articulation index* (AI). This was developed to measure speech intelligibility over a transmission channel that is assumed to be nearly linear, but, with disturbance caused by additive noise. The method assumes that the loss of articulation can be estimated by summing the AI values over 20 frequency bands, following roughly the Bark scale.
- *Percentage articulation loss of consonants* ($\%\mathrm{AL}_{\mathrm{cons}}$) (Peutz, 1971). This is a simple and relatively often used estimate of speech intelligibility in a room, auditorium, or other large space. The $\%\mathrm{AL}_{\mathrm{cons}}$ value is computed from the basic acoustic parameters of the space. The method is described in Section 17.9.3.
- *Speech transmission index* (STI). The index is based on the *modulation transfer function* (MTF), and it can be used to estimate relatively reliably the effect of reverberation and additive noise of a transmission channel on speech intelligibility. The method is described in Sections 17.7.1 and 17.7.2, and STIPA, a simplified version of STI, is discussed in Section 17.7.4.
- *Signal-to-noise ratio* (SNR). This is a traditional measure of how well a signal differentiates from background noise. It has different variants, such as frequency-weighted SNR and segmental SNR. The classical SNR can be defined as $\mathrm{SNR} = 10 \log_{10}\{\sum_n x^2(n)/\sum_n [x(n) - x_{\mathrm{d}}(n)]^2\}$, where $x(n)$ is the original signal and $x_{\mathrm{d}}(n)$ is the distorted signal after going through the communication channel. The SNR is very sensitive to all kinds of differences between $x(n)$ and $x_{\mathrm{d}}(n)$ as well as to differences that are not at all perceivable. Thus, it is a relevant measure only in some simple cases.
- *Spectral distance measures*. These measures are based on the notion that the magnitude spectrum and spectral differences reflect better the perceptual attributes of sound than signals or signal differences in the time domain. The difference between smoothed time–frequency presentations of a signal and the communicated signal often produces usable information about the distortion in the system. Several different versions of these spectral distance measures exist. The cepstrum has also been similarly used to evaluate the difference.

- *Weighted spectral slope distance measure*. This uses the slope of spectra instead of the basic magnitude spectrum to compute the spectral distance, and so it reflects certain phonetic differences between speech sounds more accurately.
- *LPC distance measures*. These measures are based on linear predicton. LPC coefficients are computed for both the original signal and the distorted signal, and a distance measure is computed from the LPC coefficients, reflecting the difference between the signals.
- *Auditory sound quality measures*. These computational methods for signal difference measurement mimic the spectro-temporal resolution of hearing in quality estimation. The sound quality measures for audio are reviewed in Section 17.5.2 and for speech in telecommunication in Section 17.8.1. These methods are widely used in the mobile telecommunication industry.

17.7 Measuring Speech Understandability with the Modulation Transfer Function

The understandability of speech is of crucial importance in public announcement systems, in telephones, and also in auditoria with or without sound reinforcement. The intelligibility can be estimated by computing the *speech transmission index* (STI), which is based on the concept of the *modulation transfer function*. STI methods are well established, and in some countries the STI values for sound systems in public spaces are regulated. Furthermore, the standard that describes sound systems for emergency purposes ISO 7240-19 (2007) defines the minimum STI value to be 0.5. Public announcement systems for public spaces are costly, and if the STI of a new building is measured to be lower than that targeted in the design process, the acoustic and electrical changes that must be made may be expensive. This is one reason why STI simulations and measurement systems are of great importance. Thus, this section makes a relatively detailed overview of the techniques used.

17.7.1 Modulation Transfer Function

The *modulation transfer function* (MTF) is an objective measure of a communication channel required to compute the *speech transmission index* described in the following section. The background of the MTF lies in the basic properties of speech spectra and their variation over time. Figure 3.6 on page 54 shows that the produced spectra change radically between phones. As discussed in Section 11.5.1, human hearing is very sensitive to *temporal changes* in magnitude spectrum, suggesting that the temporal modulations present in each auditory band are very important for successful speech reception.

When speech is communicated over an acoustic space or a technical channel, it seems to better retain understandability the more naturally the modulation in each auditory frequency band is conserved. Based on knowledge from auditory modelling, it would be logical to monitor the temporal changes in specific loudness for each critical band (see Section 10.2.5). However, such an approach has not been adopted in this context, but a slightly simpler and technically more straightforward technique, based on the MTF, is widely utilized instead.

The MTF reflects how well the modulation of the envelope of a narrowband signal is conserved when travelling from the source to the receiver (Houtgast and Steeneken, 1985; Steeneken and Houtgast, 1985). A single MTF value does not estimate the speech intelligibility,

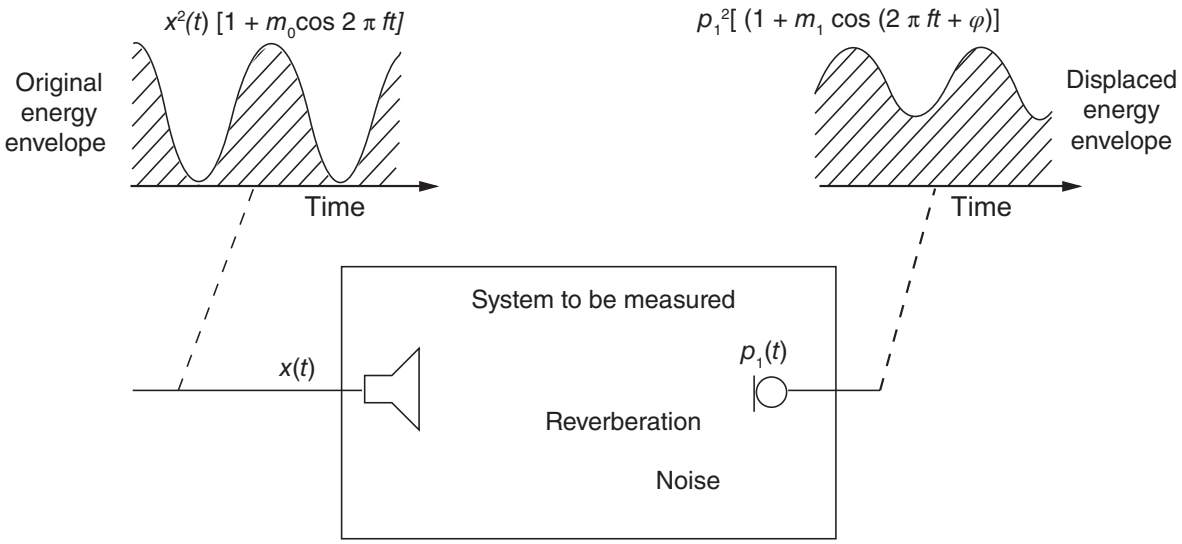

Figure 17.6 A modulated octave-band signal is reproduced in a room or over an arbitrary communication channel and recorded at the position of the listener. The modulation changes depending on reverberation and noise in the system.

but, as will be discussed in the next section, when the MTF is measured for different modulation frequencies for different narrowband signals, the speech transmission index (STI) can be computed. The STI correlates well with different subjective intelligibility measures.

When the modulation is computed from a squared pressure response, representing the energy, the effects of background noise and reverberation are made commensurate. Also, the sinusoidal modulation of a carrier signal, after the effects of noise and reverberation, is preserved as a sinusoid, though with a lower modulation depth. The principle is shown in Figure 17.6, where a modulated signal $x(t)$ is presented to an acoustic system from where the pressure signal $p_1(t)$ is captured. The envelopes related to the squared signals are shown in the figure. Note that, in general in STI literature, pressure squared is associated with 'intensity'. In this book we restrict the term intensity to mean the net flow of energy.

The MTF is most commonly measured by applying the signal

$$x(t) = \sqrt{0.5(1 + m_0 \cos(2\pi f_\mathrm{m} t))}\; s(t) \tag{17.1}$$

to the system, where $s(t)$ is an octave-band signal with centre frequency f and f_m is the modulation frequency. When no background noise is present and no reverberations or echoes exist in the room, the degree of modulation in the system is preserved, and the pressure signal

$$p_0(t) = A_0\sqrt{0.5(1 + m_0 \cos(2\pi f_\mathrm{m} t + \varphi_0))}\; s(t) \tag{17.2}$$

can be measured, where A_0 is a static gain and φ is a term related to the delay of the phase. The response shares the unmodified modulation depth m_0 in this ideal case. We are interested

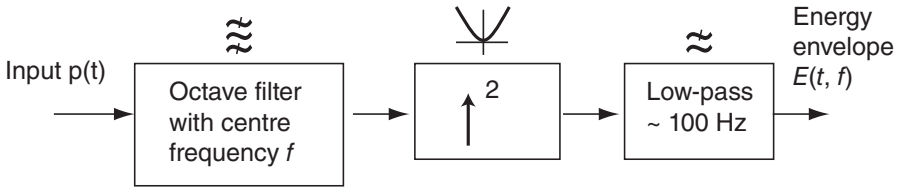

Figure 17.7 The processing of recorded pressure signals $p(t)$ to extract energy envelopes $E(t,f)$, where f is the centre frequency of the octave-band filter.

in how the modulation is transferred when noise and reverberation exist. Let us assume that $p_1(t)$ is recorded in reverberant and noisy conditions:

$$p_1(t) = A_1\sqrt{0.5(1 + m_1\cos(2\pi f_{\mathrm{m}}t + \varphi_1))}\; s(t), \tag{17.3}$$

where A_1 is a static gain, and φ_1 is a term related to the delay of the phase.

The depth of modulation m_1 is the variable that is to be measured. To this end, we may process the measured pressure signal as shown in Figure 17.7 to obtain the energy envelope signal $E_1(t,f)$ for the octave band with centre frequency f. Note the similarity of the set-up to the filter-bank-based auditory models in Section 13.2. The frequency content of $s(t)$ is chosen to be located above 100 Hz, and consequently only its envelope remains after the processing. The energy envelope $E_1(t,f)$ obtained from recorded signal $p_1(t)$ has the form

$$E_1(t,f) = E_1\left[1 + m_1\cos(2\pi f_{\mathrm{m}}t + \varphi)\right], \tag{17.4}$$

where E_1 is the energy of the signal recorded and m_1 is the depth of modulation. The value of m_1 can be calculated as

$$m_1(f,f_{\mathrm{m}}) = 2\,\frac{\sqrt{|\int_t E_1(t,f)\sin(2\pi f_{\mathrm{m}}t)\,\mathrm{d}t|^2 + |\int_t E_1(t,f)\cos(2\pi f_{\mathrm{m}}t)\,\mathrm{d}t|^2}}{\int_t E_1(t,f)\,\mathrm{d}t} \tag{17.5}$$

(IEC, 2011). The equation thus computes the magnitude of modulation at frequency f_{m} of the signal $E(t,f)$, using the set-up in Figure 17.7 normalized to values between zero and one. Finally, the modulation transfer ratio is given by

$$m(f,f_{\mathrm{m}}) = \frac{m_1(f,f_{\mathrm{m}})}{m_0(f,f_{\mathrm{m}})}, \tag{17.6}$$

which is sometimes also called 'modulation reduction'. The closer the number is to unity, the better the modulation is transferred, and values near zero imply largely lost modulation.

Figure 17.8 shows the reduction in modulation with reverberation in case A and with added noise in case B computed for a speech signal. The left-most panels show the change in the envelope of the measured signal at the octave-band noise with centre frequency 500 Hz. The reverberation smooths the change of the envelope in time, acting as a low-pass filter for the envelope and thus also for modulation. The effect of additional noise is seen from the increase in the minimum value of the envelope.

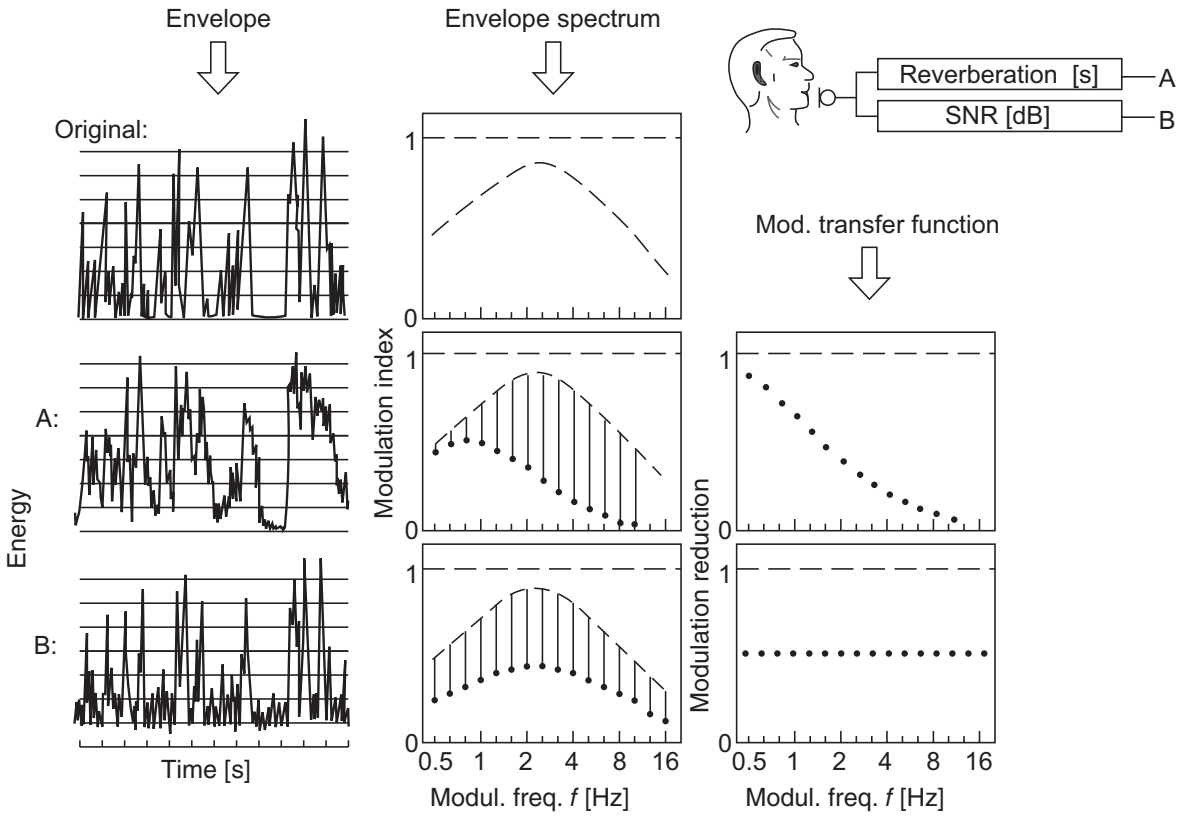

Figure 17.8 The reduction in modulation depth with reverberation in system A and with background noise in system B. The left panels show, from top to bottom, the envelope of the signal in its original form, after reverberation, and after additive noise. The envelope spectrum plots (centre) show the modulation spectra in each case, and the modulation transfer function (right) shows the reduction in modulation for each case depending on the modulation frequency. Adapted from Houtgast and Steeneken (1985). Courtesy of Bruel & Kjær.

The spectra of the modulation in the original and in the transmitted signal are shown in the centre panel at modulation frequencies from 0.5 Hz to 16 Hz. Reverberation reduces the modulation depth more at high modulation frequencies than at low modulation frequencies. Noise, in turn, reduces the modulation depth equally at all modulation frequencies. The rightmost plots show the corresponding modulation transfer functions for cases A and B. In case A, the reverberation clearly acts as a low-pass filter in the transfer function, while the noise in case B reduces the modulation evenly at all modulation frequencies.

The MTF can, in principle, be measured with signals that have energy at all audible frequencies and modulations at all modulation frequencies of interest. Speech, music, and other such signals can be used. However, the measurement can be conducted with more accuracy using signals specifically designed for the task. The MTF can also be estimated during the design of the acoustics of halls, or during the design of public address systems, if the impulse response and background noise levels can be estimated.

A straightforward method to measure the MTF is to use a loudspeaker or an artificial mouth in a room or in an auditorium in the position where the speaker would be. The source is used to emit 100% amplitude-modulated octave-band noise (m_0=1.0) at a level corresponding to the average level of speech. The measurement is repeated for each modulation frequency f_m, from 0.63 Hz to 12.5 Hz in steps of 1/3 octave, as shown in Figure 17.9.

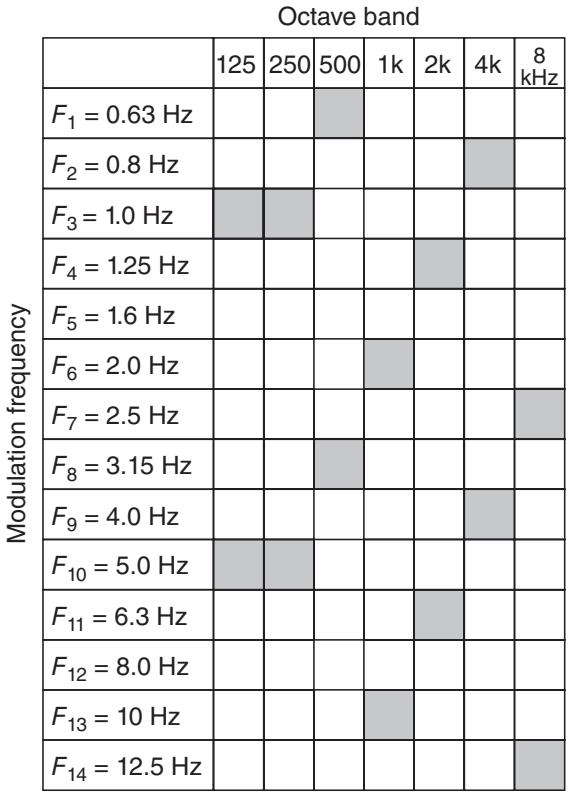

Figure 17.9 The octave bands of the carrier signal and the modulation frequencies used in the STI measurement, represented by all the squares in the matrix. The grey squares represent the corresponding values in the STIPA measurement method.

A microphone is placed in the position of the listener to measure the response, and in each case the modulation of the envelope is analysed at frequencies having octave bands of the carrier signal in the range 125 Hz–8 kHz. The reduction in modulation from the original 100% gives the value of the MTF $m(f, f_{\mathrm{m}})$, where f is the centre frequency of the octave band and f_{m} is the frequency of modulation.

The effects of reverberation and noise can, in general, be expressed as

$$m(f, f_{\mathrm{m}}) = \frac{1}{\sqrt{1 + (2\pi f_{\mathrm{m}} \frac{T_f}{13.8})^2}} \cdot \frac{1}{1 + 10^{-\mathrm{SNR}_f/10}}. \tag{17.7}$$

Here, the first term corresponds to the effect of reverberation and the second reflects the effect of background noise.

17.7.2 Speech Transmission Index STI

It was proposed that the modulation transfer function (MTF) be measured at seven frequency bands with fourteen modulation frequencies, resulting in $7 \times 14 = 98$ m values (see Figure 17.9). Ideally, the estimate of speech intelligibility should be expressed with a single value.

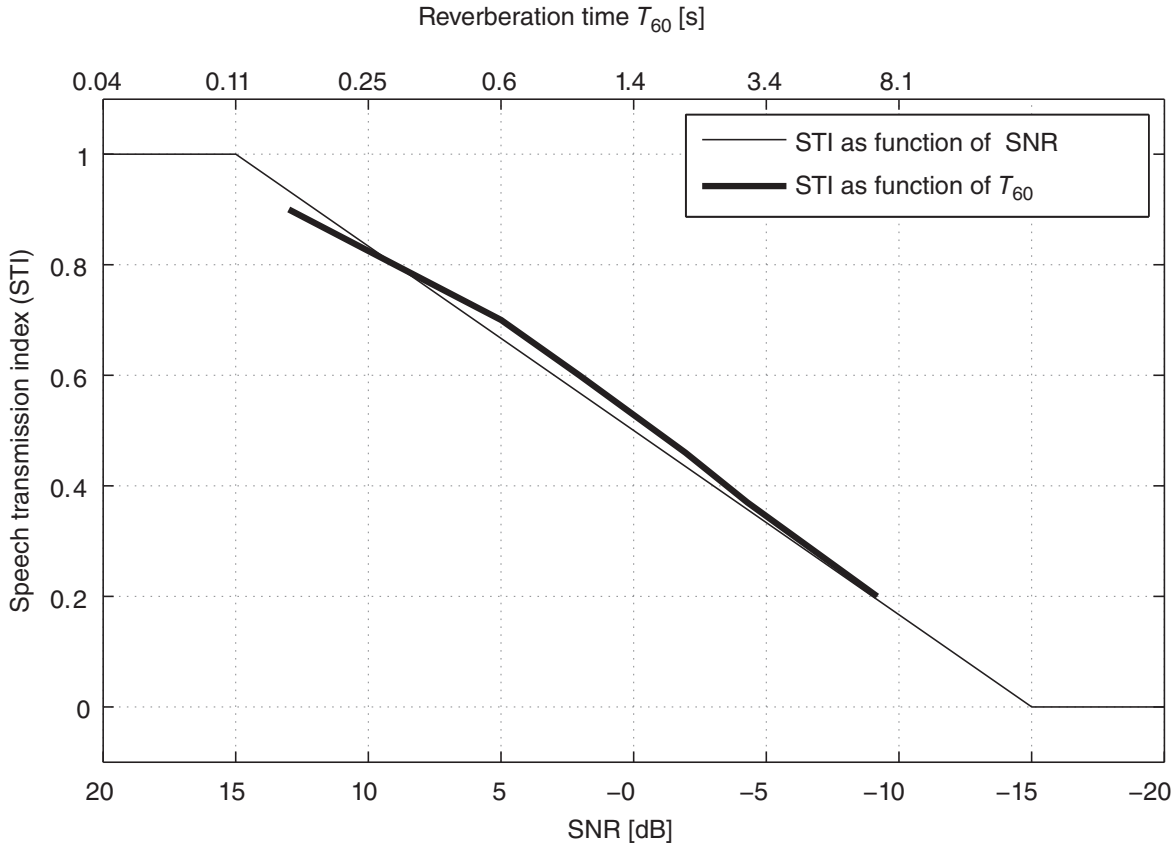

Figure 17.10 The dependency of the STI in the ideal case with noise only as a function of SNR, and with reverberation only as a function of T_{60}. Adapted from Houtgast and Steeneken (1985).

The *speech transmission index* (STI) has been found to serve this purpose. The principle is that the 98 m values are first transformed into apparent SNR values as

$$\mathrm{SNR}_{\mathrm{app}} = \max\left(-15, \min\left(15, 10\,\log\frac{m}{1-m}\right)\right), \tag{17.8}$$

where $\mathrm{SNR}_{\mathrm{app}}$ is expressed in dB on a scale from -15 to 15. The values are scaled to lie between 0 and 1, and a weighted average is calculated (IEC, 2011). The weights emphasize the octave bands most relevant for understanding speech. The value of the STI is limited to between 0.0 and 1.0, 0.0 corresponding to estimates with no speech intelligibility and 1.0 to perfect speech intelligibility.

Figure 17.10 shows the effect of the SNR on the STI and also the effect of ideal reverberation with different T_{60} values. Since the curves are on top of each other, the correspondence of the SNR to the reverberation time is easily seen.

17.7.3 STI and Speech Intelligibility

The speech transmission index (STI) cannot directly be associated with the subjective intelligibility of speech, since many other features affect perceptual intelligibility, such as the

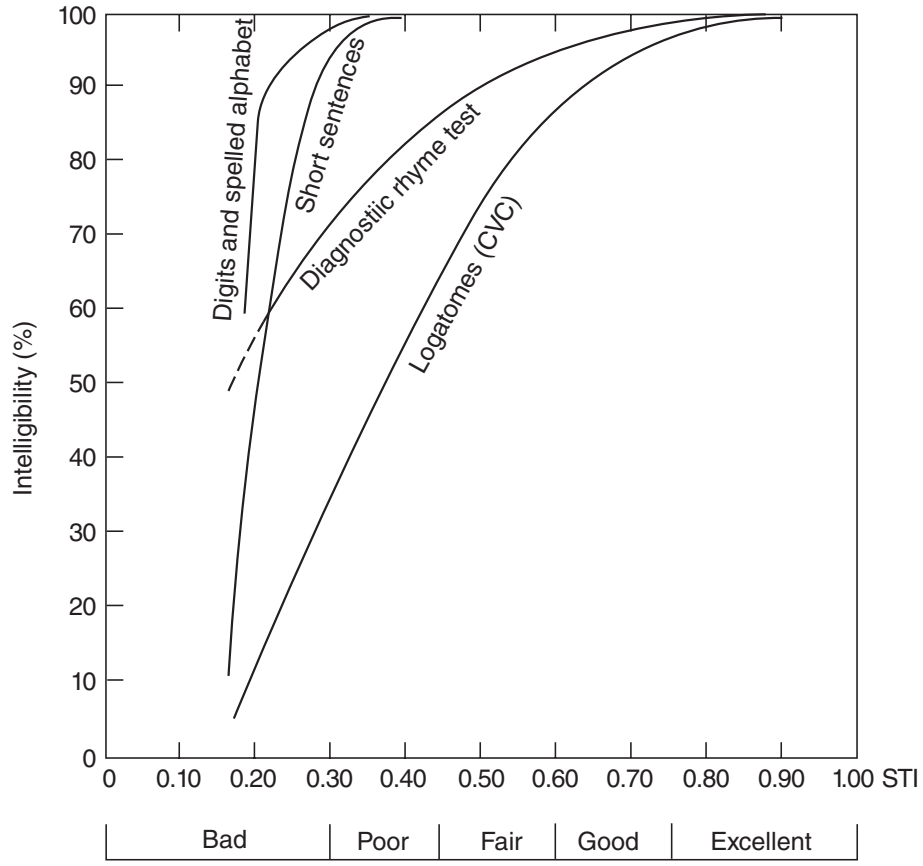

Figure 17.11 The dependence of speech intelligibility on the STI index for different test signals. Adopted from (Houtgast and Steeneken, 1985 and Steeneken and Houtgast, 2002). Courtesy of Bruel & Kjær.

speaker, familiarity with the message, and language. However, the STI has been found to have a monotonic effect on intelligibility: the higher the STI, the higher the intelligibility (Steeneken and Houtgast, 2002). Figure 17.11 shows the dependency between the type of speech and the STI. When the speech content is short words from a known small vocabulary, high intelligibility is obtained even with relatively low STI values in the range 0.2–0.4. With sentences, the linguistic context aids in understanding the message, and thus good intelligibility is obtained with STI values in the range 0.4–0.5. In more challenging cases, such as with logatomes, a considerably higher STI is needed for high intelligibility. In the case of logatomes, it would be more correct to call the measured value articulation instead of intelligibility.

The STI values are also categorized in Figure 17.11. The labels designate quality categories that can be used to interpret the measurement results. The step size for the categories is 0.15. If a measurement of STI is repeated several times, and if the standard deviation is notably smaller than 0.15, the STI measurement and its interpretation can be considered to be reliable.

17.7.4 Practical Measurement of STI

The measurement of the STI is quite a complicated procedure – 98 different $m(f,f_{\mathrm{m}})$ values must be measured individually, and this takes about 15 minutes to conduct. Often, the reverberation and background noise have such properties that the measurement can be simplified

to cover fewer combinations of modulation frequencies and octave bands. STIPA employs two modulation frequencies for each frequency band, as shown in Figure 17.9 (IEC, 2011). Using only two modulation frequencies makes it possible to present all of the octave-band signals simultaneously, in which case each octave-band signal is modulated by the factor $\sqrt{0.5(1 + m_0 \cos(2\pi f_{m1} t) + m_0 \cos(2\pi f_{m2} t))}$, where f_{m1} and f_{m2} are the modulation frequencies at the specific octave band, and the value of m_0 is about 50%. With this method, the measurement time is reduced to between 10 and 15 seconds.

There are a number of dedicated measurement devices available for STIPA. Relatively good results can also be measured using inexpensive applications on handheld devices, available for at least the IOS and Android platforms, and the interested reader is encouraged to test the applications to get hands-on experience of STI measurements.

The STI and STIPA, and the predecessor of STIPA, RASTI, have been standardized in many versions. The different versions and their usage in practical situations are discussed by Steeneken *et al.* (2011). STIPA measurements seem to provide results quite similar to STI measurements in many cases, although the measurement range has been pruned from the STI measurements. However, in some cases, significant differences occur (Mapp, 2005).

17.8 Objective Speech Quality Measurement for Telecommunication

The STI measure considers only the intelligibility of speech. This has not been found to be an adequate analysis system for modern telecommunication systems, where the speech codecs may, in the worst case, for example, make the speaker unidentifiable, although the intelligibility of speech is good. To estimate other factors in transmitted speech, more advanced objective analysis methods have been developed.

There are three main requirements for quality measurement methods:

- *Models for general speech quality* are expected to give a high MOS value only for natural-sounding and intelligible speech.
- *Methods for measuring the perceptual effect of background noise suppression* estimate the performance of the noise-suppression algorithms, and also evaluate whether the methods introduce unwanted side effects affecting the quality of the speech signal.
- *Measures for echo suppression* evaluate the capability of the telecommunication system to reduce unwanted echoes in two-way communication.

The problem definition is quite broad, as three different aspects are to be measured. Also, telecommunication devices are used in various acoustic conditions, and the network connectivity for the devices affects the sound quality. Furthermore, interfering background noise, nearby sources, room acoustics, and acoustics in hands-free use may make the acoustic conditions challenging. The transmission of data over a network may also cause unpredictable delays, jitter, and occasional loss of data. In addition, the acoustic feedback from the receiver to the sender depends on the properties of the device, on the acoustic conditions on the receiver side, and also on the properties of the connection.

A device that enters the market should thus be tested under many different conditions. Unfortunately, conducting such a set of listening tests that would cover all possible combinations of different conditions would be a tedious task. Instead of listening tests, objective methods

are used widely in industry, which is reflected in the active standardization of the methods. The methods are signal-processing structures that normally include some parts simulating the properties of human hearing. The methods may also include acoustic measurements of the devices under controlled acoustic conditions.

Quite commonly, when a device is approved by an authority or network operator, it has to achieve certain scores in standardized objective tests. The methods are relatively complicated, and they are only briefly touched on below. Interested readers can find in-depth information on the current criteria (2014) for the properties of devices and technologies for speech communication in the 3GPP measurement techniques (ETSI, 2014a) and in the specifications of the requirements for the devices (ETSI, 2014b).

17.8.1 General Speech Quality Measurement Techniques

The first recommendation of a method to measure objective speech quality (ITU-T, 1998a) was called the perceptual speech quality measure (PSQM) (Beerends and Stemerdink, 1994). The PSQM is almost identical to the PAQM model for audio quality shown in Figure 17.5, with only some minor details being different (Beerends and Stemerdink, 1994). The PSQM was primarily focused on identifying the quality impact of speech codecs. Unfortunately, the method did not turn out to be successful, and the recommendation has since been withdrawn. The PSQM was unable to estimate correctly the quality impairment due to filtering, variable delay, and distortion common in mobile communication. For example, the transmission delay may vary in mobile transmission, and the variations can be fatal in VoIP applications. Subtracting time-varying feature vectors, as shown in Figure 17.5, would estimate a large error in the case where the test case had random variations in the lengths of syllables in speech. Such small delay variations do not cause much deterioration of perceived quality, implying that the model would underestimate the speech quality in this case.

Subsequent to the PSQM, work was initiated to create an algorithm suitable for assessing the additional impact of network impairment, which resulted in a method called the perceptual evaluation of speech quality (PESQ)(Beerends *et al*., 2002; ITU-T, 1998b). The PESQ method includes processing steps such as automatic compensation of the dynamically varying jitter prior to the auditory model, and other similar steps, which are designed to directly overcome the problems that existed with the PSQM. Many of the features do not stem directly from the neural structure of the auditory system, although they are motivated perceptually. PESQ can be characterized as a signal difference measurement device, which has some features stemming from the human auditory system.

PESQ was validated with results from numerous experiments that specifically tested its performance across combinations of factors such as filtering, coding distortions, variable delay, and channel errors. It is recommended to be used for speech quality assessment of telephone-band handset telephony and narrowband speech codecs. PESQ is measured using an electrical interface (by using a connector between the device and the measurement device), which does not take into account the acoustic degradations of the signal in real listening. The TOSQA (Telecommunications Objective Speech Quality Assessment) system can also take input from an acoustic interface (TOSQA, 2003), which may have a significant effect on the perceived and measured quality. Otherwise, it shares many similar operation principles with PESQ.

The PESQ method has been extended to cover the assessment of wideband speech as well as networks and codecs that introduce time warping. The most recent version (ITU-T, 2011) is

known as POLQA (perceptual objective listening quality assessment) (Beerends *et al*., 2013; ITU-T, 2011), and again, this gives better estimates of subjective evaluations from measured signals.

A specific method has been developed to estimate objectively the speech quality perceived by hearing-impaired users wearing a hearing aid. The Hearing-Aid Speech Quality Index (HASQI) (Kates and Arehart, 2014) is based on a model of the auditory periphery that can incorporate changes due to hearing loss. The model can be used to predict the effect of signal processing in hearing aids to the perceived speech quality of either normally-hearing or hearing-impaired listeners, and it has an interesting application in the development of hearing aids. That is, the engineers often have normal hearing, which means that they cannot test the devices thoroughly themselves. Moreover, the hearing aids should be suited for different types of impairment. Thus, a method to evaluate the speech quality with different severities and types of hearing impairment is potentially helpful in product development.

17.8.2 Measurement of the Perceptual Effect of Background Noise

Telecommunication devices, such as mobile phones, often involve algorithms to suppress background noise using non-linear DSP methods with time-variant processing. An unwanted side effect is that the noise suppression algorithms may also degrade the quality of speech, especially if the background noise is non-stationary. Objective measures have been developed to measure such effects, and the most recent one is 3QUEST, defined by ETSI (2008). The method is targeted for both wide- and narrowband transmission in noisy environments, and it has been calibrated with a large set of subjective tests.

The subjective tests were conducted using a set of noisy recordings with real devices, according to ITU-T P.835 (2003). In the tests, the subjects rated three different factors in the samples:

- Speech MOS (S-MOS): the speech sample was rated 5 – not distorted, 4 – slightly distorted, 3 – somewhat distorted, 2 – fairly distorted, or 1 – very distorted.
- Subjective noise MOS (N-MOS): the background of the sample was 5 – not noticeable, 4 – slightly noticeable, 3 – noticeable but not intrusive, 2 – somewhat intrusive, or 1– very intrusive.
- Overall MOS (G-MOS) on the standard MOS scale.

The method is a relatively complex signal processing structure, which includes parts mimicking human time–frequency resolution, and ultimately uses a trained pattern recognition algorithm to produce final quality estimates. It requires three inputs to the system: a clean speech signal, the unprocessed signal, and the processed signal. The clean signal is the speech signal of a real human subject recorded in a free field. The unprocessed and processed signals are measured in realistic noisy conditions created in the laboratory. A dummy head with a mouth is placed in a noisy environment generated in a laboratory, and the device under test is placed near the dummy head, just as it would be in practice. The noisy environment is generated using an equalized loudspeaker set-up with four loudspeakers and one subwoofer around the head. The sound applied to the loudspeakers originates from recordings from noisy natural conditions.

A separate microphone positioned near the microphone of the device under test is used to capture the unprocessed signal. The processed signal is the signal captured and processed by the device, possibly also aiming to reduce the background noise. The recording of the device is performed separately with all noise signals captured in different natural conditions. The use of different noisy signals simulates the functioning of the device in different acoustic scenarios. The recordings of the processed and unprocessed signals are used to estimate the MOS. The measurement system is described ETSI (2008).

Similarly processed signals were used in listening tests to provide reference data to train the 3QUEST algorithms. Finally, the 3QUEST system estimates the S-MOS, N-MOS, and G-MOS values for any signals recorded in an identical manner from any device under test (3QUEST, 2008).

17.8.3 Measurement of the Perceptual Effect of Echoes

A common unwanted feature of two-way telecommunication is the presence of loud echoes (Appel and Beerends, 2002). In natural conditions, a speaker perceives his or her own voice from the mouth-to-ear path and from the acoustic response of the environment, providing direct feedback that is used subconsciously to control the speech production process. Interference in the feedback can influence the comfort of speaking and also the manner of speaking. A well-known effect is the raising of one's voice in the presence of loud background noise, called the Lombard effect (Lane and Tranel, 1971). In contrast, we lower the volume of our voice when we are played back our voice loudly. Delaying the echo increases the perception of discomfort. For small delays (<10 ms) and high levels, the echo interferes with the sound coming directly from the mouth of the speaker, leading to the perception of colouration due to comb filtering. Medium delays (10–30 ms) lead to the perception of hollowness in one's voice, and for larger delays ($>$ 30 ms) we perceive a clear, distinct echo. When the delay is large ($>$ 200 ms) and the level of the echo is high, subjects experience difficulty in producing words (Appel and Beerends, 2002). The difficulties may be simply because, when producing a syllable, one's own voice present in the ear canals due to the echo has the timbre of the previous syllable, which causes confusion in our neural speech production system.

In two-way communication with mobile phone, some cases the sound from the loudspeaker is captured by the microphone in it, causing the phone to send the signal back to the caller. This can happen, for example, when using the phone in the hands-free mode via a loudspeaker. In such a situation, the round trip delay is of the order of 300 ms, a result of the signal processing and other delays in both phones, and also due to transmission delays. In VoIP calls, the round trip delay is typically even longer, often more than 500 ms. Such long delays can thus cause difficulties in speaking, which motivates the implementation of echo cancellation algorithms and subsequently objective measures to measure the positive and negative effects of the algorithms.

A method to measure objectively the degradation of quality due to strong echoes was presented by Appel and Beerends (2002). Another application for the same task is the *echo quality evaluation of speech in telecommunications* test (EQUEST), which is an instrumental method for estimating the annoyance (EQUEST, 2012). EQUEST is, again, a relatively complex signal processing algorithm, which uses psychoacoustic knowledge to perform the evaluation. This time the knowledge of the psychoacoustic temporal masking characteristics of human listeners is injected into the algorithm to estimate the annoyance caused by echoes. An alternative method for this task is described in the 3GPP standard (ETSI, 2014a).

17.9 Sound Quality in Auditoria and Concert Halls

The spaces for presenting arts involving sound, such as theatres, auditoria, and especially concert halls, have a special status with respect to sound quality. The quality of sound created by verbal or music performances for an audience has been of interest for hundreds and even thousands of years (Blauert, 2013a). For this reason, the design of concert hall acoustics has been a showcase of acoustic technologies to the community at large. Concert halls have been designed and built using trial-and-error, and a relatively good consensus exists concerning about 20 halls with great 'acoustics', such as Vienna Musikverain (Vienna, Austria), Metropolitan Opera (New York, USA), Boston Symphony Hall (Boston, USA), and Concertgebow (Amsterdam, The Netherlands). During the late 1900s, a scientific approach was finally adopted for concert hall design, although what is the best method is still open to debate.

Concert halls for music performances, that is, big halls for orchestral music, opera houses, chamber music halls, and halls for electronically reinforced sound, have to be designed keeping in mind the primary use of the hall. The main difference between the halls is the reverberation: too long a reverberation at too high a level compared to the acoustics optimal for the music performed there, and the general impression of the music is degraded. On the other hand, if the hall is too 'dry' in reverberation, the loudness of music may be perceived as too low with acoustic instruments, and the general impression of the music will again be different from the way it should be. The evolution of concert halls can also be assumed to have had an impact on music. Since composers created their music to be performed in specific halls, the composition of the music, and also the composition of orchestras, was adapted to them.

The criteria for speech auditoria and drama theatres, where the target is to maximize speech quality, are different than for halls for music performances. The overall guidelines for speech intelligibility presented in Section 17.6 also hold for auditoria and concert halls, and the STI measure can be used to estimate speech intelligibility. A simpler method than STI to estimate intelligibility in an auditorium is presented in this section, where the proportion of consonants not delivered intelligibly to the listener is approximated using knowledge of reverberation time and hall geometry.

17.9.1 Subjective Measures

A considerable effort has been made to define the vocabulary to describe the main properties of concert halls (Barron, 1993; Beranek, 1996). Beranek (1996) suggests a list of 18 attributes based on his own extensive experience of listening in concert halls, which includes terms describing sound from both the audience and the stage. Lokki (2014) used descriptive sensory analysis with reproduced spatial sound of different concert halls, and ended up with a rather similar, though shorter, list. It can be assumed that Beranek's list contains some attributes with which all listeners do not agree. However, Beranek's list is presented here, as it gives the reader a general impression of the kind of attributes that are at least thought to exist in subjective attribute palettes of concert halls.

- *Intimacy* or *presence*. The hall gives an impression of a small and intimate space.
- *Reverberation* or *liveness*. A long and perceivable reverberant tail makes the hall give the impression of a 'live' hall, and, correspondingly, a short reverberant tail makes the hall 'dry'.
- *Spaciousness: Apparent source width* (ASW). This is the width of the auditory object associated with the sound source itself. The reflections and the reverberation of the hall may make the sound sources seem to be wider than they actually are.

- *Spaciousness: Listener envelopment* (LEV). This is the directional distribution of the auditory object associated with reverberant sound. LEV is judged to be high when the reverberant sound is perceived to arrive from all directions.
- *Clarity*. This attribute is related to how well the sounds generated by instruments in a musical performance stand apart from each other. It depends on the performance and also on the acoustics, according to Beranek.
- *Warmth*. A hall is said to be warm if the reverberation time is longer at low frequencies (below 350 Hz) than at higher frequencies. If the reverberation time is too long, or if the low frequencies are overly strong, the hall may be called 'dark', which is an undesirable feature.
- *Loudness*. This simply means the perceived loudness at the listening position.
- *Acoustic glare*. This is generated if the sound is reflected by flat, smooth side panels to the audience. Rough and irregularly shaped panels reduce glare.
- *Brilliance*. This is the perception when high frequencies are prominent and decay slowly.
- *Balance*. A good balance is obtained when all sound sources are audible to the listener as intended. The balance depends, naturally, on the performers, but also on the acoustics of the hall.
- *Blend*. This is defined by the 'mixing' of sounds at the listening position. With a good blend, the sounds from the instruments are perceived as intervals and chords by the listener, with the intended level of consonance and dissonance.
- *Ensemble*. This refers to the ability of the performers to synchronize their playing as intended in the music. Typically, a better ensemble is obtained when the performers hear each other clearly on the stage.
- *Immediacy of response*. This is related to how performers perceive the hall's response to the played notes. If the response contains significantly delayed and strong reflections, the playing of the performers is affected negatively.
- *Texture*. This is the temporal pattern derived from the early reflections of the hall.
- *Freedom from echo*. An echo is a reflection that is loud enough and delayed sufficiently to be perceived as a separate auditory event, as discussed in Section 12.5.1. Echoes are not desired in concert halls.
- *Dynamic range and background noise level*. The lower end of the dynamic range is, in principle, defined by the level of background noise, or, if extremely low, the hearing threshold. The upper end of the range depends on the loudness of sounds a source may generate depending on the source itself, and also on the room response.
- *Extraneous effects on tonal quality*. No extra sounds should be produced by the hall, such as rattling sounds. Beranek also mentions the shift of localization of the sources in this context, which is referred to as *image shift*.
- *Uniformity of sound*. The sound should have good tonal quality at all listening positions.

17.9.2 Objective Measures

Several studies seeking objective measurements of concert halls that correlate with subjective perception have been conducted (Bradley, 2011). The measurements should thus estimate subjective factors such as those listed in the previous section. Again, in principle, the best objective measurement method would be an auditory model responding to concert hall acoustics in the same way as a real listener would do. Some attempts to use auditory models to assess concert hall acoustics have been made (van Dorp Schuitman, 2011), although no final answers

Table 17.3 The objective measures of concert hall acoustics defined in ISO 3382-1 (2009).

Subjective level of sound	Sound strength G in decibels
Perceived reverberance	Early decay time (EDT)
Perceived clarity of sound	Clarity C_{80} in decibels
Apparent source width (ASW)	Early lateral energy fraction, J_{LF}
Listener envelopment	Late lateral sound level, L_J in decibels

exist. Unfortunately, it seems that the models currently proposed are not able to explain human sensitivity to fine a esthetic details of the responses of concert halls to instrument sounds.

In practice, the perception of concert hall acoustics has traditionally been estimated with the analysis of measured impulse responses. Different metrics have been proposed (Barron, 1993; Beranek, 1996), and a set of measurements has also been standardized (ISO 3382-1, 2009), some of which are discussed below. The measurements discussed here are only a representative subset of the complete set in the standard. The standard proposes that the acoustics of a hall can be described with a few measurements obtained by spatially averaging over several positions. Many aspects of the standard have been criticized: the algorithms to compute the parameters are imprecise the applied frequency range is too narrow compared to human perception; and a single omnidirectional source is an inadequate representation of the sound sources present in a real orchestra (Bradley, 2011; Kirkegaard and Gulsrud, 2011). Furthermore, an impulse response is only a technical measure and does not represent how a human perceives the response of the hall to continuous instrument sounds or voices.

However, because the methods are widely used, they can be assumed to deliver some useful information. Thus, we think the standardized methods based on impulse responses might be of interest to the readers of this book. The measures are outlined in Table 17.3, and the methods to compute the measures are shown below.

- *Strength*: The ratio of the energy at the listening position to the energy measured 10 m in a free field from the source is called the strength G. Mathematically,

$$G = 10 \log_{10} \frac{\int_0^\infty p^2(t)\,\mathrm{d}t}{\int_0^\infty p_A^2(t)\,\mathrm{d}t}, \tag{17.9}$$

  where $p(t)$ is the sound pressure measured at the listener's position and $p_A(t)$ is the sound pressure measured at a distance of 10 m from the source in a free field, when an omnidirectional sound source is used as the excitation.
- *Early decay time* EDT: The time required for the reverberation to decay from 0 dB to −10 dB, scaled to correspond to the decay from 0 dB to −60 dB. This is calculated from the gradient of the energy decay curve (EDC) as introduced by Schroeder (1965) and defined as:

$$\mathrm{EDC}(t) = \int_t^\infty h^2(\tau)\mathrm{d}\tau \tag{17.10}$$

  The EDC function is typically more smooth than the impulse response itself, and so it is more useful than ordinary amplitude envelopes for estimating EDT.

- *Clarity*: This measure expresses the energy ratio between the early and late responses. A strong early response is beneficial to clarity, while a strong late response is harmful. C_{80} is a commonly used measure where the boundary between the early and late responses is set at 80 ms and is defined as

$$C_{80} = 10 \log_{10} \frac{\int_0^{80\,\mathrm{ms}} p^2(t)\,\mathrm{d}t}{\int_{80\,\mathrm{ms}}^{\infty} p^2(t)\,\mathrm{d}t}. \tag{17.11}$$

- *Lateral fraction*: This measure, J_{LF}, is obtained from the impulse responses measured using a figure-of-eight microphone signal $p_8(t)$ with the null of the response pointing towards the source and an omnidirectional microphone $p(t)$. It is computed as

$$J_{\mathrm{LF}} = \frac{\int_{5\,\mathrm{ms}}^{80\,\mathrm{ms}} p_8^2(t)\,\mathrm{d}t}{\int_0^{80\,\mathrm{ms}} p^2(t)\,\mathrm{d}t}. \tag{17.12}$$

  J_{LF} reflects the ratio of lateral sound in the overall response.
- *Late lateral sound level*: Defined as

$$L_{\mathrm{J}} = 10 \log_{10} \left(\frac{\int_{80\,\mathrm{ms}}^{\infty} p_8^2(t)\,\mathrm{d}t}{\int_0^{\infty} p_{10}^2(t)\,\mathrm{d}t} \right), \tag{17.13}$$

  L_{J} has a higher level in decibels if the reverberation after 80 ms of the arrival of the direct sound has a considerable degree of laterally flowing energy.

17.9.3 Percentage of Consonant Loss

The *percentage articulation loss of consonants* ($\%\mathrm{AL}_{\mathrm{cons}}$) is a simple measure used in the design of auditoria and concert halls to estimate the understandability of speech, based on a relatively simple mathematical formulation (Davis and Patronis, 2006; Peutz, 1971):

$$\%\mathrm{AL}_{\mathrm{cons}} = 200\, r^2 (T_{60})^2 / (VQ) + \mathrm{k}, \tag{17.14}$$

where r is the distance between the speaker and the listener, T_{60} is the reverberation time, V is the volume of the room, Q is the directivity of the source, and k is a constant describing the individual hearing capabilities of the listener. In the best case k = 1.5 and in the worst k = 12.5. For distances greater than $r = 0.20\sqrt{V/RT}$, the equation becomes

$$\%\mathrm{AL}_{\mathrm{cons}} = 9\, T_{60} + \mathrm{k}. \tag{17.15}$$

The value of $\%\mathrm{AL}_{\mathrm{cons}}$ thus estimates the percentage of consonants that are not perceived correctly in an auditorium. Relatively high values, 25–30%, may be acceptable, since the redundancy in speech makes it possible to ‘guess’ the ‘lost’ phones.

17.10 Noise Quality

Noise can be defined as sound that is disturbing or annoying. In principle, this subjective definition does not exclude any sounds, since basically any sound can be disturbing depending on

many listener-related factors. Although this book mostly concerns itself with audio and speech techniques, applications in which the sounds are desired by the listener, noise is also interesting in the context of sound quality and also in the context of psychoacoustics. A basic introduction to the relation of noise to psychoacoustics is given by Marquis-Favre *et al.* (2005b).

Noise will be discussed in Chapter 19 in the context of technical audiology, where issues such as the effects of excessive SPL on the auditory system and limits for noise exposure in work environments are of interest. In the context of sound quality, the most relevant concepts are *annoyance* and *disturbance* (Guski *et al.*, 1999; Öhrström and Rylander, 1982; Ouis, 2001; Pedersen and Waye, 2004). All terms related to noise quality are negative, and as such, annoyance and disturbance should be minimized. We use the term annoyance as a general concept of noise quality, but, also to describe how noise may upset an operation or activity. The term disturbance is connected to negative feelings where the functioning of the subject is not necessarily disrupted, the capacity of the subject to perform any task is merely hampered.

The degree to which noise is annoying is studied primarily with listening tests, similarly to the quality of sound in general. All that has been said earlier in this book about psychoacoustic research methods is valid for noise as well. Often, the tests relating to the quality of noise are conducted with inexperienced listeners, and so the test design has to be simple enough. For example, a two-alternative forced choice test is often used, where the result places the sound samples in the order of annoyance.

The results from listening tests can be compared with psychoacoustic attributes computed using auditory models, such as loudness, fluctuation strength, sharpness, roughness, tonality, and impulsiveness. A common approach is to attempt to create a model from these attributes to estimate the annoyance and disturbance caused by noise, which should be in agreement with results from listening tests. If successful, such a model could be used as an objective measurement system of subjective annoyance (Marquis-Favre *et al.*, 2005a). Unfortunately, these models are usually valid only with a limited set of noise signals, and separate models have to be constructed for different types of noise signals, and in some cases the models do not explain the measured results (Waye and Öhrström, 2002).

Loudness, or loudness level, is usually one of the attributes used to explain annoyance. The sharpness of sound, that is, the high level of high-frequency components, and also the roughness of sound increase annoyance. The narrowband components of sound and certain temporal components, such as buzzing, banging, or screeching, are also perceived as more annoying.

The subjective nature of noise is clearly evident in open-plan offices, where most office desktops are located nowadays. Acoustic noise, and especially speech and laughter, is the most significant source of distraction in the physical work environment in open-plan offices (Helenius *et al.*, 2007; Jensen and Arens, 2005; Pejtersen *et al.*, 2006; Virjonen *et al.*, 2009). In contrast, sounds that are very stable in time and have a nearly constant sound pressure level, like ventilation noise, cause very little distraction. Quite interestingly, in these environments, high speech intelligibility decreases sound quality, which is just the opposite of the requirement in public spaces, drama theatres, and auditoria. Hongisto (2005) suggests that the STI value between desktops in open-plan offices should be below 0.2 to prevent the negative effects of being able to hear each other's discussions.

17.11 Product Sound Quality

Blauert and Jekosch (1997; 2012) define *product sound quality* as 'the adequacy of a sound in the context of a specific technical goal and/or task'. All products that produce a perceivable sound have their product sound quality evaluated every time they are used.

Blauert and Jekosch show that product sound quality is a broader concept than the auditory attributes evoked by a sound event. This characteristic is essential to relate the sound of the product and the subject actively using the product. The evoked auditory attributes are interpreted differently depending on the expectations of the subject. For example, the presence of a buzzing sound is generally not desirable, but when a subject uses an electric shaver, the buzzing sound communicates that the device is on and working. The subject also uses the fine structures of the sound to monitor the inner condition and quality of the device itself. The simple quantitative input–output relationship that psychoacoustics aims to measure has to be extended to cover psychological concepts such as *cognition*, *action*, and *emotion*.

The goal of product sound quality is not only pleasantness of sound, just as in noise control, minimizing the level of sound is not the only goal. A more important factor than the pleasantness of sound is often the informativeness of sound. Communication of the state of functioning of the product is often the factor determining why the perception of the sound is desirable. In particular, if the subject has been exposed many times to the sound of the product, the auditory system serves as a very sensitive indicator of the condition and state of the device producing the sound.

Some examples of product sound quality are discussed below.

- *Vehicles*. The concept of product sound has been strongly affected by the need of the automotive industry to design sounds generated by vehicles to give an impression of high quality in every aspect. The sounds generated by a car indicate certain aspects of how it functions, and these positive sounds are thus enhanced to compete with other vehicle brands.

  In the case of vehicles, besides the sound of the engine, the product sounds also include the sounds generated by the wheels and the turbulence of air. Additionally, they also include the sounds generated when using different parts of the vehicle, such as opening the window, moving the seat, and pressing the buttons. The sounds and audio-tactile interaction when entering a car are important: opening the lock, using the door handle, and closing the door generate both auditory and tactile perceptions, which create an impression of the quality of finishing of the car.

  The sound of the engine when listened to inside the cabin should be designed such that it is not disturbing, although it must be heard over other sounds in the cabin, since it gives information on the functioning of the engine. Some cars even have mode switches where the sound level of the engine can be selected to be higher in the 'sporty' mode and lower in the luxurious' mode. The change in sound level can be implemented either by opening a channel to the engine chamber or, more simply, just by reproducing the engine sound using the car audio system.
- *Household appliances*. Most household appliances produce sound, either continuously or only when used actively. Continuous sounds should typically be almost silent, and they should not cause annoyance. In a device that is used occasionally, louder sounds may be acceptable. For example, a vacuum cleaner may be thought to be less effective if its sound is very soft. However, the sound of the vacuum cleaner should not have disturbing components, such as rattling or high-level, high-frequency sounds. The vacuum cleaner is an example of a product where the sound level of the device itself is decreasing with the evolution of vacuum cleaners. The first vacuum cleaners were really noisy, and very loud sound was an indication of high power. Fortunately, nowadays vacuum cleaners are more silent, and the association between loudness and assumed power is weaker. Consumers have learned that the vacuum cleaner motor can be both powerful and silent, and they pay more attention to the sounds created by the suction of air and by particles entering the suction tube.

- *Personal devices*. Electric shavers and hairdryers are also good examples of devices where the product sound quality has been taken into account. For example, the shaver should have a 'manly' and 'powerful' sound, communicating that the device is designed for a 'real man' to cut a 'strong beard'.

Summary

This chapter has broadly introduced the reader to sound quality. In the course of history, different aspects of sound quality have been of interest at different times. A factor unifying different sound quality trends seems to be the concept of product sound quality, which can be used to investigate sounds from different devices, systems, and from information and entertainment utilities. Although the components of sound quality are different in different cases, they always stem from the properties of the auditory system.

Further Reading

The theory and measurement of speech quality can be studied further by referring to Jekosch (2006), Möller (2000), Quackenbush *et al*. (1988), Raake (2007). Further knowledge of sound quality in audio reproduction can be found in Bech and Zacharov (2006) and Toole (2012). More information on recent trends in sound quality in concert halls can be found in Blesser and Salter (2007) and Pätynen *et al*. (2014).

References

3QUEST (2008) 3-Fold Quality Evaluation of Speech in Telecommunications. Application note, HEAD acoustics.

Appel, R. and Beerends, J.G. (2002) On the quality of hearing one's own voice. *J. Audio Eng. Soc.*, **50**(4), 237–248.

Barron, M. (ed.) (1993) Auditorium Acoustics and Architectural Design, E & FN Span.

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

Beerends, J.G. and Stemerdink, J.A. (1992) A perceptual audio quality measure based on a psychoacoustic sound representation. *J. Audio Eng. Soc.*, **40**, 963–978.

Beerends, J.G. and Stemerdink, J.A. (1994) A perceptual speech-quality measure based on a psychoacoustic sound representation. *J. Audio Eng. Soc.*, **42**(3), 115–123.

Beerends, J.G., Hekstra, A.P., Rix, A.W., and Hollier, M.P. (2002) Perceptual evaluation of speech quality (PESQ) the new ITU standard for end-to-end speech quality assessment part II: Psychoacoustic model. *J. Audio Eng. Soc.*, **50**(10), 765–778.

Beerends, J.G., Schmidmer, C., Berger, J., Obermann, M., Ullmann, R., Pomy, J., and Keyhl, M. (2013) Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part I – Temporal alignment. *J. Audio Eng. Soc.*, **61**(6), 366–384.

Beranek, L. (ed.)(1996) *Concert and Opera Halls – How they Sound.* Acoustical Society of America.

Blauert, J. (2005) *Communication Acoustics*. Springer.

Blauert, J. (2013a) Conceptual aspects regarding the qualification of spaces for aural performances. *Acta Acustica United with Acustica*, **99**(1), 1–13.

Blauert, J. (2013b) *The Technology of Binaural Listening*. Springer.

Blauert, J. and Jekosch, U. (1997) Sound-quality evaluation: a multi-layered problem. *Acta Acustica United with Acustica*, **83**(5), 747–753.

Blauert, J. and Jekosch, U. (2012) A layer model of sound quality. *J. Audio Eng. Soc.*, **60**(1/2), 4–12.

Blesser, B. and Salter, L.R. (2007) *Spaces Speak, Are You Listening?* MIT Press.

Bradley, J. (2011) Review of objective room acoustics measures and future needs. *Appl. Acoust.*, **72**, 713–720.

Davis, D. and Patronis, E. (2006) *Sound System Engineering*. Taylor & Francis.

EQUEST (2012) Echo Quality Evalution of Speech in Telecommunications. Application note, HEAD acoustics.

ETSI (2008) Speech Quality performance in the presence of background noise Part 3: Background noise transmission – Objective test methods. Recommendation EG 202 396-3, European Telecommunication Standards Institute.

ETSI (2014a) Universal mobile telecommunications system (UMTS); LTE; speech and video telephony terminal acoustic test specification. Recommendation 3GPP TS 26.132, European Telecommunication Standards Institute.

ETSI (2014b) Universal mobile telecommunications system (UMTS); LTE; Terminal acoustic characteristics for telephony; Requirements. Recommendation 3GPP TS 26.131, European Telecommunication Standards Institute.

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

Guski, R., Felscher-Suhr, U. and Schuemer, R. (1999) The concept of noise annoyance: How international experts see it. *J. Sound Vibr.*, **223**(4), 513–527.

Helenius, R., Keskinen, E., Haapakangas, A., and Hongisto, V. (2007) Acoustic environment in Finnish offices – The summary of questionnaire studies. *19th Int. Congr. Acoust.*

Hongisto, V. (2005) A model predicting the effect of speech of varying intelligibility on work performance. *Indoor Air*, **15**(6), 458–468.

Houtgast, T. and Steeneken, H.J.M. (1985) The modulation transfer function in room acoustics. *B&K Techn. Rev.*, **3**, 3–12.

IEC (2011) IEC 60268-16:2011 sound system equipment – Part 16: Objective rating of speech intelligibility by speech transmission index.

ISO 3382-1 (2009) Acoustics – measurement of room acoustic parameters – Part 1: Performance spaces. Organization for Standardization.

ISO 7240-19 (2007) Design, installation, commissioning and service of sound systems for emergency purposes. Standard 7240-19, International Organization for Standardization.

ITU-R BS.1116-1 (1997) Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. Recommendation, International Telecommunication Union, Geneva, Switzerland.

ITU-R BS.1534-1 (2003) Method for the subjective assessment of intermediate quality level of coding systems. Recommendation, International Telecommunication Union, Geneva, Switzerland.

ITU-T (1998a) Objective quality measurement of telephone–band (300–3400 Hz) speech codecs. Recommendation P.861, International Telecommunication Union, Geneva, Switzerland.

ITU-T (1998b) Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Recommendation P.862, International Telecommunication Union, Geneva, Switzerland.

ITU-T (2011) Perceptual objective listening quality assessment. Recommendation P.863, International Telecommunication Union, Geneva, Switzerland.

ITU-T P.800 (1996) Methods for subjective determination of transmission quality. Recommendation, International Telecommunication Union, Geneva, Switzerland.

ITU-T P.835 (2003) Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm. Recommendation, International Telecommunication Union, Geneva, Switzerland.

ITU-T P.910 (2008) Subjective video quality assessment methods for multimedia applications. Recommendation, International Telecommunication Union, Geneva, Switzerland.

Jekosch, U. (2006) *Voice and Speech Quality Perception: Assessment and Evaluation*. Springer.

Jensen, K. and Arens, E. (2005) Acoustical quality in office workstations, as assessed by occupant surveys. *Indoor Air*, pp. 2401–2405.

Karjalainen, M. (1985) A new auditory model for the evaluation of sound quality of audio systems. *Proc. of IEEE ICASSP'85*, pp. 608–611.

Kates, J.M. and Arehart, K.H. (2014) The hearing-aid speech quality index (HASQI) version 2. *J. Audio Eng. Soc.*, **62**(3), 99–117.

Kirkegaard, L. and Gulsrud, T. (2011) In search of a new paradigm: How do our parameters and measurement techniques constrain approaches to concert hall design? *Acoustics Today*, **7**(1), 7–14.

Lane, H. and Tranel, B. (1971) The lombard sign and the role of hearing in speech. *J. Speech, Lang. Hearing Res.*, **14**(4), 677–709.

Le Callet, P., Moller, S., and Perkis, A. (eds) (2012) *Qualinet White Paper On Definitions of Quality of Experience*. Qualinet.

Liebetrau, J., Sporer, T., Kämpf, S., and Schneider, S. (2010) Standardization of PEAQ-MC: Extension of ITU-R BS.1387-1 to multichannel audio. *40th Int. Audio Eng. Soc. Conf.: Spatial Audio* AES.

Lokki, T. (2014) Tasting music like wine: Sensory evaluation of concert halls. *Physics Today*, **67**(1), 27–32.

Mapp, P. (2005) Is STIPA a robust measure of speech intelligibility performance? *Audio Eng. Soc. Convention 118* AES.

Marquis-Favre, C., Premat, E., and Aubree, D. (2005a) Noise and its effects: a review on qualitative aspects of sound. Part II: Noise and annoyance. *Acta Acustica United with Acustica*, **91**(4), 626–642.

Marquis-Favre, C., Premat, E., Aubree, D., and Vallet, M. (2005b) Noise and its effects a review on qualitative aspects of sound. Part I: Notions and acoustic ratings. *Acta Acustica United with Acustica*, **91**(4), 613–625.

Möller, S. (2000) *Assessment and Prediction of Speech Quality In Telecommunications*. Springer.

Öhrström, E. and Rylander, R. (1982) Sleep disturbance effects of traffic noise: a laboratory study on after effects. *J. Sound Vibr.* **84**(1), 87–103.

Ouis, D. (2001) Annoyance from road traffic noise: A review. *J. Environment. Psych.*, **21**(1), 101–120.

Pätynen, J., Tervo, S., Robinson, P.W., and Lokki, T. (2014) Concert halls with strong lateral reflections enhance musical dynamics. *Proc. Nat. Acad. Sci.*, **111**(12), 4409–4414.

Pedersen, E. and Waye, K.P. (2004) Perception and annoyance due to wind turbine noise: a dose–response relationship. *J. Acoust. Soc. Am.*, **116**(6), 3460–3470.

Pejtersen, J., Allermann, L., Kristensen, T., and Poulsen, O. (2006) Indoor climate, psychosocial work environment and symptoms in open-plan offices. *Indoor Air.*, **16**(5), 392–401.

Peutz, V.M.A. (1971) Articulatory loss of consonants as a criterion for speech transmission in a room. *J. Audio Eng. Soc.*, **19**, 915–919.

Quackenbush, S.R. and Clements, M.A. (eds) (1988) *Objective Measures of Speach Quality*. Prentice–Hall.

Raake, A. (2007) *Speech Quality of VoIP: Assessment and Prediction*. John Wiley & Sons.

Rumsey, F., Zieliński, S., Kassier, R., and Bech, S. (2005) On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality. *J. Acoust. Soc. Am.*, **118**(2), 968–976.

Sabine, W.C. (1922) *Collected Papers on Acoustics*. Harvard University Press.

Schroeder, M.R. (1965) New method of measuring reverberation time. *J. Acoust. Soc. Am.*, **37**(3), 409–412.

Schroeder, M.R., Atal, B.S., and Hall, J.L. (1979) Optimizing digital speech coders by exploiting masking properties of the human ear. *J. Acoust. Soc. Am.*, **66**, 1647–1652.

Sporer, T., Liebetrau, J., and Schneider, S. (2009) Statistics of MUSHRA revisited. *Audio Eng. Soc. Convention 127* AES.

Steeneken, H.J.M. and Houtgast, T. (1985) A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J. Acoust. Soc. Am.*, **77**, 1060–1077.

Steeneken, H.J. and Houtgast, T. (2002) Validation of the revised STIr method. *Speech Commun.* **38**(3), 413–425.

Steeneken, H.J., van Wijngaarden, S.J., and Verhave, J.A. (2011) The evolution of the speech transmission index. *Audio Eng. Soc. Convention 130* AES.

Thiede, T., Treurniet, W.C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J.G., and Colomes, C. (2000) PEAQ - the ITU standard for objective measurement of perceived audio quality. *J. Audio Eng. Soc*, **48**(1/2), 3–29.

Toole, F. (2012) *Sound Reproduction: The Acoustics and Psychoacoustics of Loudspeakers and Rooms*. Focal Press.

TOSQA (2003) Option TOSQA. Telecommunications objective speech quality assessment. Application note, HEAD acoustics.

van Dorp Schuitman, J. (2011) *Auditory modelling for assessing room acoustics*. PhD thesis, Technische Universiteit Delft.

Vilkamo, J., Lokki, T., and Pulkki, V. (2009) Directional audio coding: Virtual microphone-based synthesis and subjective evaluation. *J. Audio Eng. Soc.*, **57**(9), 709–724.

Virjonen, P., Keränen, J., and Hongisto, V. (2009) Determination of acoustical conditions in open-plan offices: Proposal for new measurement method and target values. *Acta Acustica United with Acustica*, **95**(2), 279–290.

von Helmholtz, H. (1954) *On the Sensation of Tone*. Dover Publications.

Waye, K.P. and Öhrström, E. (2002) Psycho-acoustic characters of relevance for annoyance of wind turbine noise. *J. Sound Vibr.*, **250**(1), 65–73.