Advanced speech quality testing of modern telecommunication equipment: An overview

H.W. Gierlich*, F. Kettler

HEAD Acoustics GmbH, Ebertstr. 30a 52134, Herzogenrath, Germany

Received 1 January 2005; received in revised form 1 April 2005; accepted 8 June 2005
Available online 14 November 2005

Abstract

Modern telecommunication equipment differs quite significantly from equipment used in the past. While in traditional networks the behaviour of each transmission channel was deterministic, modern transmission systems may behave nonlinear and time variant even during a call. Signal processing is used extensively, mostly at the end points of a connection. The impairments introduced by the transmission systems and by the signal processing used are time variant as well and need to be investigated under different conversational conditions and by simulating different network and environmental conditions.

The paper gives an overview about simulation and testing techniques to be used for telecommunication equipment. The impact of different network conditions and the acoustical environment on speech quality is discussed and appropriate testing techniques for evaluating speech quality under different conversational aspects are described.

© 2005 Elsevier B.V. All rights reserved.

Keywords: Speech quality; Testing techniques; Network simulations; Noise field simulation; Double talk performance; Background noise performance; Echo performance; Loudness

1. Introduction

In today’s telecommunication systems various types of signal processing techniques are used to accommodate for the network conditions as well as the typical conditions of use. While in traditional TDM networks the parameters of a connection are mostly deterministic, modern networks behave nonlinear and time variant even during a call. When assessing the speech quality performance different parameters have to be taken into account. The classical separation between the terminal and the network is no longer possible. The assumption that creating a set of independent parameters for the terminal and the network allows an independent assessment of terminal and network parameters is no longer valid. Due to the time variance of the system, the performance of modern telecommunication equipment must be investigated under the different conversational conditions it is used in. Different network conditions as well as environmental conditions have to be included in order to achieve a valid and realistic view on the overall performance of telecommunication equipment.
2. Parameters influencing speech quality subjectively

When focusing on the speech communication aspect from the user’s point of view, the listening situation, the talking situation, the conversational situation (including double talk) and the background noise situation have to be taken into account and need to be regarded separately. A detailed investigation of the quality entities describing the quality of a speech communication service is given in [1]. An overview of the parameters influencing speech quality is shown in Fig. 1.

In the **listening situation** the speech signal may be degraded by a variety of impairments:

**Loudness**: The loudness perceived in the listening situation mainly contributes to speech quality. The speech signal loudness should be comparable to the loudness perceived when having a conversation between two persons in 1 m distance, the “orthotelephonic reference position” [2].

**Speech sound quality**: The speech sound quality is influenced by the bandwidth, the frequency response characteristics, the S/N and the distortion introduced by the transmission system. In modern telecommunication speech coding and other types of nonlinear and time variant signal processing are widely used. Objective assessment methods can be used under controlled and limited test conditions. Under unknown or more complex environmental or conversational conditions speech quality has to be determined subjectively.

**Speech intelligibility**: Although often ignored in the subjective experiments conducted for example for speech coder evaluation the intelligibility of speech is one of the basic features of a communication system. Especially in noisy environments sufficient speech intelligibility cannot always be achieved.

In the **talking situation** speech quality is influenced by impairments typically related to the user’s own voice: sidetone coupled either acoustically or electrically (by the handset receiver) to the user’s ear and echo. The echo signal is produced either by acoustical coupling between the receiver signal and the microphone of the far end terminal, or electrically by a hybrid echo e.g. resulting from a 4 wire to 2 wire hybrid in an analogue telephone. The perception of echo depends mainly on two parameters: the echo delay and the echo attenuation. The longer the delay the more echo attenuation is required (see [3]).

In the **conversational situation** speech quality is influenced by many parameters:

**Delay**: Delay is introduced by all communication systems. Algorithmic delay may be introduced by signal processing in the terminal such as background noise reduction (noise cancellation), echo cancellation, speech coding, voice controlled attenuation and others. Packetization as used e.g. in VoIP systems may further contribute to the delay introduced in a connection. The main effect of delay is the impairment of the conversational dynamics. For highly interactive conversational tasks a delay of no more than 150 ms is recommended (see [4]). Another impact of delay is the increased audibility of echo components for increasing delay (see [3]).

**Double talk capability**: The double talk capability of a telephone system is mainly determined by its echo control and the associated switching performance. Mostly all speech echo cancellers used in a connection are equipped with additional nonlinear processors which are introduced to minimize the residual echoes not cancelled by the speech echo canceller. These algorithms may introduce switching in sending and receiving direction. Any switching may result in the partial loss of speech segments, initial syllables may be suppressed, the intelligibility in the double talk situation may decrease as well as the naturalness of the conversation. Insufficient echo loss during double talk may result in echo impairments leading to a speech sound quality degradation and an intelligibility loss due to the masking by the echo signal. Requirements for switching and echo and their relationship to the perceived quality of a telephone system have been evaluated subjectively [5] and can be found e.g. in [6].
Background noise transmission performance: Although background noise is not a desired signal, the transmission of background noise in different conversational situations highly influences the overall quality perceived by the user. Background noise may influence the transmission quality during periods where no speech signal is present, during periods with near end speech and with far end speech. Furthermore, the speech signal may be degraded by noise cancellation systems in these scenarios. The subjective (auditory) relevance of the background noise situation in comparison to the other conversational situation is still under discussion. However, it can be stated that the quality of the background noise transmission as well as the speech quality in the presence of background noise is one of the dominating quality parameters especially in mobile communication systems.

A variety of test procedures have been developed for assessing the different parameters subjectively and most of them have been standardized (see e.g. [7–10]). The subjective tests form the basis for any objective procedure used for assessing speech communication systems.

3. Simulation of acoustical environments

Since today's terminal equipment is used mainly in noisy environments the simulation of typical acoustical environments in a lab-type environment is increasingly important. For the laboratory simulation different approaches can be made. The simplest one is to use uncorrelated noise sources and create a mostly diffuse sound field in an anechoic room. Typically random noise is used, spectrally shaped in order to simulate the average power density of an office-type environment with various speakers talking. The sound field can be calibrated in level and power density (see e.g. [11]). This approach leads to a diffuse sound field, the properties of which can be controlled easily and which can be reproduced reliably at different locations. Such a simulation, however, does not take into account the temporal structure of the typical background noise signals and furthermore is limited to a diffuse sound field.

A more realistic approach is the use of a loudspeaker playing back background noise signals recorded before. The loudspeaker can be calibrated so it can be ensured that the background noise level is simulated appropriately. Generally, this procedure allows the reproduction of realistic background noise signals concerning signal level and temporal structure. However, the direction of sound incidence is not reproduced adequately. As a result, the reproduction of the sound may differ in different lab environments depending on the acoustical conditions (anechoic or office-type test rooms).

A four loudspeaker arrangement as proposed e.g. in [12] further increases the realism of the reproduction. This procedure was originally investigated to reproduce binaurally recorded signals using artificial head technology [13]. It provides an almost realistic impression of direction and distance. Four loudspeakers are typically positioned equidistantly in a square formation around a central point (listening point). The binaural recordings are played back by feeding the two left-hand loudspeakers with the left artificial head signal. The right-hand loudspeakers are fed by the right ear signal of the artificial head. The loudspeakers are equalized by using an artificial head in the centre of the four loudspeaker arrangement. It is the goal of the equalization process to achieve the same ear signals at the artificial ear of the artificial head which was present during the recording. A physically exact sound field reproduction is not possible with this approach and would require a much more complex scenario like cross talk cancellation. But this loudspeaker arrangement has clearly practical advantages in terms of realization effort and robustness against slight variations of the listening position. Thus a mostly accurate simulation of the sound impression with respect to the artificial head position can be achieved. Once equalized, this arrangement can be used to play back any background noise recorded with an artificial head using the same equalization. It should be noted that this methodology is not a true sound field reproduction, but the results reported in [12] indicate that a reasonable direction and distance perception can be achieved. This methodology is currently studied for standardization in ETSI STF 273.

The advantage of the four loudspeaker arrangement is that a realistic simulation and impression can be achieved by relatively simple arrangements. Moreover, the setup is relatively robust against slight variations of the listening position. Due to the good reproduction of the spatial cues this arrangement is also suitable for subjective testing where background noise is required.

For higher-order array techniques (beamforming) the methodology is not suitable. Adaptive beamformers will place spatial zeros in the directions
where the loudspeakers are located. So the noise suppression will be unrealistically high. For such kinds of array techniques a simulation can be realized by recording the signals of the individual microphones in the real sound field and inserting the pre-recorded signals electrically in addition to the speaker.

For background noise simulation in vehicles the method described above is already used successfully for driving simulations and as test environment for mobile communication (see [14]). However, instead of using an artificial head for the recording, one or two calibrated microphones positioned close to the hands-free microphone in a car can be used. The equalization to this position has some advantages for testing hands-free applications equipped with more sophisticated microphone paths including noise reduction algorithms or microphone array solutions, since the equalization of a four loudspeaker arrangement can be made only to a limited sphere in space. A hands-free microphone in a car is typically positioned too far away from the artificial head (driver’s position) in order to achieve a realistic simulation when equalizing to the artificial head position.

4. Network simulation

In the past the testing concepts in telecommunication were divided into tests of terminal equipment and network equipment. This concept is still valid for TDM-type networks but cannot be applied fully to IP-networks and mobile networks.

For traditional TDM networks the network simulation is based on the assumption of a time invariant system after a connection is established. Under this assumption the network can be described by a few simple parameters: attenuation of the speech signal (junction loudness rating), system noise, delay introduced by the network, echo loss as provided either by the network or by the speech echo cancellers integrated in the network. This approach is made for network transmission planning in the E-model (see [15]). For terminal tests the network is simulated by static parameters, i.e. neither the time variant behaviour of speech echo cancellers in the network nor any other nonlinear or time variant signal processing is taken into account. The terminal is connected to the network termination point (NTP) which is clearly defined for the different types of analogue or digital wireline networks. If the interaction between a terminal and e.g. an echo canceller in the network is tested, the individual devices have to be integrated in the setup.

The test setup for VoIP configurations is more complex since the classical differentiation between terminal and network is no longer possible. While traditional networks have been designed for speech transmission, IP networks originally were designed for pure data transmission. The input signal (speech signal) is segmented into packets typically of a fixed length. Typical packet lengths used are in the range of 5–60 ms. Due to the packetization a minimum delay of the packet length is introduced into the transmission. IP systems can never be regarded as linear and/or time invariant:

- The network is not synchronized. Therefore all devices are operating on their own clock. This requires all devices to compensate for the jitter introduced by the time-varying clocks.
- One speech channel is not associated to a single physical channel, there is no guarantee for a certain bandwidth. Therefore the performance measured for a speech channel depends on the traffic transported over the same IP link. The delay of the individual speech packets sent over the network may vary due to the unpredictable network traffic resulting from other speech channels and data transmission.
- The network itself does not provide any means to guarantee a specific speech quality or to compensate (speech quality) problems resulting from impairments introduced by the network (e.g. echo cancellation).
- A wide range of speech coders may be used by the devices in an IP network.
- VAD (voice activity detection) may be used in order to reduce the network load. During periods without speech activity silence packets are transmitted and the missing sequences are replaced by comfort noise in the terminal or gateway at the receiving end.

When testing IP terminals or IP gateways the different network conditions have to be taken into account since the speech quality performance of the equipment under test is highly influenced by the different network conditions. Therefore an appropriate jitter and packet loss simulation has to be used (see e.g. [16]). In order to monitor and verify the network conditions, jitter and packet loss are monitored typically and related to the impairments
measured at the speech signal. A typical test setup used e.g. in international test events like the 1st, 2nd and 3rd ETSI SQTE [17] is shown in Fig. 2.

In mobile networks the radio channel quality may vary significantly. Disturbances result in bit errors. The radio network environment is affected by the factors (see [18]):

- co-channel interference (C/I),
- noise limitations (C/N),
- mobile speed (fading frequency),
- time dispersion,
- adjacent channel interference (C/A).

Discontinuous Transmission (DTX) reduces the total interference in the network and the speech impairment due to network interference may be decreased. Speech is only transmitted during speech activity, but due to DTX speech clipping may occur. Comfort noise insertion used on the far end in order to substitute the background noise (not transmitted) may further contribute to the degradation of the overall speech quality. DTX and comfort noise generation is part of the speech codec and can therefore differ in different implementations.

5. Objective speech quality testing

In Table 1 an overview is given about the impairments perceived subjectively and the correlating objective parameters.

5.1. Parameters relevant in single talk conditions

5.1.1. Loudness

The loudness in today’s telecommunication networks is considered the main parameter used for planning of telephone networks and the main parameter which has to be determined for all devices involved in a communication chain. The basic description of the loudness rating calculation as used today is found in [2,19,20]. The loudness rating is calculated

\[
\text{Loudness Rating} = -\frac{10}{m} \log_{10} \sum_{i=1}^{N} 10^{\left(\frac{L_{UME} - L_{RME}}{10}\right)} w_i,
\]

where \(L_{UME}\) is the weighted average mouth-to-ear loss (unknown system), \(L_{RME}\) the weighted average mouth-to-ear loss (reference system [21]), \(m\) the (partly) defines the loudness growth function \(Q(Z)\) and \(w_i\) the weighting factors, specific for SLR, RLR, OLR, STMR, LSTR.

For a given telephone or transmission system the values of \(L_{UME}\) can be derived from the different sensitivities \(S_{MJ}\) (mouth-to-junction) used for calculation of the sending loudness rating (SLR) and \(S_{JE}\) (junction-to-ear) used for the calculation of the receiving loudness rating (RLR) or the overall loudness rating (OLR).

In a similar manner the sidetone paths can be calculated: STMR is the sidetone masking rating describing the perceived loudness of the user’s own voice and LSTR (listener sidetone rating) describing the perceived loudness of room noise coupled to the user’s ear.

When assessing the loudness rating of a terminal the acoustical coupling between the terminal and the artificial ear and artificial mouth is of major importance. While in the past mostly simplified ear simulators were used—ITU-T P.57 Types 1 and 3.2 artificial ears [22]—in combination with a simplified artificial mouth (see [23]), recent studies, investigations and standards use the artificial head technique (HATS—head and torso simulator [24]) in order to access the terminals in a more realistic way [25]. Doing so in combination with an appropriate artificial ear (ITU-T Types 3.4 and 3.3 artificial ears [22]), the sound field of the artificial mouth can be simulated more realistically on the sending side. In receiving the effect of acoustical leakage between a handset or headset receiver is reproduced very realistically. When an appropriate positioning is used (see [26]) the measured transmission properties of a handset can be assessed comparable to an average human user [27]. This is of increasing importance since most modern handset designs do not provide a sufficient acoustical coupling to the human ear.

![Fig. 2. Test setup for VoIP configurations.](image-url)
<table>
<thead>
<tr>
<th>Relevant subjective parameter</th>
<th>Description</th>
<th>Correlating objective parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delay</td>
<td>Delay impairment discovered in conversational situations, especially with high interaction between the conversational partners leading to reduced conversational dynamics and unintended double talk situations</td>
<td>• Delay</td>
</tr>
<tr>
<td></td>
<td>(time variant) echo loss</td>
<td>• (time variant) echo loss</td>
</tr>
<tr>
<td></td>
<td>TCL (terminal coupling loss)</td>
<td>• TCL (terminal coupling loss)</td>
</tr>
<tr>
<td></td>
<td>Switching characteristics</td>
<td>• Switching characteristics</td>
</tr>
<tr>
<td></td>
<td>Occurrence of double talk performance</td>
<td>• Occurrence of double talk performance</td>
</tr>
<tr>
<td>Quality of background noise transmission</td>
<td>Impairments introduced to the background noise signal in the send direction during</td>
<td>• Minimum activation level in send direction</td>
</tr>
<tr>
<td></td>
<td>• Idle mode</td>
<td>• Comfort noise spectral and level adaptation</td>
</tr>
<tr>
<td></td>
<td>• Far end speech active in the presence of background noise</td>
<td>• Design of noise reduction systems</td>
</tr>
<tr>
<td></td>
<td>• Near end speech active in the presence of background noise</td>
<td>• Attenuation in send direction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Sensitivity of background noise detection (activation level, absolute level, level fluctuations)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Attenuation range introducing noise modulation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Switching characteristics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• D-value (sensitivity for background noise compared to the speech signal sensitivity)</td>
</tr>
<tr>
<td>Double talk performance</td>
<td>Typically in send and receive direction:</td>
<td>• Attenuation range</td>
</tr>
<tr>
<td></td>
<td>• Loudness variation between single and double talk periods</td>
<td>• Attenuation in send and receive direction during double talk</td>
</tr>
<tr>
<td></td>
<td>• Loudness variation during double talk</td>
<td>• Switching characteristics</td>
</tr>
<tr>
<td></td>
<td>• Echo disturbances</td>
<td>• Minimum activation level to switch over from receive to send direction and from send to receive direction</td>
</tr>
<tr>
<td></td>
<td>• Occurrence of speech gaps</td>
<td>• Echo attenuation during double talk</td>
</tr>
<tr>
<td></td>
<td>• Occurrence of artefacts uncorrelated to the transmitted speech signal</td>
<td>• Spectral and time dependent echo characteristics</td>
</tr>
<tr>
<td>Echo disturbances under single talk conditions</td>
<td>Talking related disturbance introduced between receive and send direction</td>
<td>• Echo attenuation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Echo level fluctuation vs. time</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Spectral echo attenuation</td>
</tr>
<tr>
<td>Speech sound quality</td>
<td>In send and receive direction</td>
<td>• Frequency response</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Distortions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Objective speech quality measures based on hearing models</td>
</tr>
<tr>
<td>Loudness of the speech signal</td>
<td>In send and receive direction</td>
<td>• Loudness ratings in send and receive</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Noise level</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Level fluctuations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Spectral characteristics</td>
</tr>
<tr>
<td>Noise</td>
<td>In send and receive direction, perceived during silent intervals</td>
<td></td>
</tr>
</tbody>
</table>
5.1.2. Frequency response, linearity, distortion and noise

In modern telephone systems the frequency response characteristics are mainly determined by the telephone terminal. It is assumed that—except for older analogue transmission systems where the telephone line may influence strongly the overall connection quality due to frequency-dependant attenuation and group delay distortion—the network components typically do not influence the frequency response characteristics significantly. Requirements for distortion, linearity and noise can be found in many relevant standards like [11]. Again, artificial head technique is used for assessing these parameters on the acoustical side of the terminals.

5.1.3. Speech sound quality—listening speech quality

Speech sound quality is influenced by many parameters as they have been described before: frequency response, loudness rating, non-linear distortions, switching, noise and others. Since any degradation of the speech sound quality is mostly perceived in the listening situation the main focus of speech sound quality is during one way transmission (single talk condition).

Due to the use of different types of speech coders and the impairments introduced by the transmission network (packet loss, jitter, interference) the listening speech quality may be degraded. The type of distortion added to the signal is typically uncorrelated to the speech signal. This effect cannot be assessed by loudness ratings, frequency response or traditional distortion measurements. The assessment methods used under these conditions are hearing model based comparison techniques. The principle of these assessment methods is shown in Fig. 3.

All methods (see e.g. [28–30]) compare an original signal \( x(t) \) with a degraded signal \( y(t) \) that is the result of passing \( x(t) \) through a communication system. The output is a prediction of the perceived quality that would be given to \( y(t) \) by subjects in a subjective listening test. In a first step the signal \( x(t) \) is adapted to the delay and the bandwidth limitation introduced by the transmission system. The time alignment between original signal and transmitted signal is made in blocks in order to take care of time varying delays in transmission systems. It is assumed that the delay within a block is constant.

The comparison is made based on the time aligned signals \( x'(t+\tau) \) and \( y(t) \). The key to this process is transformation of both the original and the degraded signals to an internal representation that is comparable to the psychophysical representation of audio signals in the human auditory system, taking into account the frequency resolution and the loudness sensation of the human ear.

Mostly all of the procedures provide level alignment to a calibrated listening level, time-frequency mapping, frequency warping, and loudness scaling. Since not all differences found in the representations of \( x(t) \) and \( y(t) \) contribute to the perceived degradation of the processed signal, minor steady-state differences between original and degraded signal are compensated and not taken into account for the comparison. More severe effects, like e.g. rapid variations, are only partially compensated so that a residual effect remains and contributes to the overall perceptual disturbance. Segments with missing signal energy in the processed signal are weighted less than any components added to the processed signal \( y(t) \) (asymmetry weighting). By this procedure a small number of quality indicators is used to model the most important subjective effects.

The error parameters are computed in a cognitive model and combined to an internal objective listening quality score. By appropriate mapping this

---

**Fig. 3.** Principle of reference signal based objective listening speech quality assessment.
internal score is adjusted to the MOS results of the subjective listening experiments (MOS-LQS, mean opinion score-listening quality subjective, see [31]) conducted using the same types of impairments for which the objective speech quality assessment method is intended to be used. In order to distinguish this objective result from the subjective MOS-LQS results, this score is named MOS-LQO (mean opinion score-listening quality objective, see [31]).

The objective speech quality evaluation procedure recommended by ITU-T is PESQ which is described in P.862 [32]. Like many other methods this procedure is suitable for the electrical access of speech signals e.g. using interfaces provided at the network termination points. The influence of terminals e.g. degradations of the speech signals introduced by non ideal frequency responses, transducer distortions etc. are not modelled. TOSQA2001 [30,33] allows the acoustical access and takes into account the impairments introduced by the acoustical components. In order to achieve realistic conditions an artificial head (HATS—head and torso simulator [22,24]) equipped with an artificial mouth and a type 3.4 artificial ear is used for assessing the acoustical input and output signals of a terminal.

A general drawback of all hearing model-based approaches is the assumption of the same listening level in a subjective listening test under all conditions. The standard listening level assumed is 79 dB SPL and all signals are equalized to this listening level. The psychoacoustic model including the human ear sensitivity characteristics and all masking thresholds are fixed to this listening level. The wide range of different signal levels found in the networks but also at the terminal outputs is not taken into account by the models. Therefore the determination of the perceived loudness has to be made in addition to the determination of any MOS-LQO value. For higher or lower level acoustical output signals deviating from the standard listening level the predicted MOS-LQO value may be incorrect.

5.1.4. Terminal coupling loss—echo loss

The more delay is found in a connection the higher is the importance of the echo loss between the acoustical receiver and the terminal microphone. Any signal component received by the terminal and transmitted back to the far end user may be perceived as echo by the far end user. A further source of echo may be an insufficient transhybrid loss e.g. in analogue networks which is not properly compensated by the network speech echo cancellers. From the perception point of view the actual echo source is of minor importance, the requirement for the echo attenuation depending on the delay inserted in a connection is described e.g. in [3] (see Fig. 4).

The echo loss has to be provided during the entire call without any temporal change of the echo loss. Testing techniques for evaluating the convergence behaviour of echo cancellers in the networks can be found in [34] and for hands-free terminals in [6].

5.2. Testing procedures taking into account the conversational situation

Most measurements and parameters listed in Chapter 5.1 are based on the assumption that the systems under test are broadly linear and time invariant (LTI systems). In cases where the signal processing in the components do not hold these conditions (except the codec), signal processing may influence the speech transmission quite substantially, especially in the conversational situation. The signal processing procedures to be expected are voice activated switching and amplification, echo cancellation (acoustic and electric), noise reduction,
The importance of double talk performance and background noise transmission was derived by conversational tests and investigated in more detail by using specific double talk tests and listening only tests (see [5,35,36]). The signal processing components expected in non-LTI (nonlinear time variant) systems are found e.g. in small (mobile) terminals, hands-free terminals, echo cancellers [34], packetizing equipment.

When assessing the performance of telecommunication equipment in the conversational situation the test setup has to simulate the different conversational situations. Besides the listening situation the following conversational situations mainly have to be taken into account:

- double talk,
- transition periods from sending to receiving,
- transition periods from single talk to double talk,
- the presence of background noise during periods with no speech activity, with near end speech and with far end speech.

In order to simulate the different conversational situations different types of speech-like signals can be used to assess the different objective parameters reliably and reproducibly. A set of test signals to be used for advanced measurements is specifically adapted to the measurement and can be found in ITU-T Recommendations P.501 [37] and P.50. [38]. The test procedures and appropriate analyses are basically described in ITU-T Recommendation P.502 [39].

Especially for all tests which require the measurement of switching times, time variant amplification or attenuation, speech like signals are required which provide exactly defined characteristics in time and frequency. These signals simulate speech like properties to some extent but can be defined precisely.

5.2.1. Double talk performance

The double talk performance of a hands-free terminal and the resulting requirements have been investigated in great detail (see [5]). The evaluation of double talk periods led to a classification of systems which allows to characterize their performance objectively by a set of parameters (see [3,6]). Although the subjective tests were mainly based on the evaluation of hands-free systems the results can be used for other terminals and network equipment as well since (a) the impairments are mostly perceived by the far end user, (b) the signal processing algorithms of different devices are comparable and (c) the subjects were not informed about any detail of the setup in the subjective tests. Table 2 gives an overview about the required performance parameters. Their impact on the performance perceived subjectively is shown in Table 3.

The performance parameters during double talk are as follows:

**Attenuation range in send and receive direction** ($a_{Hsd}$, $a_{Hrdt}$): For the measurement of the attenuation range a time-multiplex test signal (see Fig. 5) is inserted simultaneously in send and receive direction. The test signal consists of a sequence of

Table 2

<table>
<thead>
<tr>
<th>Type 1 (full duplex)</th>
<th>Type 2 (partial duplex)</th>
<th>Type 3 (non-duplex)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TELR_{DT} (dB)</td>
<td>$\geq 37$</td>
<td>$\geq 33$</td>
</tr>
<tr>
<td>$a_{Hsd}$ (dB)</td>
<td>$\leq 3$</td>
<td>$\leq 6$</td>
</tr>
<tr>
<td>$a_{Hrdt}$ (dB)</td>
<td>$\leq 3$</td>
<td>$\leq 5$</td>
</tr>
</tbody>
</table>

TEL$_{RDT}$: talker echo loudness rating during double talk, $a_{Hsd}$: attenuation range in send during double talk, $a_{Hrdt}$: attenuation range in receive during double talk.

Table 3

<table>
<thead>
<tr>
<th>MOS</th>
<th>$\geq 4.0$</th>
<th>$4.0-3.5$</th>
<th>$3.5-3.0$</th>
<th>$3.0-2.5$</th>
<th>$2.5-2.0$</th>
<th>$\leq 2.0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>TELR_{DT} (dB)</td>
<td>$\geq 37$</td>
<td>$\geq 33$</td>
<td>$\geq 27$</td>
<td>$\geq 21$</td>
<td>$\geq 13$</td>
<td>$&lt; 13$</td>
</tr>
<tr>
<td>$a_{Hsd}$ (dB)</td>
<td>$\leq 3$</td>
<td>$\leq 6$</td>
<td>$\leq 9$</td>
<td>$\leq 12$</td>
<td>$\leq 15$</td>
<td>$&gt; 15$</td>
</tr>
<tr>
<td>$a_{Hrdt}$ (dB)</td>
<td>$\leq 3$</td>
<td>$\leq 5$</td>
<td>$\leq 8$</td>
<td>$\leq 10$</td>
<td>$\leq 12$</td>
<td>$&gt; 12$</td>
</tr>
</tbody>
</table>

(TEL$_{RDT}$, $a_{Hsd}$, $a_{Hrdt}$) as function of correlated MOS scores derived from listening only test (see [5,40]) MOS—mean opinion score (1-excellent; 5-poor).
partially overlapping composite source signals ([37,41]). The attenuation range is measured during the active parts of the signal being transmitted in the direction under test. The attenuation range is the difference between minimum and maximum level achieved during these periods, referred to the excitation signal. For the measurement a very short time constant for level measurements is necessary. Typically a time constant of about 5 ms for the level analysis should be used.

The test signal may be constructed in such a way that the level of each composite source signal burst varies from signal to signal. By this procedure different signal level combinations can be achieved and the attenuation range can be determined depending on the signal energy in sending and receiving.

**Talker echo loudness rating/echo attenuation (TELR<sub>dt</sub>):** As seen in Table 3 the talker echo contributes significantly to the perceived speech quality in the double talk situation. The talker echo assessment can be made by using a frequency-multiplexed test signal consisting of two voiced sounds inserted simultaneously in send and receive direction. By applying an appropriate comb filter technique the echo signal can be extracted from the double talk signal and the echo loss during double talk can be determined. Another possibility is to use a sequence of CSS double talk signals overlapping partially in time (see Fig. 5). When using the measurement signal shown in Fig. 5, the echo attenuation can be determined up to a certain extent as well. In this case the overlapping of the sequences must be constructed in such a way that the echo signal is present during the pauses of the double talk signal in sending direction. This measurement is not exactly a double talk examination, however, it gives a good estimation of real situations since the human ear is mainly sensitive to the echo signal during periods, where the double talk signal itself does not really mask the echo signal. The echo loss is subsequently determined using a short-term FFT analysis or a short-term level analysis (short time constant of 5 ms) of the echo signal, referring it to the excitation signal and calculating the echo loss.

**Switching characteristics, switching times:** The measurement and optimization of switching times is required for all systems which do not operate fully duplex. Such systems typically show switching behaviour, sometimes a minimum speech level is required to activate the system. The switching characteristics can be evaluated basically in a similar manner as the evaluation of the attenuation range. The definitions of switching times, etc. can be found in P.340 [6]. The analysis of the switching characteristics is always made during the active part of one channel whereas the opposite channel provides a pause. The measured signal again is referred to the excitation signal.

### 5.2.2. Background noise performance

In background noise situations the perceived speech quality is influenced by more than one parameter, the scenario is multidimensional. The human listener focuses on two signals: the speech signal and the background noise signal (see also [42]). The quality judgement is made based on the perception of the speech being transmitted together with background noise, on the perception of the transmitted background noise in the same scenario, i.e. during speech activity and finally on the perception of the transmission quality of the background noise signal during periods where no speech signal is present. Therefore, an objective method describing the quality in this scenario must take into account these different dimensions. Furthermore, subjective experiments conducted with expert listeners indicate that there exist at least two groups of “typical listeners”:

- One group is very sensitive to any impairments added to the speech signal but is not sensitive against the background noise level, the other group is less sensitive against impairments added to the speech signal but prefers a maximum of noise reduction. A complete background noise evaluation method does not exist yet but for the different situations instrumental procedures can be used.
When focusing only on the background noise transmission with no speech signal present a level variation analysis or a spectrographic analysis can be used to analyse the transmitted signal. Typically the output signal level is referred to the input signal level and the signal differences are analysed. This method can be applied only if the reference (input signal) can be assessed. Preliminary auditory investigations indicate that a level variation of $\pm 3$ dB versus time should not be exceeded. However, more information about the transmission characteristics of background noise can be found when comparing output and input signals based on a spectral analysis. The output/input signal is analysed spectrographically and the spectral difference between input and output is calculated depending on time and frequency. Again a reference signal is needed to perform the analysis. No complete validation of this method has been made yet. However, there is some indication that variations in time and/or frequency should not exceed $\pm 3$ dB. In general, the disadvantage of such simple spectral difference methods is the non-existent relationship to the human ear signal processing. There is a high probability to get misleading results due to poorly adapted frequency resolution, not taking into account the masking effects in time and frequency, disregarding the nonlinearity of the human ear signal processing and others.

Another possibility using the psychoacoustically motivated method “relative approach” seems to be promising. The basis for the analysis is a hearing model according to [43]. In contrast to all other methods the relative approach does not use any reference signal in its present form. The nonlinear relationship between sound pressure level and auditory perceived loudness is taken into account by time/frequency warping in a Bark filter bank and proper integration of the individual outputs. The filter bank is realized in the time domain. The output signals of the filter bank are rectified and integrated and an envelope is generated. The three-dimensional output of the hearing model is the basis for the relative approach. In each critical band the long term level (integration time: 2–4 s) is compared with the short-term level (2 ms). An overall value can be derived for example by applying the following equation (see [44]):

$$Q = f(N, S) + f \left( \sum_{i=1}^{24} |F_G(i - 1) - F_G(i)| \cdot w_1(iF_G(i)) + \sum_{n=1}^{T} |F_G(i, n) - F_G(i, n + 1)| \cdot w_2(i, F_G(i)) \right),$$

where $F_G(i)$ is a mean value of the critical band level over a period $T$ of 2–4 s, $F_G(0) = F_G(1)$, $F_G(i, n)$ is a mean value of the critical band level over a much shorter period (approx. 2 ms), $n$ is the current (time-dependent) value. The weighting factors $w_1(i, F_G(i))$, $w_2(i, F_G(i))$ depend on the critical band level $F_G(i)$. In addition, the overall value is influenced by the function $f(N, S)$ which describes an auditory factor, dependent on the loudness $N$ and the sharpness $S$. Fig. 6 shows the block diagram including the hearing model [43] and the calculation of the relative approach signals in the Bark domain.

The method can be applied for short-term stationary background noise signals e.g. car noise during constant driving conditions or office noise with low dynamic variations. The typical evaluations conducted with this method are the initial convergence characteristics of a background noise reduction algorithm or the match of comfort noise

![Fig. 6. Model of the “relative approach”: hearing model and calculation of the energy differences in critical bands.](image-url)
inserted during periods where the far-end speaker is active and the near-end signal is substituted by comfort noise in cases where the echo suppressor (nonlinear processor) is active. Two examples for this evaluation are shown in Figs. 7 and 8.

When focussing on the quality of the transmitted speech signal all objective methods proposed have some disadvantages. S/N based methods as described in [45] do not take into account the perceptual effects of the human ear and are limited in their application to specific types of noise reduction algorithms. More advanced methods like PESQ [32] or TOSQA2001 [30] take into account the human ear signal processing. However, it is known (see [30,46]) that such technologies do not yield sufficient results in conditions with S/N-ratios of less than about 20 dB.

The spectral representation of the relative approach is shown (the outputs of the relative approach as shown in Fig. 6 before calculating the auditory sensation). Although only the output signal is analysed, very important information can be found in this “spectral” representation. In the beginning, when the system is switched on, a broadband “onset” peak is detected which is audible and annoying. Between 1.5 and 3 s spectral structures can be found between 1 and 4 kHz which again are annoying. Expert test confirm these findings, the onset peak is a strong “click” and the spectral structures between 1.5 and 4 kHz sound as if someone would scratch a microphone. Both observations are due to processing artefacts of the background noise reduction algorithm since the background noise does not vary significantly during the analysis period. This always has to be taken into account when applying the relative approach in its present form: Since no reference signal is used, the algorithm is not able to distinguish between artefacts introduced by the signal itself and the processing. Therefore, the algorithm should be applied only for quasi stationary background noise signals.

While background noise is present in sending direction a speech signal is inserted in receiving direction (between 4.5 and 9 s). The transmitted signal in sending direction (background noise) is analysed. This situation is specifically critical for echo cancellers since the adaptive filter is disturbed by the background noise signal and may diverge. Typically the echo canceller is backed up by gain switching and comfort noise insertion in order to mask or reduce possible echo components.

Fig. 7. Convergence behaviour of a background noise during the initial call: “relative approach” analysis of the processed (output) signal; bright colours indicate audible components in time or frequency.

Fig. 8. Spectral analysis (upper) and “relative approach” analysis (lower) of the transmitted background noise signal with inserted comfort noise.
In the example shown in Fig. 8, from simple spectral comparison (upper) it is obvious that the spectrum of the transmitted background noise signal is similar but not equivalent in the time periods before and during far end speech transmission. Instead of the original signal comfort noise which is spectrally shaped is inserted. The analysis indicates a difference but no judgement is possible whether this difference is audibly significant or not. The relative approach analysis (lower), however, indicates no audible difference which can be confirmed by expert tests.

6. Summary and outlook

A variety of test methods are available for the qualification and optimization of modern telecommunication equipment. In combination with the appropriate simulation techniques for network and environmental conditions a reasonable performance overview can be given for the different conversational situations. In most cases, it is necessary to assess single parameters and evaluate single signal processing units in order to verify and improve the performance or to optimize settings. In addition, a combined, final quality number is also desired. However, there is currently no way to describe the multidimensional speech quality by a single number indicating the overall speech quality. The combination of the different single values for the different speech quality parameters to an overall speech quality number for non-LTI systems is still an open point. Further work is needed to improve the quality evaluation procedures for background noise transmission and to adapt the existing quality evaluation methods to the new signal processing used in recent telecommunication products. Additional work is required to define test signals similar to the CSS signal [7] which provide more speech-like characteristics without losing the advantages for objective testing.

References

[11] ETSI TBR - 008, Integrated Services Digital Network (ISDN); Telephony 3.1 kHz teleservice; Attachment requirements for handset terminals.


