ETSI TS 103 558 V1.1.1 (2019-11)

Speech and multimedia Transmission Quality (STQ); Methods for objective assessment of listening effort

Reference
DTS/STQ-264

Keywords
assessment, listening effort, model

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

Important notice

The present document can be downloaded from:
http://www.etsi.org/standards-search

The present document may be made available in electronic versions and/or in print. The content of any electronic and/or print versions of the present document shall not be modified without the prior written authorization of ETSI. In case of any existing or perceived difference in contents between such versions and/or in print, the prevailing version of an ETSI deliverable is the one made publicly available in PDF format at www.etsi.org/deliver.

Users of the present document should be aware that the document may be subject to revision or change of status. Information on the current status of this and other ETSI documents is available at https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

If you find errors in the present document, please send your comment to one of the following services:
https://portal.etsi.org/People/CommiteeSupportStaff.aspx

Contents

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI IPR Policy, no investigation, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

Foreword

This Technical Specification (TS) has been produced by ETSI Technical Committee Speech and multimedia Transmission Quality (STQ).

The present document describes auditory and instrumental test methodologies for the prediction of perceived speech signal in the presence of background noise of modern communication terminals. Audio bandwidths from narrowband up to super-wideband and fullband are considered.

Modal verbs terminology

In the present document "**shall**", "**shall not**", "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

1 Scope

The present document describes auditory and instrumental testing methodologies, which can be used to evaluate the perceived listening effort in the following speech communication scenarios at acoustical interfaces in the presence of acoustical near-end ambient noise.

Similar to other instrumental quality prediction methods like e.g. ETSI TS 103 281 [4] or Recommendation ITU-T P.863 [i.2] valid objective predictions can only be made based on a specific listening test design and on auditory results obtained in such tests.

The present document specifies the test design and reference conditions used to evaluate listening effort subjectively.

The objective prediction model specified are based on this test design and validated against the results of the underlying subjective tests; only normal hearing listeners are considered. The usage for hearing impaired listeners is for further study.

Several application scenarios and types of terminals are covered:

- (Mobile) Handset.
- In-car communication systems.

The following applications are for further study:

- Headset (including active noise cancelling devices).
- Group audio terminals.
- Mobile handheld hands-free.
- Vehicle hands-free.
- Fixed, mobile and IP-based networks (including impairments).

Binaural as well as monaural recording situations are covered. The listening effort prediction model utilizes binaural signals for acoustical recordings and monaural signals for electrical recordings.

2 References

2.1 Normative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

Referenced documents that are not found to be publicly available in the expected location might be found at https://docbox.etsi.org/Reference.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long-term validity.

The following referenced documents are necessary for the application of the present document.

[1] Recommendation ITU-T P.800: "Methods for subjective determination of transmission quality".

[2] Recommendation ITU-T P.835: "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm".

[3] Recommendation ITU-T P.56: "Objective measurement of active speech level".

[4] ETSI TS 103 281: "Speech and multimedia Transmission Quality (STQ); Speech quality in the presence of background noise: Objective test methods for super-wideband and fullband terminals".

[5] Recommendation ITU-T P.501: "Test signals for use in telephonometry".

[6] Recommendation ITU-T P.57: "Artificial ears".

[7] Recommendation ITU-T P.58: "Head and torso simulator for telephonometry".

[8] ITU-T Handbook: "Practical procedures for subjective testing", 2011.

[9] ITU-T Handbook: "Handbook on Telephonometry", 1992.

[10] Directive 2003/10/EC of the European Parliament and of the Council of 6 February 2003 on the minimum health and safety requirements regarding the exposure of workers to the risks arising from physical agents (noise), Official Journal; OJ L42, 15.02.2003, p.38.

[11] Recommendation ITU-T G.160: "Voice enhancement devices".

[12] Roland Sottek: "A Hearing Model Approach to Time-Varying Loudness". Acta Acustica united with Acustica, vol. 102(4), pp. 725-744, 2016.

[13] Til Aach and Volker Metzler: "Defect Interpolation in Digital Radiography - How Object-Oriented Transform Coding Helps". SPIE Vol. 4322: Medical Imaging 2001.

[14] Rui Wan, Nathaniel I. Durlach and H. Steven Colburn: "Application of a short-time version of the Equalization-Cancellation model to speech intelligibility experiments with speech maskers". The Journal of the Acoustical Society of America, Vol. 136/2, pages 768-776, 2014.

[15] Nathaniel I. Durlach: "Equalization and Cancellation Theory of Binaural Masking-Level Differences", The Journal of the Acoustical Society of America 35(8), pages 1206-1218, 1963.

NOTE: Available at http://daviddurlach.com/nat-mem/wp-content/uploads/2016/10/Durlach_JASA_1963_ECModel.pdf.

[16] J. H. Friedman: "Multivariate Adaptive Regression Splines", The Annals of Statistics, Vol 19, No. 1, pp. 1-141, 1991.

NOTE: Available at https://projecteuclid.org/download/pdf_1/euclid.aos/1176347963.

[17] ISO 389-7:2005: "Acoustics - Reference zero for the calibration of audiometric equipment - Part 7: Reference threshold of hearing under free-field and diffuse-field listening conditions".

[18] ANSI S3.5-1997: "Methods for Calculation of the Speech Intelligibility Index".

[19] IEC 61260-1:2014: "Electroacoustics - Octave-band and fractional-octave-band filters - Part 1: Specifications".

[20] IEC 61672-1:2013: "Electroacoustics - Sound level meters - Part 1: Specifications".

[21] Recommendation ITU-T P.810: "Modulated noise reference unit (MNRU)".

[22] Recommendation ITU-T P.50: "Artificial voices".

[23] Recommendation ITU-T P.Imp830: "Implementer's Guide for P.830 (Subjective performance assessment of telephone-band and wideband digital codecs)".

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long-term validity.

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

[i.1] Recommendation ITU-T P.10/G.100: "Vocabulary for performance and quality of service".

[i.2] Recommendation ITU-T P.863: "Perceptual objective listening quality assessment".

[i.3] Recommendation ITU-T P.1401: "Methods, metrics and procedures for statistical evaluation, qualifying and comparison of objective quality prediction models".

[i.4] ETSI TS 103 224: "Speech and multimedia Transmission Quality (STQ); A sound field reproduction method for terminal testing including a background noise database".

[i.5] ETSI ES 202 396-1: "Speech and multimedia Transmission Quality (STQ); Speech quality performance in the presence of background noise; Part 1: Background noise simulation technique and background noise database".

[i.6] ETSI TS 103 106: "Speech and multimedia Transmission Quality (STQ); Speech quality performance in the presence of background noise: Background noise transmission for mobile terminals-objective test methods".

[i.7] Bendat, J. S.; Piersol, A. G.: "Engineering applications of correlation and spectral analysis". New York, Wiley-Interscience, 1980.

[i.8] Alexandre Chabot-Leclerc: "PAMBOX: A Python auditory modeling toolbox". EuroScipy proceedings, Cambridge, 27-30 August 2014.

[i.9] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen: "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech". IEEE Transactions on Audio, Speech and Language Processing, Vol 19 No. 7, 2011.

[i.10] ETSI EG 202 396-3: "Speech and multimedia Transmission Quality (STQ); Speech Quality performance in the presence of background noise; Part 3: Background noise transmission - Objective test methods".

[i.11] Gheorghe Micula, Sanda Micula: "Handbook of Splines", Springer, 1999.

[i.12] J. Reimes, G. Mauer und H. W. Gierlich: "Auditory Evaluation of Receive-Side Speech Enhancement Algorithms". Proceedings of DAGA 2016, Aachen.

[i.13] Jan Reimes and Christian Lüke: "Perceived Listening Effort for In-car Communication systems". Proceedings of 13th ITG Conference on Speech Communication, Oldenburg.

[i.14] Rabea Landgraf, Johannes Köhler-Kaeß, Christian Lüke, Oliver Niebuhr, and Gerhard Schmidt: "Can you hear me now? Reducing the Lombard effect in a driving car using an in-car communication system", in Proceedings Speech Prosody, (Boston, MA, USA), June 2016.

[i.15] ETSI EG 202 518: "Speech and multimedia Transmission Quality (STQ); Acoustic Output of Terminal Equipment; Maximum Levels and Test Methodology for Various Applications".

3 Definition of terms, symbols and abbreviations

3.1 Terms

Void.

3.2 Symbols

For the purposes of the present document, the following symbols apply:

ACT	Frames in the signal(s) containing active speech
dB_{Pa}	Sound Pressure Level in dB, referenced to 1 Pa
dB_{SPL}	Sound Pressure Level in dB, referenced to 20 µPa
F_N	Noise flag, indicating if the prediction algorithm uses a noise-only reference or not
G_{FB}	Gain in dB, which is used to scale the feedback signal
G_{out}	Gain in dB, which is used to increase the output volume of an ICC system
M_A	Number of frames, which contain active speech
T_{FB}	Time between playback of a sound over an ICC system and the corresponding feedback into the system
T_{ICC}	Processing time of an ICC system

3.3 Abbreviations

For the purposes of the present document, the following abbreviations apply:

AMR	Adaptive Multi-Rate codec (narrowband)
AMR-WB	Adaptive Multi-Rate codec (WideBand)
ASL	Active Speech Level
BWE	BandWidth Extension
DRP	Drum Reference Point
DUT	Device Under Test
FB	FullBand
HATS	Head And Torso Simulator
ICC	In-Car Communication
IIR	Infinite Impulse Response
IR	Impulse Response
LE	Listening Effort
MARS	Multivariate Adaptive Regression Splines
MNRU	Modulated Noise Reference Unit
MOS	Mean Opinion Score
MOS_{LE}	Listening Effort on MOS scale
MRP	Mouth Reference Point
NB	NarrowBand
NELE	Near-End Listening Enhancement
NLMS	Normalized Least-Mean Square (adaptive filter)
NS	Noise Suppression
PC	Personal Computer
PCM	Pulse-Code Modulation
POI	Point Of Interconnect
SII	Speech Intelligibility Index
SNR	Signal-to-Noise Ratio
SPNF	Signal Processing Network Function
SQ	Speech Quality
STEC	Short-Time Equalization-Cancellation
SWB	Super-WideBand
WB	WideBand

4 Introduction

Communication in noisy environments may be extremely stressful for the person located at the near-end side. Since the background noise is originated from the natural environment, it can usually not be reduced for the listener. In addition, the perceived signal may be disturbed by other linear or non-linear signal processing. In consequence, speech intelligibility may decrease, i.e. listening effort may increase, respectively.

The present document describes an auditory test design for the assessment of perceived listening effort as well as an instrumental prediction model. Both provide MOS values based on binaural recording and listening to real speech signals in noisy conditions. The audio bandwidth of the model is fullband (20 Hz - 20 kHz) according to [i.1]. Speech signals may be presented in narrow-band, wideband, super-wideband or fullband.

In contrast to "classical" intelligibility tests, the auditory assessment of listening effort collects opinion scores instead of "measuring" the word error rate of multiple test subjects. In general, it seems difficult to compare results of these two methods, but since both metrics obviously depend on similar conditions (SNR, temporal and spectral structure of the background noise, speech degradations), a certain correlation can be expected. Annex B includes a summary of studies investigating this relationship.

5 Auditory test design

5.1 Overview

The basis of any perceptually based measure, which models the behaviour of human test persons, are auditory tests. In general, these tests are carried out with naïve test persons, who are asked to rate a certain quality aspect of a presented speech sample.

For the assessment of listening effort, a test design related to Recommendations ITU-T P.800 [1] and P.835 [2] with multiple attributes is chosen. The additional assessment of any speech quality attribute is in general optional, but is strongly recommended. It may help the test subjects to better differentiate between the ambient noise and speech-related degradations. Any speech quality results obtained with this procedure are outside the scope of the present document.

5.2 Speech material

The source speech database (far end signal) to be used for data collection and listening tests needs to consist of at least eight samples (2 male and 2 female talkers, 2 samples per talker). Appropriate test signals for multiple languages and in fullband bandwidth can be found in Recommendation ITU-T P.501 [5] or in annex E of ETSI TS 103 281 [4].

Each sentence shall be centred in a time window of 4 seconds. The minimum duration of an active speech material shall be 1 second, i.e. resulting in not more than 1,5 seconds of leading and trailing silence. The duration of the active speech material shall not exceed 3 seconds, which correspond to a minimum leading/trailing silence period of 0,5 seconds. The samples shall be concatenated to a single speech sequence for the measurement of the degraded signals.

For proper conditioning of systems including signal processing, a conditioning sequence consisting of an initial silence period followed by at least four different sentences from four different talkers is used.

The concatenated speech sequence shall always be available as in fullband. This signal is denoted as the reference signal $r(k)$ in the following clauses. Depending on the application, a pre-filtering (e.g. to narrow-band or wideband) may be necessary for the electrical insertion of the test sequence in the Device Under tesT (DUT) in receiving direction.

5.3 Background noise simulation

The presence of ambient noise is the most influencing aspect on listening effort. In order to provide an accurate sound field reproduction at the DUT and/or at the listener position, the method according to ETSI TS 103 224 [i.4] shall be used for the recording of samples. The present document includes two recording/playback procedures: head-oriented and generic sound field reproduction. Depending on the application, the most suitable recording/playback procedures shall be selected.

The number of different background noises may vary from one application to the other. For in-car communication scenarios for example, only car noise(s) is reasonable. For testing of mobile phones in handset or handheld hands-free mode, as many different noise types as possible should be selected. The consideration of silent condition (no background noise playback) is strongly recommended.

5.4 Recording procedure

5.4.1 Acoustic recordings (receiving)

The test setup is motivated by the requirement that all signals can be measured outside the device. For capturing the signals, a HATS according to Recommendation ITU-T P.58 [7] is used. The specific setup may vary from one application to another. However, the recording procedure shall always follow the guidelines described in the following.

The recording procedure is conducted in two steps:

1) The reference signal $r(k)$ is inserted to the DUT in receiving direction. The processed speech signal and the noise playback are recorded simultaneously. These signals are recorded binaurally. This binaural signal is denoted as $d(k)$ in the following.

2) In the second step, the transmission of the speech signal is deactivated; only the near-end noise is recorded as a binaural signal, which is denoted as $n(k)$. The DUT shall be active/mounted/be in the same operational mode as for the first step. No disturbing signal shall be produced by the DUT.

This measurement principle allows the extraction of a processed, but noise-free speech signal $p(k)$ from the degraded signal $d(k)$ within the prediction model.

Figure 5.1 illustrates an example measurement setup for handset testing. For this purpose, the mobile DUT is mounted at right ear of head and torso simulator (HATS) according to Recommendation ITU-T P.58 [7] with an application force of 8N. The artificial head is equipped with diffuse-field equalized type 3.3 ear simulators according to Recommendation ITU-T P.57 [6]. Then the HATS is placed into a measurement chamber. Inside this room, a playback system according to ETSI TS 103 224 [i.4] is arranged.

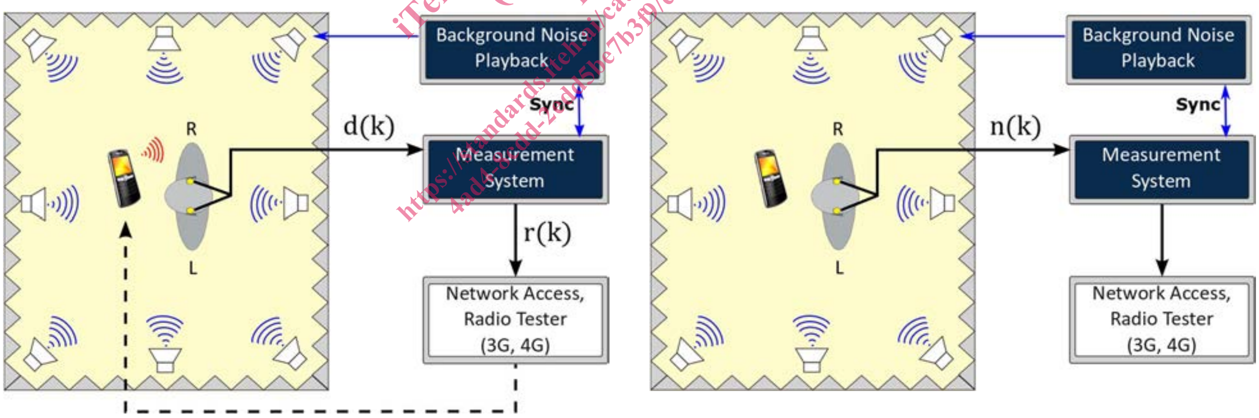

Figure 5.1: Schematic recording setup for (binaural) signal assessment

In the first measurement step, degraded speech and near-end noise are recorded by the right artificial ear (left side of figure 5.1). The left ear signal does not contain any speech signal, but is recorded as well. It is used for the auditory evaluation (binaural presentation) as well as for the instrumental listening effort assessment. In the second step, only the near-end noise (with DUT still mounted) is recorded (right side of figure 5.1).

NOTE: For the instrumental assessment of listening effort, the usage of the noise-only reference in the algorithm is optional, but recommended for higher prediction accuracy. However, in some applications, speech and noise may not be separately accessible.

5.4.2 Electrical recordings (sending)

The measurement setup records the degraded signal $d(k)$ at the electrical POI. Either acoustical (via HATS and terminal) or electrical insertion (via e.g. gateways or SPNF devices) are possible.