ETSI TS 103 558 V1.3.1 (2021-07)

TECHNICAL SPECIFICATION

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

Reference

RTS/STQ-294

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 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

Contents

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are 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 Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, 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.

DECT™, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™** and **LTE™** are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

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 telephony and other speech-based applications".

[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.

[16] Void.

[17] ISO 389-7:2019: "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/ASA 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)".

[24] Leo Breiman: "Random Forests", Statistics Department, University of California Berkeley, CA 94720, October 2001.

[25] ONNX: "Open Neural Network Exchange", Github®.

NOTE: Available at https://github.com/onnx/onnx.

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, Massachusetts, 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".

[i.16] ETSI TS 103 739: "Speech and multimedia Transmission Quality (STQ); Transmission requirements for wideband mobile wireless terminals (handset and headset) from a QoS perspective as perceived by the user".

[i.17] ETSI TS 103 740: "Speech and multimedia Transmission Quality (STQ); Transmission requirements for wideband mobile wireless terminals (handsfree) from a QoS perspective as perceived by the user".

[i.18] ETSI TS 103 557: "Speech and multimedia Transmission Quality (STQ); Methods for reproducing reverberation for communication device measurements".

[i.19] ETSI TS 126 073: "Digital cellular telecommunications system (Phase 2+) (GSM); Universal Mobile Telecommunications System (UMTS); LTE; ANSI-C code for the Adaptive Multi Rate (AMR) speech codec (3GPP TS 26.073)".

[i.20] ETSI TS 126 173: "Digital cellular telecommunications system (Phase 2+) (GSM); Universal Mobile Telecommunications System (UMTS); LTE; ANSI-C code for the Adaptive Multi-Rate - Wideband (AMR-WB) speech codec (3GPP TS 26.173 version 16.0.0 Release 16)".

[i.21] ETSI TS 126 442: "Universal Mobile Telecommunications System (UMTS); LTE; 5G; Codec for Enhanced Voice Services (EVS); ANSI C code (fixed-point) (3GPP TS 26.442)".

[i.22] T. Sainburg: "Noise reduction in python using spectral gating", Github®.

NOTE: Available at https://github.com/timsainb/noisereduce.

[i.23] E. N. Gilbert: "Capacity of a burst-noise channel", Bell System Technical Journal 39 (5), p. 1253-1265, 1960.

[i.24] E. O. Elliott: "Estimates of error rates for codes on burst-noise channels", Bell System Technical Journal, 42 (5), p. 1977-1997, 1963.

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:

ACR	Absolute Category Rating
AMR	Adaptive Multi-Rate codec (NarrowBand)
AMR-WB	Adaptive Multi-Rate codec-WideBand
ANC	Active Noise Cancellation
ASL	Active Speech Level
BGN	BackGround Noise

BW	BandWidth
BWE	BandWidth Extension
DRP	Drum Reference Point
DUT	Device Under Test
ENG	English (language)
EVS	Enhanced Voice Service
FB	FullBand
GER	German (language)
HATS	Head And Torso Simulator
HD	High Definition
HE	Headset
HF	Hands-Free
HS	Handset
ICC	In-Car Communication
IIR	Infinite Impulse Response
IR	Impulse Response
ITD	Interaural Time Difference
LE	Listening Effort
LQS	Listening Quality Subjective
MAN	Mandarin/Chinese (language)
MAX	Maximum volume setting (of e.g. a mobile phone)
maxabs	absolute maximum error
MNRU	Modulated Noise Reference Unit
MOS	Mean Opinion Score
MOS_{LE}	Listening Effort on MOS scale
MOS-LQ	Listening Quality on MOS scale
MRP	Mouth Reference Point
NB	NarrowBand
NELE	Near-End Listening Enhancement
NLMS	Normalized Least-Mean Square (adaptive filter)
NOM	Nominal volume setting (of e.g. a mobile phone)
NS	Noise Suppression
ONNX	Open Neural Network eXchange
OS	Opinion Score
PC	Personal Computer
PCM	Pulse-Code Modulation
POI	Point Of Interconnect
rmse	root-mean-square error
SII	Speech Intelligibility Index
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level
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.

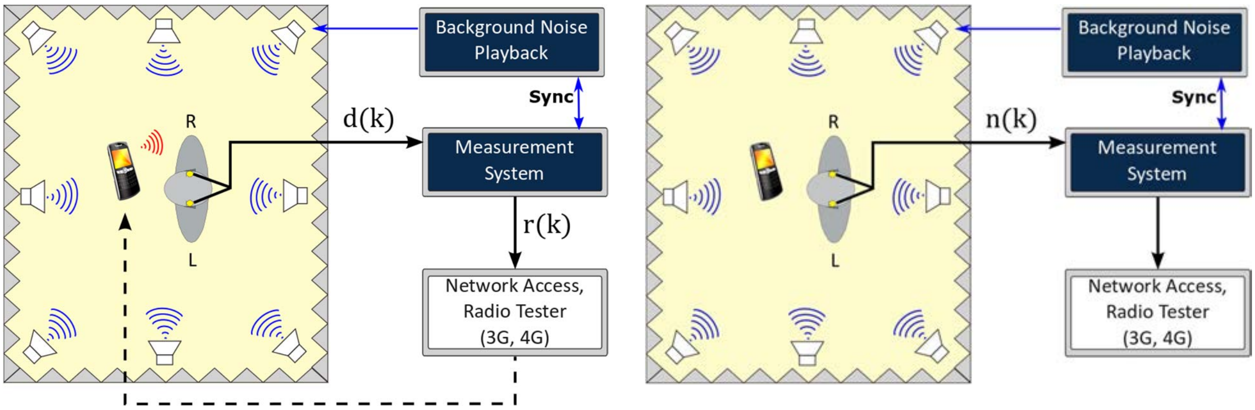


Figure 5.1: Schematic recording setup for (binaural) signal assessment

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.

5.5 Sample presentation

5.5.1 General considerations

Besides varying background noise levels, speech signals at different levels are an included use case and can be used in the listening test. To avoid any hearing impairment in the tests, the minimum health and safety requirements regarding noise exposure according to Directive 2003/10/EC [10] shall be met. Additional guidelines on maximum playback levels are provided in ETSI EG 202 518 [i.15] and should be considered as well.

A minimum speech level is not specified, since low levels (or even non-existing signals on one ear) may be a variable under test (like, e.g. volume control settings). A default and comfortable listening level of 73 dB_{SPL} or an optimum level of 79 dB_{SPL} (see also clause 6.2.3) may be considered when no specific level is considered for the evaluation itself. Whenever possible, active speech levels should be calculated and reported according to Recommendation ITU-T P.56 [3].

For the listening test, the measured sequence according to clause 5.2 is cropped into shorter samples. Either one or two sentences (duration of 4,0 s or 8,0 s) per sample can be used for presentation to the test subjects.

5.5.2 Monaural signals

If only monaural degraded signals are available (e.g. in case of single-channel electrical recordings), diotic presentation shall be used. Similar to the case of binaural recordings, diffuse-field equalized headphones shall be used for the playback of the samples. No further listening filter shall be used.

The calibration of the signals from the electrical to the acoustical domain may differ for different technologies and applications.

EXAMPLES:

- PCM signals (wave files, codec output, etc.), -26 dBov should be mapped to 73 dB_{SPL}.
- Signal captured in a network access, -18,2 dBV / -16,0 dBm0 should be mapped to 73 dB_{SPL}.

5.6 Anchor/Reference Conditions

Reference conditions are a well-established method for conducting meaningful comparisons of auditory test results from different laboratories or from the same laboratory at different times. These conditions always include a best possible (also often denoted as *clean* or *direct*) condition, as well as conditions where known, controlled degradations have been added to the speech materials. This so-called reference system also provides specific anchor points. The *direct* condition represents the very best condition that is attainable in the experiment (is not necessarily a fullband *clean* speech signal at MRP).

A reference system set of 12 conditions shall be used, which address several degrees of listening effort and speech quality. Since the field of application of auditory assessed listening effort is quite broad, it is difficult to specify a distinct set of reference system with exactly one type of controlled degradations.

The reference conditions should not be noticed by (naïve) listeners, thus the impairments simulated should include artefacts, which are similar to the ones of the test conditions. For this purpose, annex B provides prescribed procedures for appropriate reference systems, depending on the corresponding use case scenario.

5.7 Attributes and test methodology

The instructions to the test subjects shall be presented in written form in the mother tongue of the test subjects. For presentation, e.g. text printed on paper, assessment terminal/PC or projected slides may be used. Examples for listening test instructions in different languages are given in annex A.

The listening test shall include at least one attribute for the evaluation: *listening effort*. The five-point scale and corresponding categories are given in Table 5.1.

Table 5.1: Categories for listening effort (LE)

Category Description LE	Value
Complete relaxation possible; no effort required	5 (best)
Attention necessary; no appreciable effort required	4
Moderate effort required	3
Considerable effort required	2
No meaning understood with any feasible effort	1 (worst)

As a second attribute, it is recommended to include *speech quality* according to Recommendation ITU-T P.800 [1] as well. It supports the test subjects in differentiating between the near-end noise component (major impact on listening effort) of the signal and possible introduced speech degradations (minor to medium impact on listening effort), which are included in the signal-under-test. The five-point scale and corresponding categories of this attribute are given in Table 5.2.

Table 5.2: Categories for speech quality (SQ)

Category Description SQ	Value
Excellent	5 (best)
Good	4
Fair	3
Poor	2
Bad	1 (worst)

In addition, several other attributes (like e.g. coloration, discontinuity, etc.) could be added to the auditory test. Further evaluations with more than two attributes are for further study.

5.8 Requirements for the listening laboratory

The listening laboratory facilities need to comply with the recommendations provided in Recommendation ITU-T P.800 [1] and the ITU-T Handbook of subjective testing practical procedures [8].

5.9 Listening test structure

Beside the 12 reference conditions, between 12 and 60 test conditions shall be included per auditory database. This reflects a reference to overall condition ratio between 17 % and 50 %. The recommended ratio equals to 20 %, i.e. referring to 48 test conditions.

At least four different samples shall be used per condition, between eight and sixteen are recommended. Each condition shall include the same number of speech samples. In order to reduce fatigue of subjects during the test, different samples per conditions can be used.

Depending on the amount of overall conditions, not all samples may be judged by each participant due to practical limit on the total test time. In this case, the listening test shall be conducted by the principle of the "balanced block design" according to [8].

At least 12 votes per sample shall be collected, 16 are recommended. Since the number of samples per condition may vary, there is no requirement on the number of votes per condition.

5.10 Reporting of results

All titles of the samples used in an auditory database shall be reported as rows in a table. As a second column, the information about the corresponding condition number (e.g. C01, C02, etc.) should be included (if applicable). All per-sample and per-condition MOS values shall be rounded to two digits for the report.

The votes per sample and per attribute are averaged and then reported as further columns in the table. In addition, the number of votes per sample, the standard deviation and the 95 % confidence interval shall be reported as well. Thus, four columns per attribute are added to the result table. An example of the per-sample result is provided in Table 5.3.

Table 5.3: Example report of per-sample results

Sample	Condition	LE	Votes LE	STD(LE)	CI95(LE)	SQ	Votes SQ	STD(SQ)	CI95(SQ)
C01_m1s1	C01	2,94	16	0,44	0,24	2,29	14	0,91	0,53
C01_f1s1	C01	3,14	14	0,53	0,31	2,14	14	0,77	0,44
…	…	…	…	…	…	…	…	…	…
C48_m2s2	C48	2,19	16	0,75	0,40	2,88	16	1,02	0,55
C48_f2s2	C48	2,71	14	0,61	0,35	2,00	14	0,78	0,45

In addition, results shall be averaged per condition. Note that the aggregation of standard deviation and 95 % confidence interval is conducted according to the principles of Recommendation ITU-T P.1401 [i.3]. An example of the per-condition results is provided in Table 5.4.

Table 5.4: Example report of per-condition results

Condition	LE	Votes LE	STD(LE)	CI95(LE)	SQ	Votes SQ	STD(SQ)	CI95(SQ)
C01	3,69	58	0,70	0,18	2,68	58	0,70	0,18
C02	3,77	58	0,97	0,26	2,84	58	0,97	0,26
…	…	…	…	…	…	…	…	…
C47	2,83	58	0,83	0,22	3,59	58	0,83	0,22
C48	2,56	58	0,44	0,12	1,24	58	0,44	0,12

6 Instrumental Assessment

6.1 Overview

In general, the listening effort prediction algorithm requires several input signals:

- Degraded input signal $\boldsymbol{d}(k)$: By default, this signal is a diffuse-field equalized binaural recording of noisy speech. In several applications, only a single-channel signal is available or of interest, see also clause 6.8 for monaural modes.
- Noise-only signal $\boldsymbol{n}(k)$ (optional): This signal is a diffuse-field equalized binaural recording containing only the noise of the degraded signal, but no speech. It is used in order to separate speech and noise components for the further analysis. In several applications, it may not be possible to accurately differentiate between speech and noise components by the measurement procedure described in clause 5.3. In this case, this reference signal can be omitted, a noise estimate is then calculated within the prediction algorithm. However, if possible the usage of the noise-only reference is recommended for higher prediction accuracy.
- Reference signal $r(k)$: This single-channel reference signal contains the fullband clean speech signal used for the measurement of the degraded signal, as described in clause 5.1. For the instrumental assessment, typically one or two sentences within one signal are analysed.

In the following clauses, the instrumental assessment of listening effort is described. In addition to the aforementioned input signals, several naming conventions are defined:

- Time signals are in general denoted with lowercase letters and sample index k (like e.g. $d(k)$).
- Signal representations in the frequency domain vs. time are denoted with the corresponding capital letter, frame index i (different from k) and frequency band index j (like e.g. $D(i,j)$).
- By default, the input signals $\boldsymbol{d}(k)$ and $\boldsymbol{n}(k)$ are assumed to be binaural, diffuse-field equalized signals and are formatted in bold. The same formatting is applied for time-frequency representations, e.g. $\boldsymbol{D}(i,j)$.
- Binaural signals consist of two separate signals (for left and right ear). These single-channel signals are formatted normally and marked with indices L or R, if applicable (e.g. the reference signal $r(k)$ is always monaural).

- The binaural signal is defined as a tuple of both, e.g. $\boldsymbol{d}(k) = [d_L(k), d_R(k)]$. The same formatting is applied for time-frequency representations, e.g. $\boldsymbol{D}(i,j) = [D_L(i,j), D_R(i,j)]$.
- Monaural input signals are assumed to be presented diotically to the listener. For example, a monaural input signal $d(k)$ leads to the (pseudo-)binaural signal $\boldsymbol{d}(k) = [d(k), d(k)]$.
- If the noise-only signal $\boldsymbol{n}(k)$ is used as an input, a noise-compensated but signal-processed signal $\boldsymbol{p}(k)$ is introduced during the pre-processing stage.

Figures 6.1 and 6.2 illustrate the two basic operational modes of the instrumental assessment based on the introduced naming conventions.

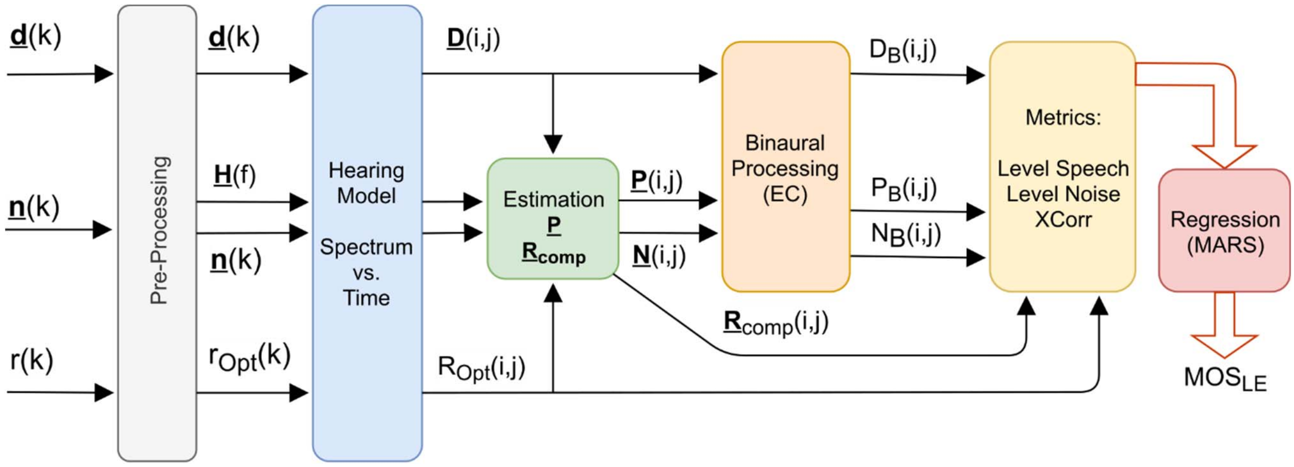


Figure 6.1: Instrumental listening effort assessment with noise-only reference

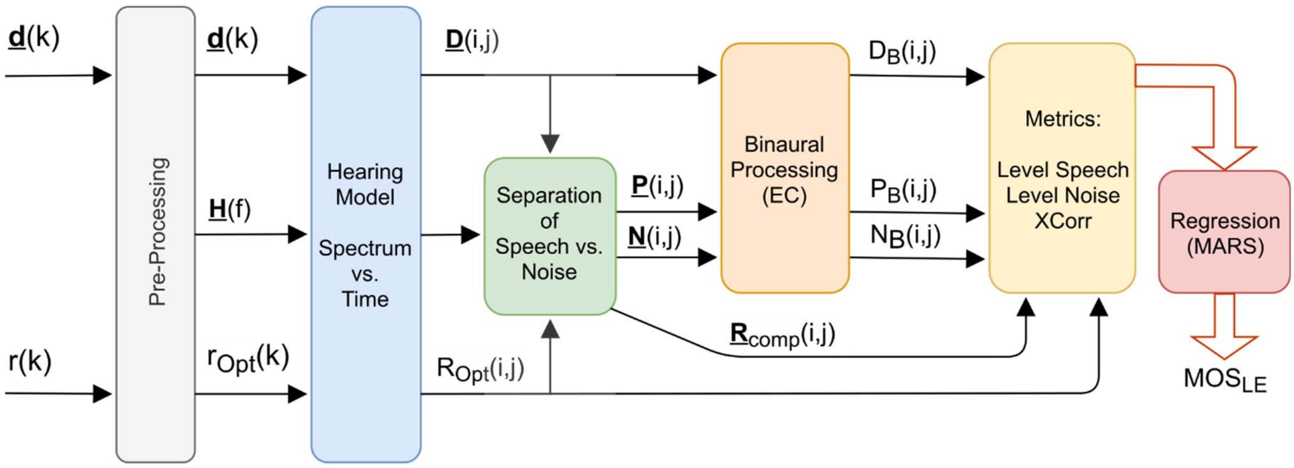


Figure 6.2: Instrumental listening effort assessment without noise-only reference

6.2 Pre-processing

6.2.1 Overview

Before performing any metric calculation and predicting listening effort scores, the input signals of the algorithm have to be prepared for the following stages.

The pre-processing of the inputs $\boldsymbol{d}(k)$, $\boldsymbol{n}(k)$ and $r(k)$ is conducted in order to compensate differences regarding temporal alignment, level offsets and spectral shaping between these signals.

The following assumptions on all input signals are made by the algorithmic pre-processing, i.e. everything different from these shall be realized by the implementer and is not specified here:

- All input signals are expected to be inserted with a sampling rate of 48 kHz into the algorithm. Resampling methodologies are not specified in the present document.
- All input signals shall have the same length, i.e. number of samples per signal. No padding or cropping strategies are specified in the present document.
- It is always assumed that the amplitudes of all signals are already calibrated to the physical unit *Pascal*. In case of electrical recordings, the (typically single-channel) signals shall be calibrated to a reasonable listening level.
- The delay between $\boldsymbol{d}(k)$ and $\boldsymbol{n}(k)$ is assumed to be zero, i.e. no time alignment between these signals is applied. This assumption is usually inherently met when using noise playback systems according to ETSI ES 202 396-1 [i.5] or ETSI TS 103 224 [i.4], which provide a high temporal reproduction accuracy.
- The recordings $\boldsymbol{d}(k)$ and $\boldsymbol{n}(k)$ are assumed to diffuse-field equalized recordings according to Recommendation ITU-T P.58 [7] obtained with a HATS.

In order to avoid singularity issues (division by zero) during e.g. delay compensation (see clause 6.2.2) or binaural processing (see clause 6.6), a white noise signal (20 - 20 000 kHz) with a level of -90 dB_{Pa} should be added to $\boldsymbol{d}(k)$ and $\boldsymbol{n}(k)$ (if applicable) in advance to all further processing steps.

6.2.2 Compensation of Delay

Similar as in speech quality prediction models like e.g. ETSI TS 103 281 [4] or Recommendation ITU-T P.863 [i.2], the reference signal shall be compensated for possible delays introduced by e.g. terminals, network or signal processing. This is indispensable for a valid comparison towards the reference signal at a later stage.

Since the signal $\boldsymbol{d}(k)$ is expected to be binaural, the determination of delay has to be conducted for both channels individually. For sake of simplicity, the following steps are only described for one single signal (left or right ear) indicated as $d(k)$ (without any subscript index).

Due to the high expected amount of noise portion in the degraded signal, the computational effort is higher than for speech quality metrics. First, the input signals are filtered with an IIR Butterworth band-pass of 6th order and a frequency range of 300 Hz - 3 300 Hz. By limiting bandwidth to this range, only the signal parts containing most speech energy are taken into account.

Then, the resulting band-pass filtered signals $d_{BP}(k)$ and $r_{BP}(k)$ are segmented in frames of $T = 131\,072$ samples with 75 % overlap, resulting in $d_{BP,m}(k)$ and $r_{BP,m}(k)$. For each frame *m*, the cross-correlation function $\Phi_{dr}(m,\tau)$ between $d_{BP,m}(k)$ and $r_{BP,m}(k)$ is calculated in the time domain according to equation assuming periodic continuation of the frames:

$$\Phi_{dr}(m,\tau) = \sum_{k=1}^{T} d_{BP,m}(k) \cdot r_{BP,m}(k+\tau) \qquad (1)$$

The envelope $E(m,\tau)$ is calculated by the Hilbert transformation $H\big(\Phi_{dr}(m,\tau)\big)$ of the cross-correlation according to equations (2) and (3):

$$H\big(\Phi_{dr}(m,\tau)\big) = \sum_{u=u_{min}}^{u=u_{max}} \frac{\Phi_{dr}(u)}{\pi(\tau-u)} \qquad (2)$$

$$E(m,\tau) = \sqrt{\left[\Phi_{dr}(m,\tau)\right]^2 + \left[H\big(\Phi_{dr}(m,\tau)\big)\right]^2} \qquad (3)$$

The per-frame envelopes are averaged over all M frames according to equation (4):

$$E(\tau) = \frac{1}{M} \cdot \sum_{m=1}^{M} E(m,\tau) \qquad (4)$$

The maximum peak $\mathrm{P}_{\mathrm{max,X}}$ of $E(\tau)$ determines the delay $\mathrm{D}_{\mathrm{ref}}$ for the compensation of the reference signal on the time abscissa. These two values are determined for both channels (left and right), which is indicated with X in equations (5) and (6) (X ∈ [L, R]):

$$\mathrm{P}_{\mathrm{max,X}} = \max E_X(\tau) \tag{5}$$

$$\mathrm{D}_{\mathrm{ref,X}} = \operatorname{argmax} E_X(\tau) \tag{6}$$

Based on the peak values $\mathrm{P}_{\mathrm{max,L}}$ and $\mathrm{P}_{\mathrm{max,R}}$, the better ear/channel B (B ∈ [L, R]) and the overall maximum peak value $\mathrm{P}_{\mathrm{max}}$ are determined by the maximum of both, as shown in equation (7):

$$\mathrm{P}_{\mathrm{max}} = \max\left(\mathrm{P}_{\mathrm{max,L}}, \mathrm{P}_{\mathrm{max,R}}\right) \tag{7}$$

The final delay $\mathrm{D}_{\mathrm{ref}}$ is then defined as the delay value determined of the better channel B:

$$\mathrm{D}_{\mathrm{ref}} = \mathrm{D}_{\mathrm{ref,B}} \tag{8}$$

NOTE 1: In case of monaural input signals, only one delay and peak calculation is carried out. In consequence, there is also no selection of a better ear anymore, i.e. the better ear is the monaural signal itself.

The alignment is conducted by adding zeros at the beginning and cropping at the end of the reference signal $r(k)$ in case of a positive determined delay. The inverse procedure is applied in case of a negative delay.

As mentioned before, it is assumed that the delay between degraded and noise signals is zero. Thus, also the noise signal $\boldsymbol{d}(k)$ are compensated with the same delay $\mathrm{D}_{\mathrm{ref}}$ as the degraded signal.

This compensation step does not affect the degraded signal $\boldsymbol{d}(k)$, i.e. the duration of all signals is maintained in all signals.

In addition, an estimate of the interaural time difference (ITD) is calculated according to equation (8a) as the absolute difference between left and right delay:

$$ITD = \left|\mathrm{D}_{\mathrm{ref,L}} - \mathrm{D}_{\mathrm{ref,R}}\right| \tag{8a}$$

NOTE 2: Other common definitions of ITD usually are calculated via cross-correlation between left and right ear, including speech and noise. The definition in the present document however only take the speech component of the perceived ear signals into account.

6.2.3 Reference Scaling

Comparisons between degraded and reference signal in later stages of the prediction algorithm are carried out based on realistic listening levels. The reference signal is intended to provide an optimum regarding clear pronunciation, frequency shaping, etc., but also regarding best-possible speech level.

According to the ITU-T Handbook [9], an active speech level about -10 $\mathrm{dB}_{\mathrm{Pa}}$ (84 $\mathrm{dB}_{\mathrm{SPL}}$) maximizes the Listening-Effort score for monaural listening. For binaural/dichotic listening, this would refer to approximately 78 - 79 $\mathrm{dB}_{\mathrm{SPL}}$. Thus, the reference signal $r(k)$ shall be calibrated to an active speech level according to Recommendation ITU-T P.56 [3] of 79,0 $\mathrm{dB}_{\mathrm{SPL}}$. This scaled version of the signal is denoted as $r_{\mathrm{Opt}}(k)$ (optimal reference) in the following clauses.

6.2.4 Speech Part Detection

In order to determine the time ranges of active speech, the classification algorithm according to Appendix II of Recommendation ITU-T G.160 [11] is applied on the reference signal $r_{\mathrm{Opt}}(k)$. The first step is to classify energy frames of 10 ms (block-wise, no overlap) according to the method described in [11]. The thresholds for the classification are defined relatively to the active speech level (in this case 79 $\mathrm{dB}_{\mathrm{SPL}}$).

As a result, each speech frame is identified either as high (H), medium (M), low (L) or uncertain (U) activity. Frames without activity are either classified as short pauses (P) or silence (S). Short speech pauses are defined as silence periods with a duration up to 400 ms.

The speech parts are finally determined as regions excluding frames of type S, i.e. including also short pauses. The information of the active time ranges is employed in several other algorithmic parts, which are introduced in the following clauses.

6.2.5 Determination of Processed Signal

In order to analyse the impact of the possibly disturbed speech components on perceived listening effort, the influence of the noise has to be cancelled out. In case of acoustic recordings with near-end noise (i.e. not transmitted or processed), the processed but noise-free signal $\boldsymbol{p}(k)$ can be easily determined according to equation (9):

$$\boldsymbol{p}(k) = \boldsymbol{d}(k) - \boldsymbol{n}(k) \tag{9}$$

NOTE: The valid subtraction of time signals requires the usage of highly accurate noise playback systems regarding reproduction and synchronization. Systems according to ETSI ES 202 396-1 [i.5] or ETSI TS 103 224 [i.4] for example meet these claims.

6.2.6 Transfer Function

In order to characterize the transmission system of the degraded signal in the frequency domain, the complex transfer function $H(\mathrm{f})$ is utilized in later stages. Similar as for the determination of delay, the high amount of noise in the degraded signal requires some more computational effort. Again, the calculation shall be carried out for both channels L/R of $\boldsymbol{d}(k)$, but for sake of simplicity, the following steps are only described for one single signal (left or right ear), indicated as $d(k)$ (without any subscript index).

Similar as for the determination of delay, the degraded signal $d(k)$ and the optimum reference $r_{0\mathrm{pt}}(k)$ are segmented by a rectangular window in short frames of 1 024 samples and 50 % overlap. For each frame *m*, the cross-spectral density $S_{dr}(m,f)$ and the auto-spectral densities $S_{rr}(m,f)$ and $S_{dd}(m,f)$ are calculated. Then, the magnitude-squared coherence $C_{dr}(\mathrm{m,f})$ is calculated for each frame according to equation (10):

$$C_{dr}(m,f) = \frac{|S_{dr}(m,f)^2|}{S_{rr}(m,f) \cdot S_{dd}(m,f)} \tag{10}$$

In a similar way, the short-time transfer function $H(\mathrm{m,f})$ is calculated for each frame according to equation (11):

$$H(m,f) = \frac{S_{dr}(m,f)}{S_{rr}(m,f)} \tag{11}$$

NOTE 1: In literature, this calculation is also known as H1 method, as e.g. described in [i.7].

A frame is considered to contribute to the overall transfer function if the coherence $C_{dr}(\mathrm{m,f})$ exceeds 5 % for all frequencies between 100 Hz and 16 kHz. All contributing frames are stored in a set A, the size of this set is M_A. Finally, the average transfer function can be calculated according to equation (12):

$$H(f) = \frac{1}{M_A} \sum_{m \in A} H(m,f) \tag{12}$$

NOTE 2: For the determination of a transfer function, the usage of the processed signal $\boldsymbol{p}(k)$ instead of $\boldsymbol{d}(k)$ seems more obvious. Since it is expected that the prediction algorithm may be updated also for applications where the noise-only reference $\boldsymbol{n}(k)$ may not be available (which is necessary to obtain $\boldsymbol{p}(k)$), the analysis was designed directly for the usage with $\boldsymbol{d}(k)$. No additional case distinction is made here.

6.3 Spectral transformation

The hearing model according to Sottek [12] is calculated for the signals degraded $\boldsymbol{d}(k)$, noise-only $\boldsymbol{n}(k)$ (if applicable) and clean speech reference $r_{0\mathrm{pt}}(k)$. The transformation includes an auditory filter bank representation of the signal and a hearing-adequate envelope determination.

Table 6.1: Filterbank frequencies (in Hz) of the hearing model

	Frequency Index								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Lower	31,5	111,3	203,2	308,9	430,4	570,2	731,1	916,1	1 128,9
Center	70,0	155,7	254,2	367,5	497,9	647,8	820,3	1 018,7	1 247,0
Upper	111,3	203,2	308,9	430,4	570,2	731,1	916,1	1 128,9	1 373,6
	10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
Lower	1 373,6	1 655,2	1 979,1	2 351,6	2 780,1	3 273,0	3 840,0	4 492,2	5 242,4
Center	1 509,5	1 811,5	2 158,8	2 558,4	3 018,0	3 546,6	4 154,7	4 854,2	5 658,8
Upper	1 655,2	1 979,1	2 351,6	2 780,1	3 273,0	3 840,0	4 492,2	5 242,4	6 105,4
	19	**20**	**21**	**22**	**23**	**24**	**25**	**26**	
Lower	6 105,4	7 098,0	8 239,7	9 553,1	11 063,8	12 801,6	14 800,4	17 099,7	
Center	6 584,3	7 648,9	8 873,4	10 282,0	11 902,3	13 766,0	15 909,8	18 375,8	
Upper	7 098,0	8 239,7	9 553,1	11 063,8	12 801,6	14 800,4	17 099,7	19 744,5	

Table 6.1 lists the centre frequencies (in Hz) as well as the bandwidth of the 26 frequency bands of the auditory filter bank. In contrast to other hearing-adequate frequency scales, the proposed method includes the whole FullBand (FB) range.

Each frequency band of the hearing model is temporally aggregated to frames of 1 ms by calculating the average across 48 output samples (no overlap).

This time-frequency representation is calculated for all pre-processed signals individually for left and right ear, resulting in the hearing model spectra vs time $\boldsymbol{D}(i,j)$ (degraded), $\boldsymbol{N}(i,j)$ (noise-only, if applicable), $\boldsymbol{P}(i,j)$ (processed, if applicable) and $R_{\mathrm{Opt}}(i,j)$ (reference scaled to optimum level).

6.4 Compensated Reference

Based on the transfer function $H(f)$ and the hearing model spectrum $R_{\mathrm{opt}}(i,j)$ of the optimum reference, a so-called compensated reference $R_{\mathrm{Comp}}(i,j)$ spectrum vs time is determined. First, a linearly interpolated version $\widetilde{\boldsymbol{H}}(j)$ is calculated from the transfer function $\boldsymbol{H}(f)$, which uses the same frequency resolution as the hearing model. The compensated reference is then calculated as per equation (13). It represents a filtered version of the reference, which has the same spectral shaping as the degraded signal, but without any further degradations (e.g. due to non-linear signal processing):

$$\boldsymbol{R}_{\mathbf{Comp}}(i,j) = \widetilde{\boldsymbol{H}}(j) \cdot R_{\mathrm{opt}}(i,j) \tag{13}$$

6.5 Separation of Speech and Noise Component

In case no noise-only signal $\boldsymbol{n}(k)$ is provided, in consequence also no processed signal $\boldsymbol{p}(k)$ is available. Thus, the corresponding hearing model spectra $\boldsymbol{N}(i,j)$ and $\boldsymbol{P}(i,j)$ are estimated based on the available inputs. Equation (14) shows the basic assumption of the composition:

$$\boldsymbol{D}(i,j) = \boldsymbol{P}(i,j) + \boldsymbol{N}(i,j) \tag{14}$$

For the decomposition, a (pseudo-)Wiener filter is used for the determination of $\boldsymbol{P}(i,j)$ as shown in equation (15):

$$\boldsymbol{P}(i,j) = \boldsymbol{D}(i,j) \cdot \boldsymbol{W}(i,j) \tag{15}$$

The Wiener gain $\boldsymbol{W}(i,j)$ is obtained according to equation (16). The compensated reference $\boldsymbol{R}_{\mathbf{Comp}}(i,j)$ and an initial noise estimation $\widetilde{\boldsymbol{N}}(i,j)$ is used:

$$\boldsymbol{W}(i,j) = \sqrt{\frac{\boldsymbol{R}_{\mathbf{Comp}}(i,j)^2}{\boldsymbol{R}_{\mathbf{Comp}}(i,j)^2 + \widetilde{\boldsymbol{N}}(i,j)^2}} \tag{16}$$

For the determination of $\widetilde{\boldsymbol{N}}(i,j)$, first a soft mask $\boldsymbol{M}(i,j)$ of active/inactive time-frequency bins is determined with $\boldsymbol{R}_{\mathbf{Comp}}(i,j)$. The speech part classification algorithm according to Appendix II of Recommendation ITU-T G.160 [11] is carried out for each frequency band of the hearing model spectrum vs. time (see also clause 6.2.4). For silent (S) and short-paused (P) frames, a mask value of 1, for high activity of speech to 0 is set. Weights for medium (M), low (L) and uncertain (U) activity frames are provided in Table 6.2. An example of this threshold-based method for one single frequency band is illustrated in Figure 6.3.

Table 6.2: Mask weights of activity

Activity	Value
H	0,0
M	0,15
L	0,4
U	0,7
P, S	1,0

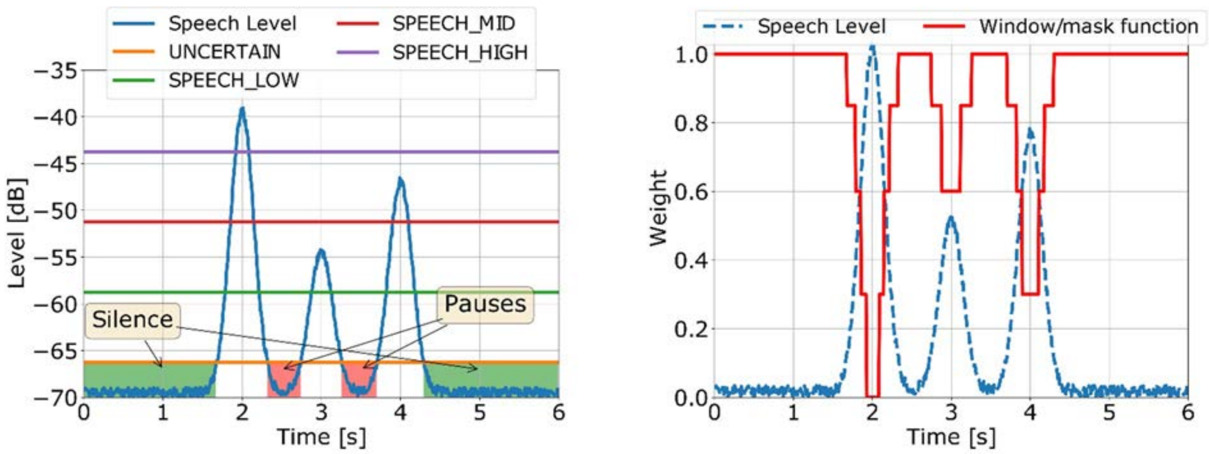


Figure 6.3: Principle of frame classification (left) and mask generation (right)

NOTE: For sake of clarity, time-frequency indices (i, j) are neglected until the end of this clause.

Then, the degraded spectra are multiplied by the masks, resulting in $\boldsymbol{M_D}$ as per equation (17). This step suppresses the active speech frames and can be considered as a windowed noise signal.

$$\boldsymbol{M_D} = \boldsymbol{D} \cdot \boldsymbol{M} \approx \boldsymbol{N} \cdot \boldsymbol{M} \tag{17}$$

With the Fourier operator $\boldsymbol{\mathcal{F}}(\cdot)$, an estimate for the noise signal per frequency band can be described as a deconvolution problem (denoted as $\{\cdot\}^{-\mathbf{1}}$), as shown in equations (18) to (20).

$$\boldsymbol{M_D} ∘\!\!-\!\!\bullet\ \boldsymbol{\mathcal{F}}(\boldsymbol{M_D}) = \boldsymbol{\mathcal{F}}(\boldsymbol{N}) * \boldsymbol{\mathcal{F}}(\boldsymbol{M}) \tag{18}$$

$$\Leftrightarrow \boldsymbol{\mathcal{F}}(\boldsymbol{N}) = \{\boldsymbol{\mathcal{F}}(\boldsymbol{M_D}) * \boldsymbol{\mathcal{F}}(\boldsymbol{M})\}^{-\mathbf{1}} \tag{19}$$

$$\Rightarrow \widetilde{\boldsymbol{N}} \approx \boldsymbol{N} = \boldsymbol{\mathcal{F}}^{-\mathbf{1}}\left(\{\boldsymbol{\mathcal{F}}(\boldsymbol{M_D}) * \boldsymbol{\mathcal{F}}(\boldsymbol{M})\}^{-\mathbf{1}}\right) \tag{20}$$

In general, multiple deconvolution algorithms are available in literature. For the current prediction model, the algorithm described in e.g. [13] is used. The estimated noise spectra $\widetilde{\boldsymbol{N}}(i,j)$ are used for the Wiener filter according to equation (15) in order to obtain finally the processed spectra $\boldsymbol{P}(i,j)$.

6.6 Binaural processing

In order to address the capability of human hearing to improve SNR compared to monaural listening, a binaural processing stage is included in the prediction model. The spectral components for left and right ears are combined by a short-term equalization-cancellation (STEC) model according to [14]. This extension of the well-known model of Durlach [15] requires the availability of the isolated speech and masker (noise-only) components, i.e. processed and noise spectra.

The STEC model is employed exactly as described in [14], with only one slight modification: an increased block size constant of 100 ms (instead of 20 ms) is used. A reference implementation can be found in [i.8].

As a result of this stage, combined and enhanced hearing model spectra vs time are available:

- $D_B(i,j)$ (degraded).
- $P_B(i,j)$ (processed).
- $N_B(i,j)$ (noise).
- $R_{B,\mathrm{comp}}(i,j)$ (compensated reference).

NOTE: The spectra of the optimum reference is not processed via the EC-model.

6.7 Instrumental Assessment

6.7.1 Metrics

6.7.1.1 Level Metrics

Since the hearing model spectra for processed speech and noise component are individually available, level metrics can be calculated in the frequency domain. The active speech level S_{act} is calculated according to equation (21) on the processed signal across all frequencies. Only active time frames (ACT) according to clause 6.2.4 are considered in this integration and K_{ACT} denotes the number of active frames:

$$S_{\mathrm{act}} = 10 \cdot \log_{10}\left(\frac{1}{K_{ACT}} \sum_{j \in ACT} \sum_i \; P_B(i,j)^2\right) \tag{21}$$

In addition, an A-weighted noise level $L_N(\mathrm{A})$ is calculated from the noise spectrum according to equation (22). The weighting $W_A(j)$ for each frequency band is calculated according to IEC 61672-1 [20]:

$$L_N(\mathrm{A}) = 10 \cdot \log_{10}\left(\frac{1}{K_{ALL}} \sum_j \sum_i \; W_A(j) \cdot N_B(i,j)^2\right) \tag{22}$$

The integration is carried out across all frequencies and all-time indices. Here K_{ALL} denotes the overall number of frames of the hearing model spectrum.

6.7.1.2 Spectral Distance Metric

In order to investigate the relation and possible masking effects of the processed ($P_B(i,j)$) and noise ($N_B(i,j)$) spectra vs time, a similar index metric as in the calculation method according to ANSI/ASA S3.5 [18] is described in the following paragraphs.

The method of [18] provides a speech intelligibility index (SII), which is intended for the usage of stationary noises in conjunction with a constant average speech spectrum. The method is adapted for time-variant speech and noise signals with the following modifications:

- The hearing model spectra vs time is analysed per identified sentence (see clause 6.2.4). Each time instance is analysed according to ANSI/ASA S3.5 [18].
- The calculation variant is based on 1/3rd octave-bands, including the corresponding band importance weights (see Table 3 of [18]). Since frequencies higher than 8 kHz are needed, 1/3rd octave-bands up to 20 kHz are generated according to IEC 61260-1 [19].

- The band importance weights are interpolated to the fullband 1/3rd octave-bands by a cubic interpolation (see below). Since the sum of the weights is not 1 after this interpolation, they are re-normalized by dividing each value by the sum of the new weights.

- The pre-processing of [18] only describes free-field-to-DRP correction (last column of e.g. Table 2 in [18]). Instead, Table 3 of Recommendation ITU-T P.58 [7] for diffuse-field-to-DRP correction is used.

- Each short-time spectrum of speech and noise of the hearing model is interpolated to 1/3rd octave-bands by a cubic interpolation (see below).

- If more than one sentence is included in the speech sample, a weighted average across the metrics per sentence is performed. The weight per sentence corresponds to its duration.

Thus, the analysis provides one single metric output, denoted as I_{SD} in the following.

For some of the above steps, a cubic interpolation method vs frequency is required. The interpolation itself is a quite common technique, as described e.g. in [i.11]. The interpolation function $f_{\mathrm{I}}(\cdot)$ depends on several input variables, as shown in equation (23):

$$y_{\mathrm{new}} = f_{\mathrm{I}}\left(x_{\mathrm{new}}, x_{\mathrm{existing}}, y_{\mathrm{existing}}\right) \qquad (23)$$

With:

x_{existing}: the existing frequency axis f, but inserted on logarithmic scale: $\log_{10}(f)$.

y_{existing}: the existing ordinate values (e.g. spectral magnitude).

x_{new}: the frequency axis to interpolate for, but inserted on logarithmic scale: $\log_{10}(f_{new})$.

y_{new}: the interpolated ordinate values (e.g. spectral magnitude).

NOTE: The logarithmic operator on the frequencies take the (approximately) logarithmic spacing between frequency bands into account.

In case y_{existing} represents a (short-time) spectrum (e.g. $Z(j)$), the interpolated version $\dot{Z}(j')$ shall have the same energy as the original one. This is ensured by scaling the interpolated version by a gain factor g_S, as shown in equations (24) and (25):

$$g_S = \sqrt{\sum_j Z(j)^2 \Big/ \sum_j \dot{Z}(j')^2} \qquad (24)$$

$$\ddot{Z}(j') = g_S \cdot \dot{Z}(j') \qquad (25)$$

Figure 6.4 illustrates the principle of interpolation to 1/3rd octave-bands and provides two examples. The graph on the left shows the interpolation for the band importance weights, while the graph on the right demonstrates the transformation of a short-time spectrum. The solid blue lines indicate data from another frequency range, the orange dashed curves show the results in 1/3rd octave-bands after interpolation.

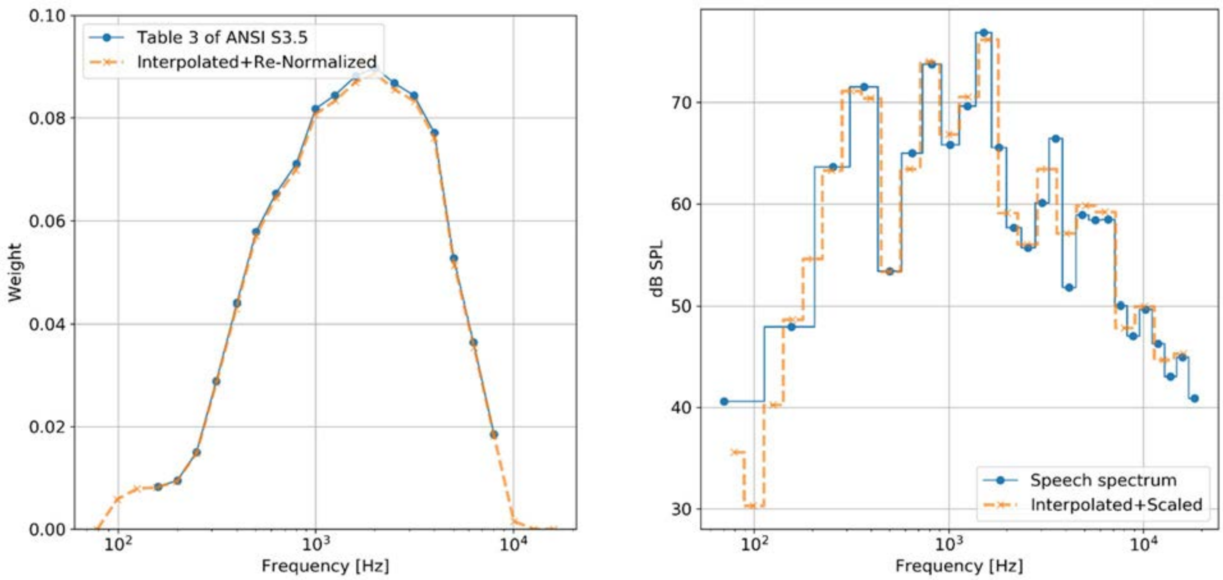


Figure 6.4: Examples of interpolation for weights (left) and spectrum (right)

6.7.1.3 Correlation Metrics

Similar as the intelligibility metric introduced in e.g. [i.9], a spectral cross-correlation is carried out in four different ways. Instead of the clipping procedure described in [i.9], the threshold of hearing $T_H(j)$ according to ISO 389-7 [17] is applied to the spectra. After this step, the non-linear loudness transformation $\mathcal{L}[\cdot]$ according to the hearing model of Sottek [12] or annex K of ETSI EG 202 396-3 [i.10]) is applied on all spectra, as shown in equations (26) to (29):

$$D'_B(i,j) = \mathcal{L}[max(D_B(i,j),\ T_H(j))] \tag{26}$$

$$P'_B(i,j) = \mathcal{L}[max(P_B(i,j),\ T_H(j))] \tag{27}$$

$$R'_{\mathrm{Opt}}(i,j) = \mathcal{L}\big[max(R_{\mathrm{Opt}}(i,j),\ T_H(j))\big] \tag{28}$$

$$R'_{B,\mathrm{Comp}}(i,j) = \mathcal{L}\big[max(R_{B,\mathrm{Comp}}(i,j),\ T_H(j))\big] \tag{29}$$

NOTE: In contrast to [i.9], no further normalization is necessary, since the data is already provided on an absolute loudness scale.

Each band is divided into sub-frames of 320 ms (L = 320 frames, only active indices, see clause 6.7.1.1) with 50 % overlap. In general, the correlation metric $d_{X,Y}(m,j)$ between two (generic) spectra $X(i,j)$ and $Y(i,j)$ is calculated per frequency band j and for the m-th sub-frame as given by equations (30) to (32):

$$d_{X,Y}(m,j) = \frac{\sum_{n\epsilon m}\big(X(n,j)-\bar{X}(m,j)\big)\big(Y(n,j)-\bar{Y}(m,j)\big)}{\sqrt{\sum_{n\epsilon m}\big(X(n,j)-\bar{X}(m,j)\big)^2 \cdot \sum_{n\epsilon m}\big(Y(n,j)-\bar{Y}(m,j)\big)^2}} \tag{30}$$

With:

$$\bar{X}(m,j) = \frac{1}{L}\sum_{l\epsilon m} X(l,j) \tag{31}$$

$$\bar{Y}(m,j) = \frac{1}{L}\sum_{l\epsilon m} Y(l,j) \tag{32}$$

Equation (33) provides the final aggregation to an overall metric $d_{X,Y}$:

$$d_{X,Y} = \frac{1}{L\cdot J}\sum_{l\epsilon m}\sum_j d_{X,Y}(m,j) \tag{33}$$

With $X(i,j)\ \in\ [D'_B(i,j), P'_B(i,j)]$ and $Y(i,j)\ \in\ \big[R'_{B,\mathrm{Comp}}(i,j), R'_{\mathrm{Opt}}(i,j)\big]$, four metrics according to Table 6.3 can be derived by this analysis.

Table 6.3: Combinations of signals for correlation metrics

Title	Spectrum #1	Spectrum #2	Description
d_{D,R_C}	$D'_B(i,j)$	$R'_{B,\mathrm{Comp}}(i,j)$	Noisy speech vs optimum reference
d_{D,R_O}	$D'_B(i,j)$	$R'_\mathrm{Opt}(i,j)$	Noisy speech vs compensated reference
d_{P,R_C}	$P'_B(i,j)$	$R'_{B,\mathrm{Comp}}(i,j)$	Noise-free/processed speech vs optimum reference
d_{P,R_O}	$P'_B(i,j)$	$R'_\mathrm{Opt}(i,j)$	Noise-free/processed speech vs compensated reference

6.7.2 Regression

In order to combine all metrics described in the previous clauses, a random forest regression according to [24] is used. Table 6.4 provides a summary of all metrics used for the regression. Beside the eight parameters directly related to the perceptual model, additional flags are used in the regression:

- F_N is a flag indicating if the noise-only reference was provided or not. In the latter case, the noise-only spectrum was internally estimated by the algorithm described in clause 6.5.
- F_B is a flag indicating if binaural processing was applied or not. In the latter case, the steps as described in clause 6.6 are skipped.

These additional bits of information support the regression in order to compensate for smaller errors in the separation algorithm.

Table 6.4: Metrics used for regression

Variable	Title	Clause	Description
x_0	S_act	6.7.1.1	Speech- and Noise Levels
x_1	$L_N(\mathrm{A})$	6.7.1.1	
x_2	I_SD	6.7.1.2	Spectral distance, similar to SII
x_3	d_{D,R_C}	6.7.1.3	Correlation-based similarity metrics
x_4	d_{D,R_O}	6.7.1.3	
x_5	d_{P,R_C}	6.7.1.3	
x_6	d_{P,R_O}	6.7.1.3	
x_7	ITD	6.2.2	Interaural time difference
x_8	F_N		Flag: equals 1, if noise-only was provided (otherwise 0)
x_9	F_B		Flag: equals 1, if binaural processing (clause 6.6) was applied (otherwise 0)

Random forests are an ensemble learning method for classification and/or regression that operate by constructing a multitude of decision trees at training time and outputting the class or mean prediction of the individual trees. Table 6.5 provides the parameters of the random forest regressor used for the training process.

Table 6.5: Parametrization of random forest regressor

Parameter	Value
Number of trees	50
Maximum depth	12
Minimum samples per leaf	7
Number of features to split	All/no limit

Based on parameter fitting according to the auditory data shown in annexes C and D, MOS_LE can be determined according to the model files provided in ONNX format [25]. The file LE-RFR-V1.3.onnx is contained in archive ts_103558v010301p0.zip which accompanies the present document.

Void. (34)

6.7.3 Post-processing

The MOS calculated in the previous clause does not take into account too loud speech levels. It is obvious that in general an increasing speech level correlates with improvement in listening effort. However, in contrast to a prediction algorithm, extremely high speech levels would not be presented to test subjects in a listening test due to hearing protection (see clause 5.5.1).

For this purpose, a speech level-dependent postprocessing is applied to the MOS. A penalty is added for too loud speech-levels S_{act} according to equations (35) to (38), leading to the post-processed value $\widetilde{\mathrm{MOS}}_{\mathrm{LE}}$:

$$S_{P,\mathrm{start}} = 94\ \mathrm{dB_{SPL}} \tag{35}$$

$$S_{P,\mathrm{max}} = 104\ \mathrm{dB_{SPL}} \tag{36}$$

$$w(S_{\mathrm{act}}) = \min(1; \max(\frac{S_{P,\mathrm{max}} - S_{\mathrm{act}}}{S_{P,\mathrm{max}} - S_{P,\mathrm{start}}}; 0)) \tag{37}$$

$$\widetilde{MOS}_{LE}(S_{\mathrm{act}}) = w(S_{\mathrm{act}}) \cdot (\mathrm{MOS}_{\mathrm{LE}} - 1) + 1 \tag{38}$$

Figure 6.5 illustrates the penalty weighting as a function of the speech level S_{act}. MOS values remain unchanged for speech levels lower than $S_{P,\mathrm{start}}$ (weight equals 1,0). Between $S_{P,\mathrm{start}}$ and $S_{P,\mathrm{max}}$, the weighting function decreases the weight down to 0,0. For this speech level and above, the post-processed value $\widetilde{\mathrm{MOS}}_{\mathrm{LE}}$ equals 1,0 (worst score).

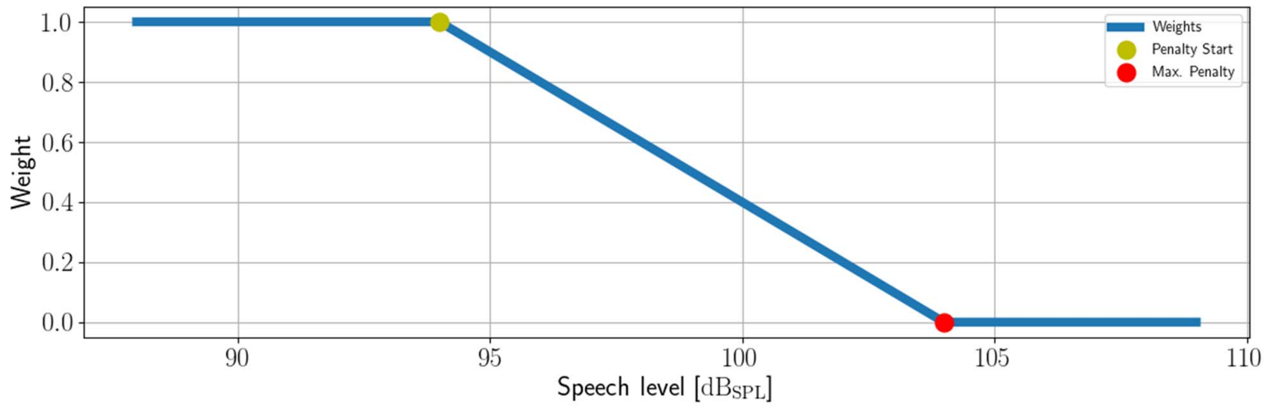


Figure 6.5: Penalty weight as function of speech level S_{act}

NOTE: The introduced clipping of MOS is not derived from subjective listening tests, but takes typical hearing protection limits into account. Subjective data for maximum acceptable presentation level (and beyond) cannot be assessed without violating directives like e.g. noise exposure according to Directive 2003/10/EC [10].

6.8 Model modes for monaural signals

For the introduced model, one or two binaural inputs (noisy speech and optionally noise-only) and one single-channel reference signal are needed. In several applications, only one ear signal is available or is of interest. In this case, the model can be run with the following simplifications:

1) Monaural mode: For terminals in handset or headset mode, the speech signal in receiving direction is only audible on one - usually right - ear. The other/left ear remains uncovered, thus no speech but noise-only is active. In this case, the noise-only recording (or estimated spectrum) of the right ear can be used as a replacement for the left ear input. The binaural processing (see clause 6.6) is skipped for the noise hearing model spectrum. The processed signal (and spectrum vs time) is set to zero in this case.

2) Diotic mode: Similar to other loudness or quality prediction models, it is assumed that single-channel signals are played back diotically, i.e. the stimulus is presented on both ears. In this case, the model can be evaluated just with two or three single-channel input signals (noisy speech, noise-only, reference), assumed to be presented on both ears. The binaural processing (see clause 6.6) is skipped for all hearing model spectra.

Annex A (informative): Translations of attributes, categories and instructions

A.1 Overview

This annex provides guidelines for listening tests according to clause 5. Example instructions to test subjects as well as labels for attributes and categories are provided in multiple languages. While categories and attributes only depend on the language, listening test instructions also may vary for different applications, i.e. these are intended to explain the acoustic scenario to the listener.

If listening tests are conducted in languages that are not specified here, a suitable translation has to be created.

A.2 English Translation

A.2.1 Attributes and categories

Instruction/questionnaire: "Effort required to understand the meanings of sentences" (Table A.1).

Table A.1: Categories for Listening Effort (LE)

Category Description LE	Value
Complete relaxation possible; no effort required	5 (best)
Attention necessary; no appreciable effort required	4
Moderate effort required	3
Considerable effort required	2
No meaning understood with any feasible effort	1 (worst)

Instruction/questionnaire: "Please mark your opinion of the speech sample you have just been listening" (Table A.2).

Table A.2: Categories for Speech Quality (SQ)

Category Description SQ	Value
Excellent	5 (best)
Good	4
Fair	3
Poor	2
Bad	1 (worst)

A.2.2 Listening test instructions

EXAMPLE 1: In-car communication:

"Imagine that you are sitting in a vehicle as a passenger or in the back seat. You try to talk to the driver, who cannot turn around while driving. Please rate in the following listening test how much you have to strain to understand the driver, respectively to follow the conversation."

EXAMPLE 2: Handset/hands-free (mobile devices):

"Imagine that you are traveling in a variety of environments, such as e.g. in a café/restaurant, at the train station or on the street. You try to make a phone call despite the many surrounding noises. Please rate in the following listening test how much you have to strain to understand the talker on the call, respectively to follow the conversation."

A.3 German Translation

A.3.1 Attributes and categories

Instruction/questionnaire: "Wie würden Sie die erforderliche Anstrengung beschreiben, um dem Gesprächspartner zu folgen?" (Table A.3).

Table A.3: Categories for Listening Effort (LE)

Category Description LE (long)	(short)	Value
Keine Anstrengung notwendig	Keine	5 (best)
Geringe Anstrengung notwendig	Gering	4
Mäßige Anstrengung notwendig	Mäßig	3
Beträchtliche Anstrengung notwendig	Groß	2
Trotz Anstrengung Bedeutung nicht verstanden	Maximal/nicht verstanden	1 (worst)

Instruction/questionnaire: "Wie würden Sie die Sprachqualität des Hörbeispiels bewerten?" (Table A.4).

Table A.4: Categories for Speech Quality (SQ)

Category Description SQ	Value
Ausgezeichnet	5 (best)
Gut	4
Ordentlich	3
Dürftig	2
Schlecht	1 (worst)

A.3.2 Listening test instructions

EXAMPLE 1: In-car communication:

"Stellen Sie sich vor, Sie sitzen in einem Fahrzeug als Beifahrer oder auf der Rückbank. Sie versuchen, sich mit dem Fahrer zu unterhalten, welcher sich aber während der Fahrt nicht zu Ihnen umdrehen kann. Bitte bewerten Sie im folgenden Hörversuch, wie sehr Sie sich anstrengend müssen, um den Fahrer zu verstehen bzw. um den Gespräch folgen zu können."

EXAMPLE 2: Handset/hands-free (mobile devices):

"Stellen Sie sich vor, Sie sind in den unterschiedlichsten Umgebungen unterwegs, wie z.B. in einem Café/Restaurant, am Bahnhof oder an der Straße. Dabei versuchen Sie, trotz der vielen Umgebungsgeräusche ein Telefonat zu führen. Bitte bewerten Sie im folgenden Hörversuch, wie sehr Sie sich anstrengend müssen, um den anderen Gesprächsteilnehmer zu verstehen bzw. um den Gespräch folgen zu können."

Annex B (normative): Reference systems for listening tests

B.1 Overview

This annex provides several reference systems, which can be considered for listening tests according to clause 5 of the present document. The designer of the test should select one of the following methods, which fits best to the considered scope of the evaluation, i.e. the types of devices, acoustic scenario, etc.; in the same way, a suitable background noise should be selected for the current application (e.g. car noise for listening test dealing with ICC). Thus, concrete background noises are not specified in the following clauses, only levels or SNRs are provided. In case of no suitable reference system can be selected, references according to clause B.2 are recommended.

NOTE 1: All information on background noises (levels and/or SNRs) refers to A-weighted levels.

NOTE 2: If not specified otherwise, the active speech level according to Recommendation ITU-T P.56 [3] (excluding background noise) is assumed as -21 dB_{Pa} / 73 dB_{SPL} for diotic and -15 dB_{Pa} / 79 dB_{SPL} for monaural presentation.

B.2 MNRU

The Recommendation ITU-T P.810 [21] describes the reference disturbance "Modulated Noise Reference Unit" (MNRU). The degradation is controlled by a factor Q, usually specified in dB. The factor describes an attenuation of a biased noise, which is multiplied to the time signal. For bandwidths extending WB (SWB or FB), speech-shaped noise according to the weighting described in Recommendation ITU-T P.50 [22] shall be used (see Recommendation ITU-T P.Imp830 [23] as well for further reference). The 12 reference conditions for combined LE and SQ evaluations are provided in Table B.1.

Table B.1: Reference conditions for MNRU

Condition	Q [dB]	SNR (A) [dB]	Comment
R01	-	-	Direct reference
R02	-	0	Lowest anchor for LE
R03	-	12	[…]
R04	-	24	[…]
R05	-	36	Second-best anchor for LE
R06	8	-	Lowest anchor for SQ
R07	24	-	[…]
R08	32	-	[…]
R09	40	-	Second-best anchor for SQ
R10	32	24	Second-best anchor overall
R11	24	12	[…]
R12	8	0	Lowest anchor overall

This reference system should be used if no other suitable system is available. The resulting single-channel signals shall be played back diotically for the presentation in the listening test.

B.3 Wiener Filter Approach

In annex D of ETSI TS 103 281 [4], a reference system for SWB and FB systems is described. For NB and WB devices, a band-limited adaptation of the method can be found in ETSI TS 103 106 [i.6]. Table B.2 provides the processing settings, which are used to obtain the degradations for the reference system.

Table B.2: Reference conditions for noise reduction application

Condition	Speech Distortion	SNR (A) [dB]	Comment
R01	-	-	Direct reference
R02	-	0	Lowest anchor for LE
R03	-	12	[…]
R04	-	24	[…]
R05	-	36	Second-best anchor for LE
R06	NS Level 1	-	Lowest anchor for SQ
R07	NS Level 2	-	[…]
R08	NS Level 3	-	[…]
R09	NS Level 4	-	Second-best anchor for SQ
R10	NS Level 3	24	Second-best anchor overall
R11	NS Level 2	12	[…]
R12	NS Level 1	0	Lowest anchor overall

This system is typically used for listening test databases where noise suppression algorithms are evaluated. The reference distortions introduced here are based on several degrees of aggressiveness of a Wiener filter, which is applied on clean speech signals. Since these kinds of processing artefacts are expected mainly for the sending direction of terminals, a frequent usage for listening tests according to clause 5 is not expected.

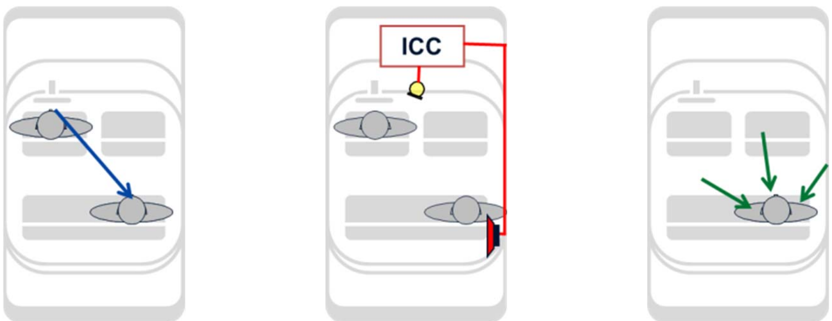
The resulting single-channel signals shall be played back diotically for the presentation in the listening test.

B.4 Reverb Artefacts

For the auditory evaluation of e.g. ICC applications, neither the Wiener filter nor the MNRU degradations sound similar to typical artefacts created by reverb and feedback cancellation of such setups and systems.

The following descriptions of artefacts are based on ICC systems, but can be generalized to other speech enhancement systems used in rooms.

In general, ICC systems should support and ease the communication between driver and passengers in the first or second row. The driver's voice is recorded via a microphone, usually the same as for the car hands-free system. The speech signal is then processed and played back over the loudspeakers close to the listener position. Thus, the perceived speech signal at the listener position is a superposition of the direct sound, the processed/reinforced signal and ambient noise. Figure B.1 illustrates the three contributions to the overall signal.



a) Direct **b) Reinforcement** **c) Noise**

Figure B.1: Signal contributions at listener position

For the generation of listening samples, a simplified ICC model according to Figure B.2 is used. As an input signal $s(k)$, monaural speech samples recorded at the MRP shall be used (like provided in e.g. [5] or [4]). For the direct path of the acoustic transmission from the driver to the listener, a binaural impulse response (IR) according to the ortho-reference condition is applied on the speech signal (see clause 1.4.1 of [9]). Even though this convolution does not represent the acoustics of a typical car cabin, at least some auditory spaciousness is introduced into the signal and binaural perception is facilitated. Finally, the direct sound is scaled to an active speech level of S_{dir} (in dB) according to Recommendation ITU-T P.56 [3] in order to simulate an acoustic loss between talker and listener.

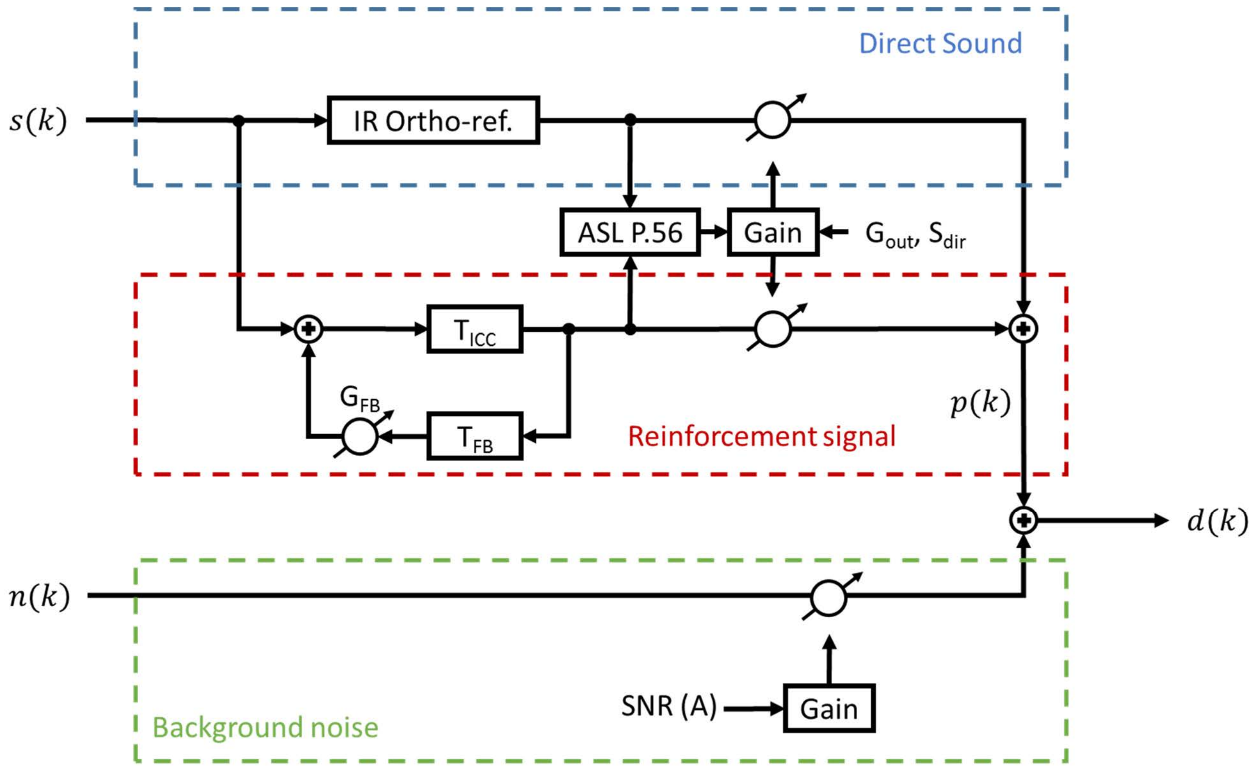


Figure B.2: Simplified ICC model for sample generation

The transmission path of the ICC system is modelled by a delay T_{ICC}, which reflects a virtual processing delay. The acoustic feedback path is approximated by the delay T_{FB} and a damping constant G_{FB} (in dB). The feedback delay T_{FB} is derived from the distance between loudspeakers to the microphone, usually in the range of 4 ms (about 1,40 m distance). The whole feedback loop can be realized as a "direct form II" filter with the coefficients specified in equations (B.1) to (B.3) (with signal sampling rate F_S); all other non-specified coefficients are set to zero:

$$a[0] = 1 \tag{B.1}$$

$$a[int(F_S \cdot (T_{ICC} + T_{FB}))] = 10^{\frac{G_{FB}}{20}} \tag{B.2}$$

$$b[int(F_S \cdot T_{ICC})] = 1 \tag{B.3}$$

The level of the reinforcement signal is then adjusted in order to provide an amplification G_{out} (in dB) compared to the direct sound. The target active speech level S_{ri} of the reinforcement signal is defined according to equation (B.4):

$$S_{ri} = 20 \cdot \log_{10}\left[(10^{\frac{G_{out}}{10}} - 1) \cdot 10^{\frac{S_{dir}}{20}}\right] \tag{B.4}$$

EXAMPLE: The active speech level of the direct sound S_{dir} is set to 67 dB_{SPL} and the desired gain of the system is +2 dB. According to equation (B.4), the active speech level S_{ri} of the reinforcement signal is calculated to 62,3 dB_{SPL}.

The processed signal $p(k)$ is calculated as the sum of the direct sound and the reinforcement signal. It contains typical ICC processing artefacts caused by the (artificial) feedback cancellation. The degree of speech degradation can be controlled by the delays T_{ICC} / T_{FB}, the gains G_{FB} / G_{out} and the level of the direct sound S_{dir}.

Finally, background noise is added in order to obtain the degraded signal $d(k)$. With the knowledge of the overall speech level (direct sound plus reinforcement), the level of the noise is scaled according to a given SNR. The noise level shall be calculated including A-weighting.

Table B.3 provides an exemplary processing set for the 12 reference conditions which should cover a wide range for speech quality and listening effort. Note that in some cases no reinforcement signal is used; in these cases, G_{out} is set to negative infinity (-inf.) and the other parameters of the feedback path are not specified.

Table B.3: Reference conditions for noise reduction application

Condition	T_{ICC} [ms]	T_{FB} [ms]	G_{FB} [dB]	G_{out} [dB]	S_{dir} [dB_{SPL}]	SNR (A) [dB]	Comment
R01				-inf.	70	-	Direct reference
R02				-inf.	61	-10	Lowest anchor for LE
R03				-inf.	64	-4	[…]
R04				-inf.	67	2	[…]
R05				-inf.	70	8	Second-best anchor for LE
R06	25	4	-2,8	3	67	-	Lowest anchor for SQ
R07	16	4	-2,8	2	67	-	[…]
R08	8	4	-3,8	2	67	-	[…]
R09	3	4	-3,8	1	67	-	Second-best anchor for SQ
R10	8	4	-3,8	1	67	2	Second-best anchor overall
R11	16	4	-2,8	2	67	-4	[…]
R12	25	4	-2,8	3	67	-10	Lowest anchor overall

Annex C (normative): Auditory Databases for Training and Validation of the model

C.1 General

This annex provides information about and references to the auditory tests, which were used to train the model (see clause 6.7.2).

C.2 Database for Handset Mode

C.2.1 Overview

Communication in noisy situations may be extremely stressful for the person located at the near-end side. Since the background noise is originated from the natural environment, it cannot be reduced for the listener. Thus, the only possibility to improve this scenario with support of digital signal processing is the insertion of speech enhancement algorithms in the downlink direction of terminals.

Some of these methods are already integrated in modern state-of-the-art mobile devices. Such algorithms target in general on the improvement of listening comfort on the near end. Methods like (artificial) BandWidth Extensions (BWE) or additional noise reduction are already quite common. Additionally, more sophisticated enhancement algorithms manipulate the speech signal with respect to the instantaneous local background noise estimation. The focus here is to improve speech intelligibility. Such methods are also known as speech reinforcement, intelligibility or Near-End Listening Enhancement (NELE).

To investigate the impact on intelligibility and quality, the combined auditory assessment of listening effort and speech quality according to clause 5 was applied on an artificially created, but realistic test corpus. In [i.12], this work was already presented in detail, thus only a brief summary is provided here.

C.2.2 Test Corpus

Figure C.1 illustrates the principle of the test corpus generation: the first stage was the acoustical noise recordings of the near-end listener. For that purpose, a mock-up device was mounted at right ear of head and torso simulator (HATS). With standard 8 N application force, a typical leakage was realized. The left ear remained uncovered for the binaural recording. A noise playback system according to ETSI TS 103 224 [i.4] with an 8-speaker-setup was then used to reproduce a realistic sound field around the HATS (left side of Figure C.1). Four standardized handset noises according to the database of ETSI TS 103 224 [i.4] were evaluated:

- Inside Car Noise - Full-size car 130 km/h.
- Public Places Noise - Cafeteria.
- Outside Traffic Street Noise - Road.
- Public Places Noise - Train station.

Each recording was played back with the realistic level. Two additional gains +6 dB and -6 dB were applied to each scenario to obtain a wider range of noise levels. Finally, silence condition (idle noise < 30 dB_{SPL} (A)) was also taken into account.

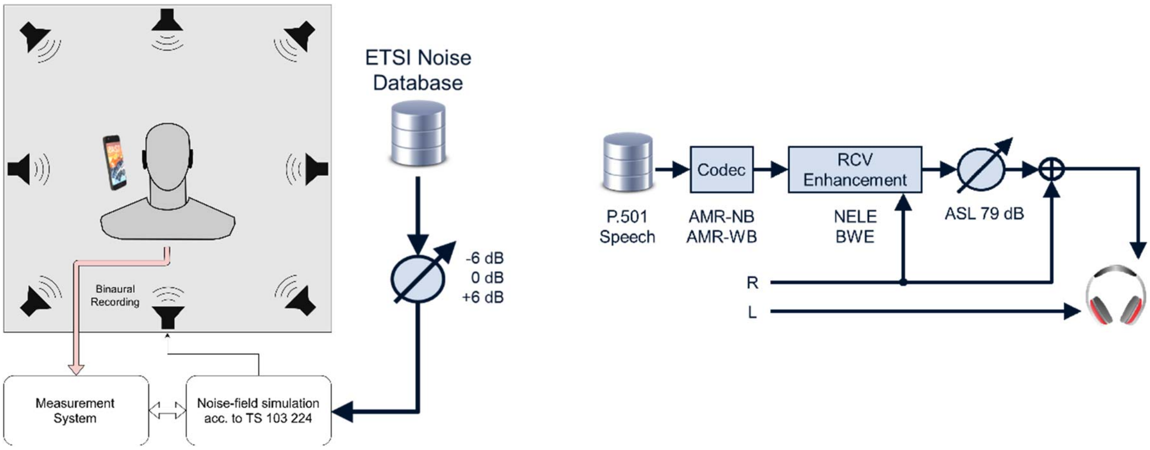


Figure C.1: Setup of binaural noise recording procedure (left) and generation of test corpus (right)

The right side of Figure C.1 shows the flow chart of the processing chain. In a first processing step, the original German speech material according to Recommendation ITU-T P.501 [5] is pre-filtered and down-sampled to narrowband and wide-band. Then, encoding and decoding of the widely used Adaptive Multi-Rate codec (AMR/AMR-WB) is applied. If applicable, the right ear noisy-only signal is used as an additional input for the speech signal enhancement (here: NELE).

After this step, the active speech level is normalized to 79 dB_{SPL} according to Recommendation ITU-T P.56 [3]. Especially common NELE algorithms utilize the maximum possible and allowed speech level. Since the focus is purely on the perception impact of sound manipulation but not on level differences all possibly occurring level differences are equalized. The resulting signals are assumed as the output of a mobile phone without further degradations, i.e. neglecting non-linear speaker distortion or any arbitrary transfer function.

Overall, nine NELE algorithms (eight for WB, one for NB), two BWE methods, two combinations of both and coding with AMR and AMR-WB only were included per background noise/gain set. Finally, the signal of the right artificial ear is mixed with the processed speech. By combining this signal with the left ear signal of the unprocessed background noise, a binaural stimulus is created for the listening test.

C.2.3 Auditory Testing

One binaurally presented sample of 8,0 s duration included two sentences of one talker. Thus, four samples per condition are obtained. In overall, 197 conditions with 788 different samples were auditory evaluated with the test design described in clause 5. In the evaluation, 56 native German speakers participated. Each participant listened to one sample per condition, which lead to 56 votes per condition or 14 votes per sample (for listening effort and speech quality).

C.3 Database for ICC

C.3.1 Overview

The in-car listening situation is often impacted by a low signal-to-noise ratio (SNR), which leads to reduced speech intelligibility and higher listening effort, respectively. This applies in particular to the communication between driver and passengers. Several ICC systems have been recently introduced in the market, aiming to improve this situation as well as to decrease driver distraction.

In order to investigate the application of perceived listening effort for ICC systems, this clause presents a comprehensive auditory experiment. It is based on binaural recordings containing realistic background noise scenarios, speech, and reinforced speech. In [i.13], this work was already presented in detail, thus only a brief summary is provided here.

C.3.2 Simulation Environment

Impulse response measurements in the cabin of two different vehicles were conducted (one mid- and one full-size car). This results in two Devices Under Test (DUTs), which are regarded in this evaluation. The talker and listener setups are identical for all conditions, the driver talks to the listener sitting directly behind him.

In order to simulate the whole ICC system offline, impulse responses from the equalized artificial mouth to the input microphones of the system as well as to the listener's ears - diffuse-field equalized Head And Torso Simulators (HATS) according to Recommendations ITU-T P.57 [6] and P.58 [7] - were measured with white noise signals. To simulate the effect of the ICC system, the impulse responses from the loudspeakers to the input microphones of the system as well as to the listener's ears were determined in a similar way. Driving noise was recorded synchronously at the ICC microphones and the listener's ears in both DUTs. The structure of the simulation environment that was used to obtain simulated binaural ear signals is shown on the left in Figure C.2.

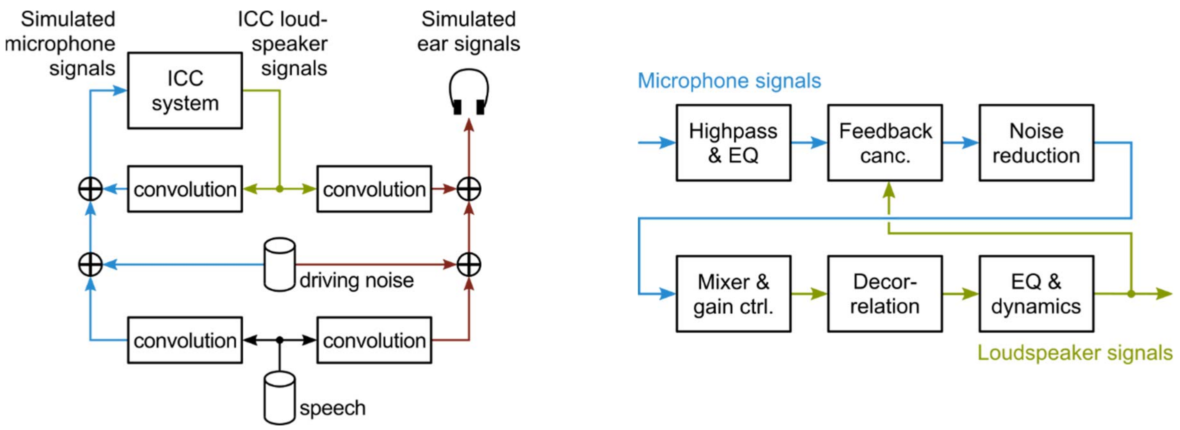


Figure C.2: Structure of the simulation environment (left) and simplified structure of the ICC system (right)

A simplified structure of the ICC system is depicted on the right in Figure C.2. The microphone signals are equalized and high-pass filtered to get rid of frequencies below the usual speech spectrum. NLMS-based feedback and echo cancellation is used to get rid of feedback and echoes from the ICC loudspeakers into the microphones. A mixer module distributes the noise-reduced signals from the talkers to the listening passengers, but also calculates and applies an appropriate gain for the present noise scenario. The loudspeaker signals are de-correlated using pitch-shifting, equalized, and their dynamic range is compressed.

The following modes of the simulated ICC system were used for the evaluation:

- ICC Off: the system is deactivated and no reinforcement is applied. This scenario is regarded as the baseline for all other settings.
- Default: the system is tuned for typical execution in the corresponding vehicle cabin in an assumed optimum/balanced setting.
- High Gain: the configuration is similar to the Default mode, but with additional output gain.
- Extra Delay 15: same as Default mode, but processing delay of the system is artificially increased by 15 ms.
- Extra Delay 25: same as Default mode, but processing delay of the system is artificially increased by 25 ms.

The ICC system in Default mode obtains a delay ΔR_x (difference between direct sound and reinforced speech) of about 5,5 ms.

C.3.3 Speech and Noise Levels

The German speech material according to ETSI TS 103 281 [4] was used for the simulation, which includes two sentences of four male and four female talkers. The speech sequence was used as a source for the simulation, representing a playback via the artificial mouth of the HATS with an Active Speech Level (ASL) according to Recommendation ITU-T P.56 [3] of -4,7 dB_{Pa} In addition, custom Lombard gains were added to each condition. Recent studies [i.14] show that the Lombard effect and its aspects are influenced by noise level, seat position, as well as by the ICC reinforcement level. The gains were manually and subjectively tuned in order to provide reasonable minimum speech levels for each noise condition.

With these figures, also the ASL of the direct path sound (without any reinforcement, but including Lombard gain) can be determined, the resulting values are shown in the upper part of Table C.1. For each DUT, two driving noises (medium and maximum speed) were binaurally recorded at the listener's position with diffuse-field equalization. The lower part of Table C.1 shows the averaged levels (left and right ear) of the background noises.

Table C.1: Levels of driving noise and speech

Level of	Noise	DUT 1	DUT 2
Speech [dB_{SPL}]	Silence	71	72
	Medium	72	74
	Maximum	74	76
Noise [dB_{SPL}(A)]	Silence	< 30	< 30
	Medium	68	75
	Maximum	74	79

The following parameters of the system were chosen for the auditory evaluation:

- Two DUTs/simulated car cabins.
- Five ICC modes.
- Three background noise scenarios (including silence).

In total, 2 × 5 × 3 = 30 test conditions were obtained by this segmentation.

C.3.4 Auditory Testing

The combined auditory assessment of listening effort and speech quality according to clause 5 was conducted in this study, including the reference conditions as defined in clause C.4.

A total of 48 naïve German test subjects participated in the auditory test, which contained 672 samples (42 conditions, 16 sentences each). Each subject listened to four blocks of 42 randomized samples (including one sample per condition). In total, 12 votes per sample and 192 votes per condition were obtained by this distribution. The stimuli were presented via diffuse-field equalized headphone playback.

C.4 Training and Validation

For the training of the model, the two databases were randomly split into a training and a validation part according Table C.2. Only test conditions are considered here (reference conditions used neither for training nor validation).

Table C.2: Levels of driving noise and speech

	Training	Validation
Database Handset	85 % (167 conditions, 668 samples)	15 % (30 conditions, 120 samples)
Database ICC	50 % (15 conditions, 240 samples)	50 % (15 conditions, 240 samples)

The validation results (performance metrics and scatter plot) for the handset database are shown in Figure C.3.

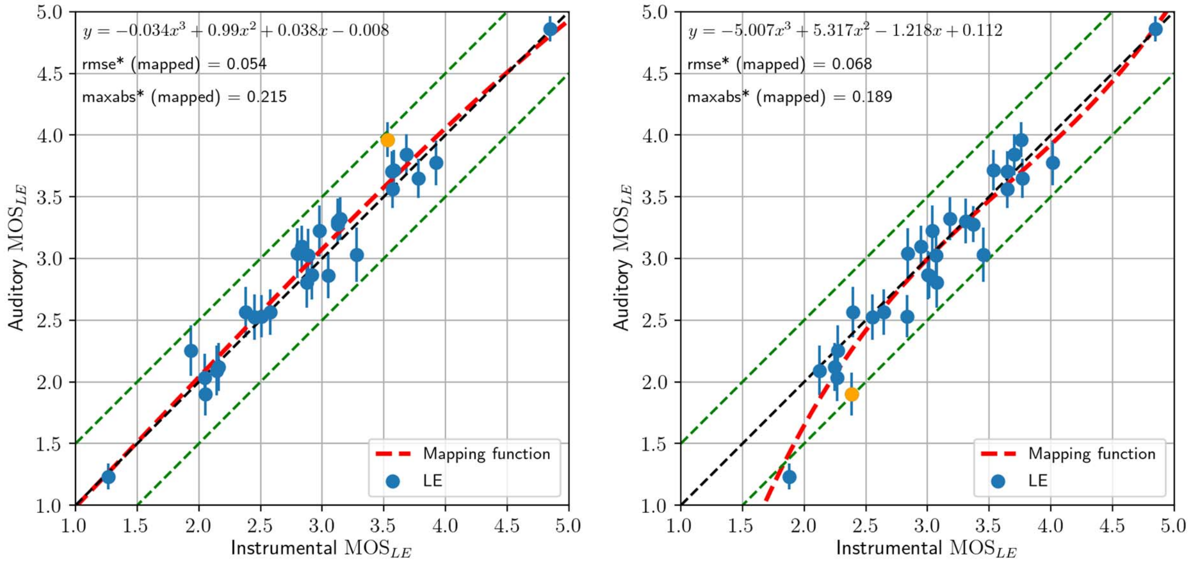


Figure C.3: Prediction results for handset database, with (left) and without (right) noise-only reference

The validation results (performance metrics and scatter plot) for the ICC database are shown in Figure C.4.

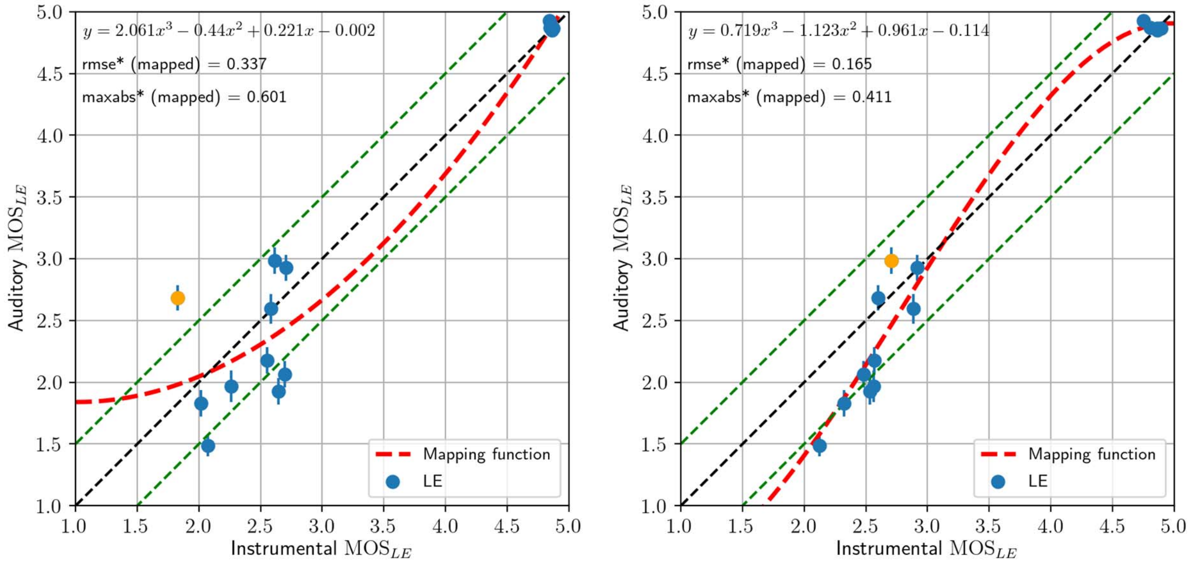


Figure C.4: Prediction results for ICC database, with (left) and without (right) noise-only reference

NOTE: For the ICC database, the vehicle interior noise is used as the noise-only recording. However, the degraded signal also contains additional processed noise, which deteriorates the prediction performance significantly (see left graph of Figure C.4). The usage without a noise-only reference leads to an accurate prediction (right graph of Figure C.4).

Annex D (normative): Assessment of Listening Effort based on subjective test databases (acoustical interfaces)

D.1 Overview

This annex provides the results of a comprehensive study, which conducted subjective tests to assess the listening effort and quality for three different applications:

- ANC (Active Noise Cancellation) Headsets.
- ICC (In-Car Communication).
- Mobile devices (handset and Hands-free terminals).

Several test conditions are defined for the different applications. One of the parameters is the background noise, which were chosen according to the typical use cases. This annex applies to the acoustical interface, which means that the listeners acoustically perceives the transmitted speech (processed by the application) and the environmental noise.

The annex describes speech material, background noises, the scenarios associated to each application, the subjective tests plan and the subjective test results. These results are split in two blocks: *TRAINING* and *VALIDATION* databases.

The *TRAINING* databases are used to train the objective model while the *VALIDATION* databases are used to validate the trained objective model with unknown data.

D.2 Test Plan

D.2.1 Overview

The test plan defines the different scenarios, including the test conditions (devices, environment noises, speech sequences in different languages and separating the databases for Model training and Model validation. Depending on the test conditions, the databases described in the following clauses are available in American English (ENG), Mandarin (MAN) and German (GER) languages. The speech samples are taken from ETSI TS 103 281 [4], annex E (up to 16 sentences).

D.2.2 Application: ANC headset

The first application is the situation where a listener is using headsets. These headsets may integrate an "Active Noise Cancellation system" that may be ON or OFF. The listener in the experiment is replaced by a HATS (Head and Torso Simulator).

Figure D.1 illustrates the three use case scenarios, which were used for the generation of acoustic HATS-based recordings with mounted ANC headset device:

- The first scenario (left of Figure D.1) records speech played back via downlink of the headset, while noise is played back via the noise field generation. This scenario is denoted as *HE* (headset usage) in the following.
- The second scenario (mid of Figure D.1) utilizes a second HATS, simulating as a second talker at left ear with distance 50 cm. This scenario is denoted as *2ndTalk* in the following.
- The third scenario (right of Figure D.1) utilizes an external loudspeaker at 50 cm above the listening artificial head (45° azimuth and elevation). This scenario simulates e.g. a public address or announcement system and is denoted as *ExtLs* in the following.

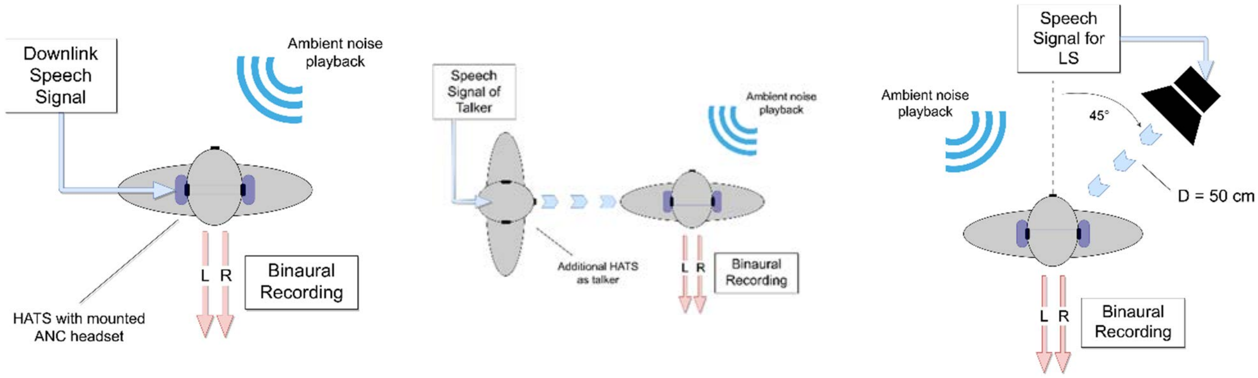


Figure D.1: Test setups for ANC use cases

Five devices under test (DUT) are defined as DUT-X (with X = A, B, C, D, E, F). Two types of headsets are selected for the experiment: two in-ear headsets and three over-ear ANC headsets. They have been evaluated for the acoustic recordings of the databases.

BGN0 to BGN3 are the names given to the background noises (or silence) in the recording room.

Table D.1: Noise Aliases

Alias	Noise type according to ETSI TS 103 224 [i.4], clause 8.3	Level Mic. #2 $dB_{SPL}(A)$	Level Mic. #7 $dB_{SPL}(A)$
BGN0	n/a (silence)	n/a	n/a
BGN1	Crossroadnoise	68,6	66,7
BGN2	Inside_Bus	66,6	66,5
BGN3	RailwayPlatform	76,8	76,4
NOTE: The level of the noises is measured close to the left/right ERPs of the HATS.			

Table D.2 and Table D.3 show the content of the two ANC headset databases for the STF575 project. Here the column "BGN" contains aliases/placeholders, which are described in Table D.1. In addition to each database, the reference conditions include the noise type Inside Airplane noise according to ETSI ES 202 396-1 [i.5].

Table D.2: ANC Headsets, Database 1 (Training)

Condition	DUT	UseCase	ANCMode	BGN	Condition	DUT	UseCase	ANCMode	BGN
C01	DUT-A	2ndTalker	on	BGN1	C25	DUT-A	ExtLS	off	BGN1
C02	DUT-A	2ndTalker	on	BGN2	C26	DUT-A	ExtLS	off	BGN2
C03	DUT-A	2ndTalker	on	BGN3	C27	DUT-A	ExtLS	off	BGN3
C04	DUT-A	2ndTalker	off	BGN1	C28	DUT-A	ExtLS	transparent	BGN1
C05	DUT-A	2ndTalker	off	BGN2	C29	DUT-A	ExtLS	transparent	BGN2
C06	DUT-A	2ndTalker	off	BGN3	C30	DUT-A	ExtLS	transparent	BGN3
C07	DUT-A	2ndTalker	transparent	BGN1	C31	DUT-B	ExtLS	on	BGN1
C08	DUT-A	2ndTalker	transparent	BGN2	C32	DUT-B	ExtLS	on	BGN2
C09	DUT-A	2ndTalker	transparent	BGN3	C33	DUT-B	ExtLS	on	BGN3
C10	DUT-B	2ndTalker	on	BGN1	C34	DUT-B	ExtLS	off	BGN1
C11	DUT-B	2ndTalker	on	BGN2	C35	DUT-B	ExtLS	off	BGN2
C12	DUT-B	2ndTalker	on	BGN3	C36	DUT-B	ExtLS	off	BGN3
C13	DUT-B	2ndTalker	off	BGN1	C37	DUT-B	ExtLS	transparent	BGN1
C14	DUT-B	2ndTalker	off	BGN2	C38	DUT-B	ExtLS	transparent	BGN2
C15	DUT-B	2ndTalker	off	BGN3	C39	DUT-B	ExtLS	transparent	BGN3
C16	DUT-B	2ndTalker	transparent	BGN1	C40	NoHeadset	ExtLS	na	BGN1
C17	DUT-B	2ndTalker	transparent	BGN2	C41	NoHeadset	ExtLS	na	BGN2
C18	DUT-B	2ndTalker	transparent	BGN3	C42	NoHeadset	ExtLS	na	BGN3
C19	NoHeadset	2ndTalker	n/a	BGN1	C43	DUT-C	HE	on	BGN2
C20	NoHeadset	2ndTalker	n/a	BGN2	C44	DUT-C	HE	on	BGN3
C21	NoHeadset	2ndTalker	n/a	BGN3	C45	DUT-C	HE	off	BGN2
C22	DUT-A	ExtLS	on	BGN1	C46	DUT-C	HE	off	BGN3
C23	DUT-A	ExtLS	on	BGN2	C47	DUT-C	HE	transparent	BGN2
C24	DUT-A	ExtLS	on	BGN3	C48	DUT-C	HE	transparent	BGN3

Table D.3: ANC Headsets, Database 2 (Validation)

Condition	DUT	UseCase	ANCMode	BGN	Condition	DUT	UseCase	ANCMode	BGN
C01	DUT-E	HE	on	BGN1	C25	DUT-F	ExtLS	on	BGN1
C02	DUT-E	HE	on	BGN2	C26	DUT-F	ExtLS	on	BGN2
C03	DUT-E	HE	on	BGN3	C27	DUT-F	ExtLS	on	BGN3
C04	DUT-E	HE	off	BGN1	C28	DUT-F	ExtLS	transparent	BGN1
C05	DUT-E	HE	off	BGN2	C29	DUT-F	ExtLS	transparent	BGN2
C06	DUT-E	HE	off	BGN3	C30	DUT-F	ExtLS	transparent	BGN3
C07	DUT-F	HE	on	BGN1	C31	DUT-E	2ndTalker	off	BGN1
C08	DUT-F	HE	on	BGN2	C32	DUT-E	2ndTalker	off	BGN2
C09	DUT-F	HE	on	BGN3	C33	DUT-E	2ndTalker	off	BGN3
C10	DUT-F	HE	off	BGN1	C34	DUT-E	2ndTalker	on	BGN1
C11	DUT-F	HE	off	BGN2	C35	DUT-E	2ndTalker	on	BGN2
C12	DUT-F	HE	off	BGN3	C36	DUT-E	2ndTalker	on	BGN3
C13	DUT-E	ExtLS	off	BGN1	C37	DUT-E	2ndTalker	transparent	BGN1
C14	DUT-E	ExtLS	off	BGN2	C38	DUT-E	2ndTalker	transparent	BGN2
C15	DUT-E	ExtLS	off	BGN3	C39	DUT-E	2ndTalker	transparent	BGN3
C16	DUT-E	ExtLS	on	BGN1	C40	DUT-F	2ndTalker	off	BGN1
C17	DUT-E	ExtLS	on	BGN2	C41	DUT-F	2ndTalker	off	BGN2
C18	DUT-E	ExtLS	on	BGN3	C42	DUT-F	2ndTalker	off	BGN3
C19	DUT-E	ExtLS	transparent	BGN1	C43	DUT-F	2ndTalker	on	BGN1
C20	DUT-E	ExtLS	transparent	BGN2	C44	DUT-F	2ndTalker	on	BGN2
C21	DUT-E	ExtLS	transparent	BGN3	C45	DUT-F	2ndTalker	on	BGN3
C22	DUT-F	ExtLS	off	BGN1	C46	DUT-F	2ndTalker	transparent	BGN1
C23	DUT-F	ExtLS	off	BGN2	C47	DUT-F	2ndTalker	transparent	BGN2
C24	DUT-F	ExtLS	off	BGN3	C48	DUT-F	2ndTalker	transparent	BGN3

D.2.3 Application: In-Car Communication (ICC)

Figure D.2 illustrates the test setup in a vehicle used for the acoustics recordings of in-car communication scenarios, where the driver is talking to the co-driver, using the In-car communication system. As for the first application the driver and co-driver are replaced by HATS. The vehicle is a convertible (compact/sports car), with only two seats in the second row. The talker HATS is positioned at the driver's seat (zone 1), the listener is located at the co-driver's seat (zone 2).

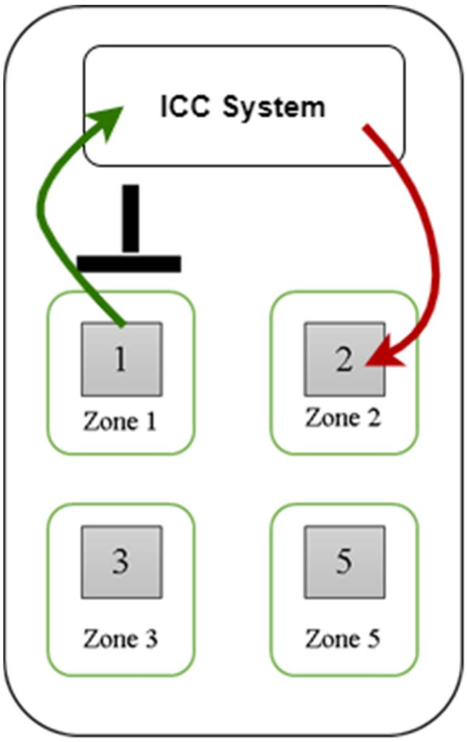


Figure D.2: Test setup for ICC

Table D.4 shows the content of the two ICC databases for the STF575 project. In addition to each database, the reference conditions include the noise type Full-size car 120 km/h according to ETSI ES 202 396-1 [i.5].

Table D.4: ICC Recordings, Database 3/4 (Training & Validation)

Condition	Speed [km/h]	ICC State	Noise Reduction	Decorrelation	Gain	Delay	Condition	Speed [km/h]	ICC State	Noise Reduction	Decorrelation	Gain	Delay
C01	0	off	n/a	n/a	n/a	n/a	C29	100	base	on	on	8	45
C02	0	base	on	on	0	15	C30	100	fb-cancel	on	on	5	15
C03	0	fb-cancel	on	on	10	15	C31	100	fb-cancel	on	on	10	15
C04	0	fb-cancel	on	on	10	45	C32	100	fb-cancel	on	on	15	15
C05	50	off	n/a	n/a	n/a	n/a	C33	100	fb-cancel	on	on	15	35
C07	50	base	on	on	5	15	C34	100	fb-cancel	on	on	20	15
C08	50	base	on	on	5	30	C35	100	fb-cancel	on	on	25	15
C09	50	base	on	on	5	45	C36	100	fb-cancel	on	off	15	15
C10	50	base	on	on	5	65	C37	100	fb-cancel	off	on	15	15
C11	50	base	on	on	8	15	C39	120	off	n/a	n/a	n/a	n/a
C12	50	base	on	on	8	45	C41	120	base	on	on	5	15
C13	50	fb-cancel	on	on	5	15	C42	120	base	on	on	5	30
C14	50	fb-cancel	on	on	10	15	C43	120	base	on	on	5	45
C15	50	fb-cancel	on	on	15	15	C44	120	base	on	on	5	65
C16	50	fb-cancel	on	on	15	35	C45	120	base	on	on	8	15
C17	50	fb-cancel	on	on	20	15	C46	120	base	on	on	8	45
C18	50	fb-cancel	on	on	25	15	C47	120	fb-cancel	on	on	5	15
C22	100	off	n/a	n/a	n/a	n/a	C48	120	fb-cancel	on	on	10	15
C24	100	base	on	on	5	15	C49	120	fb-cancel	on	on	15	15
C25	100	base	on	on	5	30	C50	120	fb-cancel	on	on	15	35
C26	100	base	on	on	5	45	C51	120	fb-cancel	on	on	20	15
C27	100	base	on	on	5	65	C52	120	fb-cancel	on	on	25	15
C28	100	base	on	on	8	15	C53	120	fb-cancel	on	off	15	15
							C54	120	fb-cancel	off	on	15	15

D.2.4 Application: Mobile Devices

The third application includes mobile devices in two different configurations: Handset mode (HS) and Hands-Free mode (HF). The positioning and mounting instructions comply with the ones described in ETSI TS 103 739 [i.16] for Handset (HS) and ETSI TS 103 740 [i.17] for Hands-Free (HF) terminals. Figure D.3 illustrates the two operational modes for a mockup device.

Figure D.3: Test setup for mobile devices: handset (left) and hands-free (right) mode

Regarding noise field simulation, the following conventions apply:

- For handset mode, the microphone array according to clause 5.4 (symmetric array) of ETSI TS 103 224 [i.4] is used for equalization and playback.
- For Hands-free mode, the microphone array according to clause 5.1 (asymmetric array) of ETSI TS 103 224 [i.4] is used for equalization and playback.

Four devices under test (DUT-01 to DUT-04) were evaluated for the acoustic recordings of the databases. Table D.6 shows the content of the two mobile phone databases for the STF575 project. Here the column "BGN" contains aliases/placeholders, which are described in Table D.5. In addition to each database, the reference conditions include the noise type Pub noise according to ETSI ES 202 396-1 [i.5].

Table D.5: Noise Aliases

Alias	Noise type according to ETSI TS 103 224 [i.4], clause 8.3 (symmetric array)	Noise type according to ETSI TS 103 224 [i.4], clause 8.2 (asymmetric array)
BGN0	n/a (silence)	n/a (silence)
BGN1	Crossroadnoise	Roadnoise
BGN2	Inside_Bus	Inside_Bus
BGN3	RailwayPlatform	TrainStation

NOTE: The difference between training and validation databases is realized by splitting the conditions into 2x24 conditions, which are then tested in different languages (see clause D.5).

Table D.6: Mobile Device Recordings, Database 5/6 (Training & Validation)

Condition	DUT	UseCase	BW-Mode	Volume	BGN	Condition	DUT	UseCase	BW Mode	Volume	BGN
C01	DUT-01	HS	NB	NOM	BGN0	C25	DUT-01	HF	NB	NOM	BGN0
C02	DUT-01	HS	NB	MAX	BGN0	C26	DUT-01	HF	NB	MAX	BGN0
C03	DUT-01	HS	NB	NOM	BGN1	C27	DUT-01	HF	NB	NOM	BGN1
C04	DUT-01	HS	NB	MAX	BGN1	C28	DUT-01	HF	NB	MAX	BGN1
C05	DUT-01	HS	NB	NOM	BGN2	C29	DUT-01	HF	NB	NOM	BGN2
C06	DUT-01	HS	NB	MAX	BGN2	C30	DUT-01	HF	NB	MAX	BGN2
C07	DUT-01	HS	SWB	NOM	BGN0	C31	DUT-01	HF	SWB	NOM	BGN0
C08	DUT-01	HS	SWB	MAX	BGN0	C32	DUT-01	HF	SWB	MAX	BGN0
C09	DUT-01	HS	SWB	NOM	BGN1	C33	DUT-01	HF	SWB	NOM	BGN1
C10	DUT-01	HS	SWB	MAX	BGN1	C34	DUT-01	HF	SWB	MAX	BGN1
C11	DUT-01	HS	SWB	NOM	BGN2	C35	DUT-01	HF	SWB	NOM	BGN2
C12	DUT-01	HS	SWB	MAX	BGN2	C36	DUT-01	HF	SWB	MAX	BGN2
C13	DUT-02	HS	NB	NOM	BGN0	C37	DUT-02	HF	NB	NOM	BGN0
C14	DUT-02	HS	NB	MAX	BGN0	C38	DUT-02	HF	NB	MAX	BGN0
C15	DUT-02	HS	NB	NOM	BGN1	C39	DUT-02	HF	NB	NOM	BGN1
C16	DUT-02	HS	NB	MAX	BGN1	C40	DUT-02	HF	NB	MAX	BGN1
C17	DUT-02	HS	NB	NOM	BGN2	C41	DUT-02	HF	NB	NOM	BGN2
C18	DUT-02	HS	NB	MAX	BGN2	C42	DUT-02	HF	NB	MAX	BGN2
C19	DUT-02	HS	SWB	NOM	BGN0	C43	DUT-02	HF	SWB	NOM	BGN0
C20	DUT-02	HS	SWB	MAX	BGN0	C44	DUT-02	HF	SWB	MAX	BGN0
C21	DUT-02	HS	SWB	NOM	BGN1	C45	DUT-02	HF	SWB	NOM	BGN1
C22	DUT-02	HS	SWB	MAX	BGN1	C46	DUT-02	HF	SWB	MAX	BGN1
C23	DUT-02	HS	SWB	NOM	BGN2	C47	DUT-02	HF	SWB	NOM	BGN2
C24	DUT-02	HS	SWB	MAX	BGN2	C48	DUT-02	HF	SWB	MAX	BGN2

D.2.5 Distribution of Languages vs Training/Validation

As indicated in the test plan, in overall six databases are provided:

- 3x Training Databases (48 conditions, 16 samples per condition, 12 votes per sample, 192 votes per condition)
- 3x Validation Databases (48 conditions, 8 samples per condition, 16 votes per sample, 128 votes per condition)

Since all databases are available in multiple languages, it would be desirable to have identical test material in two or three languages for direct comparison (of the auditory as well as for the objective results). On the other side, the training databases should cover a wide range of applications and degradations.

In order to achieve such a symmetry regarding language distribution, the databases of the handset/hands-free application are split by 2x24 conditions. Note that also for half of the database, all 12 reference conditions shall be included, which causes a slight overhead. Table D.7 provides an overview about the languages used for the training and validation databases.

Table D.7: Training & Validation databases

<table>
<tr><th></th><th>ANC</th><th>ICC</th><th>HS/HF</th></tr>
<tr><td rowspan="2">Training</td><td rowspan="2">MAN</td><td rowspan="2">ENG</td><td>50 % MAN</td></tr>
<tr><td>50 % GER</td></tr>
<tr><td rowspan="2">Validation</td><td rowspan="2">ENG</td><td rowspan="2">MAN</td><td>50 % GER</td></tr>
<tr><td>50 % MAN</td></tr>
</table>

D.3 Subjective tests

D.3.1 Overview

The subjective Laboratory performed 8 different experiments (i.e. two experiments in English, two experiments in German and 4 experiments in Mandarin). The experiment name, methodology (as defined in Recommendation ITU-T P.800 [1]) and language for each experiment is listed in Table D.8. The terms "Training" and "Validation" refer to the objective model development phase as defined in clause D.1.

Table D.8: Allocation of databases and languages

Exp.	Type	Language
1	Training ICC	ENG
2	Validation ANC	ENG
3	Training HF	GER
4	Validation HS	GER
5	Training ANC	MAN
6	Training HS	MAN
7	Validation HF	MAN
8	Validation ICC	MAN

D.3.2 Speech material

The subjective Laboratory used processed speech material delivered by the Recording Laboratory in 48-kHz sampled with 24-bit resolution stereo uncompressed wav files. The processed speech material for all tests conformed to restrictions indicated in "Practical procedures for subjective testing" [8]. No sample name blinding or anonymization has been applied as it was principally not needed.

D.3.3 Listening Environment

The tests were performed in an acoustically treated critical listening room in the Subjective Laboratory that conforms to the requirements of [1] in full. Its background noise during the tests was below than 30 $dB_{SPL}(A)$ with no peaks in audible acoustic frequency range and its reverberation time is 185 ms.

Listening equipment conformed to specified requirements [1] and [8] in full. Diffuse-field equalized Sennheiser headphones HD-650 have been used for all experiments. All used headphones have been calibrated and verified before and after performed experiments. A professional digital voting device has been used to collect the votes.

D.3.4 Test execution and sample presentation

Test duration never exceeded 1,5 hours per listening panel. Test duration comprised of up to 50 % of actual listening time and test overhead including administration, initial briefing, preliminaries, and breaks.

To avoid amplitude clipping for samples with significant peaks in time-domain, it was agreed not to stick to commonly used playout loudness calibration (73 dB_{SPL} equals to -26 dBov). Instead, the listening levels were agreed individually for each experiment type and are reported in Table D.9.

Table D.9: Listening levels for different experiment types

Type	73 dB SPL corresponds to
ANC	-36 dBov
ICC	-48 dBov
HS	-42 dBov
HF	-42 dBov

For each speech sample, both LE and MOS-LQS following ACR methodology have been assessed. The order of LE and OS-LQS questions has been balanced, i.e. the subjects were asked for their assessment of LE and consequently for OS-LQS in half of the listening sessions, and for OS-LQS and then LE in the other half. Within each session, identical question order has been kept for all samples.

D.3.5 Instructions to subjects

The instructions to subjects were available in written during the entire test sessions. During the training session of each test, the instructions were verbally explained and if needed briefly discussed by a dedicated expert person that was able to answer questions from the subjects in their native language.

The instructions used for testing in English, German and Mandarin are provided in Figure D.4, Figure D.5 and Figure D.6.

In this experiment, you will be listening to short groups of sentences via the headphones, and giving your opinion of the speech you hear.

Follow the instructions on the touchscreen in front of you, and listening to each sentence group, press the appropriate button to indicate your opinions on the following scales.

EFFORT REQUIRED TO UNDERSTAND THE MEANINGS OF SENTENCES

5 Complete relaxation possible; no effort required.

4 Attention necessary; no appreciable effort required.

3 Moderate effort required.

2 Considerable effort required.

1 No meaning understood with any feasible effort.

WHAT WAS THE QUALITY OF THE SAMPLE YOU HAVE JUST HEARD?

5 Excellent

4 Good

3 Fair

2 Poor

1 Bad

For playing subsequent sample, follow the instructions on the screen. Please do not discuss your opinions with other subjects participating in the experiment. Thank you for your help in this experiment.

Figure D.4: Instructions for test subjects (English)

In diesem Experiment werden Sie über Ihre Kopfhörer kurze Satzsequenzen hören und anschließend das Gehörte beurteilen.

Folgen Sie bitte den Anweisungen auf dem Touchscreen. Nachdem Sie sich einzelne Satzsequenzen angehört haben, bewerten Sie bitte die Hörsituation auf der folgenden Skala:

WIE ANSTRENGEND IST ES FÜR SIE, DEM INHALT DES GESPRÄCHS ZU FOLGEN?

5 absolut mühelos, keine Anstrengung notwendig

4 ein wenig Aufmerksamkeit erforderlich, geringe Anstrengung notwendig

3 mäßig anstrengend

2 beträchtliche Anstrengung

1 trotz hoher Anstrengung kein Verständnis möglich

WIE BEURTEILEN SIE DIE QUALITÄT DES SPRACHBEISPIELS?

5 ausgezeichnet

4 gut

3 ausreichend

2 dürftig

1 schlecht

Um das nächste Beispiel anzuhören, folgen Sie den Anweisungen auf dem Bildschirm. Tauschen Sie sich bitte während des Experiments nicht mit anderen Teilnehmenden aus. Vielen Dank für Ihre Mitarbeit!

Figure D.5: Instructions for test subjects (German)

在本次试验中，您将通过耳机收听简短的句组，并对听到的讲话发表自己的意见。

按照您面前的触摸屏上的指示，在听每一组句子的同时，按下下列标度中相应的按钮来表达您的意见。

您刚才听到的样品质量怎么样?

5 非常好
4 好
3 还可以
2 不太好
1 不好

需要听懂句子含义的尽力度:

5 可以完全放松;不用尽力听。
4 需要注意，但不用太尽力听
3 需要适当的尽力去听。
2 需要非常尽力地听。
1 再怎么努力都无法听懂句子的含义。

依照屏幕上的指示播放下组列句。请不要与其他受试者讨论您的意见。感谢您参与本次试验。

Figure D.6: Instructions for test subjects (Mandarin)

D.3.6 Test Subjects

For each listening test, only native speakers of the tested language were used. Only normal-hearing subjects have been used for the experiments. Their hearing normality has been verified by subject self-assessment questionnaire during the subject hiring phase and verified prior the session by expert test. Experiment required 32 (training) or 48 (validation) listeners (4 or 6 panels of 8 listeners each). The age and gender information for the set of subjects used in each listening test in each tested language is provided in Table D.10.

Table D.10: Age and gender information of subjects

Language	#females/#males	Average age (years)	Age StdDev (years)
English	1,00	33,4	9,87
German	1,00	31,8	9,14
Chinese	1,00	30,5	10,17

D.3.7 Raw data delivery

All raw voting data have been delivered to the laboratory in charge of the objective model training. They contain averages and standard deviations of LE as well as MOS-LQ for all conditions within each test.

D.4 Subjective test results of the training databases

The raw results for the training databases, i.e. for ICC (in English), ANC Headsets (in Mandarin), Mobile Handset (in Mandarin) and Mobile Handsfree (in German) can be downloaded here: https://docbox.etsi.org/STQ/Open/TS%20103%20558/Annex%20D.

D.5 Prediction Results for Validation Databases

D.5.1 Overview

The prediction results for the databases described in clause D.2.5 are provided in the following clauses. For each database up to two scatter plots are shown, corresponding to the two possible modes of the model (with and without noise-only reference, see clause 6.1). For some applications, the usage of the noise-only reference is not possible/applicable.

The following performance metrics are provided for each validation:

- rmse*: root-mean-square error per condition after 3rd order mapping according to Recommendation ITU-T P.1401 [i.3], taking the uncertainty of auditory data into account.
- maxabs*: absolute maximum error per condition after 3rd order mapping, taking the uncertainty of auditory data into account.

The raw results for the validation databases, i.e. for ICC (in Mandarin), ANC Headsets (in English), Mobile Handset (in German) and Mobile Handsfree (in Mandarin) can be downloaded here: https://docbox.etsi.org/STQ/Open/TS%20103%20558/Annex%20D.

NOTE: The prediction results presented in the following clauses are presented per condition (average across all samples) for each validation database. In each comparison between auditory and predicted listening effort, a 3rd order mapping function is provided. It should be noted that these mapping functions are only valid for the corresponding test context (combination of application, language, distribution of conditions, etc.) and are not generally applicable. There may exist language- or application-specific mapping functions; however, at this time the underlying databases do not provide sufficient information on this topic.

D.5.2 ANC headsets

The results for the application ANC in English language as described in clause D.2.2 are shown in Figure D.7.

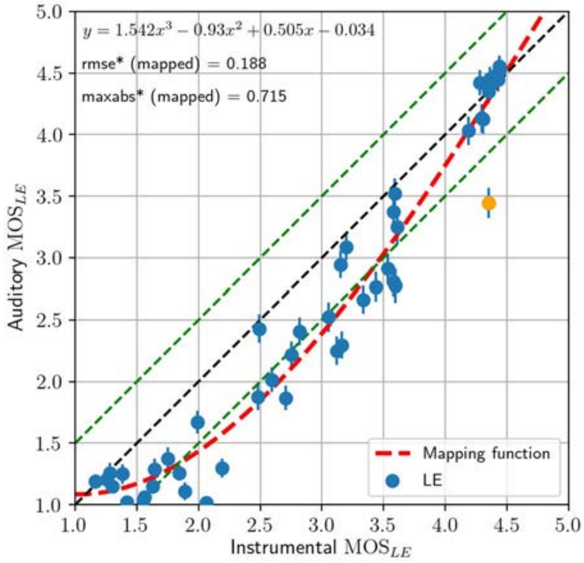


Figure D.7: Prediction results for application ANC headsets (English), without noise-only reference

D.5.3 In-Car Communication (ICC)

The results for the application In-car communication in Mandarin language as described in clause D.2.3 are shown in Figure D.8.

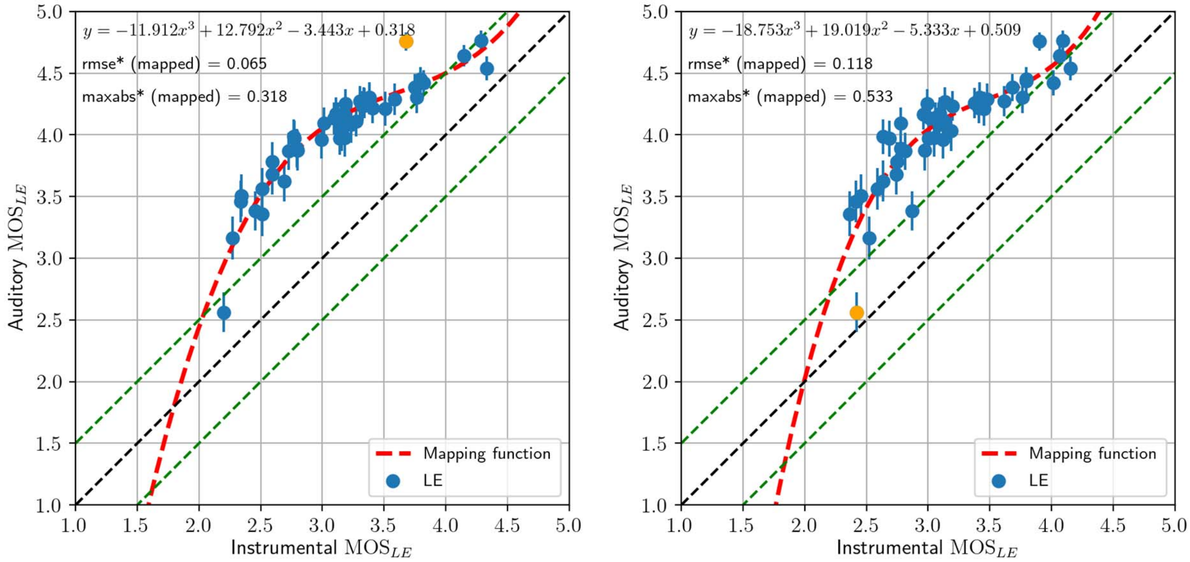


Figure D.8: Prediction results for application ICC (Mandarin), with (left) and without (right) noise-only reference

D.5.4 Mobile devices/handset

The results for the application handheld hands-free in German language as described in clause D.2.4 are shown in Figure D.9.

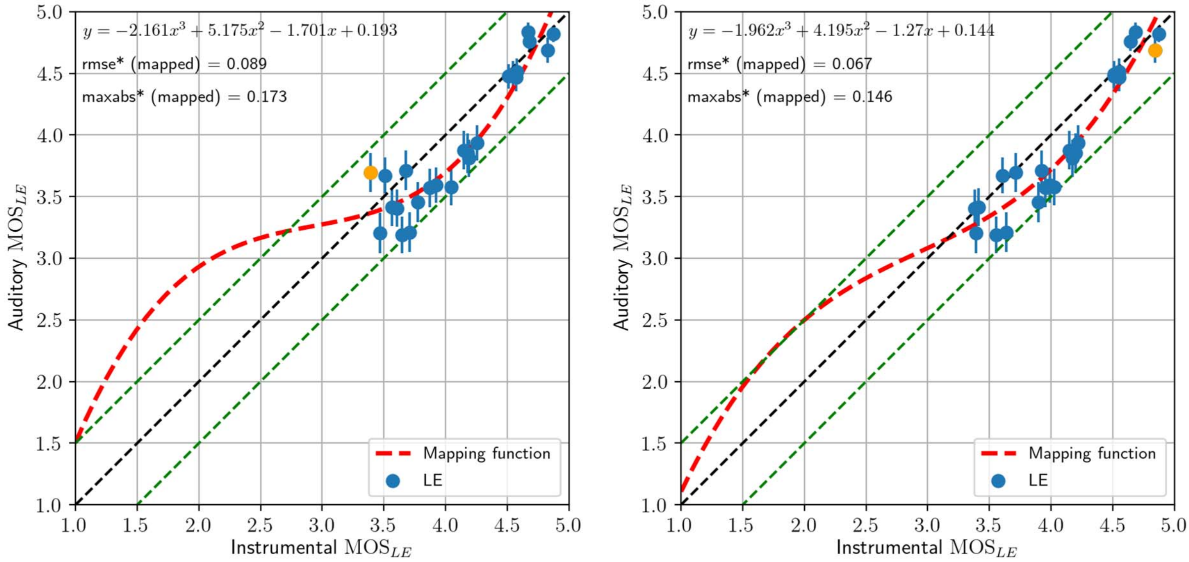


Figure D.9: Prediction results for application handset (German), with (left) and without (right) noise-only reference

D.5.5 Mobile devices/hands-free

The results for the application handheld hands-free in Mandarin language as described in clause D.2.4 are shown in Figure D.10.

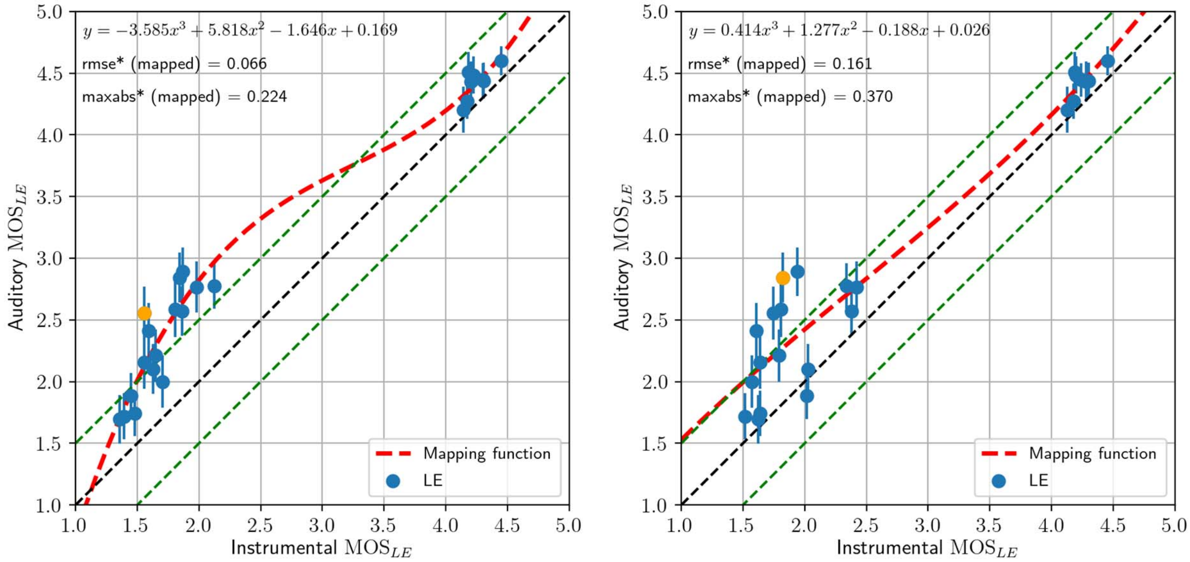


Figure D.10: Prediction results for application handheld hands-free (Mandarin), with (left) and without (right) noise-only reference

NOTE: For handheld hands-free mode, the listening effort prediction can be carried out in both modes (with and without noise-only reference signal). Even though the performance is similar in both cases, it is recommended to run the calculations with noise-only reference, especially under low SNR conditions in order to obtain a more accurate prediction performance.

Annex E (normative): Assessment of Listening Effort based on subjective test databases (electrical interfaces)

E.1 Overview

This annex provides the results of a comprehensive study, which conducted subjective tests to assess the listening effort and quality for three different scenarios.

This annex applies to the electrical interface, several test conditions are defined for the different scenarios, which cover impairments regarding acoustics (scenario #1), signal processing (scenario #2) and network (scenario #3).

The annex describes speech material, background noises, the scenarios associated to each application, the subjective tests plan and the subjective test results. These results are split in two blocks: *TRAINING* and *VALIDATION* databases.

The *TRAINING* databases are used to train the objective model while the *VALIDATION* databases are used to validate the trained objective model with unknown data.

E.2 Test Plan

E.2.1 Overview

For the recording of the following scenarios, two real mobile devices are used, which are capable of NB (AMR 12,2 kbit/s; DUT1) and SWB transmission (EVS-SWB at 24,4 kbit/s; DUT2). In order to obtain recordings with less redundancy, WB (e.g. AMR-WB codec) is excluded for this scenario (it is expected that SWB and WB will sound quite similar). Handset mode (abbreviated as HA in the following description) and handheld hands-free position (abbreviated as HH) were taken into account.

For additional offline simulation, recordings were also conducted with a measurement microphone close to the input microphones of the devices (denoted as "MeasurementMic" in the following description). These microphone recordings provide background noise with unprocessed speech signals.

For the handheld hands-free position, a simulation reverberation was applied in addition to the typical non-/low-reverberant recordings. For the simulation, the reverberation scenario "Room1" of ETSI TS 103 557 [i.18] was used (RT60 of ~600 ms). Even though the chosen environment may physically not match the noise types used for the recordings, it seems beneficial to have such a typical acoustic impairment included in the database. Conditions including reverberated speech signals are denoted with "REVERB" in the following description.

The two different recording setups are illustrated in Figure E.1 (with DUT) and Figure E.2 (with measurement microphone).

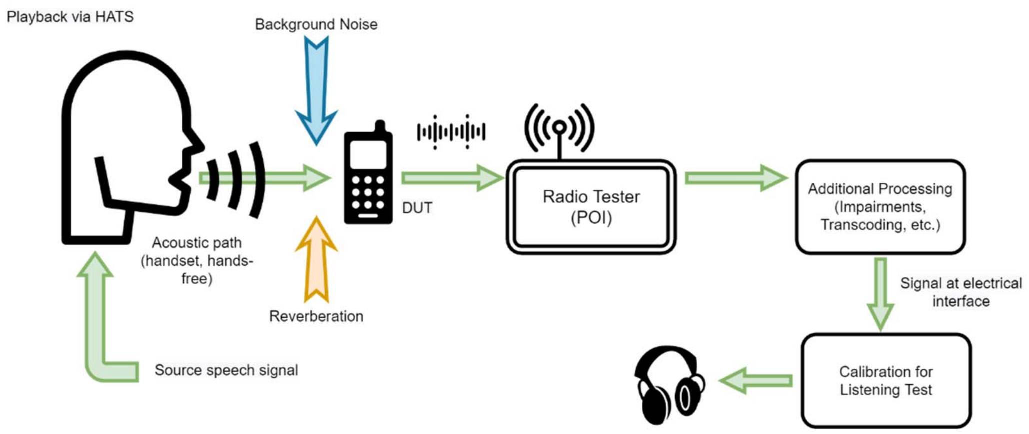


Figure E.1: Recording setup with DUT

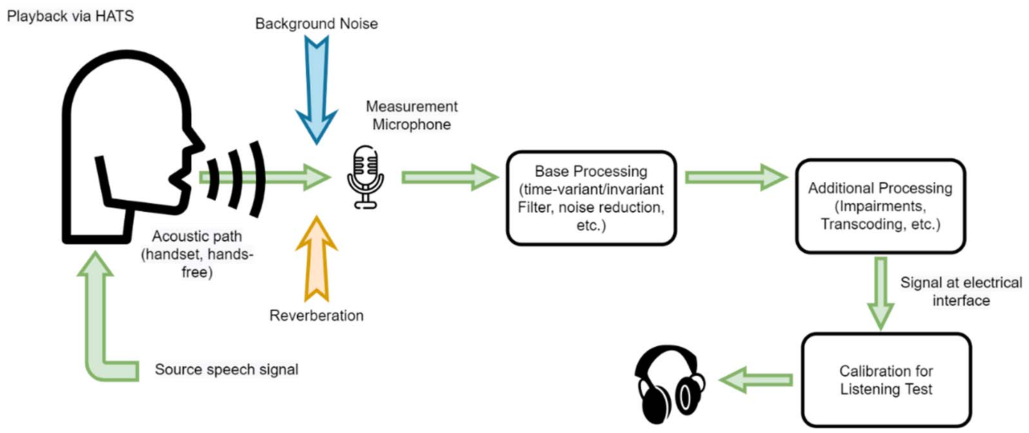


Figure E.2: Recording setup with microphone and subsequent offline processing

For the calibration of samples for the listening test, the same approach as described in clause 9.5 of ETSI TS 103 281 [4] was chosen:

- For each mode (handset, handheld hands-free) and each impairment (None, DOWN+LB, REVERB), a silent recording with the aforementioned speech level is carried out.
- For each of this silence recording, the active speech level according to Recommendation ITU-T P.56 [3] is calculated. The difference (in dB) to the target level of -21 dB_{Pa} / 73 dB_{SPL} is calculated and stored.
- Each noisy recording is calibrated with the same value as the corresponding silence recording.

E.2.2 Scenario #1

In addition to the aforementioned acoustical impairments, recordings in two positions were made in handset mode,. Beside the traditional standard handset position, the so-called *DOWN* position as described in ETSI TS 103 739 [i.16] was evaluated. In addition, the lowest bitrate possible was chosen for the transmission (4,75 kbit/s for AMR, 9,6 kbit/s for EVS-SWB). For this reason, the impairment is denoted as *DOWN+LB* in the following.

Additional encoding and decoding of signals was carried out with the reference implementation for EVS [i.21] and AMR [i.19] codecs.

The 48 test conditions for scenario#1 are provided in Table E.1.

Table E.1: Condition list for scenario#1

Condition ID	DUT	Band-width	UseCase	Impair-ment	Noises	Processing
C01	DUT1	NB	HA	DOWN+LB	FullSizeCar_130	
C02	DUT1	NB	HA	DOWN+LB	Pub	
C03	DUT1	NB	HA	DOWN+LB	Silence	
C04	DUT1	NB	HA	None	FullSizeCar_130	
C05	DUT1	NB	HA	None	Pub	
C06	DUT1	NB	HA	None	Silence	
C07	DUT1	NB	HH	None	Roadnoise	
C08	DUT1	NB	HH	None	Silence	
C09	DUT1	NB	HH	None	TrainStation	
C10	DUT1	NB	HH	REVERB	Roadnoise	
C11	DUT1	NB	HH	REVERB	Silence	
C12	DUT1	NB	HH	REVERB	TrainStation	
C13	DUT2	SWB	HA	DOWN+LB	FullSizeCar_130	
C14	DUT2	SWB	HA	DOWN+LB	Pub	
C15	DUT2	SWB	HA	DOWN+LB	Silence	
C16	DUT2	SWB	HA	None	FullSizeCar_130	
C17	DUT2	SWB	HA	None	Pub	
C18	DUT2	SWB	HA	None	Silence	
C19	DUT2	SWB	HH	None	Roadnoise	

Condition ID	DUT	Band-width	UseCase	Impair-ment	Noises	Processing
C20	DUT2	SWB	HH	None	Silence	
C21	DUT2	SWB	HH	None	TrainStation	
C22	DUT2	SWB	HH	REVERB	Roadnoise	
C23	DUT2	SWB	HH	REVERB	Silence	
C24	DUT2	SWB	HH	REVERB	TrainStation	
C25	MeasurementMic		HA	None	FullSizeCar_130	EVS-SWB-244
C26	MeasurementMic		HA	None	Pub	EVS-SWB-244
C27	MeasurementMic		HA	None	Silence	EVS-SWB-244
C28	MeasurementMic		HA	None	FullSizeCar_130	LEVEL-6;EVS-SWB-244
C29	MeasurementMic		HA	None	Pub	LEVEL-6;EVS-SWB-244
C30	MeasurementMic		HA	None	Silence	LEVEL-6;EVS-SWB-244
C31	MeasurementMic		HH	None	Roadnoise	EVS-SWB-244
C32	MeasurementMic		HH	None	Silence	EVS-SWB-244
C33	MeasurementMic		HH	None	TrainStation	EVS-SWB-244
C34	MeasurementMic		HH	REVERB	Roadnoise	EVS-SWB-244
C35	MeasurementMic		HH	REVERB	Silence	EVS-SWB-244
C36	MeasurementMic		HH	REVERB	TrainStation	EVS-SWB-244
C37	MeasurementMic		HA	None	FullSizeCar_130	AMR-122
C38	MeasurementMic		HA	None	Pub	AMR-122
C39	MeasurementMic		HA	None	Silence	AMR-122
C40	MeasurementMic		HA	None	FullSizeCar_130	LEVEL-6;AMR-122
C41	MeasurementMic		HA	None	Pub	LEVEL-6;AMR-122
C42	MeasurementMic		HA	None	Silence	LEVEL-6;AMR-122
C43	MeasurementMic		HH	None	Roadnoise	AMR-122
C44	MeasurementMic		HH	None	Silence	AMR-122
C45	MeasurementMic		HH	None	TrainStation	AMR-122
C46	MeasurementMic		HH	REVERB	Roadnoise	AMR-122
C47	MeasurementMic		HH	REVERB	Silence	AMR-122
C48	MeasurementMic		HH	REVERB	TrainStation	AMR-122
NOTE 1: Names of noise types according to clause 8 of ETSI TS 103 224 [i.4]. NOTE 2: Column "Processing" describes additionally applied processing steps. EVS-SWB-244 corresponds to encoding/decoding with EVS-SWB codec at 24,4 kbit/s, AMR-122 to encoding/decoding with AMR codec at 12,2 kbit/s. LEVEL-6 indicates that the whole signal level was decreased by 6 dB.						

E.2.3 Scenario #2

The recordings of scenario#1 already cover most realistic use cases of noise insertion into the network, including non-existing noise reduction (simulated). In this scenario#2, several time-variant impairments are introduced, which can occur in the presence of ambient noise in case the device is e.g. malfunctioning or not optimized yet for daily use.

In contrast to scenario#1, all conditions were simulated. As input to this simulation, speech and noise were separately recorded at a typical handset and/or hands-free position (including reverb simulation) with the measurement microphone introduced in scenario#1. Then additional post-processing steps were applied.

The first artificially introduced impairment are two types of time-variant filters:

1) Time-Variant Low-Pass (TVLP): The frequency of a low-pass filter is periodically varied versus time between 800 Hz and 16 kHz.

2) Time-Variant High-Pass (TVHP): The frequency of a high-pass filter is periodically varied versus time between 80 and 500 Hz.

An example result of this processing is shown in Figure E.3 for three consecutive samples of 4,0 s. The period time is set to 5,0 s in order to impair different positions of a sentence (onset, offset, etc.). Note that the filtering is applied only to the speech signal, not to the noise component. This type of degradation simulates a varying positioning of the device by the user, or incorrect operation of signal processing in the device.

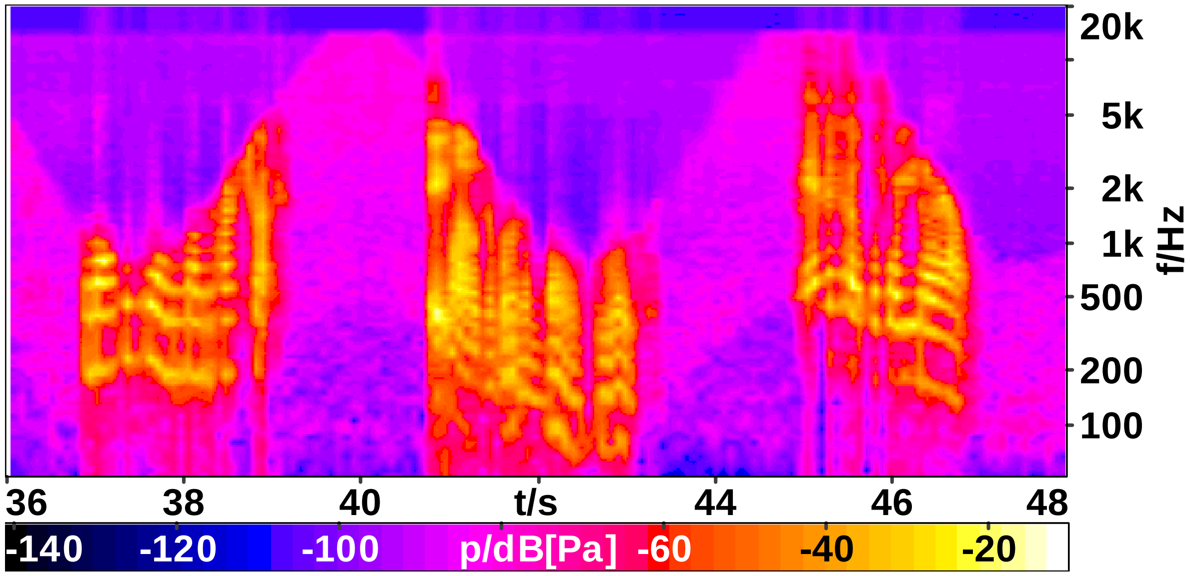


Figure E.3: Example of TVLP and TVHP

The second time-variant speech processing is a noise reduction based on spectral gating [i.22]. In contrast to other existing offline noise reductions is that this method only needs the degree of reduction (between 0 % and 100 %) as an input parameter (beside the noisy speech signal). Note that this algorithm was not developed for the reduction of noise in mobile phone applications. Especially for low SNRs, the processing introduces artefacts known as "musical tones". Figures E.4 to E.6 illustrate the noise reduction capabilities with an example of two sentences in the time-frequency domain. For all processing of scenario#2, the degree of reduction was set to 100 % (denoted as NR100 in the following).

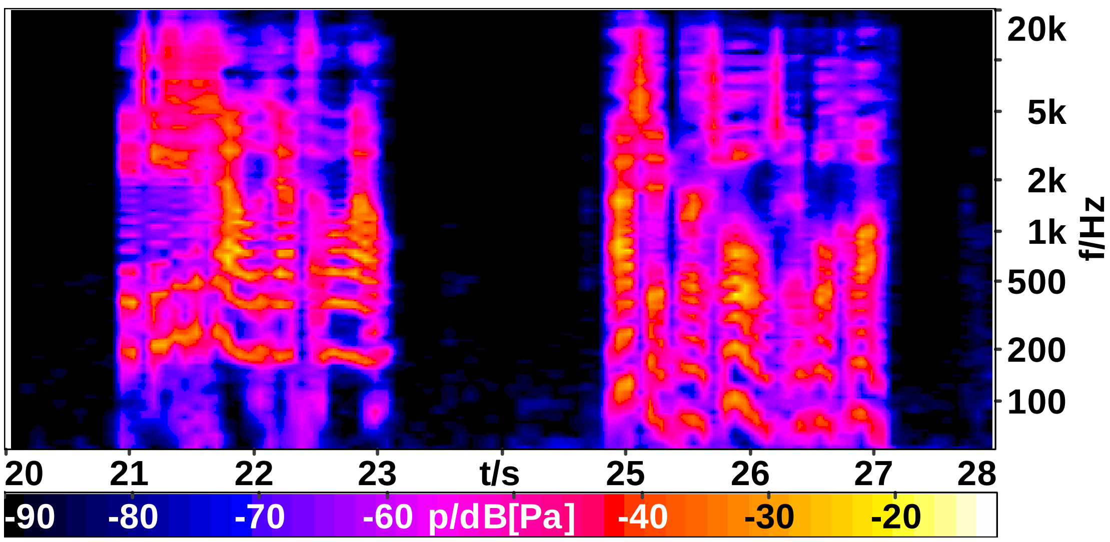


Figure E.4: Two sentences of clean speech

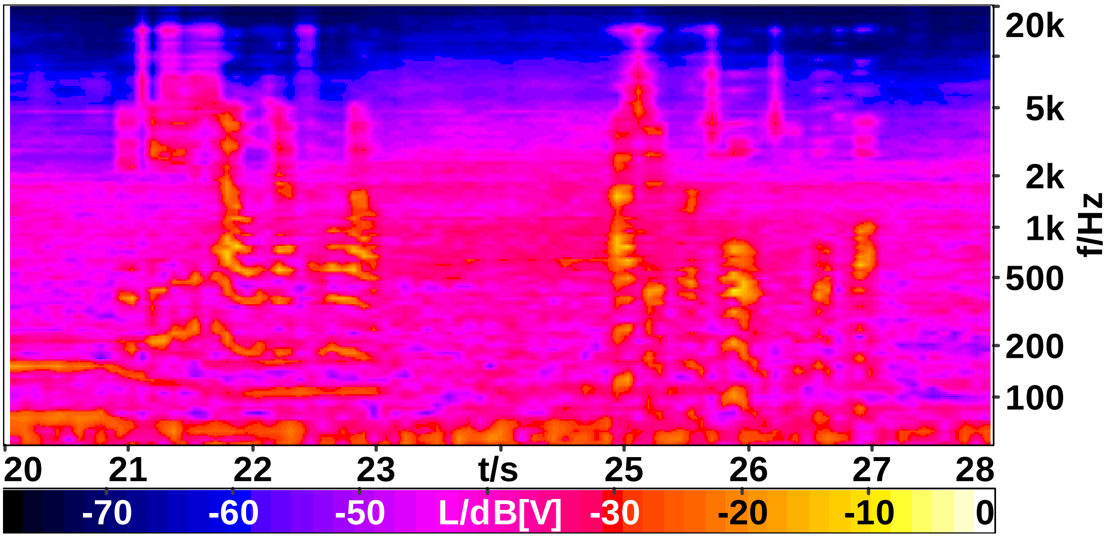


Figure E.5: Same samples with Crossroad noise

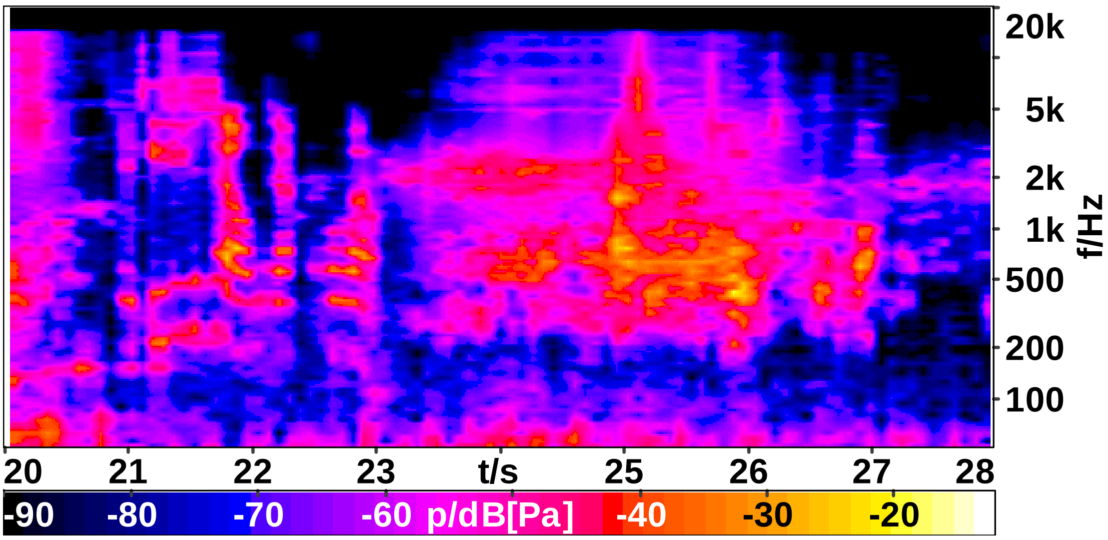


Figure E.6: Same samples with Crossroad noise and applied noise reduction

For NB and SWB transmissions, the lowest possible bitrates were chosen (4,75 kbit/s for AMR and 9,6 kbit/s for EVS-SWB). The encoding and decoding of signals was carried out with the reference implementations for AMR [i.19] and EVS [i.21] codecs.

The 48 test conditions for scenario#2 are provided in Table E.2.

Table E.2: Condition list for scenario#2

Condition ID	Use Case	Impairment	Noises	Processing	Codec
C01	HA	None	Silence	TVLP+TVHP	EVS-SWB-96
C02	HA	None	Crossroadnoise	TVLP+TVHP	EVS-SWB-96
C03	HA	None	SalesCounter	TVLP+TVHP	EVS-SWB-96
C04	HA	None	TrainStation	TVLP+TVHP	EVS-SWB-96
C05	HA	None	Silence	TVLP+TVHP	EVS-SWB-96
C06	HA	None	Pub	TVLP+TVHP	EVS-SWB-96
C07	HH	None	SalesCounter	TVLP+TVHP	EVS-SWB-96
C08	HH	None	TrainStation	TVLP+TVHP	EVS-SWB-96
C09	HH	REVERB	Silence	TVLP+TVHP	EVS-SWB-96
C10	HH	REVERB	Pub	TVLP+TVHP	EVS-SWB-96
C11	HH	REVERB	SalesCounter	TVLP+TVHP	EVS-SWB-96
C12	HH	REVERB	TrainStation	TVLP+TVHP	EVS-SWB-96
C13	HA	None	Silence	NR100	EVS-SWB-96

Condition ID	Use Case	Impairment	Noises	Processing	Codec
C14	HA	None	Crossroadnoise	NR100	EVS-SWB-96
C15	HA	None	SalesCounter	NR100	EVS-SWB-96
C16	HA	None	TrainStation	NR100	EVS-SWB-96
C17	HA	None	Silence	NR100	EVS-SWB-96
C18	HA	None	Pub	NR100	EVS-SWB-96
C19	HH	None	SalesCounter	NR100	EVS-SWB-96
C20	HH	None	TrainStation	NR100	EVS-SWB-96
C21	HH	REVERB	Silence	NR100	EVS-SWB-96
C22	HH	REVERB	Pub	NR100	EVS-SWB-96
C23	HH	REVERB	SalesCounter	NR100	EVS-SWB-96
C24	HH	REVERB	TrainStation	NR100	EVS-SWB-96
C25	HA	None	Silence	TVLP+TVHP	AMR-475
C26	HA	None	Crossroadnoise	TVLP+TVHP	AMR-475
C27	HA	None	SalesCounter	TVLP+TVHP	AMR-475
C28	HA	None	TrainStation	TVLP+TVHP	AMR-475
C29	HA	None	Silence	TVLP+TVHP	AMR-475
C30	HA	None	Pub	TVLP+TVHP	AMR-475
C31	HH	None	SalesCounter	TVLP+TVHP	AMR-475
C32	HH	None	TrainStation	TVLP+TVHP	AMR-475
C33	HH	REVERB	Silence	TVLP+TVHP	AMR-475
C34	HH	REVERB	Pub	TVLP+TVHP	AMR-475
C35	HH	REVERB	SalesCounter	TVLP+TVHP	AMR-475
C36	HH	REVERB	TrainStation	TVLP+TVHP	AMR-475
C37	HA	None	Silence	NR100	AMR-475
C38	HA	None	Crossroadnoise	NR100	AMR-475
C39	HA	None	SalesCounter	NR100	AMR-475
C40	HA	None	TrainStation	NR100	AMR-475
C41	HA	None	Silence	NR100	AMR-475
C42	HA	None	Pub	NR100	AMR-475
C43	HH	None	SalesCounter	NR100	AMR-475
C44	HH	None	TrainStation	NR100	AMR-475
C45	HH	REVERB	Silence	NR100	AMR-475
C46	HH	REVERB	Pub	NR100	AMR-475
C47	HH	REVERB	SalesCounter	NR100	AMR-475
C48	HH	REVERB	TrainStation	NR100	AMR-475
NOTE 1: Names of noise types according to clause 8 of ETSI TS 103 224 [i.4]. NOTE 2: Column "Codec" describes the applied codec. EVS-SWB-96 corresponds to encoding/decoding with EVS-SWB codec at 9,6 kbit/s, AMR-475 to encoding/decoding with AMR codec at 4,75 kbit/s.					

E.2.4 Scenario #3

While scenario #1 and #2 addressed mainly aspects of the sending side (positioning, reverberation, user behaviour, signal processing), scenario#3 investigates network-related impairments. In particular, transcoding and/or packet loss are simulated here.

In mobile applications, burst packet losses are more typical than randomly distributed losses. To simulate such a burst pattern, a simple Gilbert-Elliot-Model [i.23] and [i.24] was implemented. This widely used two-state model provides one "good" and one "bad" channel state with corresponding packet loss rates p_G and p_B. The transition probabilities r (from good to bad state) and q (from bad to good state) can be used to configure the model. Figure E.7 graphically illustrates the model.

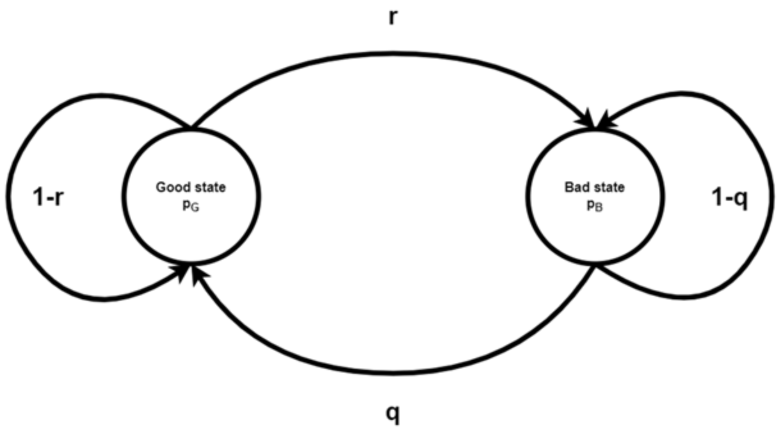


Figure E.7: Principle of Gilbert-Elliot-Model for burst simulation

For the current burst simulation, the model was configured for two impairments IMP1 and IMP2 according to Table E.3.

Table E.3: Configuration for Gilbert-Elliot-Model for burst simulation

Parameter	IMP1	IMP2
r	1 %	2 %
q	4 %	4 %
p_G	0 %	0 %
p_B	40 %	40 %
Overall packet loss rate	5,8 %	10,8 %

In addition, several transcoding steps were applied to introduce additional degradations. The encoding and decoding of signals was carried out with the reference implementations for AMR [i.19], AMR-WB [i.20] and EVS [i.21] codecs.

For the processing of recordings conducted with the measurement microphone, the noise reduction introduced for scenario#2 was applied with a factor of 50 % (NR50).

The 48 test conditions for scenario#3 are provided in Table E.4.

Table E.4: Condition list for scenario#3

Condition ID	DUT	Bandwidth	UseCase	Noises	Processing
C01	DUT1	NB	HA	Silence	AMR-475-IMP1
C02	DUT1	NB	HA	Cafeteria	AMR-475-IMP1
C03	DUT1	NB	HA	Callcenter2	AMR-475-IMP1
C04	DUT1	NB	HH	Silence	AMR-475-IMP1
C05	DUT1	NB	HH	Cafeteria	AMR-475-IMP1
C06	DUT1	NB	HH	FullSizeCar_130	AMR-475-IMP1
C07	DUT2	SB	HA	Silence	AMRWB-1265-IMP1
C08	DUT2	SB	HA	Cafeteria	AMRWB-1265-IMP1
C09	DUT2	SB	HA	Callcenter2	AMRWB-1265-IMP1
C10	DUT2	SB	HH	Silence	AMRWB-1265-IMP1
C11	DUT2	SB	HH	Cafeteria	AMRWB-1265-IMP1
C12	DUT2	SB	HH	FullSizeCar_130	AMRWB-1265-IMP1
C13	DUT2	SB	HA	Silence	EVS-SWB-96-IMP1
C14	DUT2	SB	HA	Cafeteria	EVS-SWB-96-IMP1
C15	DUT2	SB	HA	Callcenter2	EVS-SWB-96-IMP1
C16	DUT2	SB	HH	Silence	EVS-SWB-96-IMP1
C17	DUT2	SB	HH	Cafeteria	EVS-SWB-96-IMP1
C18	DUT2	SB	HH	FullSizeCar_130	EVS-SWB-96-IMP1
C19	DUT2	SB	HA	Silence	AMRWB-660;EVS-SWB-96
C20	DUT2	SB	HA	Cafeteria	AMRWB-660;EVS-SWB-96
C21	DUT2	SB	HA	Callcenter2	AMRWB-660;EVS-SWB-96

Condition ID	DUT	Bandwidth	UseCase	Noises	Processing
C22	DUT2	SB	HH	Silence	AMRWB-660;EVS-SWB-96
C23	DUT2	SB	HH	Cafeteria	AMRWB-660;EVS-SWB-96
C24	DUT2	SB	HH	FullSizeCar_130	AMRWB-660;EVS-SWB-96
C25	MeasurementMic		HA	Silence	NR50;EVS-SWB-96;AMRWB-660;AMR-475
C26	MeasurementMic		HA	Cafeteria	NR50;EVS-SWB-96;AMRWB-660;AMR-475
C27	MeasurementMic		HA	Callcenter2	NR50;EVS-SWB-96;AMRWB-660;AMR-475
C28	MeasurementMic		HH	Silence	NR50;EVS-SWB-96;AMRWB-660;AMR-475
C29	MeasurementMic		HH	Cafeteria	NR50;EVS-SWB-96;AMRWB-660;AMR-475
C30	MeasurementMic		HH	FullSizeCar_130	NR50;EVS-SWB-96;AMRWB-660;AMR-475
C31	MeasurementMic		HA	Silence	NR50;AMR-122-IMP2
C32	MeasurementMic		HA	Cafeteria	NR50;AMR-122-IMP2
C33	MeasurementMic		HA	Callcenter2	NR50;AMR-122-IMP2
C34	MeasurementMic		HH	Silence	NR50;AMR-122-IMP2
C35	MeasurementMic		HH	Cafeteria	NR50;AMR-122-IMP2
C36	MeasurementMic		HH	FullSizeCar_130	NR50;AMR-122-IMP2
C37	MeasurementMic		HA	Silence	NR50;AMRWB-2385-IMP2
C38	MeasurementMic		HA	Cafeteria	NR50;AMRWB-2385-IMP2
C39	MeasurementMic		HA	Callcenter2	NR50;AMRWB-2385-IMP2
C40	MeasurementMic		HH	Silence	NR50;AMRWB-2385-IMP2
C41	MeasurementMic		HH	Cafeteria	NR50;AMRWB-2385-IMP2
C42	MeasurementMic		HH	FullSizeCar_130	NR50;AMRWB-2385-IMP2
C43	MeasurementMic		HA	Silence	NR50;EVS-SWB-244-IMP2
C44	MeasurementMic		HA	Cafeteria	NR50;EVS-SWB-244-IMP2
C45	MeasurementMic		HA	Callcenter2	NR50;EVS-SWB-244-IMP2
C46	MeasurementMic		HH	Silence	NR50;EVS-SWB-244-IMP2
C47	MeasurementMic		HH	Cafeteria	NR50;EVS-SWB-244-IMP2
C48	MeasurementMic		HH	FullSizeCar_130	NR50;EVS-SWB-244-IMP2

NOTE 1: Names of noise types according to clause 8 of ETSI TS 103 224 [i.4].
NOTE 2: The column "Processing" contains one or more processing steps, which are separated by a semicolon (";").
NOTE 3: EVS-SWB-X corresponds to encoding/decoding with EVS-SWB codec at X/10 kbit/s.
NOTE 4: AMRWB-X corresponds to encoding/decoding with AMR-WB codec at X/100 kbit/s.
NOTE 5: AMR-X corresponds to encoding/decoding with AMR codec at X/100 kbit/s.
NOTE 6: Impairments are always applied in the context of an encoding/decoding processing step. For packet loss concealment, the behaviour of the reference implementations in [i.19], [i.20] and [i.21] were not modified.

E.2.5 Distribution of Languages vs Training/Validation

As indicated in the test plan, in overall six databases are provided:

- 3x Training Databases (48 conditions, 16 samples per condition, 12 votes per sample, 192 votes per condition).
- 3x Validation Databases (48 conditions, 8 samples per condition, 16 votes per sample, 128 votes per condition).

E.3 Subjective tests

E.3.1 Overview

The subjective Laboratory performed 6 different experiments (i.e. two experiments in English, two experiments in German and two experiments in Mandarin). The experiment name, methodology (as defined in Recommendation ITU-T P.800 [1]) and language for each experiment is listed in Table E.5. The terms "Training" and "Validation" refer to the objective model development phase as defined in clause E.1.

Table E.5: Allocation of databases and languages

Exp.	Type	Language
1	Training DB01	ENG
2	Validation DB03	ENG
3	Training DB03	GER
4	Validation DB02	GER
5	Training DB02	MAN
6	Validation DB01	MAN

E.3.2 Speech material

The subjective Laboratory used processed speech material delivered by the Recording Laboratory in 48-kHz sampled with 24-bit resolution stereo uncompressed wav files. The processed speech material for all tests conformed to restrictions indicated in "Practical procedures for subjective testing" [8]. No sample name blinding or anonymization has been applied as it was principally not needed.

E.3.3 Listening Environment

The tests were performed in an acoustically treated critical listening room in the Subjective Laboratory that conforms to the requirements of [1] in full. Its background noise during the tests was below than 30 dB_{SPL}(A) with no peaks in audible acoustic frequency range and its reverberation time is 185 ms.

Listening equipment conformed to specified requirements [1] and [8] in full. Diffuse-field equalized Sennheiser headphones HD-650 have been used for all experiments. All used headphones have been calibrated and verified before and after performed experiments. A professional digital voting device has been used to collect the votes.

E.3.4 Test execution and sample presentation

Test duration never exceeded 1,5 hours per listening panel. Test duration comprised of up to 50 % of actual listening time and test overhead including administration, initial briefing, preliminaries, and breaks.

To avoid amplitude clipping for samples with significant peaks in time-domain, the playout loudness calibration of -36 dBov corresponding to 73 dB_{SPL} was used in all experiments.

For each speech sample, both LE and MOS-LQS following ACR methodology have been assessed. The order of LE and OS-LQS questions has been balanced, i.e. the subjects were asked for their assessment of LE and consequently for OS-LQS in half of the listening sessions, and for OS-LQS and then LE in the other half. Within each session, identical question order has been kept for all samples.

E.3.5 Instructions to subjects

The instructions to subjects were available in written during the entire test sessions. During the training session of each test, the instructions were verbally explained and if needed briefly discussed by a dedicated expert person that was able to answer questions from the subjects in their native language.

The instructions used for testing in English, German and Mandarin were the same as in the experiments described in annex D. They are provided in Figure D.4, Figure D.5 and Figure D.6.

E.3.6 Test subjects

For each listening test, only native speakers of the tested language were used. Only normal-hearing subjects have been used for the experiments. Their hearing normality has been verified by subject self-assessment questionnaire during the subject hiring phase and verified prior the session by expert test. Experiment required 32 (training) or 48 (validation) listeners (4 or 6 panels of 8 listeners each). The age and gender information for the set of subjects used in each listening test in each tested language is provided in Table E.6.

Table E.6: Age and gender information of subjects

Language	#females/#males	Average age (years)	Age StdDev (years)
English	1,00	32,9	10,40
German	1,00	33,8	10,37
Chinese	1,00	30,1	8,65

E.3.7 Raw data delivery

All raw voting data have been delivered to the laboratory in charge of the objective model training. They contain averages and standard deviations of LE as well as MOS-LQ for all conditions within each test.

E.4 Subjective test results of the training databases

The raw results for the training databases, i.e. DB01 in English, DB02 in Mandarin and DB03 in German can be downloaded here:
https://docbox.etsi.org/STQ/Open/TS%20103%20558/Annex%20E.

E.5 Prediction Results for Validation Databases

E.5.1 Overview

The prediction results for the databases described in clause E.2.5 are provided in the following clauses. For each database one scatter plot is shown. In all applications, the usage of the noise-only reference (see clause 6.1) is not possible/applicable.

The following performance metrics are provided for each validation:

- rmse*: root-mean-square error per condition after 3rd order mapping according to Recommendation ITU-T P.1401 [i.3], taking the uncertainty of auditory data into account.

- maxabs*: absolute maximum error per condition after 3rd order mapping, taking the uncertainty of auditory data into account.

The raw results for the validation databases, i.e. for DB01 (in Mandarin), DB02 (in German) and DB03 (in English) can be downloaded here:
https://docbox.etsi.org/STQ/Open/TS%20103%20558/Annex%20E.

NOTE: The prediction results presented in the following clauses are presented per condition (average across all samples) for each validation database. In each comparison between auditory and predicted listening effort, a 3rd order mapping function is provided. It should be noted that these mapping functions are only valid for the corresponding test context (combination of application, language, distribution of conditions, etc.) and are not generally applicable. There may exist language- or application-specific mapping functions; however, at this time the underlying databases do not provide sufficient information on this topic.

E.5.2 Scenario #1

The results for scenario #1 in Mandarin language as described in clause E.2.2 are shown in Figure E.8.

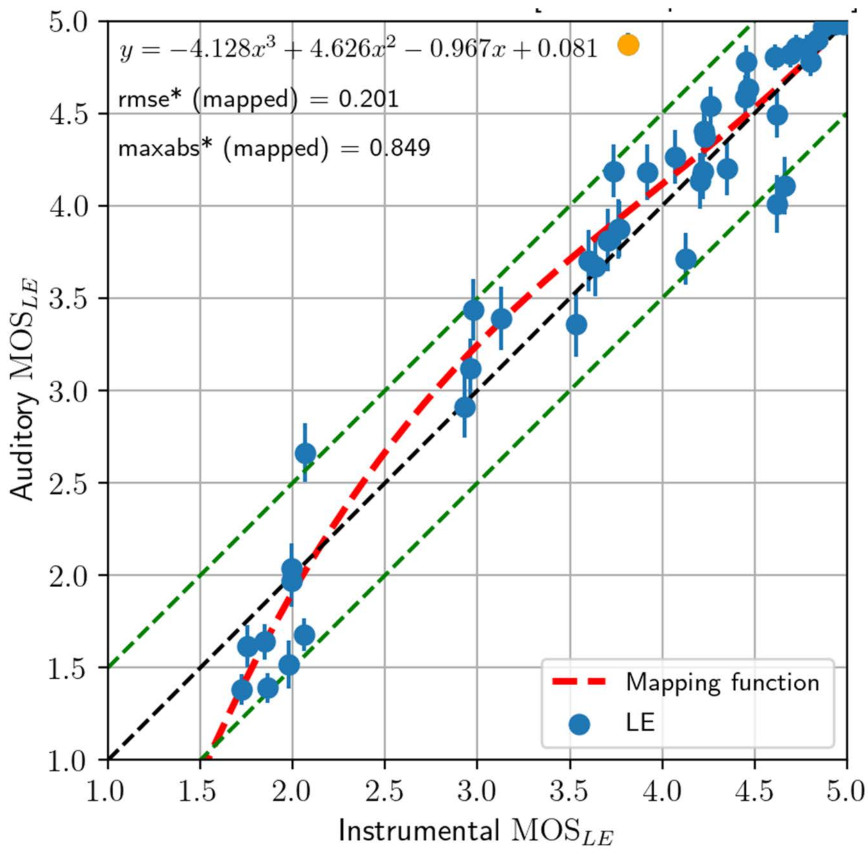


Figure E.8: Predicted validation results for scenario #1 in Mandarin

E.5.3 Scenario #2

The results for scenario #2 in German language as described in clause E.2.3 are shown in Figure E.9.

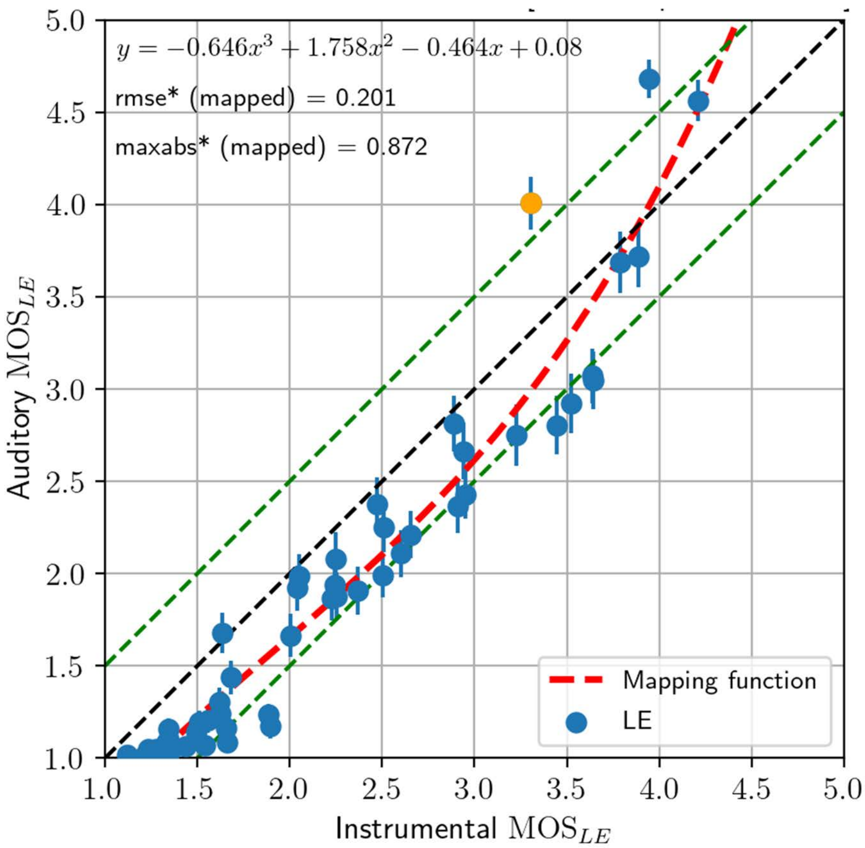


Figure E.9: Predicted validation results for scenario #2 in German

E.5.4 Scenario #3

The results for scenario #3 in English language as described in clause E.2.4 are shown in Figure E.10.

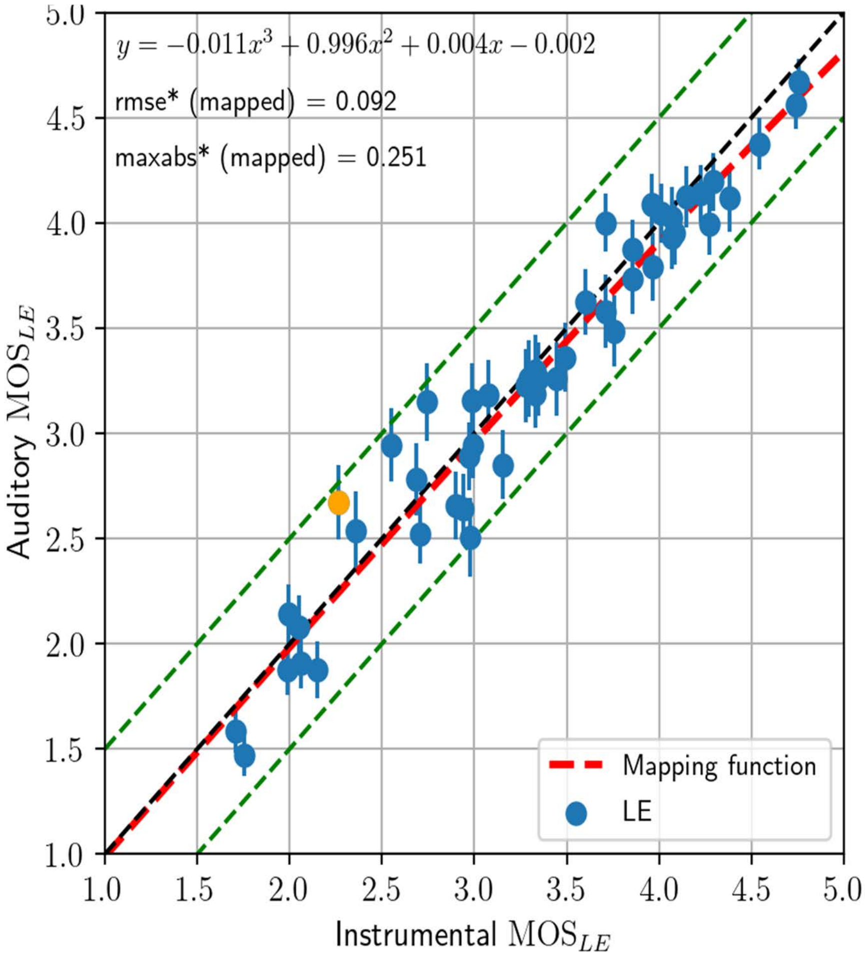


Figure E.10: Predicted validation results for scenario #3 in English

History

Document history		
V1.1.1	November 2019	Publication
V1.2.1	August 2020	Publication
V1.3.1	July 2021	Publication