ETSI TS 103 802 V1.2.1 (2025-03)

Speech and multimedia Transmission Quality (STQ);
Objective test method for
the evaluation of echo impairments

Reference
RTS/STQ-319

Keywords
conversation, echo, impairment, instrumental,
listening quality, model, objective, prediction, test

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 4
Foreword 4
Modal verbs terminology 4
Introduction 4
1 Scope 6
2 References 6
2.1 Normative references 6
2.2 Informative references 7
3 Definition of terms, symbols and abbreviations 7
3.1 Terms 7
3.2 Symbols 8
3.3 Abbreviations 8
4 Measurement procedure 9
4.1 Overview 9
4.2 Positioning of devices 9
4.3 Test signals 10
4.4 Echo signal recording 10
5 Instrumental Prediction Model 11
5.1 Overview 11
5.2 Signal Acquisition 12
5.2.1 Overview 12
5.2.2 End-to-end measurement setup 12
5.2.3 Measurement setup with POI 13
5.2.4 Sample duration 13
5.3 Conventions on delay 13
5.3.1 Overview 13
5.3.2 Artificial network delay 14
5.3.3 External echo delay 14
5.4 Calculation of echo delay and linearity measure 15
5.5 Classification of time ranges 17
5.6 Hearing model transformation 17
5.7 Self-masking of echo 18
5.8 Aggregate over time and frequency 21
5.9 Regression 22
5.10 Indication for absence of an echo signal 22
5.11 Compensation for idle noise 22
Annex A (normative): Self-masking weights determined by HATS setup 24
Annex B (informative): Subjective experiments for training and validation 25
History 28

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 IPR online database.

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™**, **LTE™** and **5G™** logo 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).

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.

Introduction

In speech communication devices of all kinds, echo artefacts and double talk impairments can occur. These might dramatically degrade a conversation between users, i.e. the quality of experience in general. With an increasing usage of hands-free terminals (e.g. motor vehicle, handheld or desktop devices) and new types of devices supporting voice services (e.g. smart home devices or wearables), the cancellation of echo and providing duplex communication at the same time is still a challenging task for signal processing components.

The objective assessment of degradations caused by echo and/or poor double talk performance is already covered in several specifications, but mainly based on simple analyses in level or spectrum. The impact on the conversation as perceived by the user is typically rarely investigated.

The auditory evaluation of a conversation between two human test subjects in a laboratory may be quite cumbersome. For that purpose, a comprehensive description for a Third-Party Listening Test (TPLT) design was specified in ETSI TS 103 801 [1], which provides a fair trade-off between a simulated conversation as close as possible to human perception and the usage of recorded signals for further analysis.

The present document introduces a prediction model, which is able to assess the perceived degradation in communication quality due to echo impairments and utilizes the same pre-recorded signals as in the aforementioned TPLT design. The instrumentally determined quality degradation is reported in terms of Mean Opinion Score (MOS).

1 Scope

The present document provides an instrumental model for the signal-based prediction of echo artefacts that may occur in telecommunication devices of all kinds. This prediction model assesses audible degradations of a virtual end-to-end communication scenario, including partial masking via sidetone as described in ETSI TS 103 801 [1]. The predicted Mean Opinion Score (MOS) corresponds to the Degradation Category Rating (DCR) according to Recommendation ITU-T P.800 [i.1].

The prediction model described in the present document is applicable to handset, headset or hands-free terminals and is not limited to a certain audio bandwidth, i.e. applications from narrowband to fullband are supported.

The following use cases are out of scope for the prediction model described in the present document:

- Ambient noise at the device-side (near-end).
- Ambient noise at the reference-side (far-end).
- Talker at near-end (double talk).
- Network impairments other than constant delay.

The prediction model assumes monaural listening at the reference-side under silent condition and high-quality handsets or headsets with low leakage. Assessment of binaurally perceived echo artefacts at the reference-side is out of scope of the present document.

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 which are not found to be publicly available in the expected location might be found in the ETSI docbox.

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] ETSI TS 103 801: "Speech and multimedia Transmission Quality (STQ); Subjective test methodologies for the evaluation of echo control systems".

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

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

[4] Recommendation ITU-T P.64: "Determination of sensitivity/frequency characteristics of local telephone systems".

[5] Recommendation ITU-T P.380: "Electro-acoustic measurements on headsets".

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

[7] ETSI ES 202 738: "Transmission requirements for narrowband VoIP loudspeaking and handsfree terminals from a QoS perspective as perceived by the user".

[8] ETSI ES 202 740: "Transmission requirements for wideband VoIP loudspeaking and handsfree terminals from a QoS perspective as perceived by the user".

[9] Recommendation ITU-T P.1100: "Narrowband hands-free communication in motor vehicles".

[10] Recommendation ITU-T P.1110: "Wideband hands-free communication in motor vehicles".

[11] Recommendation ITU-T P.501: "Test signals for use in telephony and other speech-based applications".

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

[13] ECMA 418-2: "Psychoacoustic metrics for ITT equipment — Part 2 (models based on human perception)", 2nd edition, December 2022.

[14] ISO 532-1:2017: "Acoustics — Methods for calculating loudness — Part 1: Zwicker method".

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

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.800: "Methods for subjective determination of transmission quality".

[i.2] Recommendation ITU-T P.10/G.100: "Vocabulary for performance, quality of service and quality of experience".

[i.3] M. Lepage, F. Kettler, J. Reimes: "Scalable Perceptual Based Echo Assessment Method for Aurally Adequate Evaluation of Residual Single Talk Echoes", 13th International Workshop on Echo Cancelation and Noise Control (IWAENC) 2012, Aachen, Germany.

[i.4] J. Reimes, H.W. Gierlich, F. Kettler, S. Poschen, M. Lepage: "The Relative Approach Algorithm and its Applications in New Perceptual Models for Noisy Speech and Echo Performance", Acta Acustica united with Acustica 97(2):325-341, March 2011.

[i.5] S. Bleiholder, F. Kettler: "Auditory Assessment of Super-Wideband Echo Disturbances", DAGA 2017, Kiel, Germany.

[i.6] J. Holub, O. Tomiska: "Non-monotonicity in Perceived Quality of Delayed Talker Echo", International conference on Measurement of speech, audio and video quality in networks, January 2007.

[i.7] Recommendation ITU-T P.1401: "Methods for subjective determination of transmission quality".

[i.8] 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".

3 Definition of terms, symbols and abbreviations

3.1 Terms

For the purposes of the present document, the terms given in Recommendation ITU-T P.10/G.100 [i.2] and the following apply:

device-side: end-point of a telecommunication connection, which is dedicated to and operated by a device under test

echo impairment: artefact generated by the signal processing in sending direction of the device-side (e.g. due to linear/non-linear coupling of signal components from receiving to sending direction of the device under test)

NOTE: It is triggered in talking phases of the reference receive filter.

reference receive filter: digital filter addressing the transmission path from POI to diffuse-field equalized DRP in receive direction of the reference-side

NOTE: Diffuse-field correction is specified in Table 3 of Recommendation ITU-T P.58 [3].

reference send filter: digital filter addressing the transmission path from MRP to POI in sending direction of the reference-side

reference-side: end-point of a telecommunication connection, which is operated by a reference device or gateway in order to capture stimuli for a TPLT

NOTE: This may be realized either electrically or acoustically with a HATS.

signal under test: signal transmitted from device-side to reference receive filter

NOTE: This may contain echo artefacts and/or double talk impairments caused by signal processing of DUT.

source signal: signal originated from reference receive filter and transmitted to device-side

NOTE: This may also be inserted electrically at POI to the DUT.

3.2 Symbols

For the purposes of the present document, the symbols given in Recommendation ITU-T P.10/G.100 [i.2] and the following apply:

ΔT	Duration of delay introduced in the transmission path
dB	Decibel
dB_{SPL}	Sound Pressure Level in dB, referenced to 20 µPa
dB_{Pa}	Sound Pressure Level in dB, referenced to 1 Pa
dB_{V}	Voltage in dB, referenced to 1 Volt
$dB_{Pa/V}$	Sensitivity in receiving direction (Pascal per Volt), expressed in dB
$dB_{V/Pa}$	Sensitivity in sending direction (Volt per Pascal), expressed in dB
e(k)	Echo signal recorded electrically at POI
$e_{RCV}(k)$	Acoustic representation of e(k) at the (virtual) listener
maxabs*	Maximum Error per subset considering the uncertainty of auditory data
Pa	Pascal (pressure)
RCV-Ref	Receive reference filter
rmse*	Root-Mean-Square Error considering the uncertainty of auditory data
SND-Ref	Send reference filter
x(k)	Speech source file at MRP of talker/listener at reference-side
$x_{SND}(k)$	Downlink signal sent to Device-side, based on x(k)
$x_{ST}(k)$	Sidetone signal, based on x(k)

3.3 Abbreviations

For the purposes of the present document, the abbreviations given in Recommendation ITU-T P.10/G.100 [i.2] and the following apply:

ASL	Active Speech Level
DF	Diffuse-Field
DUT	Device Under Test
FB	FullBand

NOTE: Audio bandwidth from 20 Hz to 20 kHz.

FF	Free-Field
MOS	Mean Opinion Score

MOS-TQO$_f$ MOS for Talking Quality Objective (in fullband context)
NB NarrowBand

NOTE: Audio bandwidth from 300 Hz to 3 400 kHz.

POI Point Of Interconnection
RCV Receiving Direction
SND Sending Direction
SPL Sound Pressure Level
SWB Super-WideBand

NOTE: Audio bandwidth from 50 Hz to 14 kHz.

TPLT Third-Party Listening Test
WB WideBand

NOTE: Audio bandwidth from 100 Hz to 7 kHz.

4 Measurement procedure

4.1 Overview

Figure 1 illustrates the virtual communication scenario, which is the basis for the TPLT design as described in ETSI TS 103 801 [1]. The DUT is located on the left (device-side) and the virtual talker/listener on the right (reference-side). For the application of the prediction model, measurement signals have to be inserted and captured at the POI in a similar way as for the auditory evaluation.

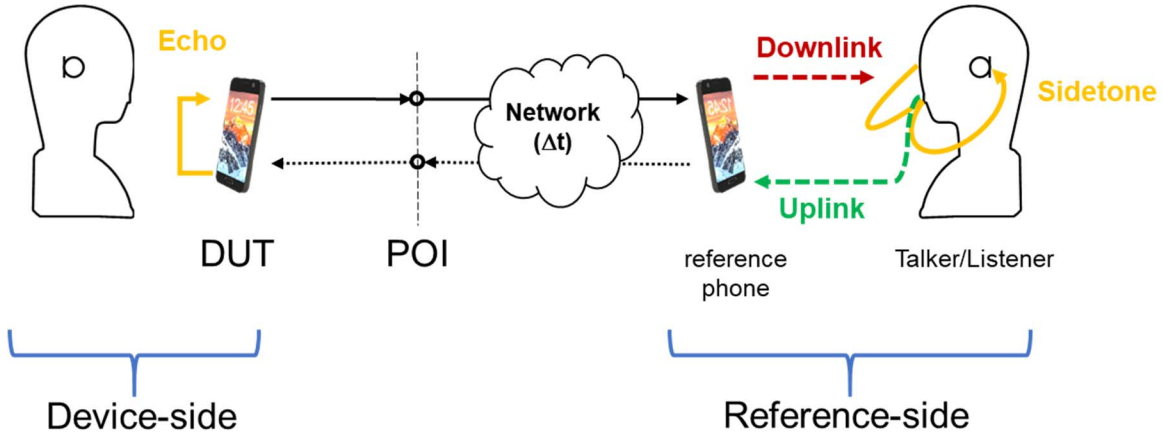

Figure 1: Communication scenario for TPLT according to ETSI TS 103 801 [1]

4.2 Positioning of devices

If positioning instructions are not provided by an underlying test specification, the terminal shall be positioned in a typical manner and best possible reproducibility:

- For handset terminals, a HATS according to Recommendation P.58 [3] equipped with ear simulators of type 3.3, 4.3, or 4.4 according to Recommendation ITU-T P.57 [2] shall be used. If not specified otherwise, the device shall be mounted according to Annexes E and F of Recommendation ITU-T P.64 [4].
- Considering the guidelines in Recommendation ITU-T P.380 [5], headset terminals shall be mounted on a HATS according to Recommendation ITU-T P.58 [3] equipped with ear simulators of type 3.3, 4.3, or 4.4 according to Recommendation ITU-T P.57 [2]. Due to the uncertainty of headset positioning, it is recommended to repeat measurements at least three times. Between each repetition, the device should be removed and mounted again.

- Positioning for desktop and handheld hands-free terminals are provided in ETSI ES 202 738 [7] or ETSI ES 202 740 [8]. Test setups for vehicle hands-free terminals are specified in Recommendation ITU-T P.1100 [9] and/or Recommendation ITU-T P.1110 [10].

Since the echo path may strongly depend on the location of and distance between device microphone(s) and loudspeaker(s), reproducibility of the test setup is crucial and shall always be reported complementary to the prediction results of the model.

4.3 Test signals

The source speech material to be used for the measurements shall comply with clause 5.3 of ETSI TS 103 801 [1]. The single and double talk sequences described in clauses 7.3.5 (British English) and 7.4.1.3 (Mandarin) of Recommendation ITU-T P.501 [11] comply in general with these requirements, but further cropping into shorter samples for analysis may be required.

The nominal active speech level according to Recommendation ITU-T P.56 [6] of the talker at the reference-side shall be calibrated (or correspond) to -4,7 dB_{Pa}. Different source levels use at the reference side (i.e. simulating softer or louder talkers) are out of scope. For the measurement setup as well as for the prediction model, this source signal is subsequently referred to as x(k) below.

4.4 Echo signal recording

For the measurement, a downlink signal to the DUT has to be inserted at the POI in the same way as specified for the TPLT design of ETSI TS 103 801 [1]. This test signal shall reflect the transmission from the MRP of the talker at the reference-side to the POI and shall be generated according to one of the three following options:

1) Playback of the source speech signal x(k) over the artificial mouth, acoustic insertion into a reference device (e.g. a known phone providing high quality).

2) Acoustic insertion is simulated by filtering the source speech signal x(k) with a sending frequency response of a known reference device.

3) Acoustic insertion is simulated by filtering the source speech signal x(k) with a default frequency response (flat transfer function in the bandwidth-specific frequency range), as specified in annex B of ETSI TS 103 801 [1] for NB, WB, SWB and FB applications.

Figure 2 is a modified representation of Figure 1 and illustrates the generation of the downlink test signal, which is subsequently referred to as $x_{SND}(k)$ below. See clause 5.4.3 of ETSI TS 103 801 [1] for more details on the three options. The send reference filter is denoted as SND-Ref. If not specified otherwise, the third method shall be used by default for measurements.

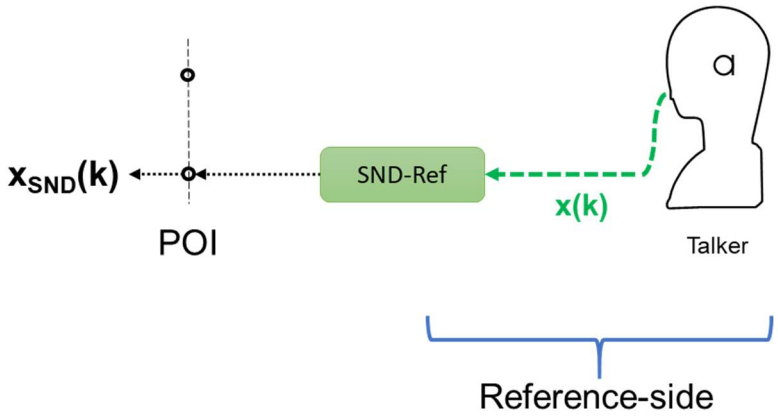

Figure 2: Generation of test signal for downlink of DUT

The recording of the echo signal is shown in Figure 3, which is again a modified representation of Figure 1. The signal $x_{SND}(k)$ is applied in receive direction of the DUT, where it is played back via device loudspeakers and possibly causes echo impairments. The echo signal e(k) is captured at the POI in send direction of the DUT. Apart from the delays introduced by the measurement equipment, neither send nor receive delays (roundtrip delay) of the DUT shall be compensated.

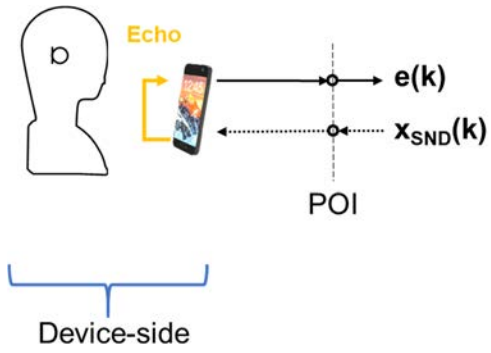

Figure 3: Recording of echo signal

5 Instrumental Prediction Model

5.1 Overview

Figure 4 provides an overview on the different signals used in the measurement (see clause 4) and in the prediction model. The virtual listener at the reference-side perceives a superposition of two signal components:

- The binaural self-hearing of the listener's own voice, denoted as $x_{ST}(k)$. It is determined by filtering the source speech signal x(k) with a suitable sidetone frequency response.
- The acoustic representation of the electrical echo signal e(k), which is presented via a reference device or filter (RCV-Ref) to the listener (including diffuse-field correction), denoted as $e_{RCV}(k)$.

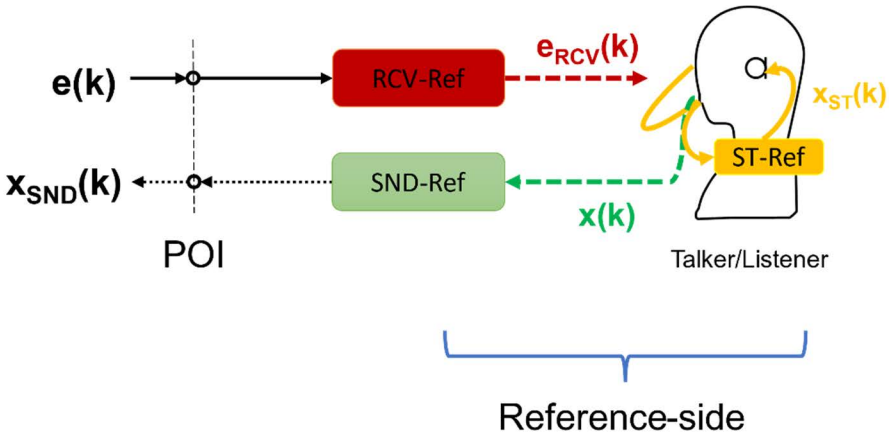

Figure 4: Signals involved in the prediction model

The acoustical sidetone signal $x_{ST}(k)$ represents a self-masking signal that decreases the perception of the acoustical echo signal $e_{RCV}(k)$. For realistic evaluation of echo artefacts, it is crucial that levels of both signals are calibrated correctly. The superposition $x_{ST}(k) + e_{RCV}(k)$ of both signals at the virtual listener should be very close to the situation of a talker perceiving his/her own voice as echo from the far-end side.

Figure 5 illustrates the basic principle of the prediction model. The echo signal e(k) captured at the POI in send direction of the DUT and the source speech signal x(k) are the main input signals. The filtering (receive frequency response for echo, sidetone filter for self-hearing) are part of the pre-processing stage of the prediction model.

Both pre-processed signals are then transformed into the time-frequency domain using the hearing model of Sottek for time-varying loudness. The resulting representations $E_{RCV}(l,m)$ and $X_{ST}(l,m)$ are then weighted and subtracted in order to simulate the perceptual masking of the listener's own voice. The zero-clipped masked echo R(l,m) is then aggregated over time and frequency and passed into a regression model to finally obtain the MOS-TQO_f.

The predicted MOS-TQO_f shall always be limited to the range between 1,0 (worst degradation) and 4,9 (no degradation). The maximum value of 4,9 is based on observations of several listening test results, in which the theoretical maximum of 5,0 typically is never achieved.

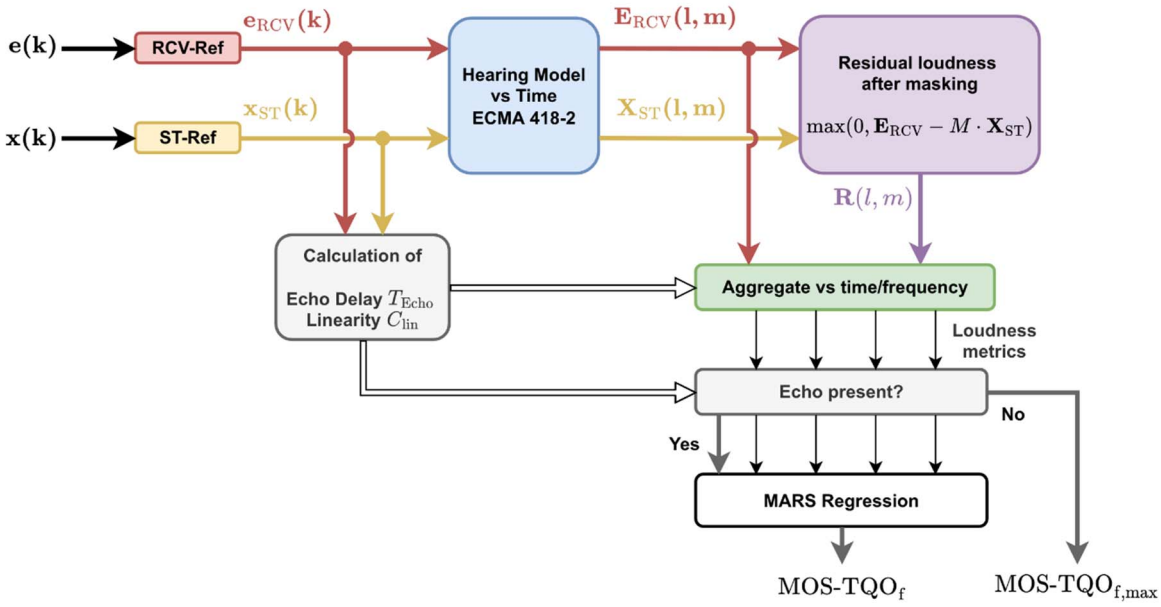

Figure 5: Flow chart of prediction model

5.2 Signal Acquisition

5.2.1 Overview

The prediction model utilizes these two acoustic signal components $x_{ST}(k)$ and $e_{RCV}(k)$ for the prediction of quality degradation. The two acoustic signal components $x_{ST}(k)$ and $e_{RCV}(k)$ are only available in a complete end-to-end scenario, but not in a typical one-way measurement setup (from/to POI). Here only the electrical echo signal e(k) and the source speech sequence x(k) are available. An initial pre-processing stage transforms the available input signals e(k) and x(k) to the acoustical representations $x_{ST}(k)$ and $e_{RCV}(k)$, if necessary.

The following clauses describe how the model input signals $x_{ST}(k)$ and $e_{RCV}(k)$ are determined for these two types of measurement setups. Both signals shall always be provided with a sampling rate of 48 kHz. Any possible re-sampling is out of scope of the present document.

5.2.2 End-to-end measurement setup

In an end-to-end setup, it is possible to measure the signals $x_{ST}(k)$ and $e_{RCV}(k)$ directly, without access to the POI. However, the signals have to be separately available, and the following measurement steps are required:

1) $x_{ST}(k)$ is directly be measured with a mouth-to-ear measurement at the DRP, including diffuse-field correction according to Recommendation ITU-T P.58 [3]. The absence of any echo signal shall be ensured by e.g. muting the send or receive direction of the device at the reference side.

2) Reference device and DUT are operating normally, and the echo signal $e_{RCV+ST}(k)$ is recorded at the DRP, including diffuse-field correction according to Recommendation ITU-T P.58 [3]. This signal includes also the sidetone path signal $x_{ST}(k)$.

3) The perceived echo signal $e_{RCV}(k)$ is determined via time-domain signal subtraction: $e_{RCV}(k) = e_{RCV+ST}(k) - x_{ST}(k)$.

4) In case $e_{RCV+ST}(k)$ is recorded with a binaural or diotic reference device (e.g. headset or hands-free device), only the dominant ear signal shall be considered. The resulting perceived echo signal $e_{RCV}(k)$ shall be amplified by 6 dB to address the typical difference in level between monaural and diotic listening.

5.2.3 Measurement setup with POI

In case POI access to the DUT is available, the usage of a real device at the reference side is not necessary and can be completely simulated in the same way as described in clause 5.4.5 of ETSI TS 103 801 [1]:

- ST-Ref: The default sidetone filter according to clause 5.4.3.2 of ETSI TS 103 801 [1] shall be applied to the source speech signal x(k) to obtain $x_{ST}(k)$.
- RCV-Ref: The default calibration (for assumed monaural listening) according to clause 5.4.5 of ETSI TS 103 801 [1] shall be applied to the electrical echo signal e(k) to obtain $e_{RCV}(k)$.

NOTE: In case access to the POI is available in an end-to-end setup, signals $x_{ST}(k)$ and $e_{RCV}(k)$ should preferably be used instead of using generic filters.

5.2.4 Sample duration

In accordance with clause 5.3 of ETSI TS 103 801 [1], the duration of the input signals shall not exceed 12,0 s and shall not contain more than four single sentences. It is recommended that analysis samples contain one or two single sentences and a duration of approximately 8 s to 10 s. Leading and trailing silence periods per samples as described in clause 5.3 of ETSI TS 103 801 [1] should be considered as well.

Longer sequences with more sentences should be first cropped into multiple shorter samples, which are then individually analysed by the prediction model. By default, the aggregated MOS-TQO_f for such sequences can be reported as the average across all samples. However, since echo artefacts are in general not desirable at all, an average across multiple samples might lead to too optimistic results. Thus, aggregation methods that takes one or more bad per-sample results (e.g. maximum or percentile) into account may be considered here as well.

5.3 Conventions on delay

5.3.1 Overview

The overall delay T_{Echo} between talking and listening at the reference-side is one of the most important quality indicators of residual echo and/or artefacts. In general, higher delays cause an increased degradation regarding perception of echo. Figure 6 provides an overview of the different delays that contribute to the overall delay.

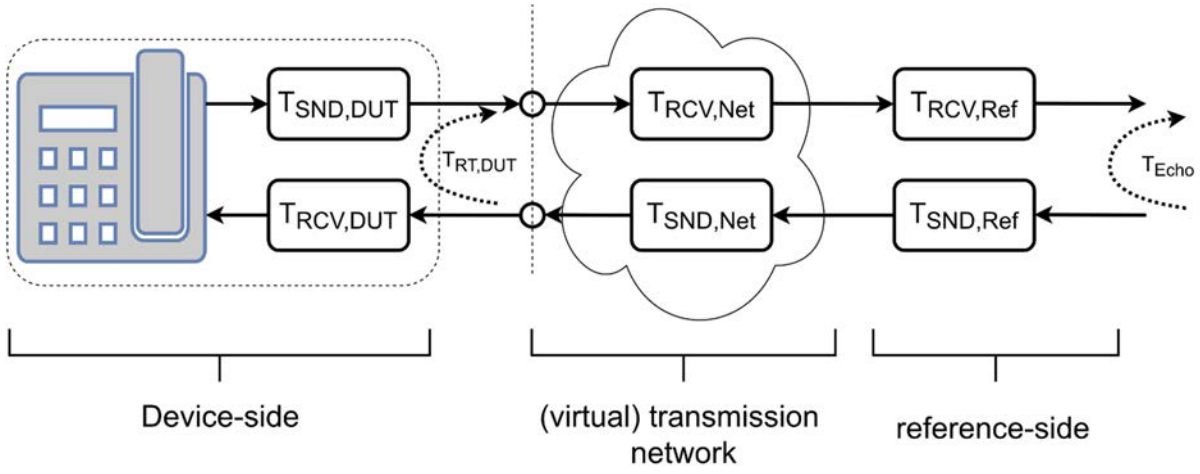

Figure 6: Contributing delays in an end-to-end communication scenario

The most relevant delay for perception of echo is the difference in time between talking at the reference-side and listening to the transmitted and received echo caused by the DUT is given as T_{Echo} in equation (1) and consists of different components:

$$T_{Echo} = T_{SND,Ref} + T_{SND,Net} + T_{RCV,DUT} + T_{SND,DUT} + T_{RCV,Net} + T_{RCV,Ref} \quad (1)$$

When evaluating a DUT with access to the POI, the roundtrip delay $T_{RT,DUT}$ as given by equation (2) has the most impact on the overall echo delay T_{Echo}:

$$T_{RT,DUT} = T_{RCV,DUT} + T_{SND,DUT} \quad (2)$$

NOTE 1: The delay that originates from the acoustic path (loudspeaker output to microphone input) at the DUT is neglected here for sake of simplicity, as it is typically much lower than 1 ms. In general, it is assumed that this delay is part of the DUT and is included in $T_{RT,DUT}$.

In case the reference-side devices are simulated by linear filters, the delays $T_{SND,Ref}$ and $T_{RCV,Ref}$ should not contribute much to the overall delay, in order to ensure that only the DUT roundtrip delay $T_{RT,DUT}$ is considered in the prediction model. It shall be ensured that $T_{SND,Ref}$ and $T_{RCV,Ref}$ introduce less than 1 ms of additional delay. Any additional delay that is introduced by the test equipment in send and receive direction shall also be compensated, as it will otherwise contribute to the overall echo delay T_{Echo}.

NOTE 2: In case of an end-to-end measurement setup, a real device with certain delays in send and receive is used. If these delays are known, it is recommended to compensate the echo signal $e_{RCV}(k)$ by these values, in order to only assess the impact of the device-side. If these delays cannot be determined, the prediction can only be carried out on the recorded echo signal that includes these reference-side delays.

5.3.2 Artificial network delay

The transmission network may also introduce additional delays $T_{SND,Net}$ and $T_{RCV,Net}$ to the echo signal $e_{RCV}(k)$. In a measurement setup with network simulators, these delays are usually known and compensated, in order to only evaluate the perceptual impact of the DUT roundtrip delay $T_{RT,DUT}$. Thus, $T_{SND,Net}$ and $T_{RCV,Net}$ are set to zero by default.

However, since $T_{SND,Ref}$ and $T_{RCV,Ref}$ are expected to be rather short, it might in some cases be of interest to consider a certain or fixed network delay in order to obtain results from a realistic end-to-end transmission with non-ideal delays. In this case, a delay T_{Net} can be inserted in the beginning the echo signal $e_{RCV}(k)$ by zero-padding. T_{Net} is considered as the sum of $T_{SND,Net}$ and $T_{RCV,Net}$ and shall be reported in conjunction with the predicted value.

NOTE: As shown in Figure 6, the network delays $T_{SND,Net}$ and $T_{RCV,Net}$ are in general applied in different directions. However, for the contribution of these components to T_{Echo}, it is not relevant in which order they are applied (see equation (1)). For this reason, only the sum T_{Net} is considered for sake of simplicity, and not the two individual directions.

5.3.3 External echo delay

The perceived echo delay T_{Echo} corresponds to the delay between the two input signals $x_{ST}(k)$ and $e_{RCV}(k)$ (see clause 5.1). In principle, it can be determined algorithmically and is also an integral part of the prediction model (see clause 5.4). However, in case the model is applied in the context of a larger performance evaluation, there are typically more specific measurement methods available to determine T_{Echo} of the DUT, by e.g. measuring send and receive delay separately. Thus, instead of using the signal-based calculation, T_{Echo} should preferably be provided to the prediction model as an external value.

NOTE: If the externally provided delay T_{Echo} originates from two separate measurements in send and receive direction as given by equation (2), these usually exclude the acoustic path delay, i.e. from mouth to input microphone (send) and from the loudspeaker output to the ear (receive). In consequence, also the acoustic echo path from loudspeaker output to microphone input is not included. As mentioned in clause 5.3.1, this delay can be neglected and does not need to be added to the sum of send and receive delays.

5.4 Calculation of echo delay and linearity measure

The perceived echo delay T_{Echo} is an important parameter, which is used in the further stages of the prediction algorithm. In case it is not provided as an external value (see clause 5.3.3), it shall be calculated by means of the extended cross-correlation analysis introduced below, which is also used to obtain a linearity measure.

First, the sidetone signal $x_{ST}(k)$ and the echo signal $e_{RCV}(k)$ are segmented into overlapping frames of length N and step size S, as defined in equations (4) and (5). The step size S depends on the percentage of frame overlap ovl and the number of samples N of the analysis window, as defined in equation (3):

$$S = [(1 - ovl) \cdot \mathrm{N}] \tag{3}$$

$$x_{ST}(l, k') = x_{ST}(\mathrm{l} \cdot \mathrm{S} + k') \tag{4}$$

$$e_{RCV}(l, k') = e_{RCV}(\mathrm{l} \cdot \mathrm{S} + k') \tag{5}$$

(with $0 \leq k' \leq N - 1$).

The length N of the analysis window limits the maximum detectable delay; thus, it shall be chosen sufficiently large. On the other hand, too large values for N can lead to unnecessarily long calculation time. Values for N and S shall be chosen as provided in Table 1. Values for ovl are rounded to two decimal digits and are provided for information. Higher values for N, i.e. larger delays than 1,36 s cannot be adequately predicted by the model.

Table 1: Values for analysis windows length N and resulting overlap percentages

N	Max. Delay T/2 (at 48 kHz)	ovl	Step S [(1 - ovl) · N]
16 384	170,7 ms	85,35 %	2 400
32 768	341,3 ms	92,68 %	2 400
65 536	682,7 ms	96,34 %	2 400
131 072	1 365,3 ms	98,17 %	2 400

NOTE 1: N is always a power of two, which allows a faster implementation of the cross-correlation calculation by means of the fast Fourier transformation.

NOTE 2: The values for N and ovl in Table 1 correspond to a resulting step size of approximately 50 ms at 48 kHz. For sake of simplicity, values for ovl and step are rounded to integer values.

The energy normalization for an arbitrary signal s(k) of length L is performed by means of Z-score as defined in equation (6):

$$Z\big(s(k)\big) = \frac{\mathrm{s(k)} - \mu}{\sigma} \tag{6}$$

(μ represents the mean and σ the standard deviation of s(k)).

The m-th frame of the normalized short-time cross-correlation function $\Phi_{xe}(\mathrm{l}, \tau)$ between sidetone signal $x_{ST}(l,k')$ and echo signal $e_{RCV}(l,k')$ is calculated in the time domain according to equation (7) for each lag τ:

$$\Phi_{xe}(l, \tau) = \frac{1}{\sqrt{\sum Z(x_{ST}(l,k'))^2 + Z(e_{RCV}(l,k'))^2}} \sum_{k'=0}^{N-\tau-1} Z(x_{ST}(l, k')) \cdot Z(e_{RCV}(l, k' + \tau)) \tag{7}$$

Then, the envelope $P(l,\tau)$ according to equation (8) of the cross-correlation function $\Phi_{xy}(i,\tau)$ is calculated by the Hilbert transformation $H\{\Phi_{xe}(l,\tau)\}$ of the cross-correlation, as shown in equation (9):

$$\mathrm{H}\{\Phi_{xe}(l, \tau)\} = \frac{1}{\pi} \sum_{u=-\infty}^{+\infty} \frac{\Phi_{xe}(l,u)}{\tau - u} \tag{8}$$

$$\mathrm{P}(l, \tau) = \sqrt{[\Phi_{xe}(l, \tau)]^2 + [\mathrm{H}\{\Phi_{xe}(l, \tau)\}]^2} \tag{9}$$

In case the echo delay was externally provided to the prediction model (see clause 5.3.3), the linearity measure C_{lin} is determined by evaluating $P(l,\tau)$ at the corresponding lag index as the 90th percentile across all frames (equation (10)):

$$C_{\mathrm{lin}} = \mathrm{percentile}_{90}\ (\mathrm{P}(l, \tau = \mathrm{T}_{\mathrm{Echo}})) \tag{10}$$

In case the echo delay was not externally provided to the prediction model, it is estimated from P(l,τ) for all lags larger or equal than zero (assuming that the echo delay cannot be negative). For each frame, the maximum and the corresponding lag index is determined according to equations (11) and (12):

$$T_{\mathrm{Echo}}(l) = \underset{\tau}{\operatorname{argmax}} |\mathrm{P}(l, \tau \geq 0)| \tag{11}$$

$$\mathrm{P}_{\max}(l) = \max_{\tau} |\mathrm{P}(l, \tau \geq 0)| \tag{12}$$

The maximum E_{max}(l) provides an indicator for the reliability of each frame. A threshold of 10 % is used to discard unreliable frames, i.e. only frames indices l' providing $E_{\max}(i) \geq 10\,\%$ are regarded in the following calculation steps. Finally, the delay $\mathrm{T_{Echo}}$ is determined as the 90th-percentile of the remaining frames, as shown in equation (13):

$$T_{\mathrm{Echo}} = \mathrm{percentile}_{90}\ (T_{\mathrm{Echo}}(l')) \tag{13}$$

The linearity measure $\mathrm{C_{lin}}$ is determined by the 90th-percentile of $\mathrm{P_{max}}(l')$ (equation (14)):

$$C_{\mathrm{lin}} = \mathrm{percentile}_{90}\ (\mathrm{P}_{\max}(l')) \tag{14}$$

If less than 10 frames meet the condition $\mathrm{P_{max}}(i) \geq 10\,\%$, $\mathrm{T_{Echo}}$ cannot be determined reliably from E_{max}(l). However, this case does not imply the complete absence of a perceivable echo. Signal variations of x(k) that are mostly uncorrelated (artefacts like e.g. non-linear components or temporal clipping) might still clearly be audible.

In this case, a default value of $\mathrm{T_{Echo}}$ = 800 ms is assumed. According to [i.6], this value for echo delay covers the range in which non-linear/-monotonic dependencies between delay and quality were observed. The linearity measure $\mathrm{C_{lin}}$ is then determined by the 90th-percentile of $\mathrm{P_{max}}(l)$ (equation (15)), i.e. aggregated across the whole signal:

$$C_{\mathrm{lin}} = \mathrm{percentile}_{90}\ (\mathrm{P}_{\max}(l)) \tag{15}$$

NOTE 3: The linearity measure $\mathrm{C_{lin}}$ cannot reach a result of 10 % or higher in order to be consistent with the method described in this clause.

In general, the status of how echo delay and linearity measure were determined may differ between measurements and (in case of low correlation) might even be unclear in advance of the method. Complementary to $\mathrm{C_{lin}}$ and $\mathrm{T_{Echo}}$, also the status of these values shall be reported:

- "Estimated" - the cross-correlation between echo and sidetone signal was used to determine $\mathrm{T_{Echo}}$.
- "Default" - due to low correlation between echo and sidetone signal, a default value of $\mathrm{T_{Echo}}$ is used.
- "External" - $\mathrm{T_{Echo}}$ was provided externally to the prediction model.

The possible ways of determining $\mathrm{C_{lin}}$, $\mathrm{T_{Echo}}$ and their status as described in the text above are illustrated in Figure 7.

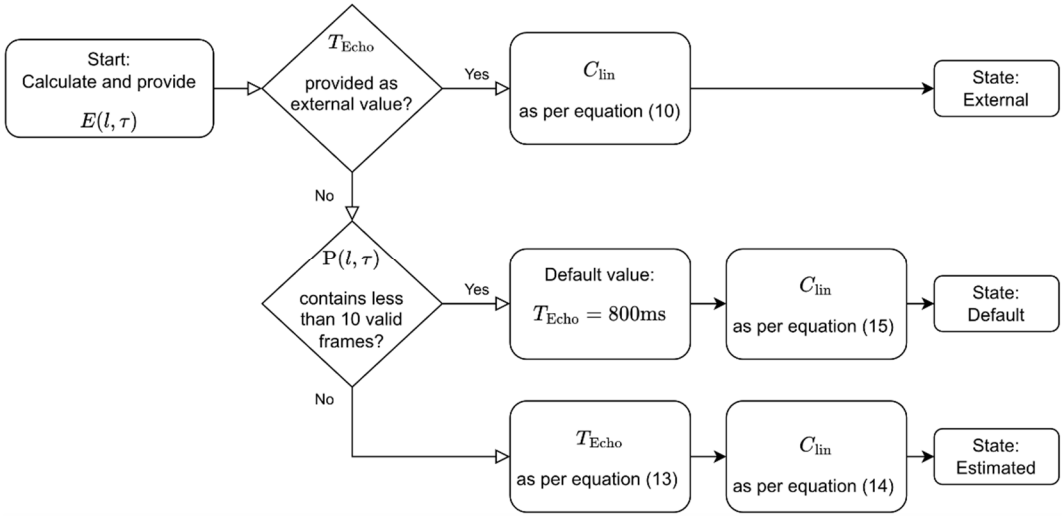

Figure 7: Determination of echo delay and linearity measure

5.5 Classification of time ranges

In order to determine the time ranges of active speech and echo, the classification algorithm according to Appendix II of Recommendation ITU-T G.160 [12] is applied on the sidetone signal $x_{ST}(k)$. The first step is to classify energy frames of 10 ms (block-wise, no overlap) according to the method described in [12]. The thresholds for the classification are defined relatively to the active speech level according to Recommendation ITU-T P.56 [6], which has to be calculated in advance.

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 resulting time instances for start and end of the n-th active speech range (i.e. the n-th sentence of a sample under test) are subsequentially referred to as $t_S(n)$ and $t_e(n)$. The duration of the n-th active range is denoted as $T_d(n) = t_e(n) - t_S(n)$.

5.6 Hearing model transformation

The hearing model according to clause 5 of ECMA 418-2 [13] is used to model the psychoacoustic signal representations of $x_{ST}(k)$ and $e_{RCV}(k)$. After applying the basis specific loudness versus time, the corresponding time-frequency matrices are denoted as $X_{ST}(l,m)$ and $E_{RCV}(l,m)$. The default configuration according to [13] is that time instances (*l)* are provided with an output sampling rate of 187,5 Hz and that 53 frequency bands (*m*) are calculated. The following parameters are modified for the prediction model:

- The outer ear filter (clause 5.1.3 in [13]) assumes Free-Field (FF) equalized input signals. The difference of the target frequency responses as defined in Table 3 and Table 2 in Recommendation ITU-T P.58 [3] are used as an additional correction for Diffuse-Field (DF).
- The step size Δz to determine the Bark band resolution (clause 5.1.4 in [13]) is decreased from 0,5 to 0,2. This results in a higher resolution in frequency (130 overlapping Bark bands).

Figure 8 shows an example of $X_{ST}(l,m)$ and $E_{RCV}(l,m)$, which also indicates the perceived echo delay T_{Echo} (about 400 ms) as well as the detected start and end times t_s and t_e of the talker's voice according to clause 5.5 (as there is only one active speech range in this sample, the sentence index n is left out here for sake of clarity).

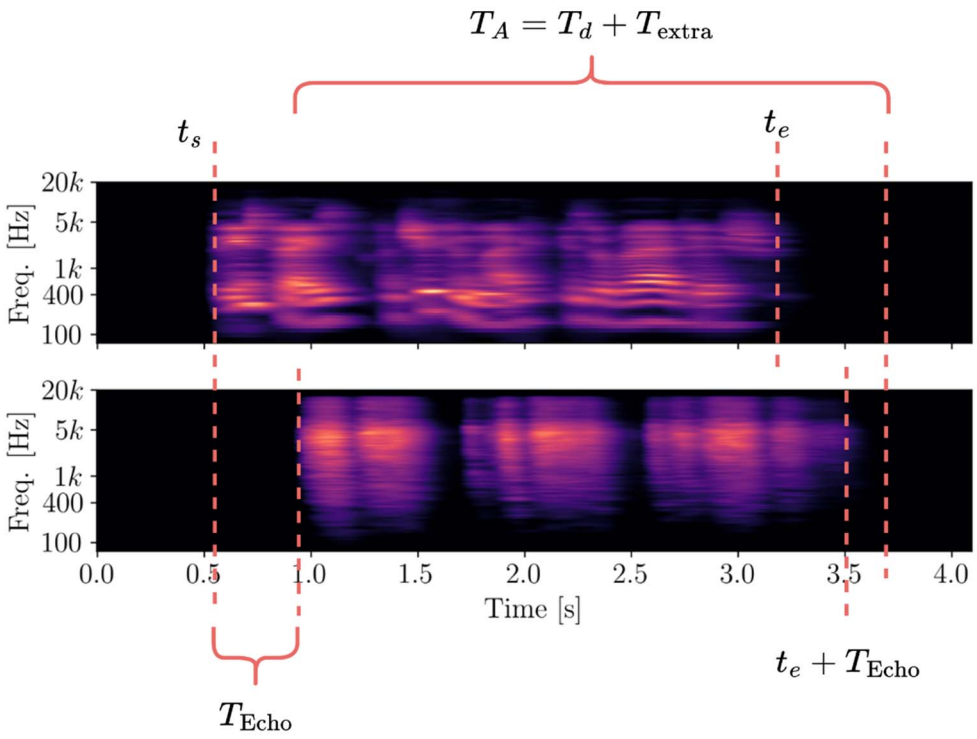

Figure 8: Example of hearing model representations of $X_{ST}(l,m)$ (side-tone, upper) and $E_{RCV}(l,m)$ (echo signal, lower)

This example also illustrates the importance of a robustly determined echo delay: possible echo artefacts can only occur within the time range $t_S + T_{Echo}$ to $t_e + T_{Echo}$. To consider possible slight misalignments of the time instances/ranges, the analysis time range T_A starts at $t_S + T_{Echo}$ and equals the sentence duration T_d plus an extra margin T_{extra} of 200 ms to consider artifacts like e.g. ringing, time warping or larger time-variant delays. T_{extra} is not applied in case the delay calculation status was determined to "Default", i.e. if T_{Echo} was set to 800 ms. The analysis range T_A shall be at least 250 ms.

Hearing model spectra vs time inside this analysis window are referred to as $X_{ST}(l',m)$ and $E_{RCV}(l',m)$. Time ranges outside this analysis window are not considered for the overall metric calculation.

5.7 Self-masking of echo

As indicated in Figure 8, most parts of the echo component $E_{RCV}(l,m)$ at the reference-side are perceived while the talker's voice is still active. As a consequence, a certain portion of this degradation is masked by the talker's own voice. This residual echo R(l,m) is modelled by weighting the side-tone $X_{ST}(l,m)$ with a masking threshold M(m) from the echo, as given in equation (16):

$$R(l,m) = \max(0, E_{\text{RCV}}(l,m) - \max(0, X_{\text{ST}}(l,m) \cdot M(m))) \quad (16)$$

The frequency-dependent masking weights M(m) is provided in Table A.1.

NOTE 1: For the determination of these weights, it was assumed that TPLT experiments are used as the basis for the prediction. In such listening tests, the sidetone is realized by HATS measurements (MRP-to-DRP, see clause 5.4.3.2 in ETSI TS 103 801 [1]), which might be different from the actual perception of real talker's own voice. Masking weights based on real human perception (including effects like e.g. bone conduction) are for further study.

With this approach, the negative impact of delay is implicitly handled as well. The echo signal exceeding the talker's speech activity is not masked at all and has a higher contribution to the overall degradation than the parts during speech.

This principle is illustrated in Figure 9, in which a modified version of Figure 8 is depicted. The first row shows the side-tone, the second one the perceived echo and the third one the residual echo after masking. The residual echo is most pronounced around 3,0 to 3,5 s, i.e. after the speech activity of the talker. This becomes even more obvious when (artificially) increasing the delay of the same echo by 500 ms (fourth row). The residual echo (last row) remains unmasked between 3,0 and 4,0 s. In general, larger delays shift the echo beyond the effective masking threshold, making it more audible.

NOTE 2: The masking during active speech of the talker depends on the time/frequency structure of the superposed components (talker's voice and echo) and the delay between both. In some cases, even a slightly higher delay might lead to a more masked echo, which of course is not a recommended way to optimize echo performance.

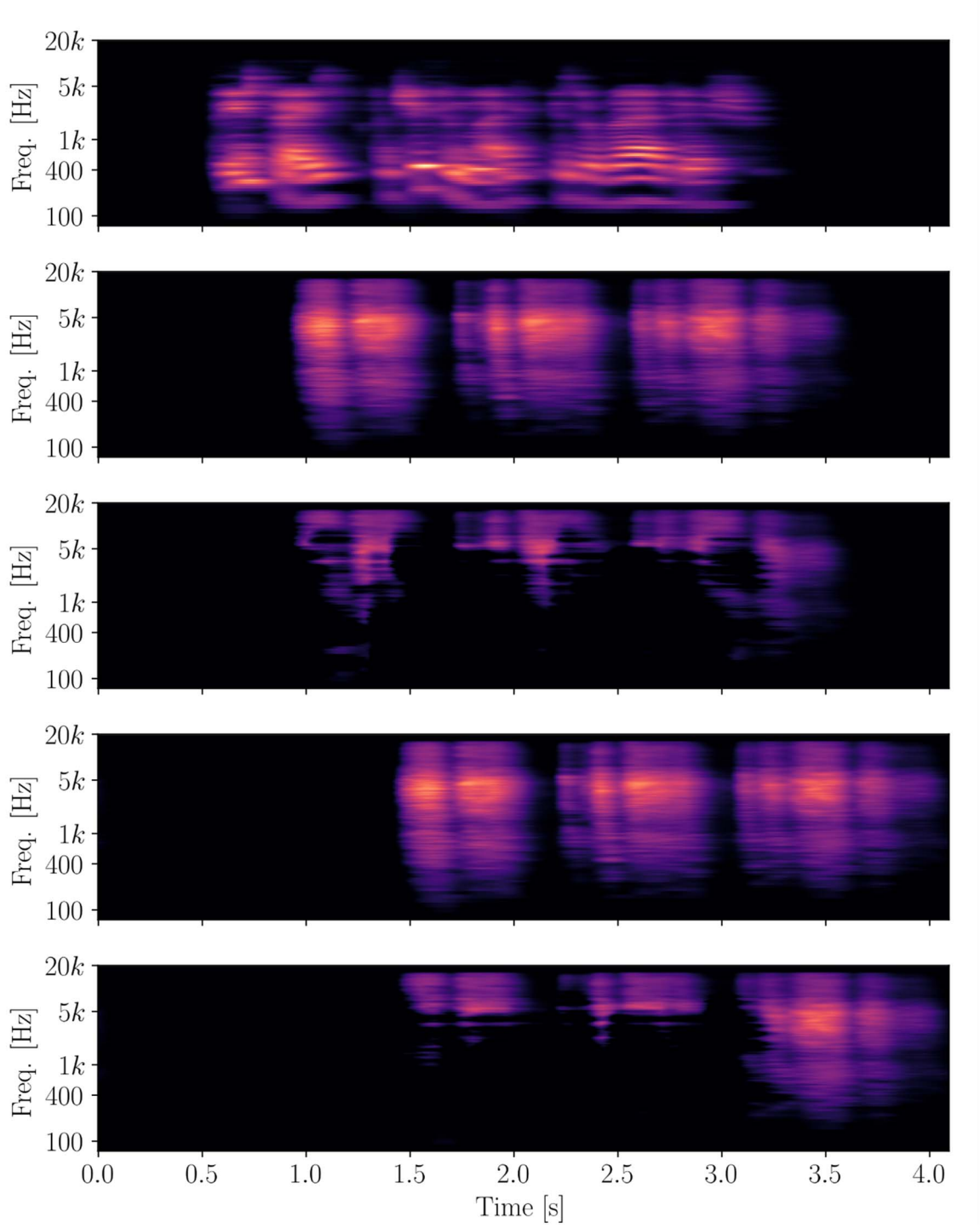

Figure 9: Example of masked echo signal: talker's voice (top row), unmasked (second row) and masked/residual echo (third row) with T_{Echo} of 400 ms; in addition unmasked (fourth row) and masked/residual echo (last row) with T_{Echo} of 900 ms

5.8 Aggregate over time and frequency

The time-frequency representations of the masked residual echo R(l,m) and the (unmasked) echo E_{RCV}(l,m) are first aggregated to loudness versus time by summing specific loudness versus frequency, as given by equations (17) and (18). Note that the modified step size of the Bark band resolution (Δz = 0,2) has to be considered here as well (see clause 5.6):

$$\hat{\mathrm{R}}(l) = \sum_m \mathrm{R}(l,m) \cdot \Delta z \tag{17}$$

$$\hat{E}_{\mathrm{RCV}}(l) = \sum_m E_{\mathrm{RCV}}(l,m) \cdot \Delta z \tag{18}$$

Based on these intermediate results, average and 95 %-percentile loudness values for masked and unmasked echo are calculated according to equations (19) to (22) individually for each sentence identified by the analysis described in clause 5.5. The time indices $l'(n)$ of the loudness versus time values represent the ones within the analysis range T_A(n), i.e. of the *n*-th sentence. $K(n)$ is the amount of time indices within analysis range T_A(n):

$$L_{M,avg}(n) = \frac{1}{K(n)} \sum_{l'} \hat{\mathrm{R}}\left(l'(n)\right) \tag{19}$$

$$L_{U,avg}(n) = \frac{1}{K(n)} \sum_{l'} \hat{E}_{\mathrm{RCV}}(l'(n)) \tag{20}$$

$$L_{M,P95}(n) = \mathrm{percentile}_{95}\left(\hat{\mathrm{R}}(l'(n))\right) \tag{21}$$

$$L_{U,P95}(n) = \mathrm{percentile}_{95}\left(\hat{E}_{\mathrm{RCV}}(l'(n))\right) \tag{22}$$

Overall loudness values are determined as the weighted average across all samples. The weights are determined by the duration T_d(n) of each sample (see clause 5.5), as shown in equations (23) to (27):

$$w(n) = {T_d(n)}\Big/{\sum_n T_d(n)} \tag{23}$$

$$L_{M,avg} = \sum_n w(n) \cdot L_{M,avg}(n) \tag{24}$$

$$L_{U,avg} = \sum_n w(n) \cdot L_{U,avg}(n) \tag{25}$$

$$L_{M,P95} = \sum_n w(n) \cdot L_{M,P95}(n) \tag{26}$$

$$L_{U,P95} = \sum_n w(n) \cdot L_{U,P95}(n) \tag{27}$$

Each loudness *L* (in unit Sone) is then converted to the corresponding loudness level $\mathcal{L}$ (in unit Phon) according to clause 5.3 in [14]. The function is provided in equation (28):

$$\mathrm{sone2phon}\,(L) = \begin{cases} 40 + 33{,}22 \cdot \log(L) & L > 1{,}0 \\ 40 \cdot (L + 0{,}0005)^{0{,}35} & L \leq 1{,}0 \\ 0 & \text{else} \end{cases} \tag{28}$$

The resulting four loudness level values are provided by equations (29) to (32):

$$\mathcal{L}_{M,avg} = \mathrm{sone2phon}\,(L_{M,avg}) \tag{29}$$

$$\mathcal{L}_{U,avg} = \mathrm{sone2phon}\,(L_{U,avg}) \tag{30}$$

$$\mathcal{L}_{M,P95} = \mathrm{sone2phon}\,(L_{M,P95}) \tag{31}$$

$$\mathcal{L}_{U,P95} = \mathrm{sone2phon}\,(L_{U,P95}) \tag{32}$$

5.9 Regression

In order to combine the loudness level metrics described in the previous clauses, Multivariate Adaptive Regression Splines (MARS) according to [15] are used. The MARS regression is a weighted sum of one or more so-called hinge functions as defined in equation (33):

$$h(a; b) = \max(0, a - b) \tag{33}$$

The prediction of MOS-TQO$_f$ in equation (34) is determined as the sum of terms that are given in Table 2:

$$\text{MOS-TQO}_f = \sum_j y_j \tag{34}$$

Table 2: Terms of coefficients and hinge functions

Term		Coefficient		Hinge function(s)
y_0	=	2,38646	·	1,0
y_1	=	-0,0148569	·	h($\mathcal{L}_{M,avg}$; 36,2625)
y_2	=	0,0611339	·	h(36,2625; $\mathcal{L}_{M,avg}$)
y_3	=	-0,0387319	·	h(44,3437; $\mathcal{L}_{U,avg}$)
y_4	=	0,0013305	·	h($\mathcal{L}_{M,P95}$; 16,0865) · h(36,2625; $\mathcal{L}_{M,avg}$)
y_5	=	-0,00196979	·	h(16,0865; $\mathcal{L}_{M,P95}$) · h(36,2625; $\mathcal{L}_{M,avg}$)
y_6	=	-0,00956064	·	h($\mathcal{L}_{M,P95}$; 39,3408)
y_7	=	0,077563	·	h(39,3408; $\mathcal{L}_{M,P95}$)
y_8	=	0,24637	·	h($\mathcal{L}_{M,avg}$; 33,5417) · h(36,2625; $\mathcal{L}_{M,avg}$)

5.10 Indication for absence of an echo signal

In several cases, DUTs may not produce a typical echo signal at all, and only idle noise is captured at the POI in send direction. Parameters that were regarded as important for the assessment of echo (like, e.g. the overall delay T_{Echo}, see clause 4 of ETSI TS 103 801 [1]), become perceptually irrelevant in this case. Independent of the residual noise level or possible other artefacts that may introduce unrelated degradations, the prediction model can be skipped and return the maximum score of 4,9 MOS-TQO$_f$, if no echo can be detected in $e_{RCV}(k)$.

The absence of echo is indicated in case the following two criteria are met:

- $\mathcal{L}_{U,P95}$ is less than 10 phon.
- C_{lin} is less than 10 %.

5.11 Compensation for idle noise

Idle noise introduced in the send direction can vary significantly across different DUTs and may impact the loudness level calculations as per equations (29) to (32). To address this, an average idle noise floor $N_{\text{RCV}}(m)$ is calculated using equation (35). All time instances of $E_{\text{RCV}}(l, m)$ are then compensated according to equation (36).

$$N_{\text{RCV}}(m) = \text{percentile}_{10}\ (E_{\text{RCV}}(l, m)) \tag{35}$$

$$E'_{\text{RCV}}(l, m) = \max(E_{\text{RCV}}(l, m) - N_{\text{RCV}}(m),\ 0) \tag{36}$$

Without compensation, the presence of idle noise may artificially increase the measured loudness levels, which leads to worse performance than it is. By subtracting the noise floor, test results reflect the actual echo signal more accurately and are more comparable across different devices. The idle noise compensation should always be activated when using the prediction model in performance/qualification testing.

NOTE 1: The idle noise estimation is carried out over the whole hearing model spectrum, including time ranges inside and outside the analysis window T_A (see Figure 8).

NOTE 2: The noise estimate in equation (35) is based on a percentile analysis and targets at tracking the lower loudness magnitudes versus time. For a robust and reliable result, at least 10 % of the available time frames should be inactive, either individually in each frequency band, or globally, by considering sufficiently leading and/or trailing speech pauses. The test signals specified in clause 4.3 contain this minimum activity in general, but the amount of silence periods might be changed in case shorter samples are cropped from a longer recording.

NOTE 3: In case the compensation is used, it would be applied after clause 5.6, which would then result in the use of $E'_{\mathrm{RCV}}(l,m)$ (compensated echo signal) instead of $E_{\mathrm{RCV}}(l,m)$ (uncompensated echo signal) in clause 5.7 and following. For sake of clarity, no distinction is made between compensated and uncompensated echo signal, and $E_{\mathrm{RCV}}(l,m)$ is used synonymously in these clauses.

Annex A (normative): Self-masking weights determined by HATS setup

The weighting coefficients as provided in Table A.1 and illustrated in Figure A.1 were determined by investigating the listening test databases in annex B, which were obtained by the test procedure specified in ETSI TS 103 801 [1]. The experiments were conducted as a TPLT, where the sidetone was simulated by a transfer function between MRP and DRP of a HATS - see clause 5.4.3.2 in ETSI TS 103 801 [1] for more details. In consequence, the self-masking weighting described in this annex can only be considered as valid for this specific test design. Note that the procedure to identify the weights was conducted on integer Bark bands and then linearly interpolated to the frequency resolution of the hearing model used in the present document.

Table A.1: Self-masking weights versus frequency

Freq. [Hz]	Mask weight	Freq. [Hz]	Mask weight	Freq. [Hz]	Mask weight	Freq. [Hz]	Mask weight	Freq. [Hz]	Mask weight
82,3	0,9149	597,5	0,1545	1 561,3	0,2790	3 697,2	0,0183	8 608,2	0,3142
98,9	0,9149	623,2	0,1487	1 615,3	0,3038	3 819,9	0,0271	8 891,8	0,3303
115,7	0,9149	649,5	0,1430	1 670,9	0,3285	3 946,6	0,0360	9 184,8	0,3464
132,6	0,9149	676,5	0,1372	1 728,3	0,3532	4 077,4	0,0449	9 487,3	0,3625
149,6	0,9149	704,2	0,1314	1 787,5	0,3779	4 212,6	0,0538	9 799,7	0,3786
166,7	0,9149	732,7	0,1257	1 848,6	0,4026	4 352,1	0,0626	10 122,5	0,3947
184,1	0,9149	761,9	0,1199	1 911,6	0,4273	4 496,2	0,0715	10 455,8	0,4109
201,6	0,9149	791,9	0,1142	1 976,6	0,4520	4 645,0	0,0804	10 800,1	0,4270
219,4	0,9149	822,7	0,1084	2 043,7	0,4767	4 798,6	0,0893	11 155,7	0,4431
237,3	0,9149	854,4	0,1027	2 112,9	0,5014	4 957,3	0,0981	11 523,0	0,4592
255,6	0,8702	887,0	0,0969	2 184,3	0,4730	5 121,2	0,1070	11 902,3	0,4622
274,0	0,8255	920,5	0,0912	2 258,0	0,4446	5 290,4	0,1159	12 294,1	0,4651
292,8	0,7807	955,0	0,0854	2 334,1	0,4162	5 465,2	0,1248	12 698,8	0,4681
311,9	0,7360	990,5	0,0797	2 412,6	0,3877	5 645,7	0,1336	13 116,8	0,4710
331,3	0,6913	1 027,0	0,0739	2 493,7	0,3593	5 832,1	0,1425	13 548,5	0,4740
351,1	0,6465	1 064,6	0,0682	2 577,3	0,3309	6 024,7	0,1514	13 994,4	0,4770
371,2	0,6018	1 103,3	0,0624	2 663,7	0,3025	6 223,5	0,1603	14 455,0	0,4799
391,7	0,5571	1 143,2	0,0567	2 752,8	0,2741	6 428,9	0,1691	14 930,7	0,4829
412,6	0,5123	1 184,3	0,0814	2 844,9	0,2456	6 640,9	0,1853	15 422,0	0,4858
434,0	0,4676	1 226,6	0,1061	2 939,9	0,2172	6 860,0	0,2014	15 929,5	0,4888
455,8	0,4229	1 270,2	0,1308	3 038,0	0,1888	7 086,2	0,2175	16 453,6	0,4918
478,1	0,3781	1 315,1	0,1555	3 139,2	0,1604	7 319,8	0,2336	16 995,0	0,4947
500,9	0,3334	1 361,4	0,1802	3 243,8	0,1319	7 561,2	0,2497	17 554,2	0,4977
524,2	0,2887	1 409,1	0,2049	3 351,8	0,1035	7 810,4	0,2658	18 131,8	0,5006
548,1	0,2439	1 458,3	0,2296	3 463,2	0,0751	8 067,8	0,2819	18 728,4	0,5036
572,5	0,1992	1 509,0	0,2543	3 578,3	0,0467	8 333,7	0,2981	19 344,6	0,5066

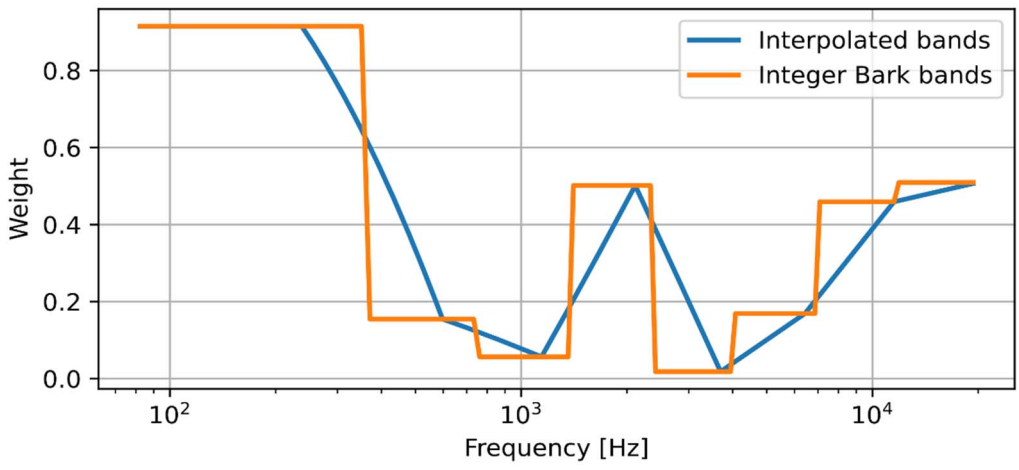

Figure A.1: Self-masking weights versus interpolated and integer Bark bands

Annex B (informative): Subjective experiments for training and validation

For training and validation of the prediction mode, three databases were used, which all follow the listening test procedure specified in ETSI TS 103 801 [1]:

- In [i.3] and [i.4] two auditory databases for NB and WB terminals were introduced.
- In [i.5], a comprehensive auditory database divided into three sub-experiments for (simulated) SWB/FB terminals was described.

The relevant information for each of these three experiments is provided in Table B.1, which also indicates the split of data between training and validation (about 1/3 of the available data was used for validation purposes).

Table B.1: Experiments used for training and validation of the model

Experiment	Number of conditions	Samples per condition	Sentences per Sample	Votes per condition	Conditions used for ...	
					Training	Validation
NB	73	2	2	40	49	24
WB	75	2	2	40	51	24
SWB/FB-1	65	4	1	92	65	0
SWB/FB-2	65	4	1	92	65	0
SWB/FB-3	50	4	1	92	0	50

The prediction performance for all training and validations sets are provided in Figure B.1 to Figure B.4. The following performance metrics are complementary provided in the figures:

- rmse*: Root-mean-square error per condition after 3rd order mapping according to Recommendation ITU-T P.1401 [i.7]. The uncertainty of the auditory data, which is represented by the 95 % Confidence Interval (CI95) is considered in the calculation.
- maxabs*: Maximum absolute error per condition after 3rd order mapping of the whole subset, again considering the uncertainty of the auditory data. See clause 8 of ETSI TS 103 281 [i.8] for further explanation.

CI95 per condition are necessary for the metric calculation and were determined according to Appendix III of Recommendation ITU-T P.1401 [i.7] based on CI95 values per-sample.

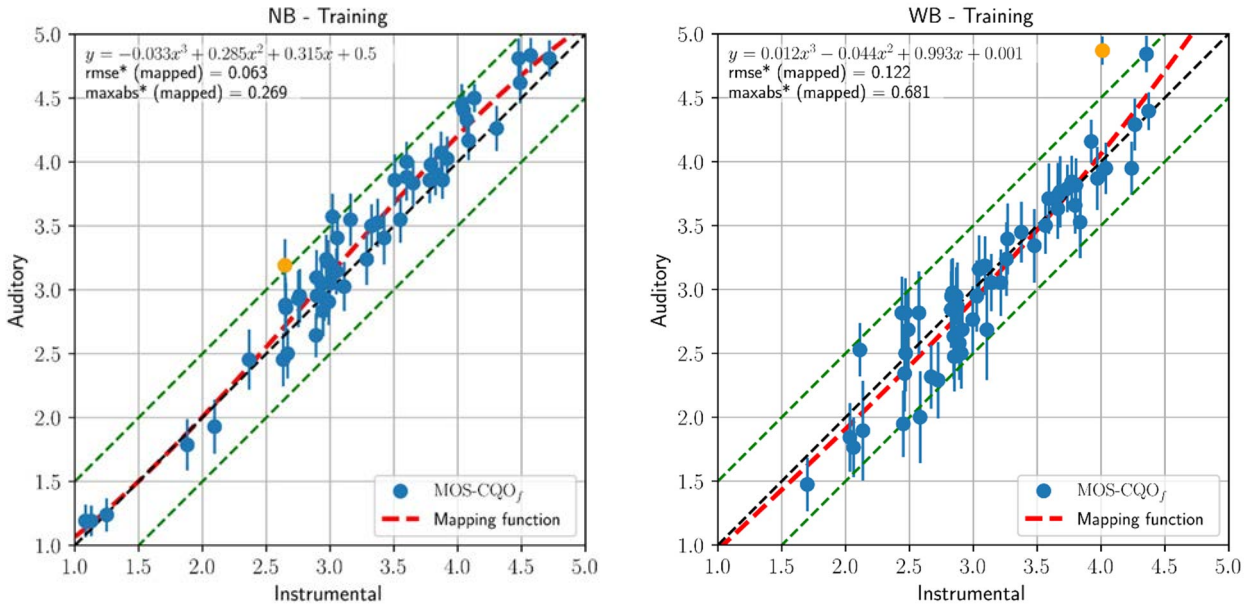

Figure B.1: Prediction results for the training data (subsets of databases NB and WB)

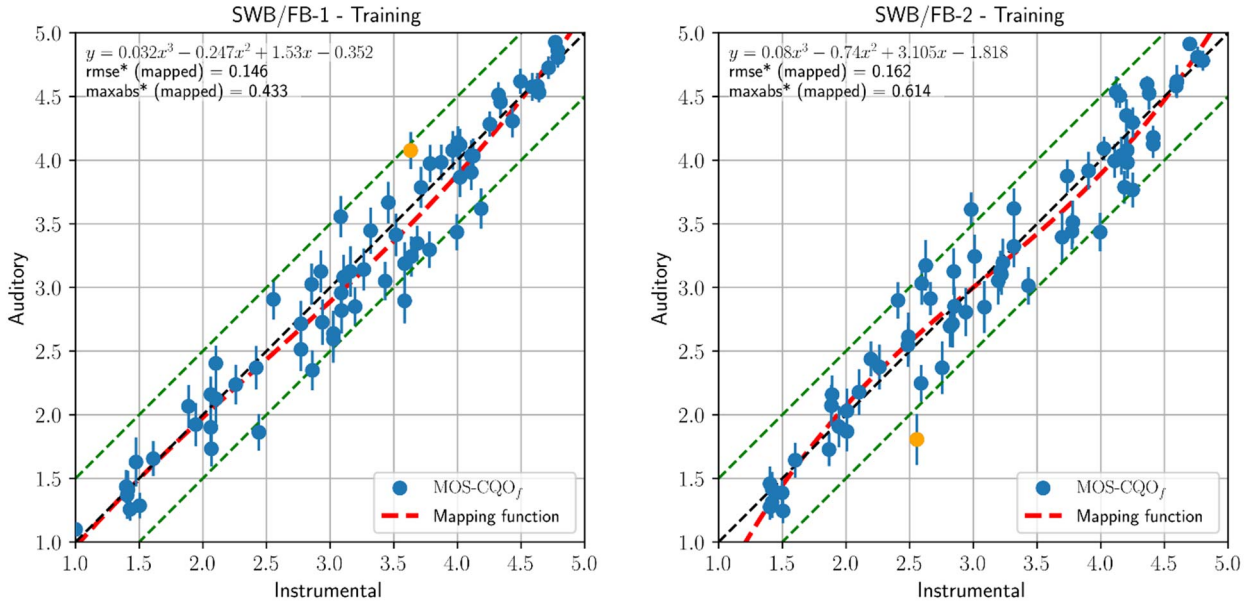

Figure B.2: Prediction results for the training data (SWB/FB-1 and SWB/FB-2)

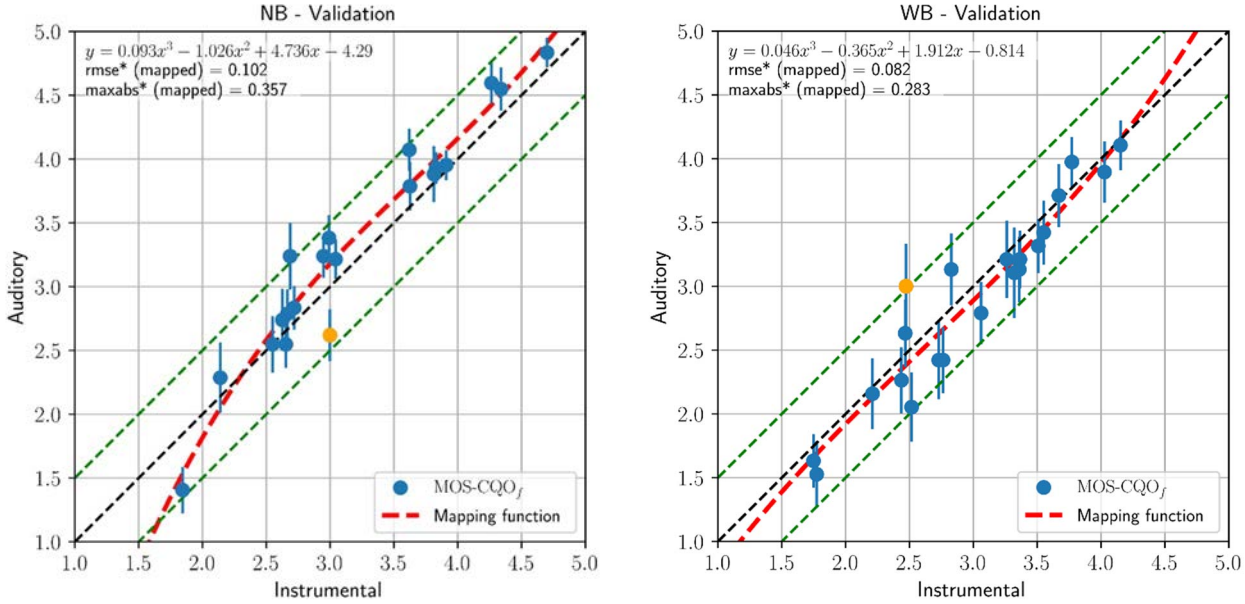

Figure B.3: Prediction results for the validation data (subsets of databases NB and WB)

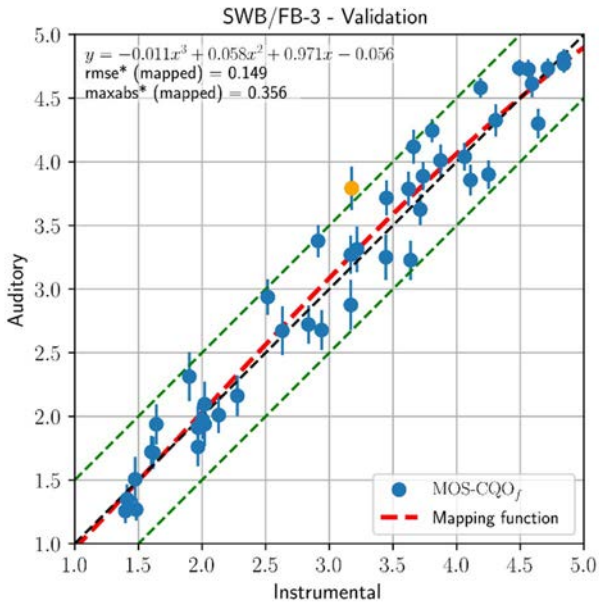

Figure B.4: Prediction results for the validation data (SWB/FB-3)

History

Document history		
V1.1.1	July 2023	Publication
V1.2.1	March 2025	Publication