ETSI TS 103 281 V1.3.1 (2019-05)

**Speech and multimedia Transmission Quality (STQ);
Speech quality in the presence of background noise:
Objective test methods for super-wideband and
fullband terminals**

Reference
RTS/STQ-269

Keywords
noise, quality, speech, testing, transmission

ETSI

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

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

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

Contents

Intellectual Property Rights

Essential patents

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

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

Trademarks

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

Foreword

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

The present document is to be used in conjunction with:

- ETSI ES 202 396-1 [i.1]: "Background noise simulation technique and background noise database"; and
- ETSI TS 103 224 [i.19] series: "A sound field reproduction method for terminal testing including a background noise database".

The present document describes an objective test method for super-wideband and fullband in order to provide a good prediction of the uplink speech quality in the presence of background noise of modern mobile terminals in hand-held and hands-free.

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 testing methodologies which can be used to objectively evaluate the performance of super-wideband and fullband mobile terminals for speech communication in the presence of background noise.

Background noise is a problem in mostly all situations and conditions and needs to be taken into account in terminal design. The present document provides information about the testing methods applicable to objectively evaluate the speech quality of mobile terminals (including any state-of-the-art codecs) employing background noise suppression in the presence of background noise. The present document includes:

- The method which is applicable to objectively determine the different parameters influencing the speech quality in the presence of background noise taking into account:
  - the speech quality;
  - the background noise transmission quality;
  - the overall quality.
- The model results in comparison with the underlying subjective tests used for the training of the objective model. The underlying languages are: American English, German, Chinese (Mandarin).
- The model validation results.

The present document is to be used in conjunction with:

- ETSI ES 202 396-1 [i.1] which describes a recording and reproduction setup for realistic simulation of background noise scenarios in lab-type environments for the performance evaluation of terminals and communication systems.
- ETSI TS 103 224 [i.19] which describes a sound field reproduction method for terminal testing including a background noise database with background noise scenarios to be used in lab-type environments for the performance evaluation of terminals and communication systems.
- American English speech sentences as enclosed in 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 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.

Not applicable.

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] 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.2] 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.3] 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.4] ETSI TS 126 441: "Universal Mobile Telecommunications System (UMTS); LTE; Codec for Enhanced Voice Services (EVS); General overview (3GPP TS 26.441)".

[i.5] Recommendation ITU-T P.835: "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm".

[i.6] Internet Engineering Task Force, Request for Comments 6716: "Definition of the Opus Audio Codec", 09/2012.

[i.7] Recommendation ITU-T P.56: "Objective measurement of active speech level".

[i.8] Recommendation ITU-T P.1401: "Methods, metrics and procedures for statistical evaluation, qualifying and comparison of objective quality prediction models".

[i.9] Recommendation ITU-T G.160 Appendix II, Amendment 2: "Voice enhancement devices: Revised Appendix II - Objective measures for the characterization of the basic functioning of noise reduction algorithms".

[i.10] Recommendation ITU-T P.501: "Test Signals for Use in Telephonometry".

[i.11] Recommendation ITU-T P.58: "Head and Torso simulator for telephonometry".

[i.12] Recommendation ITU-T P.57: "Artificial ears".

[i.13] Recommendation ITU-T P.800: "Methods for subjective determination of transmission quality".

[i.14] ETSI TS 126 132: "Universal Mobile Telecommunications System (UMTS); LTE; Speech and video telephony terminal acoustic test specification (3GPP TS 26.132)".

[i.15] Recommendation ITU-T TD 477 (GEN/12): "Handbook of subjective test practical procedures" (temporary document) - Geneva, 18-27 January 2011.

[i.16] AH-11-029: "Better Reference System for the P.835 SIG Rating Scale", Q7/12 Rapporteur's meeting, 20-21 June 2011, Geneva, Switzerland.

[i.17] 3GPP, Tdoc S4(16)0397: "DESUDAPS-1: Common subjective testing framework for training and validation of SWB and FB P.835 test predictors".

[i.18] Recommendation ITU-T P.64: "Determination of sensitivity/frequency characteristics of local telephone systems".

[i.19] ETSI TS 103 224: "Speech and multimedia Transmission Quality (STQ); A sound field reproduction method for terminal testing including a background noise database".

[i.20] Sottek R.: "Modelle zur Signalverarbeitung im menschlichen Gehör", PHD thesis RWTH Aachen, 1993.

[i.21] Sottek R.: "A Hearing Model Approach to Time-Varying Loudness", Acta Acustica united with Acustica, vol. 102(4), pp. 725-744, 2016.

[i.22] Byrne D. et al.: "An international comparison of long-term average speech spectra", The Journal of the Acoustical Society of America, Vol. 96, No. 4, 1994.

[i.23] IEC 61672-1:2013: "Electroacoustics - Sound level meters - Part 1: Specifications", 2003.

[i.24] Recommendation ITU-T P.863: "Methods for subjective determination of transmission quality".

[i.25] Côté N.: "Integral and Diagnostic Intrusive Prediction of Speech Quality", PHD thesis TU Berlin, 2010.

[i.26] Zwicker E. Fastl H.: "Psychoacoustics: Facts and Models", 1990.

[i.27] Falk, T. and Chan, W.-Y.: "A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech", IEEE Transactions on Audio, Speech, and Language Processing, Volume: 18, Issue: 7, September 2010.

[i.28] Breiman L.: "Random Forests", Machine Learning (journal), Volume 45, Issue 1, pp 5-32, October 2001.

[i.29] Berger, J.: "Instrumentelle Verfahren zur Sprachqualitätsschätzung", PhD thesis, 1998.

[i.30] W. Lu & D. Sen: "Extraction of cochlear processed formants for prediction of temporally localized distortions in synthetized speech", ICASSP 2009.

[i.31] Christian Giguere & Philip C. Woodland: "A computational model of the auditory periphery for speech and hearing research. I. Ascending path", JASA 1994, 95(1), pp 331-342.

[i.32] Wenliang Lu; Sen, D.: "Extraction of cochlear processed formants for prediction of temporally localized distortions in synthetized speech", Acoustics, Speech and Signal Processing. ICASSP 2009. IEEE International Conference on , vol., no., pp.3977-3980, 19-24 April 2009.

[i.33] W. Lu & D Sen: "Tolerance and sensitivity of various parameters in the prediction of temporally localized distortions in degraded speech", ICA 2010, Sydney, Australia, 23-27 August 2010.

[i.34] D. Talkin: "A Robust Algorithm for Pitch Tracking (RAPT)" in "Speech Coding & Synthesis", W B Kleijn, K K Paliwal eds, Elsevier ISBN 0444821694, 1995.

[i.35] Brookes Mike: "Voicebox: Speech processing toolbox for matlab" Software, available March 2011.

NOTE: Available at www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[i.36] Puria, S. & Allen, J.B.: "A parametric study of cochlear input impedance", JASA 1991, 89(1), pp 287-319.

[i.37] D. Sen and J. Allen: "Benchmarking a two-dimensional cochlear model against experimental auditory data" in Proceedings of MidWinter Meeting on Association for Research in Otolaryngology (ARO '01), February 2001.

[i.38] D. Sen and J. B. Allen: "Functionality of cochlear micromechanics-as elucidated by upward spread of masking and two tone suppression", Acoustics Australia, vol. 34, no. 1, pp. 37-42, 2006.

[i.39] J. B. Allen and M. Sondhi: "Cochlear macromechanics: time domain solutions", The Journal of the Acoustical Society of America, vol. 66, no. 1, pp. 123-132, 1979.

[i.40] Recommendation ITU-T P.380 (11/2003): "Electro-acoustic measurements on headsets".

[i.41] Recommendation ITU-T P.1120 (03/2017): "Super-WideBand (SWB) and FullBand (FB) stereo hands-free communication in motor vehicles".

[i.42] Recommendation ITU-R BS.708 (06/1990): "Determination of the electro-acoustical properties of studio monitor headphones".

[i.43] IEC 60268-7:2010: "Sound system equipment - Part 7: Headphones and earphones".

[i.44] ETSI TR 126 931: "Universal Mobile Telecommunications System (UMTS); LTE; Evaluation of Additional Acoustic Tests for Speech Telephony (3GPP TR 26.931)".

3 Definition of terms, symbols and abbreviations

3.1 Terms

Void.

3.2 Symbols

Void.

3.3 Abbreviations

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

AMR	Adaptive Multi-Rate (narrowband speech codec)
AMR-WB	Adaptive Multi-Rate Wideband (wideband speech codec)
AS	Analysis Serial
ASL	Active Speech Level
BAK	Background Noise Component
BGN	Background Noise
BM	Basilar Membrane
CM	Cochlear Model
COSM	Cochlear Output Statistic Metrics
CP	Characteristic Place
dB SPL	Sound Pressure Level re 20 µPa in dB
DB	Data Base
DES	Database Enumeration
DNN	Deep Neural Network
DRP	Drum Reference Point
EVS	Enhanced Voice Services
EVS-FB	Enhanced Voice Services - Fullband
FB	Fullband
FFT	Fast Fourier Transformation
G-MOS	Global MOS (related to the overall quality)
HATS	Head and Torso Simulator
HE	Headset
HHHF	Hand-Held Hands-Free
HS	Handset
IHC	Inner Hair Cell
IIR	Infinite Impulse Response
ITU	International Telecommunication Union
ITU-R	International Telecommunication Union - Radiocommunication sector
ITU-T	International Telecommunication Union - Telecommunication sector
LQO_{fb}	Listening Quality Objective (related to fullband scale)

LQS_{fb}	Listening Quality Subjective (related to fullband scale)
MOS	Mean Opinion Score
MRP	Mouth Reference Point
NB	Narrowband
N-MOS	Noise MOS (related to the noise intrusiveness)
NS	Noise Suppression
OHC	Outer Hair Cell
OVRL	Overall (speech + noise) Component
PCA	Principal Component Analysis
PCM	Pulse Code Modulation
RAPT	Robust Algorithm for Pitch Tracking
RMS	Root Mean Square
RMSE	Root Mean Square Error
RMSE*	epsilon insensitive Root Mean Square Error
SFP	Salient Formant Points
SIG	SIGnal component
SLR	Send Loudness Rating
S-MOS	Speech MOS (related to the speech distortion)
SNR	Signal to Noise Ratio
SNR(A)	Signal to Noise Ratio (A-weighted)
SPL	Sound Pressure Level
SWB	Super-wideband
SWB/FB	Super-Wideband/Fullband
TCP	Track Center Points
TM	Tectorial Membrane
VAD	Voice Activity Detection
WB	Wideband

4 Introduction

The present document describes models for the objective prediction of speech-, background-noise- and overall quality for super-wideband and fullband terminals and systems used in background noise in uplink on a fullband scale.

The models are intended to be used for modern terminals including e.g. different bitrates of EVS [i.4] and other state-of-the-art coding technologies. The current models were trained and validated with EVS-SWB, EVS-FB, Opus [i.6], AMR, AMR-WB, PCM including typical packet loss and jitter conditions and recordings in handset, headset, hands-free and car hands-free mode.

5 Underlying speech databases and preparations

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 evaluation of processed and transmitted noisy speech, the Recommendation ITU-T P.835 [i.5] is a state-of-the-art method for the assessment of speech and noise quality. The listening test procedure described in [i.5] is also the basis for the prediction model.

It is necessary to note that the Recommendation ITU-T P.835 [i.5] uses a slightly different nomenclature for the different quality attributes. For the speech distortion scale, SIG (signal = speech) is used instead of S-MOS-LQS, BAK (background noise) instead of N-MOS-LQS and OVRL (overall) instead of G-MOS-LQS. Whenever these abbreviations are used in the present document, this always indicates that auditory results are addressed.

In addition to Recommendation ITU-T P.835 [i.5], several details of auditory testing were specified in [i.17]. These more detailed descriptions focus on the recording and creation of the test and reference stimuli. An update of the reference processing to SWB/FB mode was introduced as an extension of the procedures described in [i.3]. This revised subjective test framework is required in order to minimize variations between subjective tests performed in different listening laboratories. A summary of this work is provided in annex D.

6 Model descriptions

6.1 Introduction

The prediction models described in the following clauses are full-reference models. Such a predictor compares the degraded signal under test against a reference signal. Audible disturbances between these two signals are assumed to highly correlate with the results of auditory tests conducted in the development phase. Two models are provided in the present document for this purpose.

6.2 Common definitions

Even though both model variants internally work differently regarding the processing steps, inputs and outputs are common. The input time signals are denoted as x(k) for the reference signal and y(k) for the degraded signal, which is evaluated either by the instrumental or the auditory assessment. Each prediction model provides three output values:

- S-MOS-LQO$_{fb}$: instrumentally assessed SIG component (speech distortion).
- N-MOS-LQO$_{fb}$: instrumentally assessed BAK component (noise intrusiveness).
- G-MOS-LQO$_{fb}$: instrumentally assessed OVRL component (global quality).

6.3 Model A

6.3.1 Introduction

In general, the model consists of several stages and calculation steps which finally conclude in the assessment of instrumental S-, N- and G-MOS. Figure 6.1 provides an overview about the structure of the method. Clauses 6.3.2 to 6.3.8 provide detailed descriptions of each processing block.

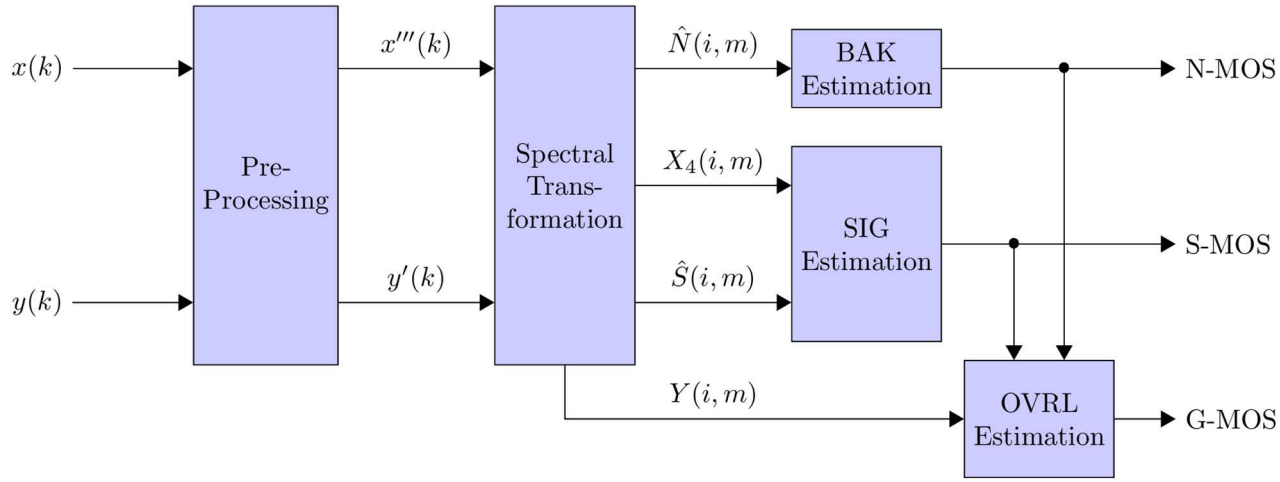


Figure 6.1: Block diagram of model A

6.3.2 Pre-Processing

The pre-processing of the inputs x(k) and y(k) is conducted to compensate differences regarding temporal alignment and level offsets between the signals. An overview of the pre-processing is given in the block diagram shown in figure 6.2.

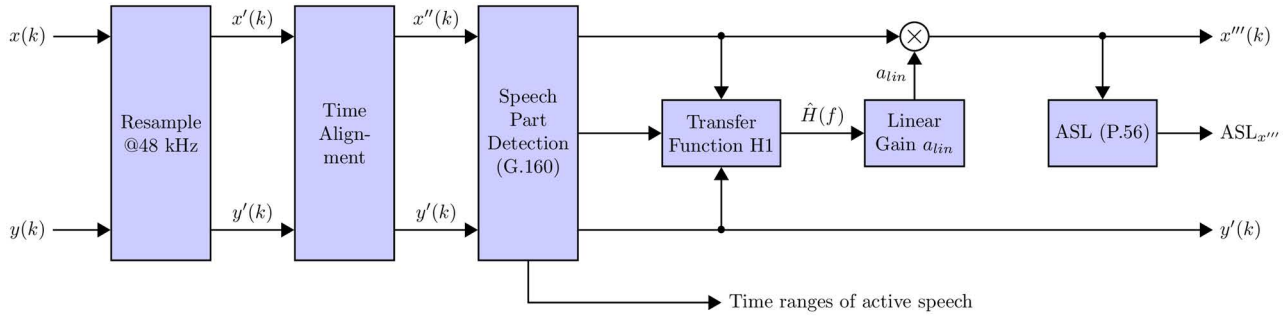


Figure 6.2: Block diagram of model A

The first block ensures that both input signals provide the same sampling rate of 48 kHz. The outputs are denoted as x'(k) and y'(k).

The delay compensation between processed and clean speech signal is applied in a similar way as in the method according to ETSI EG 202 396-3 [i.2]. The signals x'(k) and y'(k) are filtered with an IIR band-pass of 6th order and a frequency range of 300 Hz - 3 300 Hz. Limiting to this range, only the signal parts containing most speech energy are taken into account. Then, the cross-correlation $\Phi_{xy}(\tau)$ between the pre-filtered input signals x'(k) and y'(k) is calculated, followed by an envelope operation according to equation (1).

$$E(\tau) = \sqrt{\left[\Phi_{xy}(\tau)\right]^2 + \left[H\left(\Phi_{xy}(\tau)\right)\right]^2} \quad (1)$$

The envelope is calculated using the Hilbert transformation according to equation (2).

$$H\left(\Phi_{xy}(\tau)\right) = \sum_{u=u_{min}}^{u=u_{max}} \frac{\Phi_{xy}(u)}{\pi(\tau-u)} \quad (2)$$

The maximum peak of $E(\tau)$ determines the delay to compensate on the time abscissa.

The alignment is conducted by adding zeros at the beginning and cropping at the end of signal x'(k) in case of a positive determined delay. The inverse procedure is applied in case of a negative delay. This compensation step results in x''(k) and does not affect the degraded signal y'(k), i.e. the duration of y'(k) is maintained in both output signals.

The next block extracts the active speech parts from the signal x''(k). For this analysis, the first step is to classify energy frames of 10 ms (block-wise, no overlap) according to the method described in [i.9]. The thresholds for the classification are defined relatively to the active speech level [i.7]. 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). The speech parts are finally determined as regions excluding frames of type S. The information of the active time ranges is employed in several other blocks which are introduced in the following clauses.

The last stage performs an initial level calibration of the reference signal x''(k). For this purpose, the complex transfer function is determined by equation (3).

$$H(\mathrm{f}) = \frac{S_{x''y'}(f)}{S_{x''x''}(f)} \quad (3)$$

This calculation is also known as *method H1* in literature, where noise is located at the output of a system. Here $S_{x''y'}(f)$ denotes the cross-power spectral density between x''(k) and y'(k), $S_{x''x''}(f)$ represents the power spectral density of x''(k). The analysis is carried out only for the active speech segments determined previously. The gain a_{lin} required for the level calibration of x''(k) is obtained by averaging the magnitude H(f) over the entire frequency range. The scaled version of the reference signal is finally denoted as x'''(k). For later application, the active speech level of x'''(k) according to Recommendation ITU-T P.56 [i.7] is calculated as $ASL_{x'''}$.

6.3.3 Spectral transformation

Figure 6.3 depicts the flow chart of the spectral transformation and further processing steps, which are performed for the instrumental assessment of S-MOS, N-MOS and G-MOS.

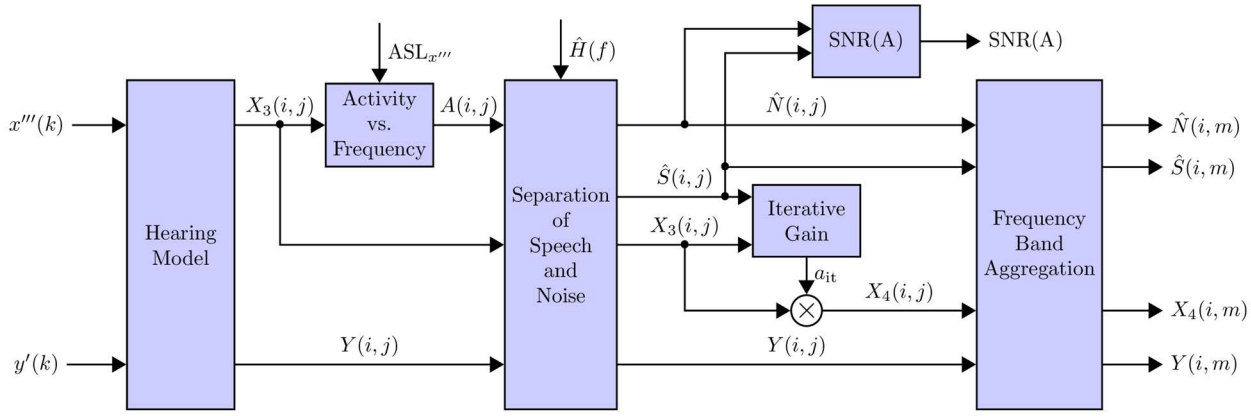


Figure 6.3: Block diagram of model A

For the consideration of a hearing-adequate signal representation, the model according to [i.20] and [i.21] is used in the spectral analysis of the pre-processed time signals. At this stage, only outer and middle ear filtering, the auditory filter bank, rectification, low-pass filtering and downsampling are applied, which result in a frame size of 8 ms. The bandwidth of the auditory filterbank is specified by equation (4).

$$\Delta f(f_m) = \Delta f(f = 0) + c \cdot f_m \tag{4}$$

The initial bandwidth Δf(f=0) is set to 50 Hz and the factor c equals 0,14. This results in M = 33 bands up to 20 kHz. In addition to the specified basis filters, two intermediate filters are inserted between the neighbouring center frequencies of the 33 frequency bands. This results in an overlap of 66 % in the frequency domain, leading to a finer resolution of J = 99 bands. For further reference, the complete hearing model is described at full length in [i.20] and [i.21].

This intermediate representation results in a hearing model spectrum versus time, but still in the physical unit Pascal. The time signals x'''(k) and y'(k) are transformed into the corresponding time-frequency representations X_3(i,j) and Y(i,j), respectively.

The next stage is the estimation of the speech activity versus time and frequency. The thresholds for activity in each band are defined by a long-term average speech spectrum according to [i.22]. Such a spectrum is created by interpolating the given frequency vector and then scaling it to the overall energy of the previously determined $\mathrm{ASL}_{x'''}$. The limits for low (L), mid (M), high (H), uncertain (U) and silence (P) for each frequency band are then determined again by the offsets described in [i.9]. As a result, an activity spectrum A(i,j) is obtained.

One of the most important stages of the spectral analysis is the separation of the degraded spectrum Y(i,j) according to equation (5) into a speech and noise component.

$$Y(i,j) = \hat{S}(i,j) + \hat{N}(i,j) \tag{5}$$

Listening tests according to Recommendation ITU-T P.835 [i.5] usually contain stimuli with at least a residual amount of background noise. Due to the specific questionnaire of this auditory test method, test subjects are urged to differentiate between disturbances caused by speech distortions (SIG scale) or by noise intrusiveness (BAK scale). In consequence, even in case of very noisy stimuli low speech distortion judgments are observed. Obviously, participants in such an auditory test are capable of separating speech and noise in the presented sounds. To address this ability of human perception and/or cognition in the instrumental assessment as well, the spectral representation Y(i,j) of the degraded signal is separated into the two components speech-only ($\hat{S}(i,j)$) and noise-only ($\hat{N}(i,j)$).

Similar to noise reduction techniques, the method for the separation is based on a Wiener filter in the frequency domain according to equation (6).

$$\hat{S}(i,j) = W(i,j) \cdot Y(i,j) \tag{6}$$

The Wiener gain W(i,j) is defined according to equation (7) and usually needs an accurate estimate of the real noise and speech component.

$$W(i,j) = \frac{S^2(i,j)}{S^2(i,j)+N^2(i,j)} \quad (7)$$

As an estimate for the speech component $S^2(i,j)$, the previously determined transfer function H(f) is applied to the reference spectrum $X_3(i,j)$. Linear distortions in the degraded signal can be explained by taking this filter into account and thus provide a better speech-only estimation.

A simple estimate of the noise component in dB can be derived from the previously used spectra according to equation (8). Here just the difference between degraded and filtered reference spectrum is evaluated.

$$N_0(i,j) = 20 \cdot \log_{10}(\max(0, Y(i,j) - H(j) \cdot X_3(i,j))) \quad (8)$$

However, due to strong non-linear processing in a terminal, this estimate may be rather unprecise - especially in time- frequency bins with high speech activity. As a refinement, the previously determined activity spectrum A(i,j) is taken into account. First, a mask M(i,j) is derived from the activity and the values provided in table 6.1.

Table 6.1: Mask values

Activity class	Mask value
Silence (S)	1,0
Uncertain (U)	0,4
Low (L)	0,2
Mid (M)	0,0
High (H)	0,0

The mask value indicate reliable (silence, value = 1), unreliable (mid and high activity, value = 0) or intermediate time-frequency bins for the noise estimation.

To exclude the impact of this mask function, an iterative spectral deconvolution according to [i.20] is conducted for each frequency band. The proposed deconvolution algorithm originally was intended for the use with audio signals. The problem to be solved is defined by equation (9) for a fixed frequency j', i.e. the deconvolution in the FFT domain. Thus, the method reconstructs windowed signal parts based on an interpolation of the dominating frequency lines in the spectrum.

$$N_0(i,j') \cdot M(i,j') \rightarrow FFT\big(N_0(i,j')\big) * FFT(M(i,j')) \quad (9)$$

The deconvoluted noise spectrum $N_1(i,j)$ is then used for the Wiener filter gain, which can now be specified according to equation (10).

$$W(i,j) = \frac{(H(j)\cdot X_3(i,j))^2}{(H(j)\cdot X_3(i,j))^2 + N_1^2(i,j)} \quad (10)$$

By providing both, speech and noise spectrum separately according to equations (5) and (6), a simple SNR calculation can be performed. First, $\hat{S}(i,j)$ and $\hat{N}(i,j)$ are averaged versus time (speech: active parts only), which results in two spectra, $\hat{S}(j)$ and $\hat{N}(j)$. To roughly address the human frequency-dependent loudness perception, an A-weighting function A(j) according to [i.23] is applied to the noise spectrum. The modified SNR(A) in dB is calculated finally by equation (11).

$$SNR(A) = 20 \log_{10} \sum_{j=1}^{33} \frac{\hat{S}(j)}{A(j)\hat{N}(j)} \quad (11)$$

As described in the previous clause 6.3.2, the input signal x'''(k) includes already a level calibration in order to achieve a similar speech level as the degraded signal y'(k). Since only linear distortions and uncorrelated additive components are considered by the transfer function H(f), a final adaption of the reference spectrum $X_3(i,j)$ is applied by employing the estimated speech spectrum $\hat{S}(j)$. Starting with an offset of a_{it} = 0 dB, an iterative bisection method (initial stepsize 3 dB) is applied to achieve a suitable correction factor. In each iteration, the number n_{above} of time-frequency bins of $[a_{it} \cdot X_3(i,j)]$ are counted, which contain a higher amplitude than the separated speech spectrum $\hat{S}(j)$. Then, a ratio $r = n_{above}/n_{total}$ is calculated by normalizing to the total amount of bins. The iteration scheme is repeated until $r \approx 30$ % and r is not changed anymore by increasing/decreasing a_{it}. This method of counting is applied again only to the active signal parts.

Figure 6.4 exemplarily illustrates the iterative method for one time instance. After the first iteration, the reference level is too low (0 % of bins above). After setting a higher scaling factor a_{it}, the reference level is too high (62 % of bins above). By applying some more iteration, the spectrum has the right scaling.

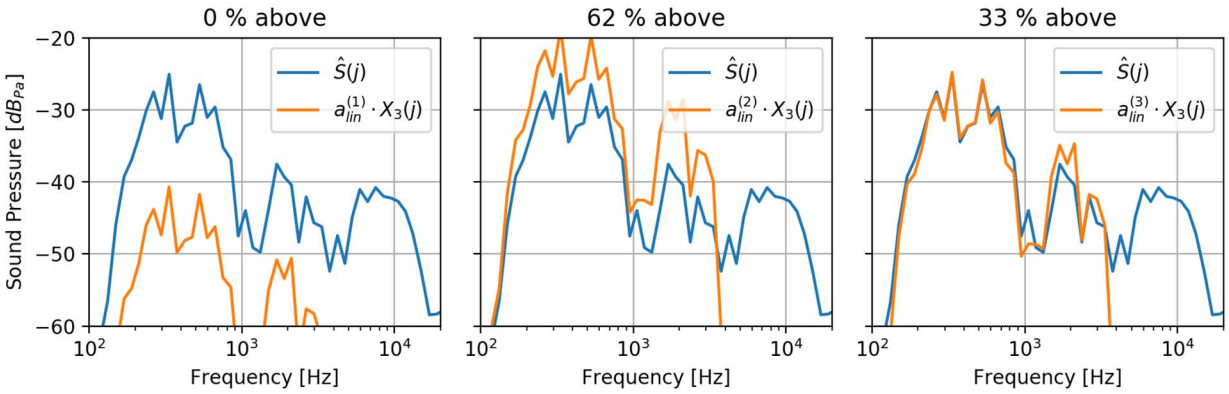


Figure 6.4: Iterative level alignment

Finally, all spectral representations are aggregated back to the basic M = 33 hearing-adequate frequency bands. This is conducted simply by the quadratic mean of three consecutive bands. This includes a change of the frequency index from j to m (e.g. Y(i,j) results in Y(i,m) after aggregation).

6.3.4 Non-linear loudness transformation

The determined spectral representations are transformed with a hearing-adequate filterbank, but are still provided in the physical unit Pascal. The real loudness as perceived by humans is not yet considered. For this purpose, the aforementioned hearing model introduces a non-linear transformation to address the compressed loudness perception at increased input levels. A detailed description of the transformation can be found in [i.20] and [i.21].

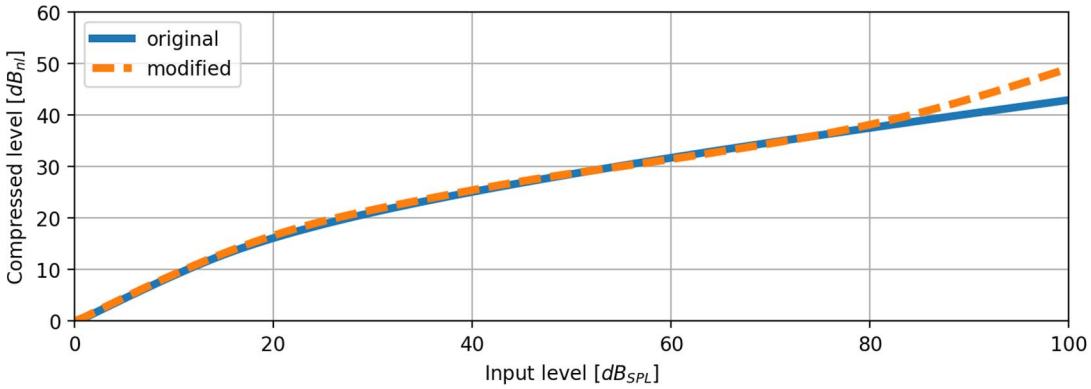


Figure 6.5: Non-linear sound pressure transformations

Figure 6.5 shows two possible non-linearity functions, derived from several psycho-acoustic loudness investigations. The original function (blue curve of figure 6.5) is described in [i.20] and in annex K of ETSI EG 202 396-3 [i.2]. Recent developments proposed also a modified version according to equation (26) of [i.21] (orange curve of figure 6.5). Both functions are used in the further stages of the model.

6.3.5 Instrumental assessment of N-MOS

6.3.5.1 Introduction

The flow chart for the instrumental assessment of N-MOS is depicted in figure 6.6. Several metrics are derived from the spectral representation and then mapped versus the auditory BAK results. Only the separated noise estimation $\hat{N}(i,m)$ is evaluated for the determination of these metrics.

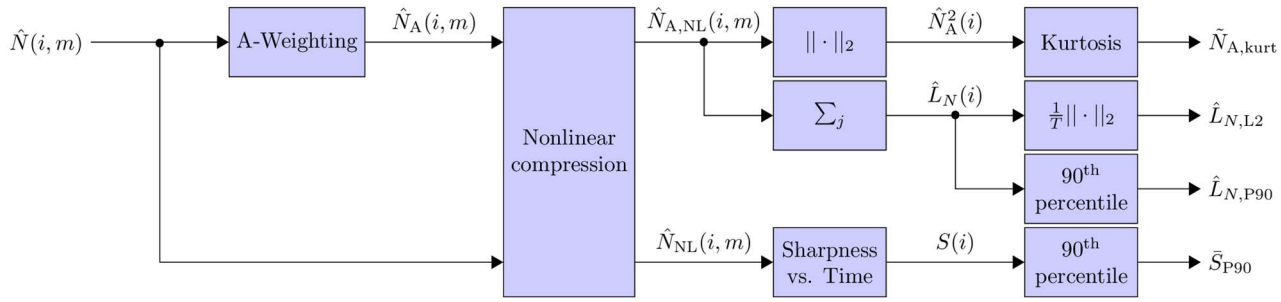


Figure 6.6: Instrumental assessment of N-MOS

6.3.5.2 Loudness-based features

The first step of the calculation is to apply an A-weighting function [i.23] to the noise-only spectrum (output of spectral transformation stage), which results in a weighted version $\widehat{N}_A(i,m)$. Both, weighted and unweighted noise spectra are then transformed according to the original non-linear transformation of sound pressure (see clause 6.3.4), providing the spectra $\widehat{N}_{A,NL}(i,m)$ and $\widehat{N}_{NL}(i,m)$.

The weighted version of the transformed noise spectrum is used for three metrics. The first one applies a L2-norm to the spectrum vs. frequency, i.e. aggregating to an average noise versus time ($\widehat{N}^2{}_A(i)$). In order to obtain a single value, the Kurtosis is calculated from the time series as $\widetilde{N}_{A,kurt}$ to describe the temporal peakedness/variance of the noise.

The next two parameters are calculated in a similar way. First, a summation over frequency is performed to provide a loudness vs. time $\hat{L}_N(i)$. An estimator for the overall loudness of the noise $\hat{L}_{N,L2}$ is then obtained by applying an averaged L2-norm versus time. To provide additional information about the maximum loudness within the noise, the parameter $\hat{L}_{N,P90}$ is calculated as the 90 % percentile of $\hat{L}_N(i)$.

6.3.5.3 Sharpness-based feature

The previously introduced metrics do not explicitly consider the higher frequency range, which is even more attenuated by the A-weighting function. For this purpose, the unweighted version of the transformed noise spectrum is used for a sharpness metric. This well-known psycho-acoustic measure especially addresses high-frequency components in a signal. According to [i.26], it can be calculated as the weighted first moment of the specific loudness, i.e. the non-linear transformed sound pressure in this case. Sharpness versus time is then determined according to equation (12). The weighting function g(m) according to equation (13) from [i.26] is used here.

$$S(i) = \frac{\int_{m=1}^{M} \widehat{N}_{NL}(i,m) \cdot g(m) \cdot m \cdot dz}{\int_{m=1}^{M} \widehat{N}_{NL}(i,m) \cdot dz} \tag{12}$$

$$g(b) = \max(1;\, 0{,}066 \cdot e^{0.171 \cdot b}) \tag{13}$$

Note that the center frequencies of the current hearing model are converted into fractional Bark band indices (b) to obtain a certain weight. Figure 6.7 illustrates the weighting curve versus real frequencies. For the aggregate versus time, 90 % percentile is used to obtain the single value $\bar{S}_{P90}$.

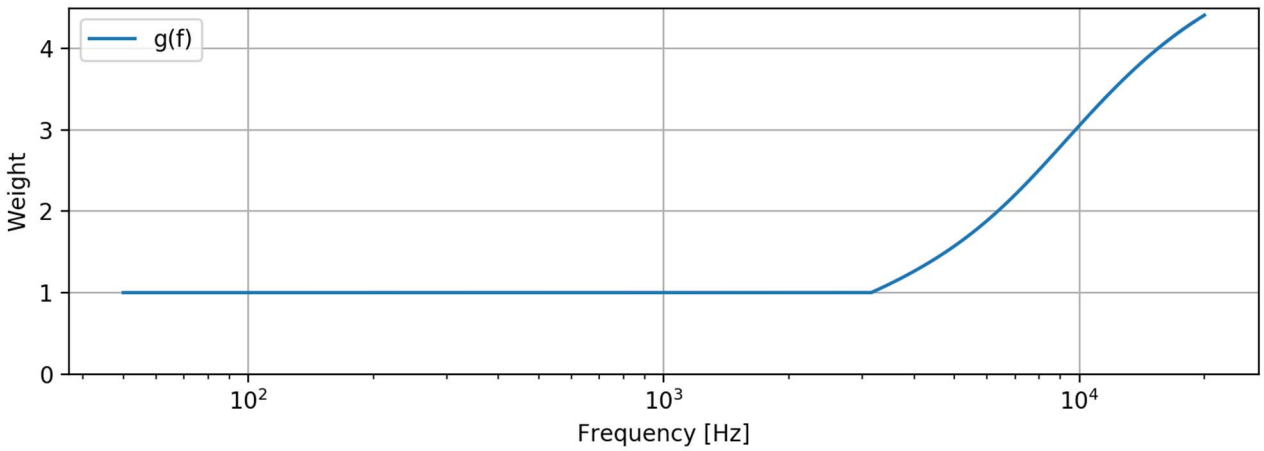


Figure 6.7: Frequency weighting for sharpness analysis

To combine the four determined features to the instrumental N-MOS, it is a common way to utilize machine learning techniques for training. According to figure 6.8, a random forest regressor as described in [i.28] is used in this case. This regression model originates from the classification domain and provides a good prediction in case of input variables which can be seen more as "labels". For example, an increased sharpness measure ($\bar{S}_{P90}$) does not necessarily lead to a decreased auditory score. Only in combination with e.g. increased loudness indicator ($\hat{L}_{N,L2}$) judgements of BAK degrade.

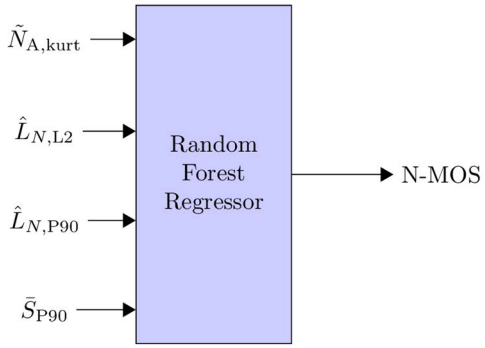


Figure 6.8: Combination of features to instrumental N-MOS

Table 6.2 provides the parameters of the random forest regressor which were used for the training process.

Table 6.2: Parametrization of random forest regressor for N-MOS

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

6.3.6 Reference optimization and asymmetry

6.3.6.1 Introduction

A common processing step in speech quality prediction algorithms (e.g. [i.24] or [i.25]) is the so-called reference optimization. When comparing spectral representations of a degraded and a reference signal e.g. by a subtraction, often differences occur which are not perceived by test subjects. The aim of this optimization stage is to manipulate the reference in a way that spectral differences which do not contribute to perceived differences are compensated in advance for later stages. Figure 6.9 illustrates the different blocks of the procedure, which can either be applied on the separated speech $\hat{S}(i,j)$ or on the noisy degraded spectrum $Y(i,j)$.

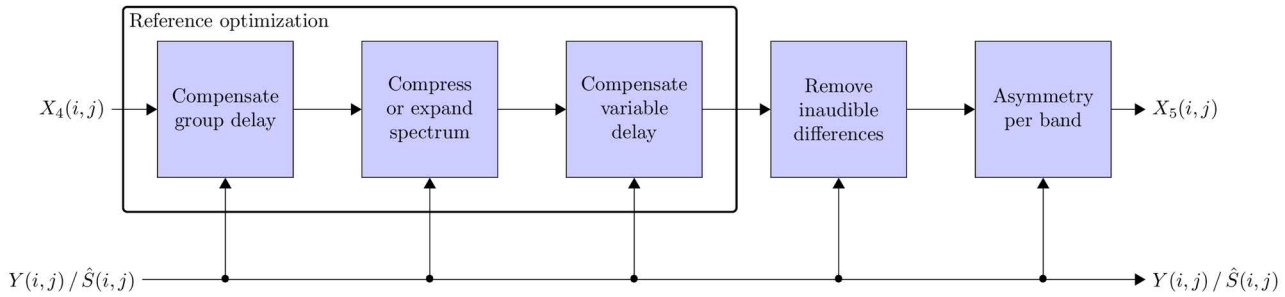


Figure 6.9: Reference Optimization and asymmetry

6.3.6.2 Reference optimization

The reference optimization compensates inaudible signal modifications between reference and degraded. All optimizations are only conducted on the reference. This processing is divided into three parts:

- Compensation of the group delay: differences in group delays between reference and degraded signal are almost inaudible, but can lead to systematic offsets in instrumental assessment. To compensate this effect, in each frequency band the magnitude versus time is first upsampled by a factor of four. Then, the delay per band is determined and compensated. Finally, downsampling by a factor of four obtains the previous time base.

- Compression/expansion of spectrum: several signal processing algorithms and codecs perform manipulation of the fundamental and/or harmonic speech frequencies. Up to a certain degree, this impact is not audible. For the compensation of this effect, each spectrum at a time index is first interpolated by a factor of four. To address the logarithmic frequency resolution, the distance between each frequency band is assumed as an equal distance. For a range of ±3 neighbouring interpolated subbands, the mean-square-error between reference and degraded spectrum at this time instance is determined. The best matching shift (in terms of minimum error) is applied for each time frame. Finally, the spectrum is downsampled back to the original domain.

- Compensation of variable delay: drift (also known as clock skew) of audio signals between reference and degraded signals is also an inaudible effect which may impact instrumental assessment with difference-based measures. To compensate this, the whole spectrum is first upsampled versus time by a factor of four. Similar to the previous optimization step, now the mean-square-error between reference and degraded spectrum of the neighbouring ±3 time sub-instances are evaluated. The best matching temporal shift (in terms of minimum error) is selected for each original frame and is inserted directly into the output spectrum (no further downsampling).

6.3.6.3 Masking of inaudible differences

So far, reference optimization only considered effects of the signal processing of the degraded signal which are inaudible to human listeners. Beside this, there are also psycho-acoustic masking effects which can impact the audible differences between degraded and reference signal. In this stage, the masking model described in [i.25] is used. Basically the method detects the maximum level per frequency band for each time instance and derives a simple masking of the high frequency range.

6.3.6.4 Asymmetry

The final block of this stage is the application of the so-called "asymmetry". This principle was already introduced in several instrumental speech quality assessment methods, like e.g. [i.29], [i.25] or [i.24]. In auditory tests, additional signal components are much more annoying than components which are removed. A simplified asymmetry is applied to time-frequency bins of an arbitrary difference $\Delta(i,m)$ as shown in equation (14). The factor α controls the amount of asymmetry and is defined as a value larger or equal 0. Lower factors increase the asymmetry effect.

$$\Delta(i,m) = \begin{cases} \Delta(i,m), & \Delta(i,m) \leq 0 \\ (1+\alpha)\cdot\Delta(i,m), & \Delta(i,m) > 0 \end{cases} \tag{14}$$

6.3.7 Instrumental assessment of S-MOS

6.3.7.1 Introduction

For the instrumental assessment of S-MOS, two different kinds of features are extracted from the spectral representations. The first type of metrics is intended to detect disturbances at certain modulation frequencies and detects degradations versus frequency. The second one is intended to determine spectral differences versus time, which is a well-known principle, as already used in e.g. [i.25] or [i.24].

6.3.7.2 Modulation-based features

Figure 6.10 provides a flow chart of the calculation of the two modulation-based features. The spectral representations of the separated speech $\hat{S}(i,j)$ and the pre-processed, but non-optimized reference $X_4(i,j)$ are used as the inputs of this block.

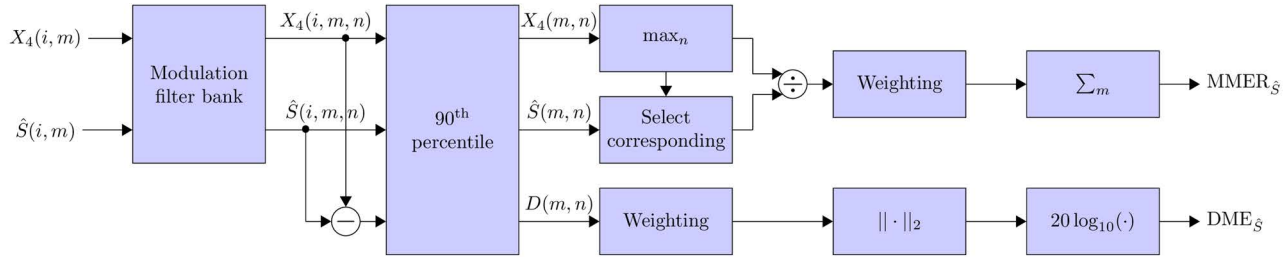


Figure 6.10: Structure of modulation-based feature extraction

In a first analysis, both spectra are transformed by a modulation filter bank according to [i.27]. This analysis provides N = 8 modulation sub-bands per frequency index m. The modulation center frequencies $f_{c,mod}(n)$ range from 4 to 128 Hz. Due to the output frame step size of 8ms (125 Hz), the spectrum is upsampled vs. time by a factor of four, resulting in a sampling rate of 500 Hz and time index i'. In addition to the two spectra, the difference between $\hat{S}(i',m)$ and $X_4(i',m)$ is evaluated as $D(i',n)$. The 4D-representations are then first aggregated versus time. To address the active time ranges, the 90 % percentile is used. As an intermediate result, three average spectra versus frequency index m and modulation index n are obtained.

To emphasize critical frequencies on the one hand and slightly depreciate low and high frequencies, a modified A-weighting function is derived according to equation (15). $A(m)$ denotes the default A-weighting from [i.23].

$$A_{clip}(m) = \max(-10\ dB, A(m)) \tag{15}$$

This weighting is applied on the difference spectrum $D(m,n)$, followed by an L2-norm and converting into dB-value. This metric $DME_{\hat{S}}$ in general describes the energy difference across all frequencies and modulation bands.

In another branch, the second measure is derived from the principle described in [i.27]. For each frequency m in the reference spectrum $X_4(m,n)$, the maximum magnitude of all modulation bands is determined. The corresponding modulation bands n_{max} are then selected in the degraded speech spectrum $\hat{S}(m,n)$ as well. The maximum modulation energy ratio $MMER_{\hat{S}}$ is then given by equation (16).

$$MMER_{\hat{S}} = 20 \cdot \log_{10} \sum_{m=1}^{M} A_{clip}(m) \cdot \frac{\hat{S}(n_{max},m)}{X_4(n_{max},m)} \tag{16}$$

6.3.7.3 Spectral difference features

A more classic approach for the instrumental assessment of speech degradation is the evaluation of spectral differences of loudness. The differences of the non-linearly compressed sound pressure level are used here. Figure 6.11 illustrates the flow chart for the determination of several features.

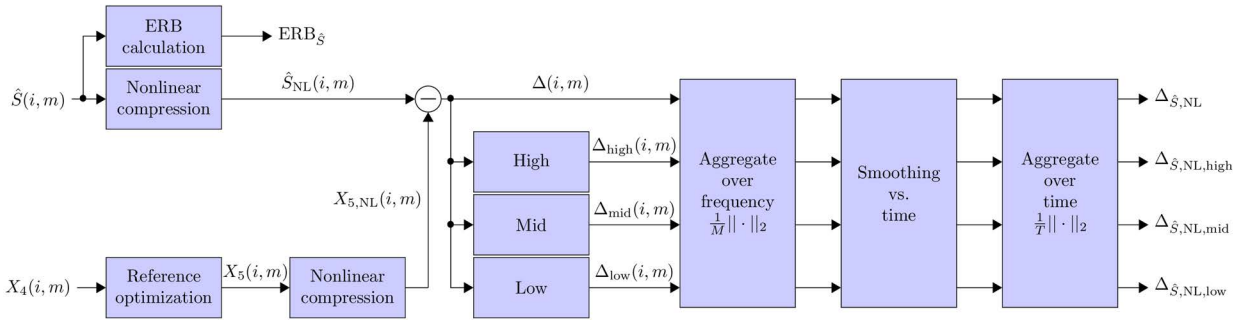


Figure 6.11: Structure of spectral difference features

In a first step, the non-linearity is applied on both input spectra $\hat{S}(i,m)$, resulting in the compressed version $\hat{S}_{NL}(i,m)$. $X_5(i,m)$ is obtained by the reference optimization according to the previous clause and is then compressed in the same way to $X_{5,NL}(i,m)$. The difference $\Delta(i,m)$ between these two spectra is then separated into three frequency bands (low/mid/high), which contribute in different ways to speech quality perception. Table 6.3 provides an overview over this division, here also different asymmetry factors are evaluated.

Table 6.3: Definition of frequency bands

Range	Frequency start/end [Hz]	Asymmetry factor α
Low	50 - 3 800	0,25
Mid	3 800 - 7 800	0,40
High	7 800 - 20 000	0,50

All four delta spectra are processed in the same way: First, an aggregate versus frequency is performed by an L2-norm and averaging over the amount of corresponding frequencies. Note that this normalization was not performed in the instrumental assessment of N-MOS.

The next step is the smoothing of the delta curve versus time. Here a median filter with a window size of 40 ms is used. Finally, the single values according to figure 6.11 are obtained by again applying an L2-norm versus time, including normalization over the active time frames (see clause 6.3.3).

6.3.7.4 Control parameters

In addition to the psycho-acoustically motivated metrics of the previous clauses and the aforementioned SNR(A), two so-called control parameters are introduced. These single values do not necessarily correlate directly with perceived speech quality, but provide a weighting for the other metrics.

The equivalent rectangular bandwidth $ERB_{\hat{S}}$ of $\hat{S}(i,m)$ is calculated according to the description in [i.25]. For this purpose, the transfer function $\widehat{H}_{\hat{S}}(j)$ is estimated in the logarithmic domain according to equation (17). Only time frames of active speech (uncertain, low, mid, high, see clause 6.3.3) are taken into account.

$$\widehat{H}_{\hat{S}}(j) = 20 \cdot \log_{10}\left(\frac{1}{N_{active}}\sum_{i\in U,L,M,H}\hat{S}(i,m)\right) - 20 \cdot \log_{10}\left(\frac{1}{N_{active}}\sum_{i\in U,L,M,H}X_4(i,m)\right) \quad (17)$$

According to [i.25], an intermediate spectrum $G_{\hat{S}}$ is calculated according to equation (18).

$$G_{\hat{S}}(j) = \max\left(\widehat{H}_{\hat{S}}(j) + 45\ dB;\ 0\ dB\right) \quad (18)$$

The single value $ERB_{\hat{S}}$ in Hz is then provided as the ratio between the area determined below $G_{\hat{S}}(j)$ and the maximum according to equation (19).

$$ERB_{\hat{S}} = \frac{area\{\,G_{\hat{S}}(j)\,\}}{\max\{\,G_{\hat{S}}(j)\,\}} = \frac{\sum_j\ G_{\hat{S}}(j)\cdot\Delta f(j)}{\max\{\,G_{\hat{S}}(j)\,\}} \quad (19)$$

This metric represents an ideal rectangular filter which has the same perceptual characteristics as the original transfer function $\widehat{H}_{\hat{S}}(j)$. It provides an estimate of the overall bandwidth loss and the coloration of the degraded speech.

A second control parameter is the active speech level $ASL_{\hat{S}}$.of the separated degraded speech signal. Since this signal is not available in the time domain, the method according to Recommendation ITU-T P.56 [i.7] cannot be applied. The value is thus calculated in the frequency domain, followed by averaging again over the active time frames as shown in equation (20). This parameter provides an information about the loudness of the speech signal without the impact of residual noise.

$$ASL_{\hat{S}} = 20 \cdot \log_{10}\left(\frac{1}{N_{active}} \sum_{i \in U,L,M,H} \sum_{j} \sqrt{\hat{S}^2(i,m)}\right) \quad (20)$$

6.3.7.5 Combination of features

As for the N-MOS determination, the derived features again are combined by a random forest regression model. Table 6.4 provides the parameters for the regressor used for the training.

Figure 6.12 illustrates the inputs which are found to be crucial for the instrumental assessment of speech distortion component (SIG). These input features can be divided into three groups:

- Modulation-based features: metrics as described in clause 6.3.7.2.
- Spectral difference features: metrics as described in clause 6.3.7.3.
- Control parameters as described in clause 6.3.7.4.

Table 6.4: Parametrization of random forest regressor for S-MOS

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

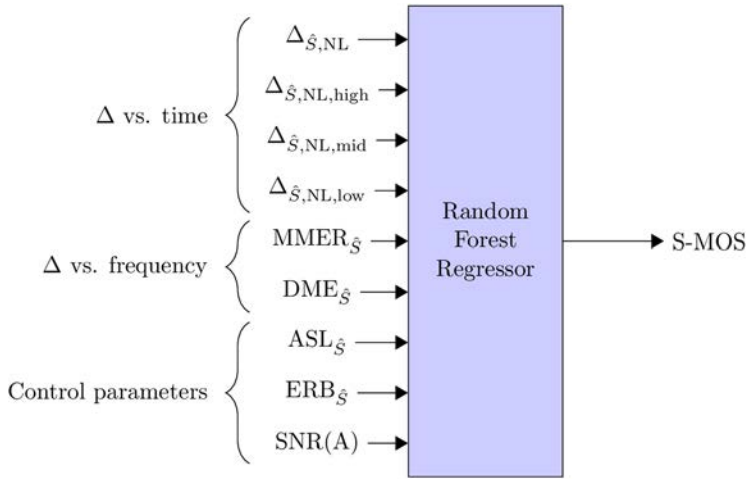


Figure 6.12: Combination of features to instrumental S-MOS

6.3.8 Instrumental assessment of G-MOS

The instrumental assessment of G-MOS is based on three input features. Similar to the prediction model in [i.2], the previously determined values of S-MOS and N-MOS are used here as well. Obviously, this usage is close to the judgement of test subjects in an auditory test. Several investigations already indicated that results of the OVRL scale can be estimated by the votes of SIG and BAK.

If only one single listening test database would be included for the training of the model, these two inputs would provide a sufficient prediction accuracy. However, due to the set of quite inhomogeneous training databases, the composition of SIG and BAK to OVRL is not always a constant mapping. To address this behaviour, $\Delta_{Y,NL}$ is used as an indicator for this issue. This value is calculated in exactly the same as $\Delta_{\hat{S},NL}$ according to figure 6.11, but using $Y(i,m)$ as an input to the calculation instead of $\hat{S}(i,m)$.

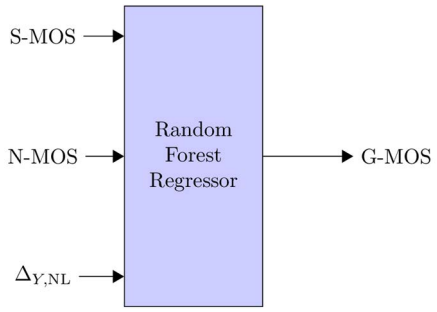


Figure 6.13: Combination of features to instrumental G-MOS

The principle for the determination of G-MOS and its corresponding input features are shown in figure 6.13. As for N-MOS and S-MOS, a random forest regression model is used for the prediction. Table 6.5 provides the parameters for the regression model used for the training process.

Table 6.5: Parametrization of random forest regressor for G-MOS

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

6.4 Model B

6.4.1 Overview

The SWB/FB Cochlear model (CM) based Prediction algorithm compares a reference signal x(k) with a signal under test y(k), and estimates the results of a listening-only test on the three speech quality attributes. The three speech quality attributes and type of listening test estimated are according to Recommendation ITU-T P.835 [i.5].

Table 6.6 provides an overview of the dimension estimators for the model. The following are the input and intermediate variables used in the model:

- x(k) is the input;
- y(k) is the output of transmission system;
- x'(k) and y'(k) are the corresponding outputs of the pre-processing steps (Alignment, ASL, VAD, etc.).

Table 6.6: Overview of the dimension estimators for the prediction model

Dimension	Input	Output	Reference estimator
S-MOS	x'(k) and y'(k); time aligned, VAD; background and voiced section	(i) Salient Formant Points (SFP); (ii) CM feature set s1; i+ii -> mapped to S-MOS	[i.32] [i.33] [i.30]
N-MOS	y'(k); VAD; background section	(ii) CM feature set s2-> mapped to N-MOS	
G-MOS	y'(k); VAD; background and voiced section	(iii) CM feature set s3 -> mapped to G-MOS	

In an initial pre-processing stage, the reference speech x (k) and degraded speech y (k) are re-sampled to 48 kHz, adjusted to a reference loudness and time-aligned. A Voice Activity Detection (VAD) step is then performed, classifying the reference and degraded signals into voiced, unvoiced and background sections. This initial stage results in the intermediate signals x'(k) and y'(k).

After the pre-processing stage, an Auditory Model processing stage is applied to the intermediate signals x'(k) and y'(k). The Auditory Model is based on a Hydro-Mechanical Cochlear Model (CM). The CM takes the pre-processed Reference and Degraded Speech and produces a cochlear response for each signal.

Following the Auditory Model processing stage, a Feature Extraction stage is performed. Features are extracted from the difference between the Reference and Degraded CM response. Each input speech pair will yield a set of Features. The Feature Extraction process uses some characteristics of the CM output and, therefore, is highly dependent on the CM module.

After the Feature Extraction stage, a Mapping stage is performed. The Mapping stage maps the several Features to the three quality dimensions of Recommendation ITU-T P.835 [i.5]. The mapping is performed based on a-priori training of the model from a large set of P.835 listening test results.

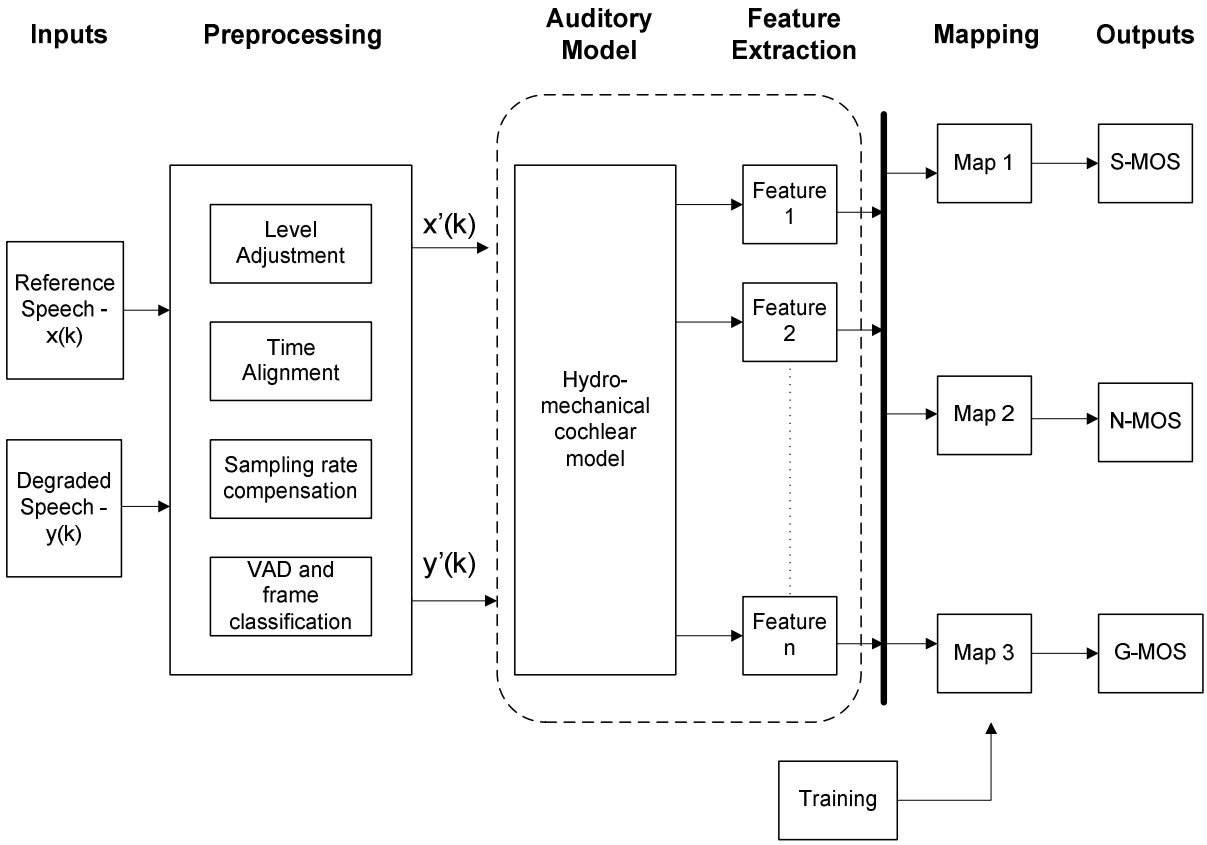


Figure 6.14: High level block diagram for the prediction model

6.4.2 Operational Modes

The model algorithm supports a single fullband operational mode and can predict scores for narrowband, wideband, super-wideband and fullband degraded speech signals. All the test conditions are re-sampled to 48 kHz prior to input into the Auditory Model. Regardless of the bandwidth of signals under test, the model algorithm always predicts scores in a fullband context. I.e. a fullband 48 kHz source signal is always present in the reference set during the listening tests.

The source signal shall be in 48 kHz format and it is recommended that the degraded signal be in 48 kHz format as well. Re-sampling of the degraded signal is performed by the algorithm as necessary.

6.4.3 Temporal Alignment

The Temporal Alignment module is currently using a subsample alignment algorithm [i.33]. This algorithm detects the delay between original and degraded wave speech by interpolation of the peak of cross correlation. A parabola curve is used for the three-point peak interpolation. The method has been verified to perform with significant amounts of background noise present.

6.4.4 Voice Activity Detection (VAD) and segment classification

The VAD module classifies speech into voiced/unvoiced/background (noise only) sections that will contribute to different distortions. The background (with background noise only) section will affect background dimensions (N-MOS). The voiced sections are the main part that decide the speech quality dimensions (S-MOS). The VAD module is a modified version of RAPT pitch tracker [i.34] from Voicebox toolbox [i.35] which tracks the Larynx frequency along time. Frames with a valid pitch estimation are marked as voiced sections.

6.4.5 Auditory Model

6.4.5.1 Introduction

Speech is converted into perceptual domain with the help of a Cochlear Model. The purpose of the Cochlear Model is to compute the Inner Hair Cell (IHC) response caused by sound impinging at the input to the Ear Canal.

The model can be conceptually divided into five parts: Ear Canal model, Middle Ear model, Hydro-mechanical model, Hair Cell Transduction model and Outer Hair Cell Motility model.

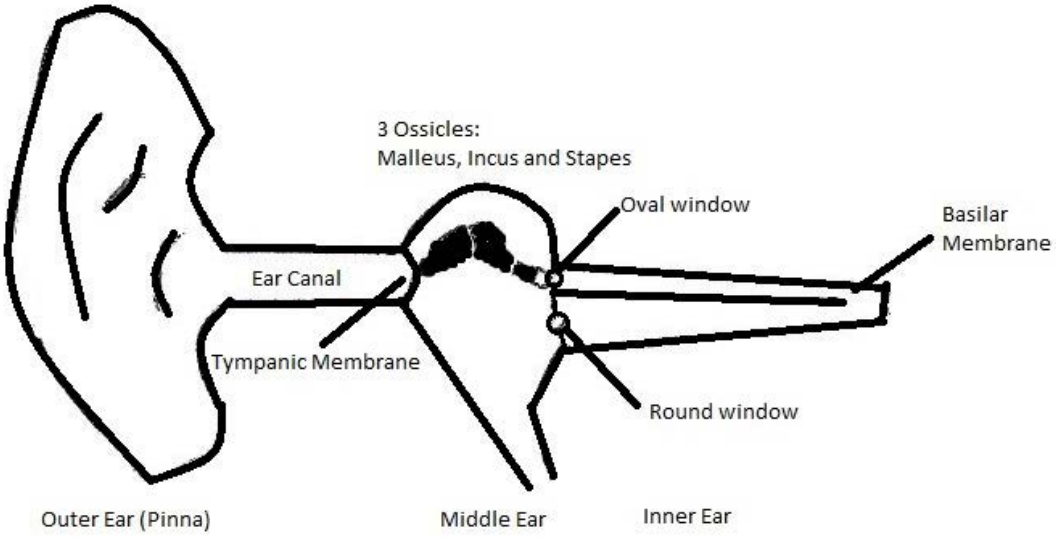
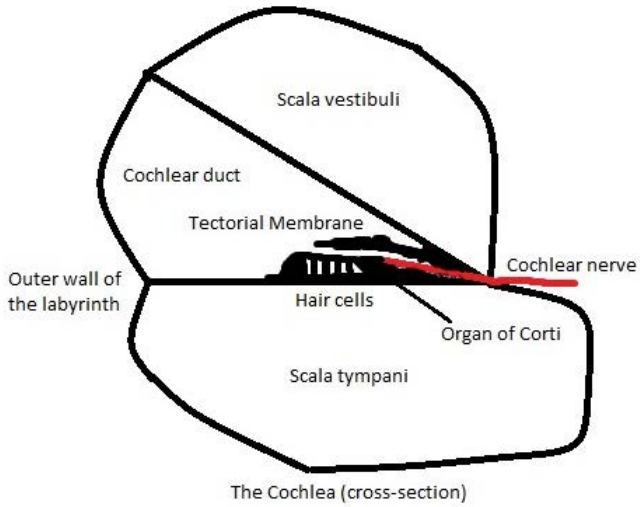


Figure 6.15: Schematic view of the Ear, with an unwrapped cochlea



NOTE: The basilar membrane is a part of the Organ of Corti.

Figure 6.16: Cross-section of the Cochlea

6.4.5.2 Ear Canal model

The ear canal model is a linear model with a primary resonance at 2 600 Hz. It models the transfer function between pressure incident at the Ear Canal opening to the pressure on the Tympanic membrane (see figure 6.15). As such, it is essentially an acoustic model of a single ended cylindrical tube [i.31]. The length of the tube is the length of a typical human ear canal. The model is implemented as a discrete IIR filter as follows:

$$P_{tm}[n] = h[n] * s[n]\,, \tag{21}$$

where $s[n]$ is the pressure signal incident at the opening of the Ear Canal and $P_{tm}[n]$ is the pressure at the Tympanic Membrane. The filter $h[n]$, is given by:

$$h[n] = \frac{\sum_{i=0}^{6} b_i[n-i]}{1+\sum_{i=1}^{6} a_i[n-i]}, \tag{22}$$

where the filter coefficients for a sampling rate of 288 kHz (see note in clause 6.4.5.3) are given by:

$b_i = \{0.4810, -1.8446, 2.2117, 0.0722, -2.2488, 1.7724, -0.4439\}$ and

$a_i = \{-5.5360, 12.8779, -16.1201, 11.4586, -4.3882, 0.7079\}$.

6.4.5.3 Middle Ear model

The middle ear model emulates the transfer function between pressure incident at the Tympanic membrane and the Volume Velocity of the Stapes (see figure 6.15). The model can be viewed as a transformer in series with the impedance of the Middle Ear and the input impedance of the Cochlea. The Middle Ear impedance is given by the mass, stiffness and damping of the combined Ossicles. The Cochlear terminating impedance is derived from [i.36]. The transformer gain models three physiological factors:

i) the length difference between the Malleus and the Incus;

ii) the difference in area between the Tympanic membrane and the Stapes footplate; and

iii) the buckling factor due to the conical shape of the Tympanic membrane. The final transfer function is given by the following:

$$U_s[n] = h[n] * P_{tm}[n]\,, \tag{23}$$

where $U_s[n]$ is the Volume Velocity of the Stapes. The filter $h[n]$, is given by:

$$h[n] = \frac{\sum_{i=0}^{4} b_i[n-i]}{1+\sum_{i=1}^{4} a_i[n-i]}, \tag{24}$$

where the filter coefficients for a sampling rate of 288 kHz are given by:

$b_i = \{0.2056, -0.4957, 0.2742, 0.1161, -0.1003\} \times 10^{-6}$ and

$a_i = \{-3.8758, 5.6312, -3.6350, 0.8795\}$.

NOTE: The model is implemented completely in the time domain. Due to discretization methods used in the model, as well as noise considerations inherent in nonlinear feedback systems, stability of the model is guaranteed only when it is run at a sampling rate considerably above the Nyquist sampling rate [i.39]. To adhere to this requirement, 48 kHz sampled acoustic stimuli used is required to be up-sampled by a factor of six (to 288 kHz) before being processed by the cochlear model. Input to the cochlear model is on a sample by sample basis. Thus, for every sample into the model there is effectively a frame of 512 points of spatial data at the output. Every five out of six frames are discarded, which has the effect of temporal down-sampling back to 48 kHz. A drawback of the use of the cochlear model is that it is highly redundant - due to the fact that the output is a 512 times oversampled relative to the input stimuli. This necessitates dimensionality reduction and to achieve this, distinct features are extracted from the model response. In particular, features which correspond to the perception of the temporal and frequency localized distortions are isolated.

6.4.5.4 Hydro-mechanical cochlear model

The cochlear model (CM) is a spatially two-dimensional hydro-mechanical model, which computes various electrical and mechanical responses in the cochlea. In particular, the model can be used to calculate Basilar Membrane (BM) and Inner Hair Cell (IHC) response as a function of time and space. Detailed aspects of the cochlear model are available in the literature [i.37], [i.38] and [i.39]. Various benchmarks comparing the model output to physiological and psychophysical data have been carried out to verify the performance of the model [i.37].

The macro-mechanical model is concerned with the dynamics of the fluid filled scalae and the Organ of Corti along the length of the cochlea. Of particular relevance is the travelling wave type mechanical response of the basilar membrane (BM). A Green's function [i.39] is used to numerically solve (in the time domain) the differential equations that result from assumptions of continuity (or conservation of fluid mass), inviscid and incompressible cochlear fluid loaded by the mass/stiffness and damping of the fluid and structures along the length of the cochlea. Spatial sampling is achieved by linearly discretizing the cochlea at 512 points along the 3,5 cm length of the cochlea.

The micromechanical model described in [i.38] is concerned with the cilia (submerged between the tectorial membrane and the BM) and the associated Inner (IHC) and Outer (OHC) Hair Cells. The movement of the cilia is modelled as the direct result of the shear force created within the subtectorial space as a result of the relative movement of the BM to the Tectorial membrane (TM). The TM is modelled as a transmission line, terminated by the cilia. The phenomenological result of the micromechanical model is a cilia response that reflects an attenuated BM response basal to the Characteristic Place (CP). The mechanical cilia displacements are rectified and low-passed to derive the OHC and IHC receptor potentials.

6.4.5.5 Hair Cell transduction model

As mentioned above, the OHC and IHC outputs are derived from the rectified and low-pass filtered cilia displacement. This is shown in figure 6.17 and figure 6.18. The IHC and OHC models are thus alike except for a high-pass filter that precedes the IHC model to account for the fact that the IHC cilia are not attached to the TM, but are driven by viscous fluid drag. The IHC response from the model are reflective of receptor potentials, however no attempt is made to normalize them to units of Volts. The IHC responses are used as the output of the cochlear model and referenced as the cochlear model response.

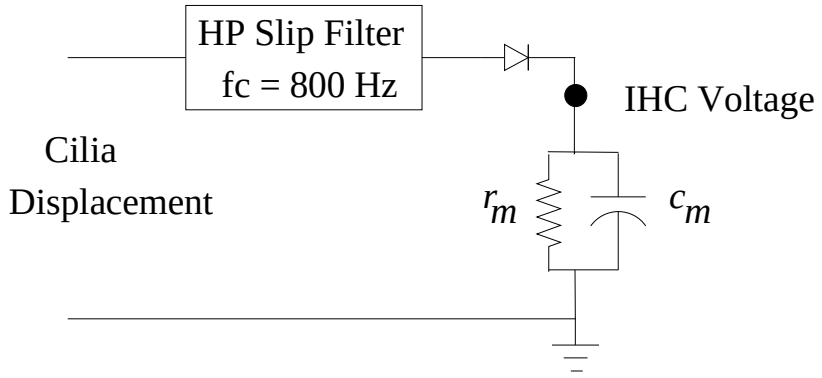


Figure 6.17: Inner Hair Cell (IHC) model

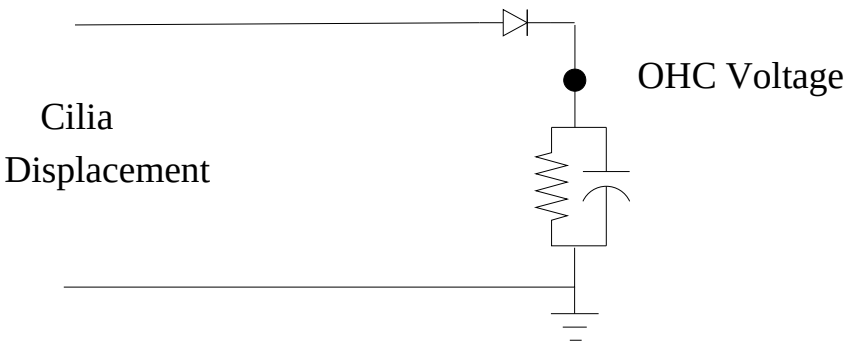


Figure 6.18: Outer Hair Cell (OHC) model

6.4.5.6 Outer Hair motility model

The cochlear non-linearity imposed by OHC motility is modelled as mechanical feedback from the OHC, which modifies the macro-mechanical impedance. This is shown in figure 6.19 as "Slow Acting Active Feedback".

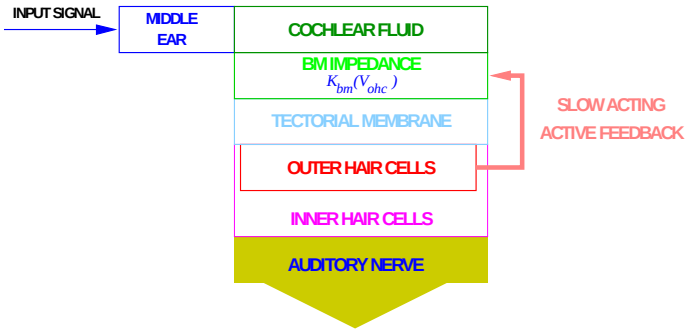


Figure 6.19: OHC motility

6.4.6 Feature Extraction

6.4.6.1 Introduction

Perceptual features are extracted from the CM output before mapped to objective prediction scores. There are two main features that are used. The Salient Formant Points (SFP) are used for time-related distortion measurement. The Cochlear Output Statistical Metric (COSM) are used for frequency related distortion (S-MOS) and noisy distortion (N-MOS). Further details are explained in clauses 6.4.6.2 and 6.4.6.3.

6.4.6.2 Salient Formant Points (SFP) feature extraction

The aim of the SFP feature set is to isolate and predict the temporally localized distortions such as single 'clicks' and 'pops', but also more temporally dense distortions which produce the perception of 'harshness'. Generally speaking, and in line with Principal Component Analysis (PCA), the distortions can be classified into a 'slow' and 'fast' mode. The methodology for their extraction is shown in figure 6.20.

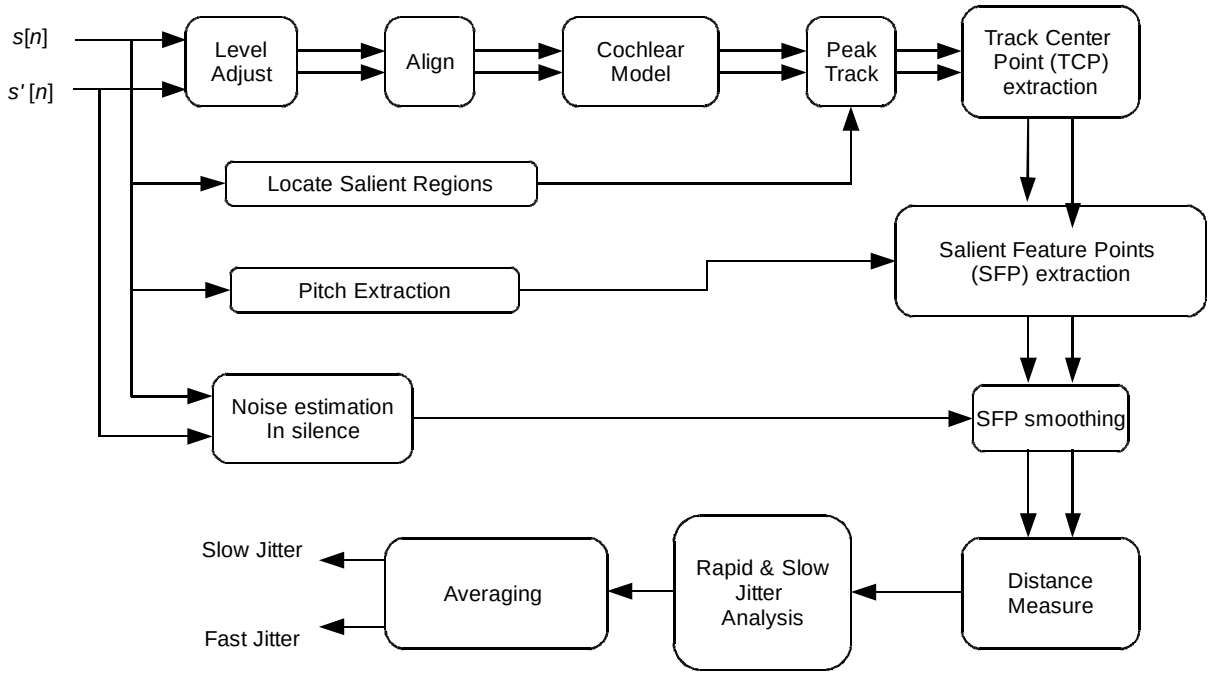


Figure 6.20: Model steps to extract the two types of temporally localized distortions (slow and fast jitter)

A detailed description of the SFP Feature Extraction methodology can be found in [i.33]. A brief description follows.

i) The original and degraded speech are level adjusted to -26 dBov and time aligned as described in clause 6.4.4.

ii) The signals are subsequently classified in voiced/unvoiced and background sections before being passed through the Cochlear Model (CM) which produces a three dimensional output (versus time, place (frequency) and IHC).

iii) A peak-tracking algorithm is used to determine the peaks over time and place. The peak tracks along with the CM output are shown in figure 6.21.

iv) A temporal 'center of mass' computation over pitch periods on the tracks from the previous step produces "Track Center Points" (TCP), following the equation:

$$TCP\ (t) = \frac{\sum_{i=1}^{N} t_i * w_i}{\sum w_i} \quad (25)$$

v) The TCPs are further subjected to a spatial 'center of mass' computation over pitch periods to extract the so-called Salient Feature Points (SFPs), here the time aspect in above equation is replaced by a frequency aspect.

vi) The SNR is computed based on the noise level in the background sections of the degraded speech signal.

vii) The SFPs are pitch independent and display robust alignment properties [i.33] (between original and degraded speech signals). The SFPs are further smoothed as a function of SNR. If the SNR is high, no smoothing is carried out. If the SNR is low (such as when high background noise is present) the SFPs are smoothed to emulate functionality of higher auditory pathways, which provides the robust tracking of salient auditory information under harsh conditions.

viii) Distance between the original and degraded SFPs are calculated, to analyse a 'slow' and a 'fast' time-varying distortion.

$$Distance = 20 * log10\left[\frac{degraded}{original}\right] \quad (26)$$

ix) The Distance results from all voiced sections are averaged and used as the predicted outputs.

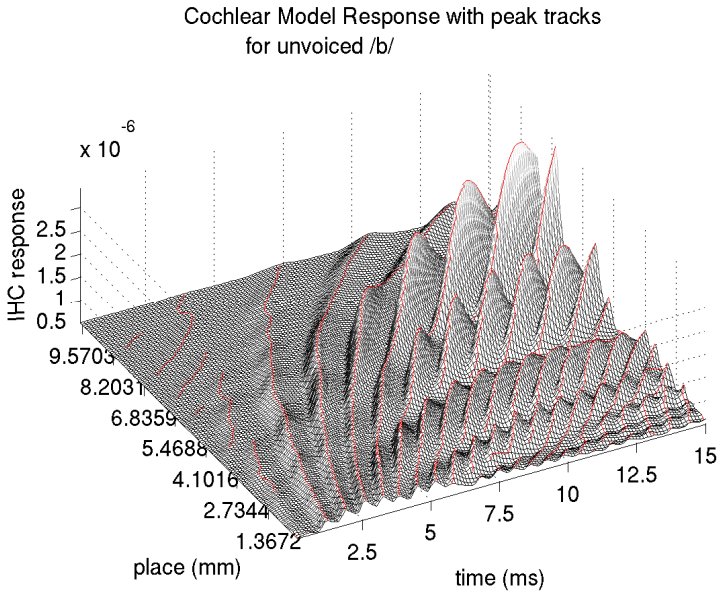


Figure 6.21: Cochlear Model output and peak tracks (red lines)

6.4.6.3 COSM (Cochlear Output Statistic Metric) feature extraction

The COSM feature set is extracted from:

1) voiced sections, which are salient for determining foreground distortions (S-MOS); and

2) background sections (where only background noise exists) for determining background distortions (N-MOS).

The COSM feature extraction operates on the output of the physiological Cochlear Model $IHC[p, t]$ as a function of place, *p*, and time, *t*, and is comprised of the following steps (see also figure 6.22):

i) Along place, analysing sections of $IHC[p, t]$ by dividing it into 4 sub-regions determined by the following positions along the human cochlear length: 3,50 cm, 3,08 cm, 1,89 cm, 1,03 cm, 0 cm. These positions correspond to the 20 Hz, 150 Hz, 1,4 kHz, 4,9 kHz and 24 kHz cut-off frequencies for stimuli at threshold level.

ii) Along time, analysing sections of $IHC[p, t]$ by dividing it into smaller time frames *j* (with up to 20 ms length for foreground voiced sections, and up to 100ms length for background sections).

iii) Extracting time-based vectors for each section as follows:

a) For each sub-region *k (*corresponding to *K1*, *K2*, *K3*, *K4* in figure 6.22), time based vectors of the CM output are calculated for each time frame *j*, by averaging $IHC[p, t]$ over place, *p*, as follows:

$$\overline{CM_{j,k}[t]} = \frac{1}{K_{j,k,H} - K_{j,k,L}} \sum_{p=K_{j,k,L}}^{K_{j,k,H}} IHC[p, t], \ where\ k = 1,2,3,4 \tag{27}$$

where $K_{j,k,L}$ and $K_{j,k,H}$ represent the lower and higher spatial indices for the kth sub-region respectively.

b) Convert the amplitude of the time-based vectors to a logarithmic scale:

$$AS_{ref(j,k)}[t] = log_{10}\left(\overline{CM_{j,k}^{ref}[t]}\right) \tag{27a}$$

$$AS_{mod(j,k)}[t] = log_{10}\left(\overline{CM_{j,k}^{mod}[t]}\right) \tag{27b}$$

c) Compare the logarithm of the time-based vectors for the modified signal to the logarithm of the time-based vectors for the reference signal to determine the Analysis Serial difference time-based vectors $AS_{diff(j,k)}[t]$.

$$AS_{diff(j,k)}[t] = \begin{cases} AS_{ref(j,k)}[t] - AS_{mod(j,k)}[t], \text{if } |AS_{diff(j,k)}[t]| > 10^{-32}, where\ k = 1,2,3,4 \\ 10^{-32}, \text{if } |AS_{diff(j,k)}[t]| < 10^{-32} \end{cases} \tag{28}$$

iv) Extracting place-based vectors for each section as follows:

a) For each sub-region *k (*corresponding to *K1*, *K2*, *K3*, *K4* in figure 6.22), place based vectors of the CM output are calculated for each time frame *j*, by averaging $IHC[p, t]$ over time, *t*, as follows:

$$\overline{CM_{j,k}[p]} = \frac{1}{T_{j,k,1} - T_{j,k,N}} \sum_{t=T_{j,k,1}}^{T_{j,k,N}} IHC[p, t], \ where\ k = 1,2,3,4, \tag{28a}$$

where $T_{j,k,1}$ and $T_{j,k,N}$ represent the lower and higher temporal indices for the k^{th} sub-region respectively.

b) Convert the amplitude of the place-based vectors to a logarithmic scale as shown in figure 6.22.

$$AS_{ref(j,k)}[p] = log_{10}\left(\overline{CM_{j,k}^{ref}[p]}\right) \tag{28b}$$

$$AS_{mod(j,k)}[p] = log_{10}\left(\overline{CM_{j,k}^{mod}[p]}\right) \tag{28c}$$

c) Compare the logarithm of the place-based vectors for the modified signal to the logarithm of the place-based vectors for the reference signal to determine the Analysis Serial place-based vectors $AS_{diff(j,k)}[p]$. Determine a positive part of the Analysis Serial place-based vectors, $AS_{diff+(j,k)}[p]$, and a negative part of the Analysis Serial place-based vectors, $AS_{diff-(j,k)}[p]$.

$$AS_{diff(j,k)}[p] = \begin{cases} AS_{ref(j,k)}[p] - AS_{mod(j,k)}[p], \text{if } |AS_{diff(j,k)}[p]| > 10^{-32}, where\ k = 1,2,3,4 \\ 10^{-32}, \text{if } |AS_{diff(j,k)}[p]| < 10^{-32} \end{cases} \quad (28d)$$

$$AS_{diff+(j,k)}[p] = \max\left(0, AS_{diff(j,k)}[p]\right), where\ k = 1,2,3,4 \quad (28e)$$

$$AS_{diff-(j,k)}[p] = \min\left(0, AS_{diff(j,k)}[p]\right), where\ k = 1,2,3,4 \quad (28f)$$

v) The (1) average, (2) median, (3) standard deviation, (4) Harmonic mean, (5) Mean Absolute Delta, and (6) Geometric shift. are then calculated for each of the following Analysis Serial vectors:

voiced sections:

$AS_{diff(j,k)}[p]$ for *k*=1,2,3,4;

$AS_{diff+(j,k)}[p]$ for *k*=1,2,3,4;

$AS_{diff-(j,k)}[p]$ for *k*=1,2,3,4;

$AS_{ref(j,k)}[t]$ for *k*=2;

$AS_{ref(j,k)}[p]$ for *k*=2;

$AS_{diff(j,k)}[t]$ for *k*=2;

background sections:

$AS_{mod(j,k)}[p]$ for *k*=1,2,3;

$AS_{mod(j,k)}[t]$ for *k*=1,2,3,4;

The six statistical metrics are defined below:

Average:

$$\overline{AS_{p}}_{j,k} = \frac{\sum_{p=K_{j,k,L}}^{K_{j,k,H}} AS_{j,k}[p]}{K_{j,k,H} - K_{j,k,L} + 1} \quad (28g)$$

$$\overline{AS_{t}}_{j,k} = \frac{\sum_{t=T_{j,2,1}}^{T_{j,2,N}} AS_{j,2}[t]}{T_{j,2,N} - T_{j,2,1} + 1} \quad (28h)$$

Median:

$$median\ (AS_{j,k}[p]) \quad (28i)$$

$$median\ (AS_{j,2}[t]) \quad (28j)$$

Standard deviation:

$$\sigma_{AS_{p}\,j,k} = \sqrt{\frac{\sum_{p=K_{j,k,L}}^{K_{j,k,H}} (AS_{j,k}[p] - \overline{AS_{p}}_{j,k})}{K_{j,k,H} - K_{j,k,L}}} \quad (28k)$$

$$\sigma_{AS_{t}\,j,2} = \sqrt{\frac{\sum_{t=T_{j,2,1}}^{T_{j,2,N}} (AS_{j,2}[t] - \overline{AS_{t}}_{j,2})}{T_{j,2,N} - T_{j,2,1}}} \quad (28l)$$

Harmonic mean:

$$H_{AS_{p_{j,k}}} = \frac{1}{\sum_{p=K_{j,k,L}}^{K_{j,k,H}} \left(\frac{1}{AS_{j,k}[p]}\right)} \tag{29}$$

$$H_{AS_{t_{j,2}}} = \frac{1}{\sum_{t=T_{j,2,1}}^{T_{j,2,N}} \left(\frac{1}{AS_{j,2}[t]}\right)} \tag{29a}$$

Mean Absolute Delta:

$$\Delta_{AS_{p_{j,k}}} = \frac{\sum_{p=K_{j,k,L}}^{(K_{j,k,H})-1} |AS_{j,k}[p+1] - AS_{j,k}[p]|}{K_{j,k,H} - K_{j,k,L}} \tag{30}$$

$$\Delta_{AS_{t_{j,2}}} = \frac{\sum_{t=T_{j,2,1}}^{(T_{j,2,N})-1} |AS_{j,2}[t+1] - AS_{j,2}[t]|}{T_{j,2,N} - T_{j,2,1}} \tag{30a}$$

Geometric shift:

$$G_{AS_{p_{j,k}}} = \begin{cases} 0, if \ \sum AS_{j,k}[p] = 0 \\ \frac{\sum_{p=K_{j,k,L}}^{K_{j,k,H}} AS_{j,k}[p] * p}{\sum_{p=K_{j,k,L}}^{K_{j,k,H}} p} - \frac{\sum_{p=K_{j,k,L}}^{K_{j,k,H}} p}{K_{j,k,H} - K_{j,k,L} + 1} \end{cases} \tag{31}$$

$$G_{AS_{t_{j,2}}} = \begin{cases} 0, if \ \sum AS_{j,2}[t] = 0 \\ \frac{\sum_{t=T_{j,2,1}}^{T_{j,2,N}} AS_{j,2}[t] * t}{\sum_{t=T_{j,2,1}}^{T_{j,2,N}} t} - \frac{\sum_{t=T_{j,2,1}}^{T_{j,2,N}} t}{T_{j,2,N} - T_{j,2,1} + 1} \end{cases} \tag{31a}$$

vi) Each of the resulting statistical metrics is further averaged over the frame index *j*, resulting in 90 statistical metrics for the voiced sections and 42 statistical metrics for the background sections:

Table 6.7: COSM Feature set

	COSM FEATURE SET						
	Voiced Sections					Background sections	
	Diff signal (time)	**Reference signal (place and time)**	**Diff signal (place)**	**Diff signal (place) Positive Part**	**Diff signal (place) Negative Part**	**Modified signal (time)**	**Modified signal (place)**
Average	$\overline{AS_{diff\,t_2}}$	$\overline{AS_{ref\,p_2}}$ $\overline{AS_{ref\,t_2}}$	$\overline{AS_{diff\,p_1}}$ $\overline{AS_{diff\,p_2}}$ $\overline{AS_{diff\,p_3}}$ $\overline{AS_{diff\,p_4}}$	$\overline{AS_{diff+\,p_1}}$ $\overline{AS_{diff+\,p_2}}$ $\overline{AS_{diff+\,p_3}}$ $\overline{AS_{diff+\,p_4}}$	$\overline{AS_{diff-\,p_1}}$ $\overline{AS_{diff-\,p_2}}$ $\overline{AS_{diff-\,p_3}}$ $\overline{AS_{diff-\,p_4}}$	$\overline{AS_{mod\,t_1}}$ $\overline{AS_{mod\,t_2}}$ $\overline{AS_{mod\,t_3}}$ $\overline{AS_{mod\,t_4}}$	$\overline{AS_{mod\,p_1}}$ $\overline{AS_{mod\,p_2}}$ $\overline{AS_{mod\,p_3}}$
Median	median $(AS_{diff(t_2)})$	median $(AS_{ref(p_2)})$ median $(AS_{ref(t_2)})$	median $(AS_{diff\,p_1})$ median $(AS_{diff\,p_2})$ median $(AS_{diff\,p_3})$ median $(AS_{diff\,p_4})$	median $(AS_{diff+\,p_1})$ median $(AS_{diff+\,p_2})$ median $(AS_{diff+\,p_3})$ median $(AS_{diff+\,p_4})$	median $(AS_{diff-\,p_1})$ median $(AS_{diff-\,p_2})$ median $(AS_{diff-\,p_3})$ median $(AS_{diff-\,p_4})$	median $(AS_{mod\,t_1})$ median $(AS_{mod\,t_2})$ median $(AS_{mod\,t_3})$ median $(AS_{mod\,t_4})$	median $(AS_{mod\,p_1})$ median $(AS_{mod\,p_2})$ median $(AS_{mod\,p_3})$
Standard Deviation	$\sigma_{AS_{diff\,t_2}}$	$\sigma_{AS_{ref\,p_2}}$ $\sigma_{AS_{ref\,t_2}}$	$\sigma_{AS_{diff\,p_1}}$ $\sigma_{AS_{diff\,p_2}}$ $\sigma_{AS_{diff\,p_3}}$ $\sigma_{AS_{diff\,p_4}}$	$\sigma_{AS_{diff+\,p_1}}$ $\sigma_{AS_{diff+\,p_2}}$ $\sigma_{AS_{diff+\,p_3}}$ $\sigma_{AS_{diff+\,p_4}}$	$\sigma_{AS_{diff-\,p_1}}$ $\sigma_{AS_{diff-\,p_2}}$ $\sigma_{AS_{diff-\,p_3}}$ $\sigma_{AS_{diff-\,p_4}}$	$\sigma_{AS_{mod\,t_1}}$ $\sigma_{AS_{mod\,t_2}}$ $\sigma_{AS_{mod\,t_3}}$ $\sigma_{AS_{mod\,t_4}}$	$\sigma_{AS_{mod\,p_1}}$ $\sigma_{AS_{mod\,p_2}}$ $\sigma_{AS_{mod\,p_3}}$
Harmonic mean	$H_{AS\,diff\,t_2}$	$H_{AS\,ref\,p_2}$ $H_{AS\,ref\,t_2}$	$H_{AS\,diff\,p_1}$ $H_{AS\,diff\,p_2}$ $H_{AS\,diff\,p_3}$ $H_{AS\,diff\,p_4}$	$H_{AS\,diff+\,p_1}$ $H_{AS\,diff+\,p_2}$ $H_{AS\,diff+\,p_3}$ $H_{AS\,diff+\,p_4}$	$H_{AS\,diff-\,p_1}$ $H_{AS\,diff-\,p_2}$ $H_{AS\,diff-\,p_3}$ $H_{AS\,diff-\,p_4}$	$H_{AS\,mod\,t_1}$ $H_{AS\,mod\,t_2}$ $H_{AS\,mod\,t_3}$ $H_{AS\,mod\,t_4}$	$H_{AS\,mod\,p_1}$ $H_{AS\,mod\,p_2}$ $H_{AS\,mod\,p_3}$
Mean Absolute Delta	$\Delta_{AS_{diff\,t_2}}$	$\Delta_{AS_{ref\,p_2}}$ $\Delta_{AS_{ref\,t_2}}$	$\Delta_{AS_{diff\,p_1}}$ $\Delta_{AS_{diff\,p_2}}$ $\Delta_{AS_{diff\,p_3}}$ $\Delta_{AS_{diff\,p_4}}$	$\Delta_{AS_{diff+\,p_1}}$ $\Delta_{AS_{diff+\,p_2}}$ $\Delta_{AS_{diff+\,p_3}}$ $\Delta_{AS_{diff+\,p_4}}$	$\Delta_{AS_{diff-\,p_1}}$ $\Delta_{AS_{diff-\,p_2}}$ $\Delta_{AS_{diff-\,p_3}}$ $\Delta_{AS_{diff-\,p_4}}$	$\Delta_{AS_{mod\,t_1}}$ $\Delta_{AS_{mod\,t_2}}$ $\Delta_{AS_{mod\,t_3}}$ $\Delta_{AS_{mod\,t_4}}$	$\Delta_{AS_{mod\,p_1}}$ $\Delta_{AS_{mod\,p_2}}$ $\Delta_{AS_{mod\,p_3}}$
Geometric Shift	$G_{AS_{diff\,t_2}}$	$G_{AS_{ref\,p_2}}$ $G_{AS_{ref\,t_2}}$	$G_{AS_{diff\,p_1}}$ $G_{AS_{diff\,p_2}}$ $G_{AS_{diff\,p_3}}$ $G_{AS_{diff\,p_4}}$	$G_{AS_{diff+\,p_1}}$ $G_{AS_{diff+\,p_2}}$ $G_{AS_{diff+\,p_3}}$ $G_{AS_{diff+\,p_4}}$	$G_{AS_{diff-\,p_1}}$ $G_{AS_{diff-\,p_2}}$ $G_{AS_{diff-\,p_3}}$ $G_{AS_{diff-\,p_4}}$	$G_{AS_{mod\,t_1}}$ $G_{AS_{mod\,t_2}}$ $G_{AS_{mod\,t_3}}$ $G_{AS_{mod\,t_4}}$	$G_{AS_{mod\,p_1}}$ $G_{AS_{mod\,p_2}}$ $G_{AS_{mod\,p_3}}$

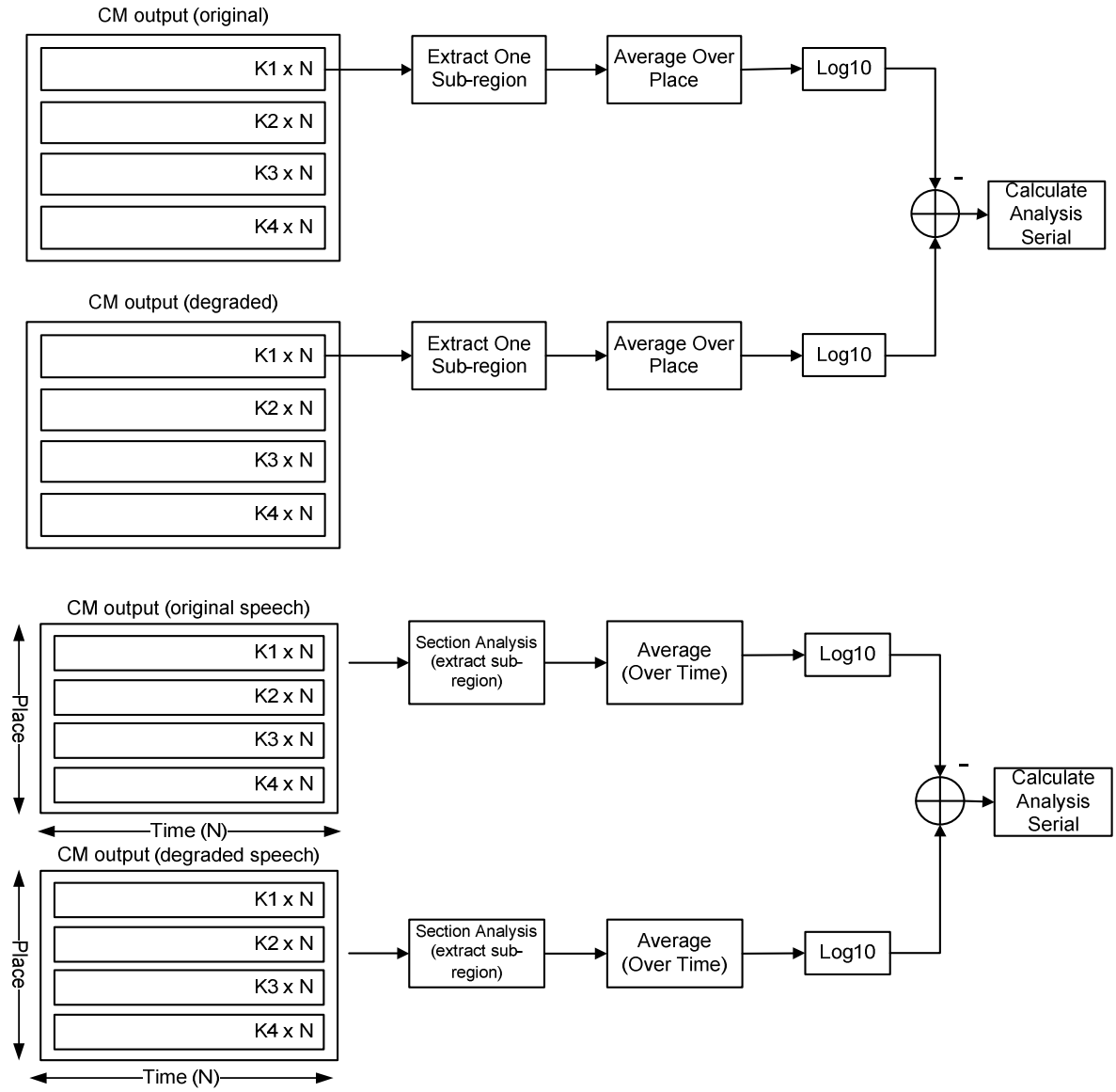


Figure 6.22: Flowchart prepared for Analysis Serial calculation

6.4.7 Training and mapping

The SFP and COSM features sets are converted into the mean opinion objective quality scores through a mapping procedure. An a-priori training process based on deep neural network (DNN) is used to determine suitable weighting factors for each of the features. The a-priori training is based on a large set of listening test databases. The weighting factors are applied with a deep neural learning network procedure for S-MOS, N-MOS and G-MOS to determine the multidimensional predicted quality scores.

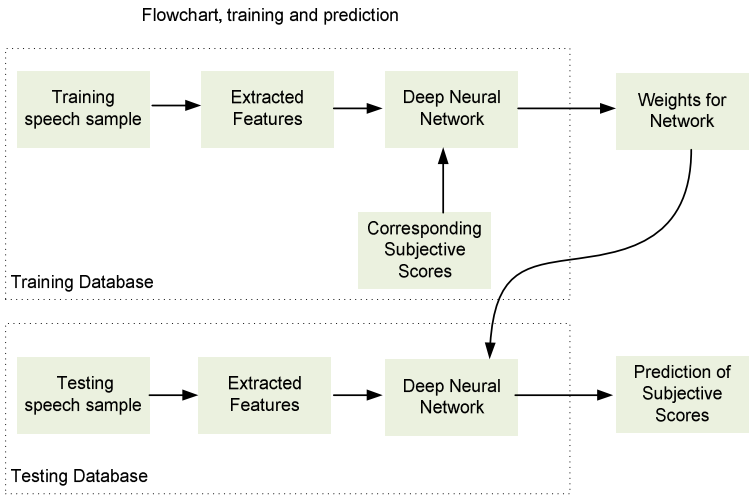


Figure 6.23: Flowchart for training of the model

The neural networks use a back-propagation algorithm which has 2 hidden layers (with fully connected nodes); each layer has an activation layer. Gaussian noise (with a standard deviation of 0,4) is added to the inputs to assist with regularization and reduce overfitting. For the training process, 400 epochs were required to reach stability. The networks for S-MOS, N-MOS and G-MOS are trained independently in separate DNN processes, with each producing its own set of coefficients.

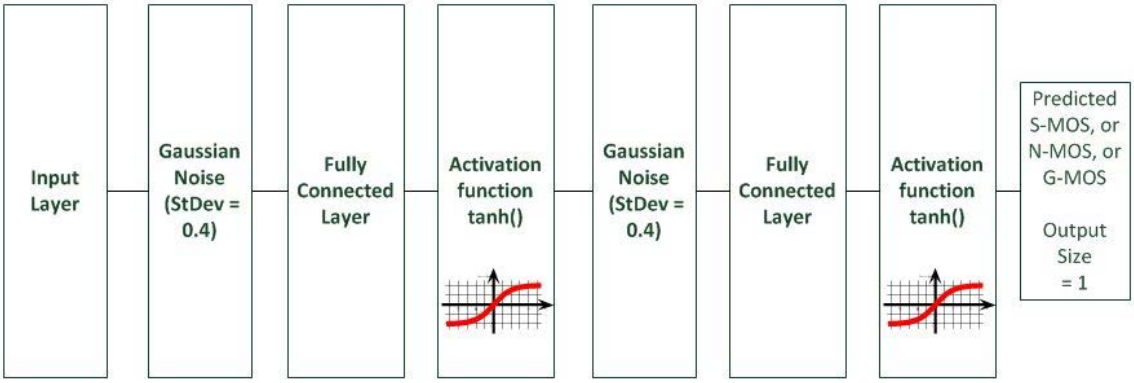


Figure 6.24: DNN sequencing; separate processes for each of S-MOS, N-MOS and G-MOS as described

6.5 Mapping of model outputs

As shown in annex A, the prediction models were trained with a large number of databases including a wide range of qualities. Even though reference conditions according to [i.17] were used in all databases, each subjective experiment has its own context, i.e. a certain average, minimum or maximum quality. This effect could e.g. lead to different scores for an identical sample/condition, which was placed into two different listening tests due to the context of the whole database.

The compensation of this influence can be conducted in two steps:

- First, one or more auditory databases including a "desired context" (e.g. representing a specific group of terminals) shall be available. Then, a 3rd order mapping function between the instrumental and auditory results of this data is determined. This mathematical transformation can be applied either to one or more attributes (SIG, BAK and/or OVRL).

- After this step, the instrumental prediction results for each new signal-under-test are transformed according to this determined mapping function. Any possible check against requirements is conducted after this mapping.

This type of correction depends on a concrete context (e.g. derived from one or more auditory experiments). Thus any specific transformation procedure is out of scope of the present document. An example for a suitable transformation is given in clause 5.5 of ETSI TR 126 931 [i.44], based on the validation databases DES-25 and DES-26 (see clause 8).

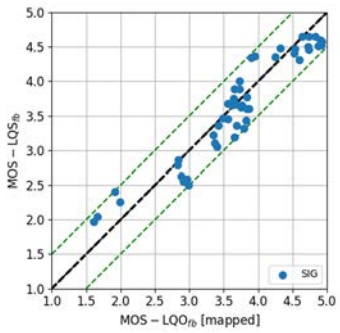
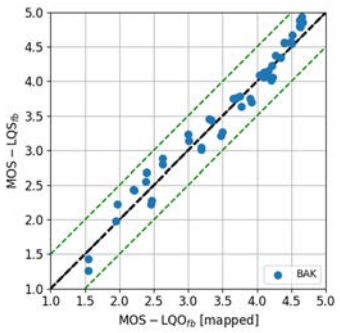
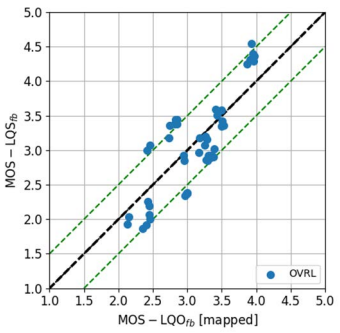
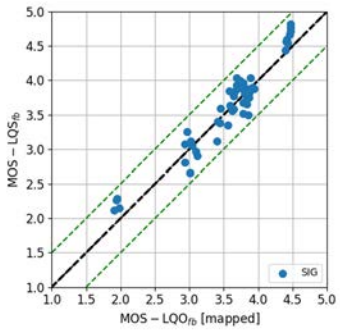
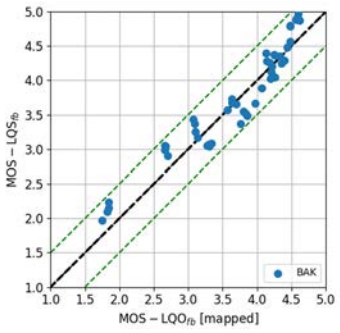
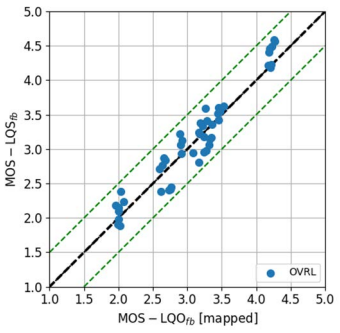
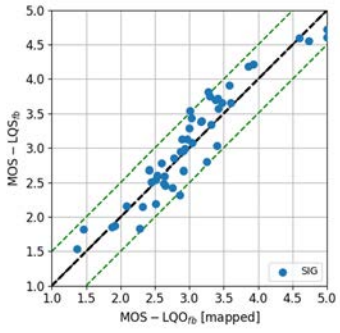
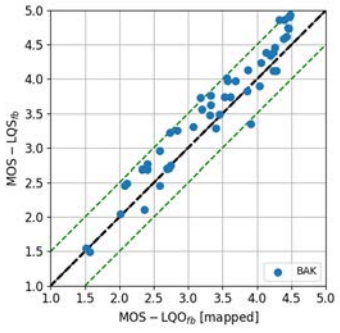
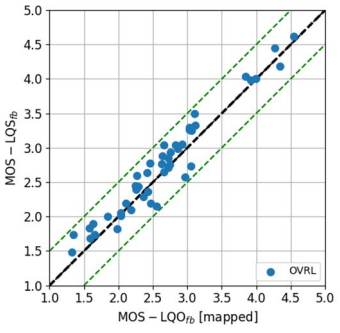
7 Comparison of objective and subjective results after the training process

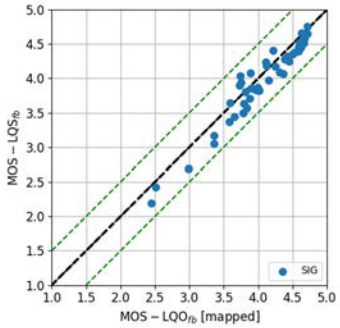
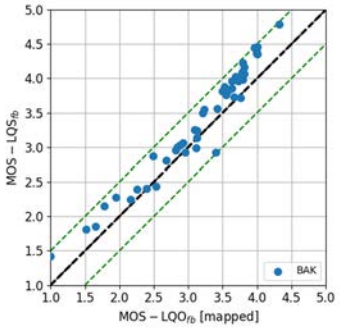
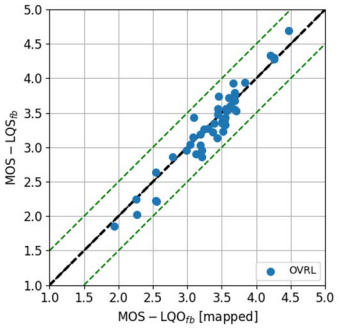
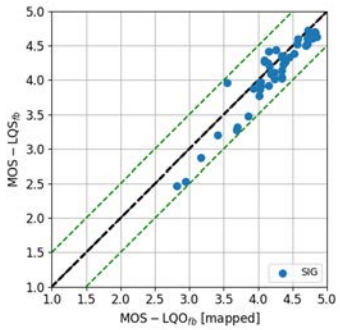
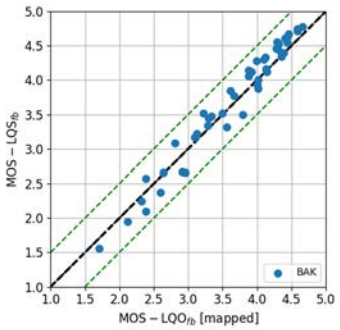
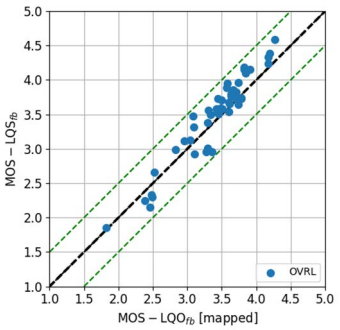
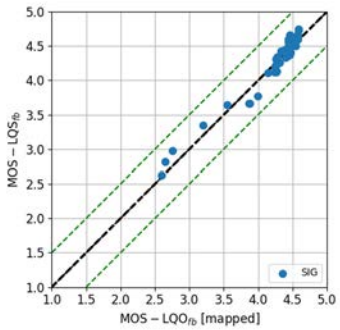
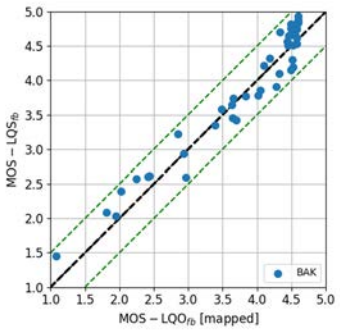
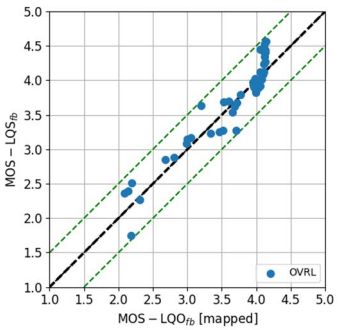
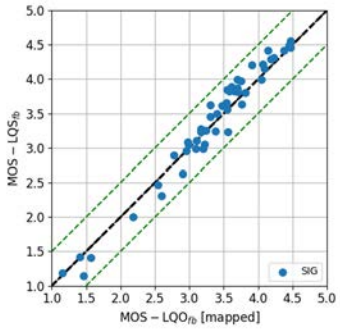
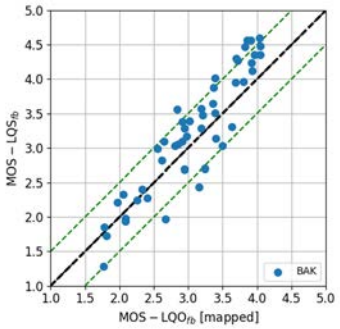
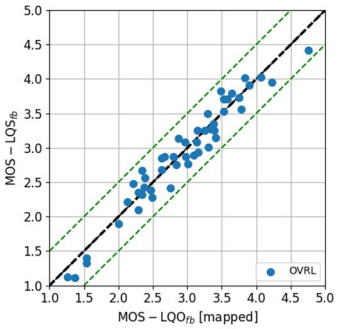
7.1 Introduction

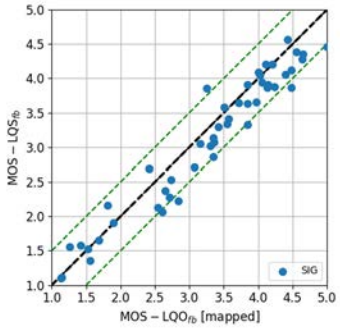
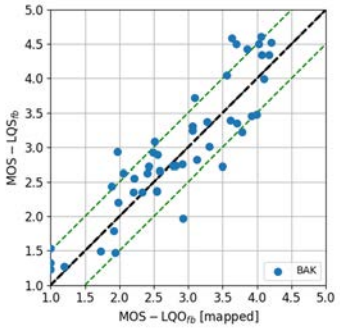
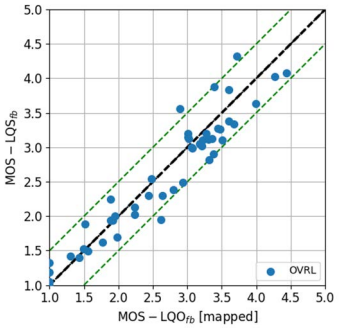
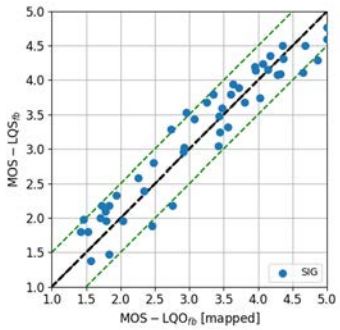
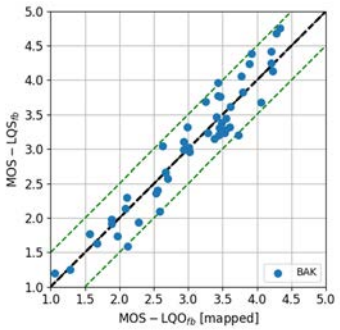
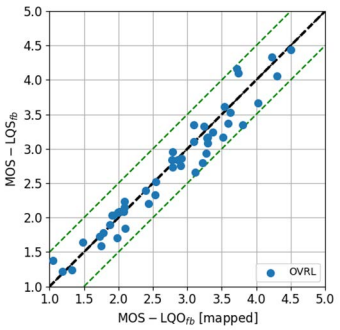
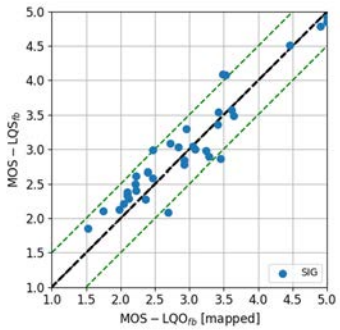
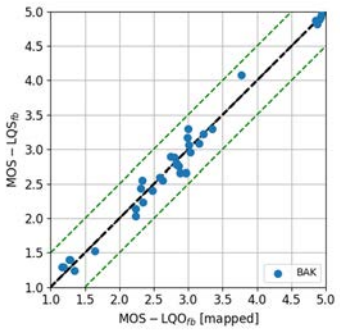
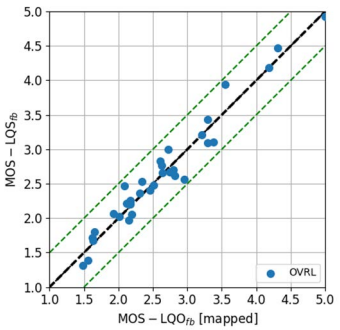
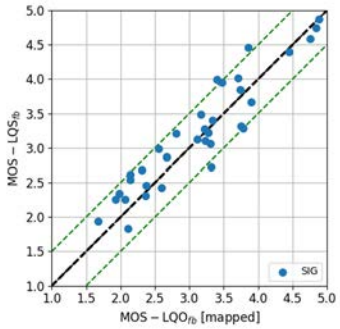
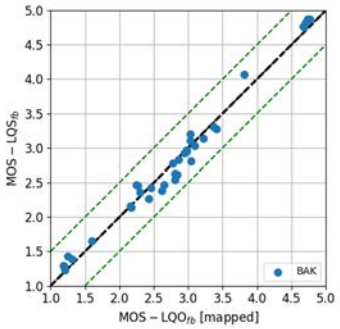
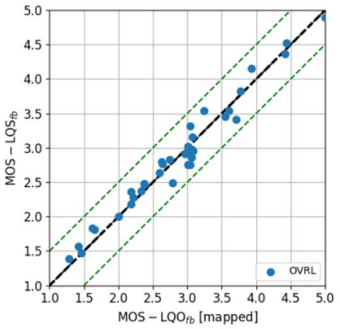
The prediction model is based on a large set of training listening test databases. Each database represent a certain aspect of upcoming SWB and FB telephony, including also a wide range of speech and noise quality. A summary of the databases and the conditions used for retraining is given in annex B.

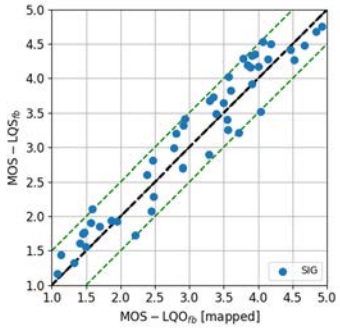
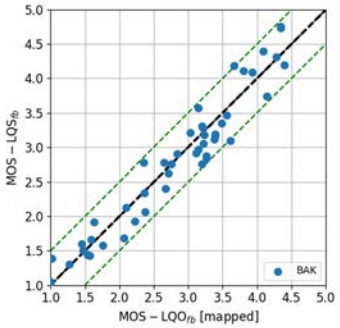
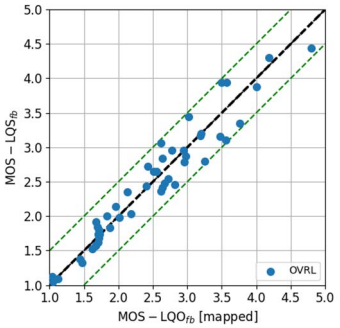
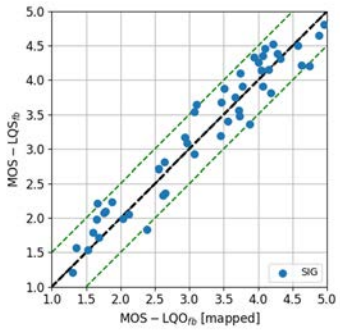
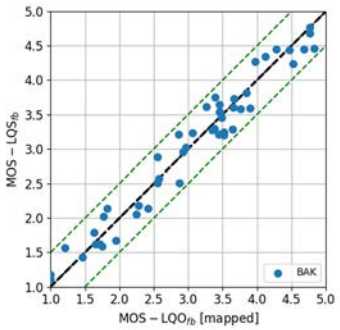
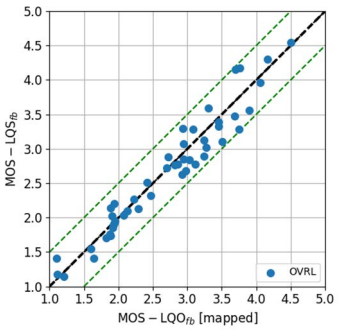
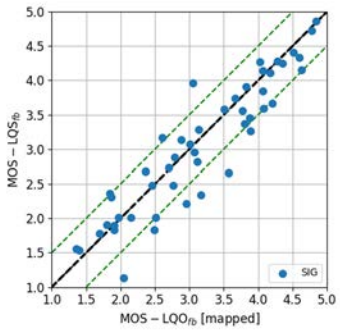
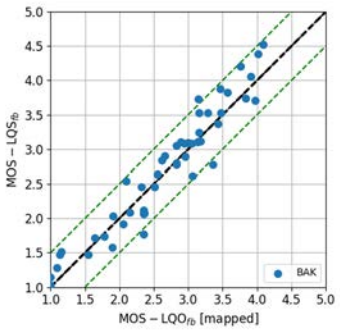
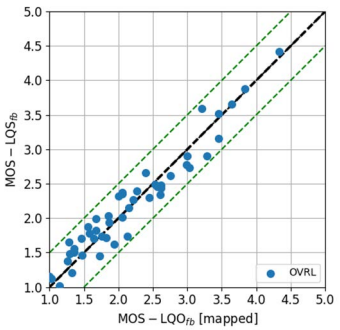
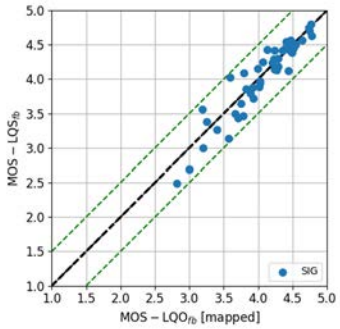
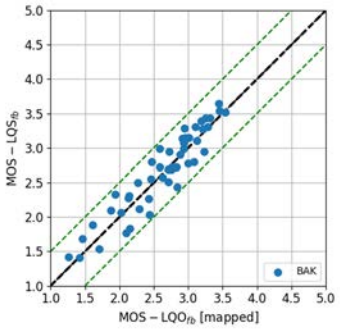
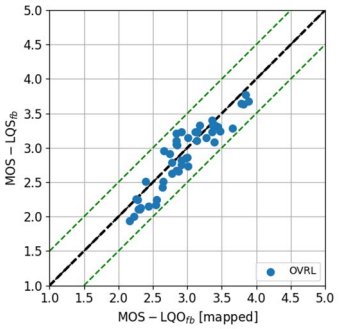
7.2 Results for Model A

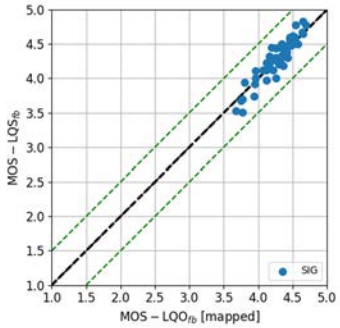
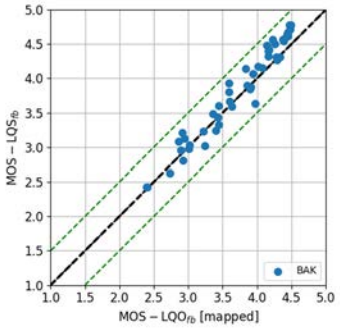
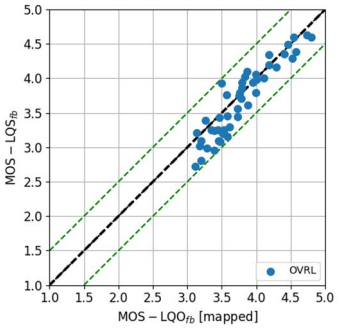
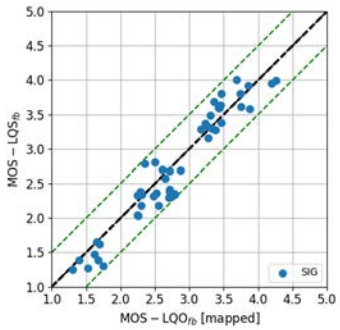
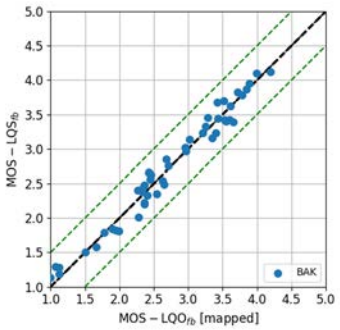
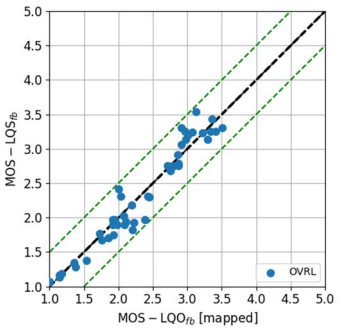
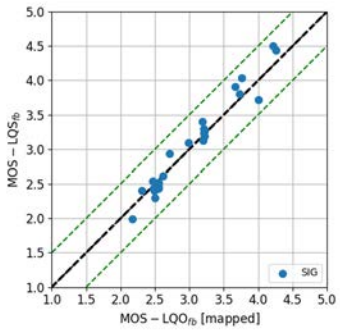
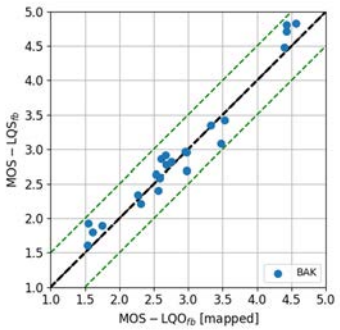
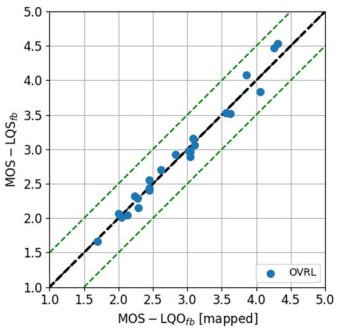
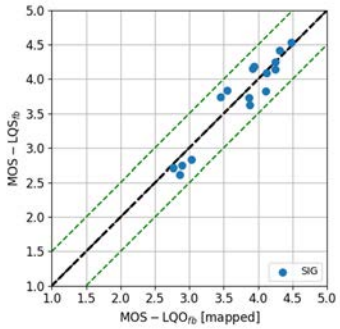
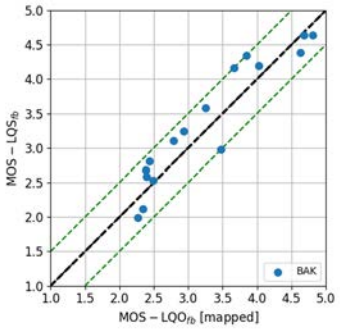
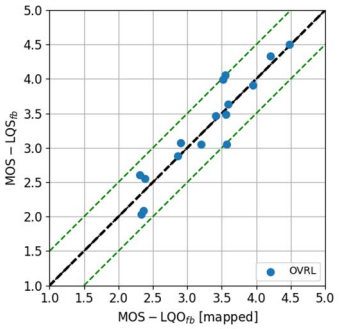
Table 7.1: Training results for model A after 3rd order mapping

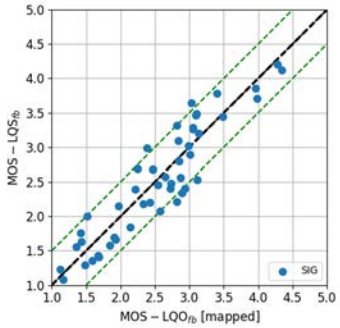
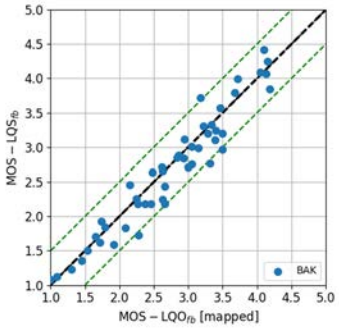
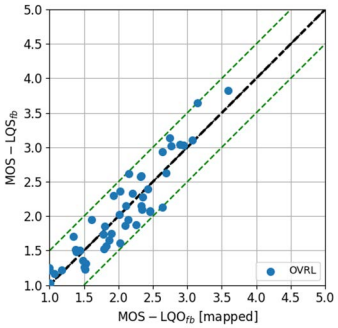
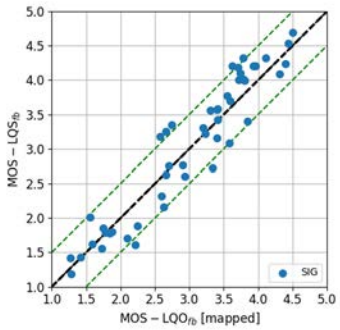
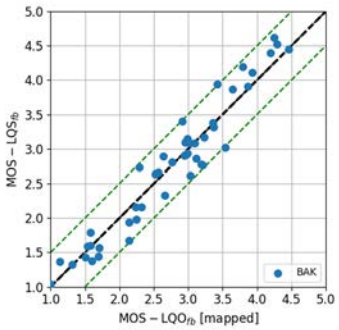
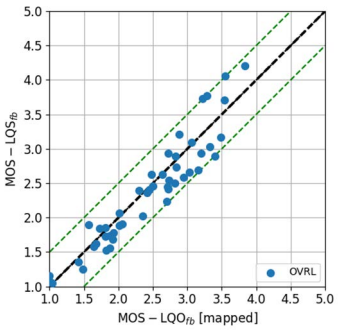
Feature vs. Database	S-MOS/SIG	N-MOS/BAK	G-MOS/OVRL
DES-01			
DES-02			
DES-03			

Feature vs. Database	S-MOS/SIG	N-MOS/BAK	G-MOS/OVRL
DES-04			
DES-05			
DES-06			
DES-07			

Feature vs. Database	S-MOS/SIG	N-MOS/BAK	G-MOS/OVRL
DES-08			
DES-09			
DES-10			
DES-11			

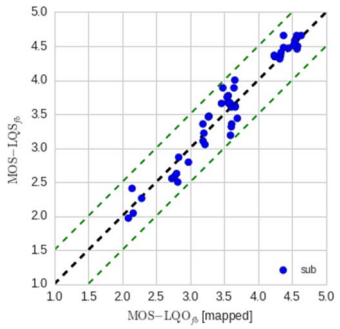
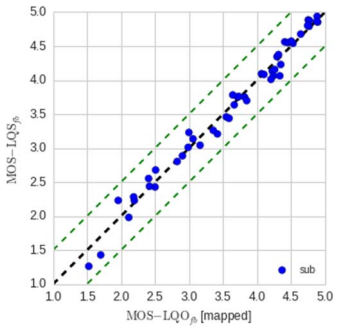
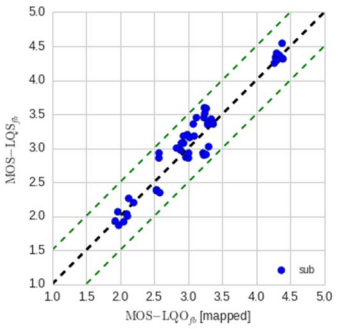
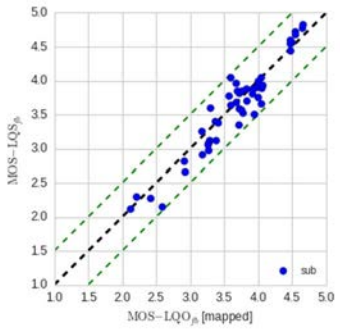
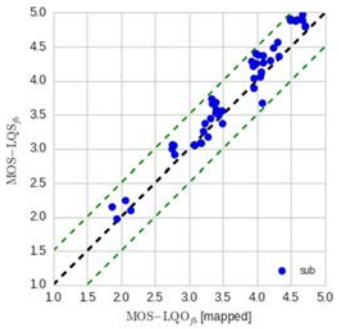
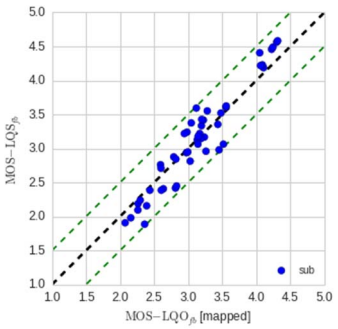
Feature vs. Database	S-MOS/SIG	N-MOS/BAK	G-MOS/OVRL
DES-12			
DES-13			
DES-14			
DES-15			

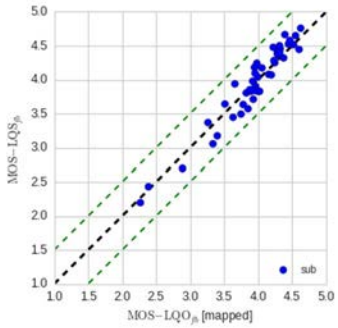
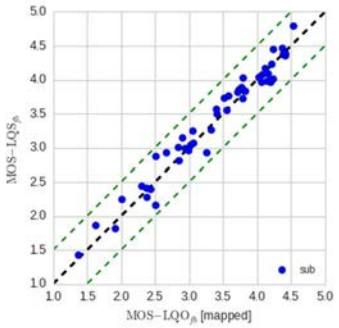
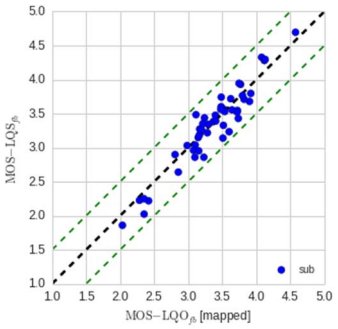
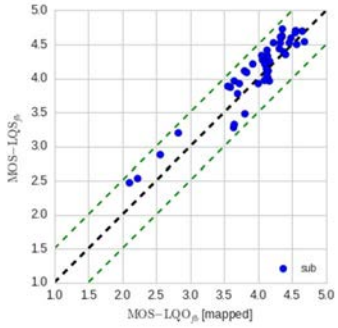
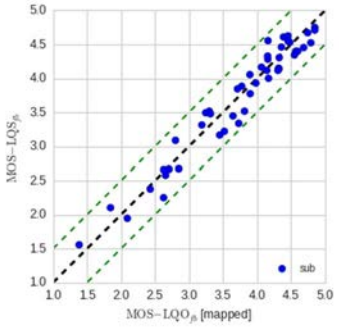
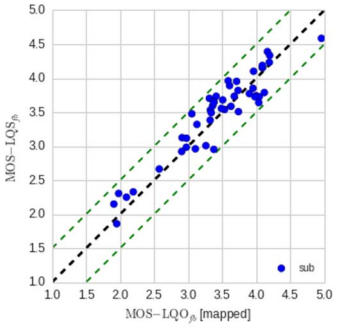
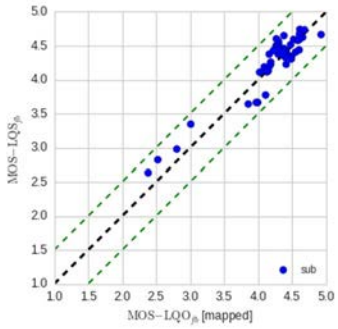
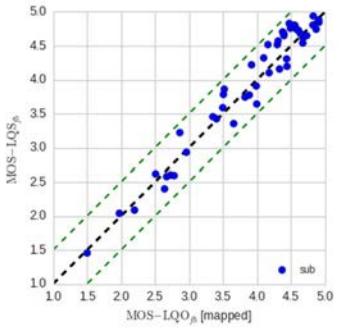
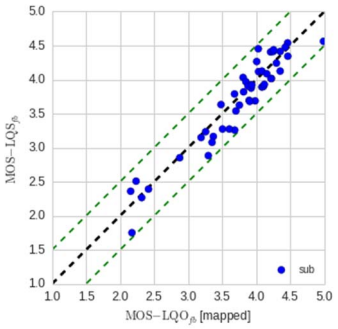
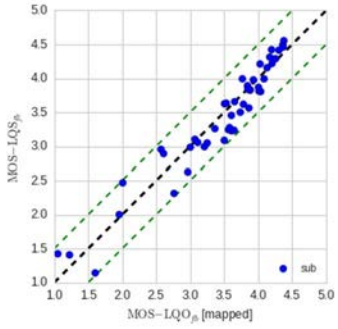
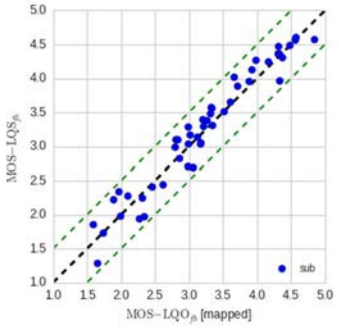
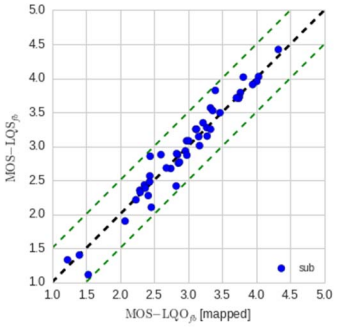
Feature vs. Database	S-MOS/SIG	N-MOS/BAK	G-MOS/OVRL
DES-16			
DES-17			
DES-19			
DES-22			

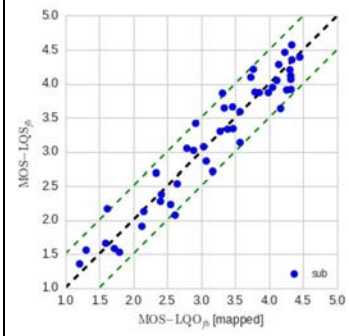
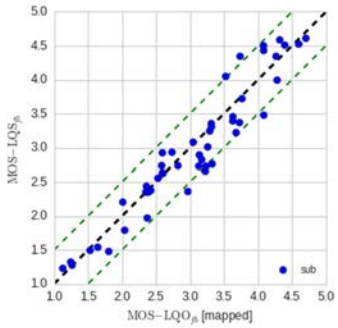
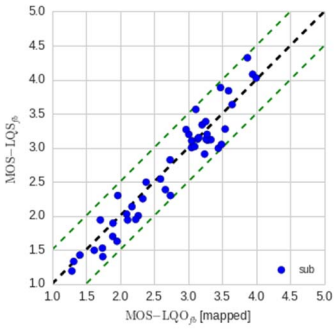
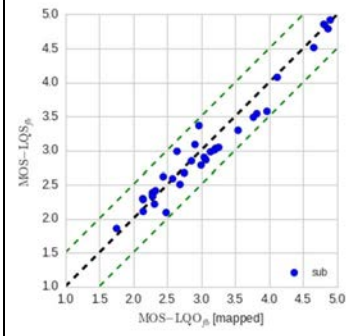
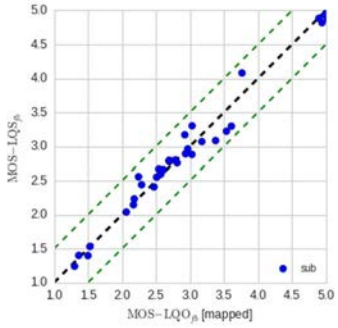
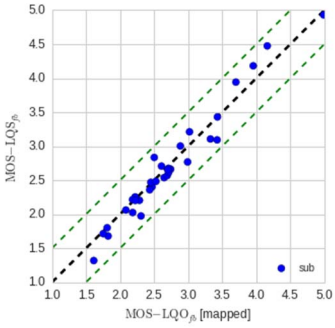
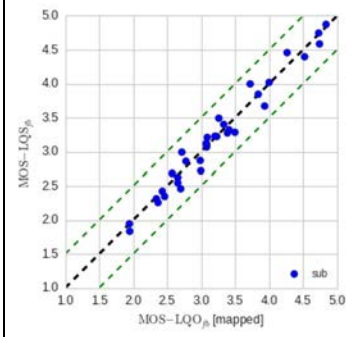
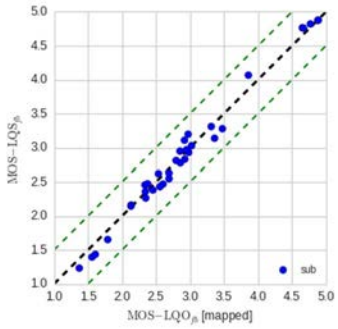
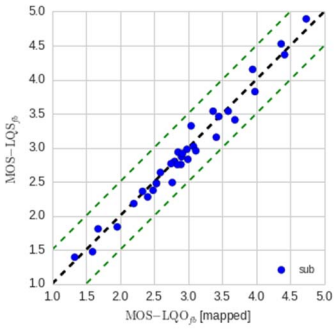
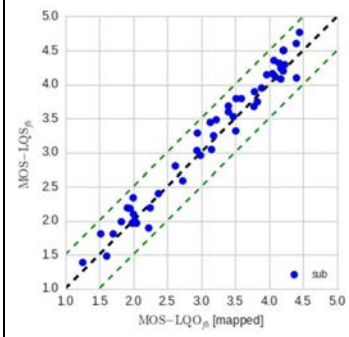
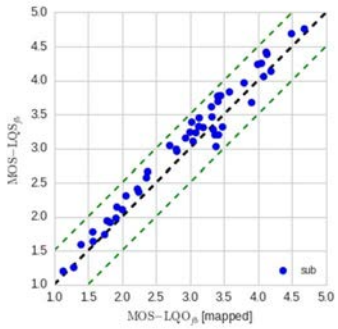
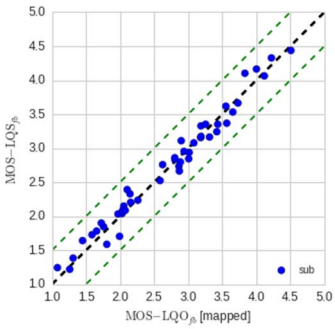
Feature vs. Database	S-MOS/SIG	N-MOS/BAK	G-MOS/OVRL
DES-23			
DES-24			

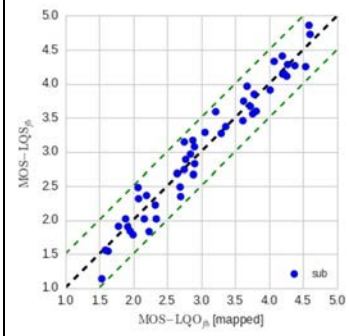
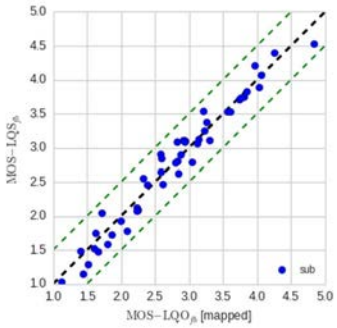
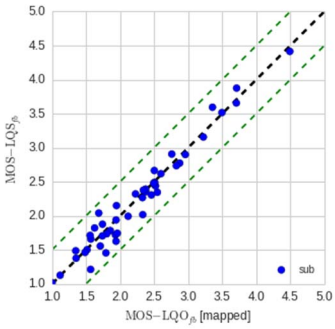
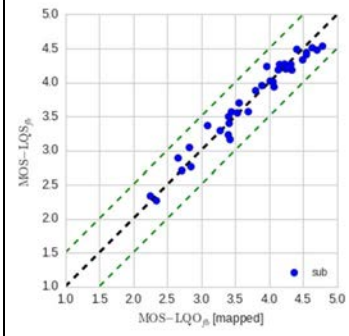
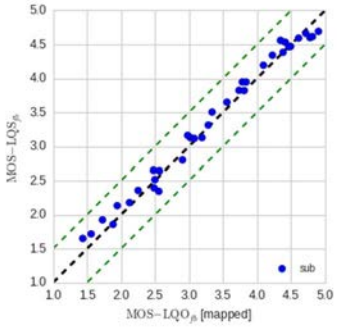
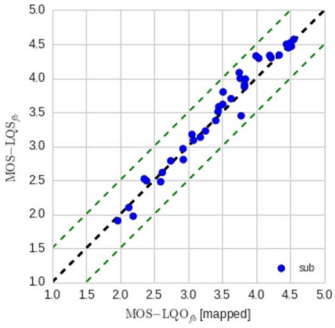
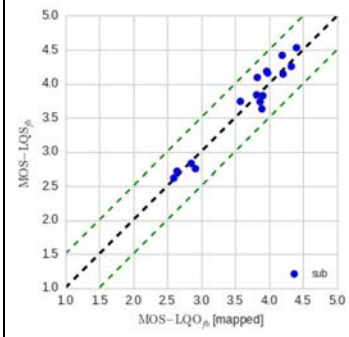
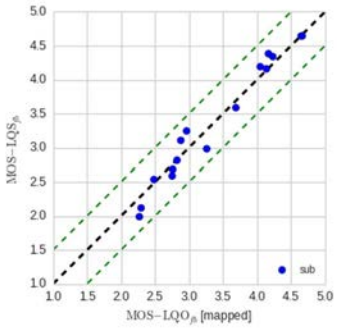
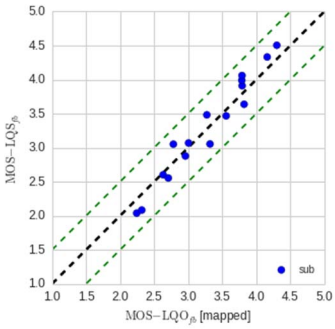
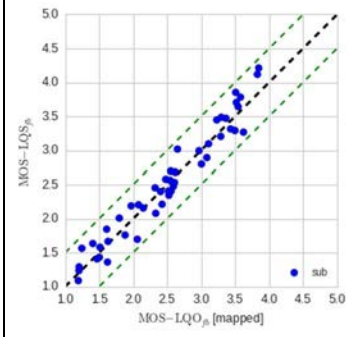
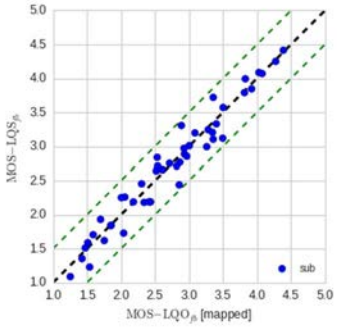
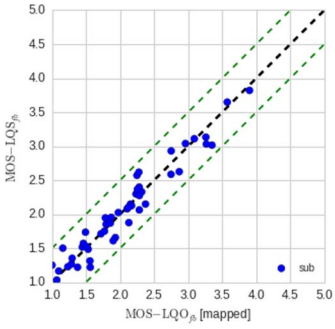
7.3 Results for Cochlear Prediction Model (Model B)

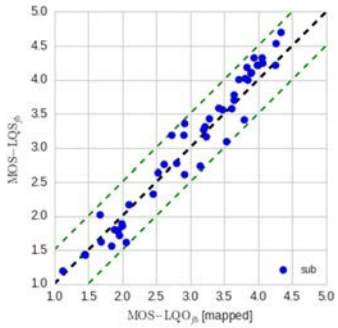
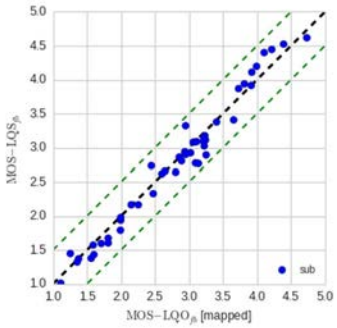
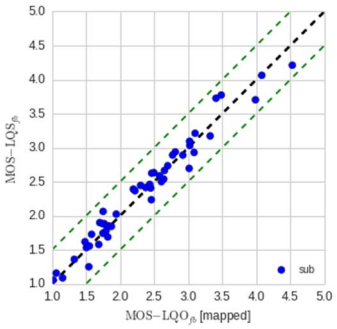
Table 7.2: Training results for model B after 3rd order mapping

Feature vs. Database	S-MOS/SIG	N-MOS/BAK	G-MOS/OVRL
DES-01			
DES-02			

Feature vs. Database	S-MOS/SIG	N-MOS/BAK	G-MOS/OVRL
DES-04			
DES-05			
DES-06			
DES-07			

Feature vs. Database	S-MOS/SIG	N-MOS/BAK	G-MOS/OVRL
DES-08			
DES-10			
DES-11			
DES-13			

Feature vs. Database	S-MOS/SIG	N-MOS/BAK	G-MOS/OVRL
DES-14			
DES-20			
DES-22			
DES-23			

Feature vs. Database	S-MOS/SIG	N-MOS/BAK	G-MOS/OVRL
DES-24			

8 Validation results

8.1 Introduction

For the validation of the models different databases were provided. The databases included different types of conditions of different terminals and simulations. The details of the validation databases are described separately for each validation database. As validation database 1 was also one of the training databases for model A, it is not used for validation of model A. Similarly, validation database 2 was also one of the training databases for model B, so it is not used for validation of model B. Validation databases 3, 4 and 5 are used for validation of both models.

As a performance metric for the prediction accuracy, rmse* according to [i.8] is calculated after 3rd order mapping. This measure takes the uncertainty of the auditory data into account. Instrumental scores predicted inside the 95 % confidence interval of the corresponding auditory value are assumed as error 0. In a similar way, the maximum absolute error* (maxabs*) is obtained. By compensating the magnitude of the confidence interval of the maximum error after 3rd order mapping, a measure for the worst-case outlier is provided.

8.2 Validation database 1 (DES-17)

8.2.1 Database description

Validation database 1 was created using an acoustic mockup with multiple microphones mounted in locations typical of contemporary handsets. The acoustic set-up was consistent with the handset and hand-held handsfree conditions described in ETSI TS 126 132 [i.14], with the acoustic mockup mounted on, and in front of, HATS, respectively.

The speech source material was full-band German, consisting of four samples from each of four male and four female talkers, for a total of 32 sentences, with six samples used per condition to collect 96 votes per condition.

In handset mode, the speech was presented from a properly equalized HATS mouth at a level of -1,7 dB Pa at MRP, while in HHHF mode, the level was +1,3 dB Pa at MRP.

The background noise was reproduced using a system according to ETSI TS 103 224 [i.19], using the noise types listed in table 8.1. Levels for these noise types can be found in ETSI TS 103 224 [i.19].

Table 8.1: Background noises used for Validation database 1

Description	Filename according to ETSI TS 103 224 [i.19]	
	Handset	Hands-free
Silence	-	-
Cafeteria	Cafeteria_handset	Cafeteria_handsfree
Crossroads	Crossroadnoise_handset	Crossroadnoise_handsfree
Full-size Car 130 km/h	FullSizeCar_130_handset	FullSizeCar_130_handsfree
Pub	Pub_handset	Pub_handsfree
Road	Roadnoise_handset	Roadnoise_handsfree
Train Station	TrainStation_handset	TrainStation_handsfree

The noise reduction was applied using typical state-of-the art noise suppression algorithms, with four tunings applied, two for each of handset and HHHF, providing varying trade-offs between reducing noise and preserving a certain amount of speech signal.

After noise reduction processing, the fixed point reference version of the EVS codec [i.4] was applied, operating in super-wideband mode at 32 kHz sample rate and 13,2 kbit/sec.

Full-band reference conditions were used, processed according to DESUDAPS-1 [i.17]. Presentation was binaural at 73 dB SPL, using headphones equalized to diffuse field.

8.2.2 Validation database 1: Results for model B

Results are shown as scatter plots, comparing instrumental predicted ratings to subjective ratings. Results from model B on validation database 1 for each of the three ratings, SIG, BAK, and OVRL are shown in figure 8.1. For each rating (rows), two scatter plots are shown, one before a monotonic mapping is applied (right column) and one after a monotonic mapping is applied (left column).

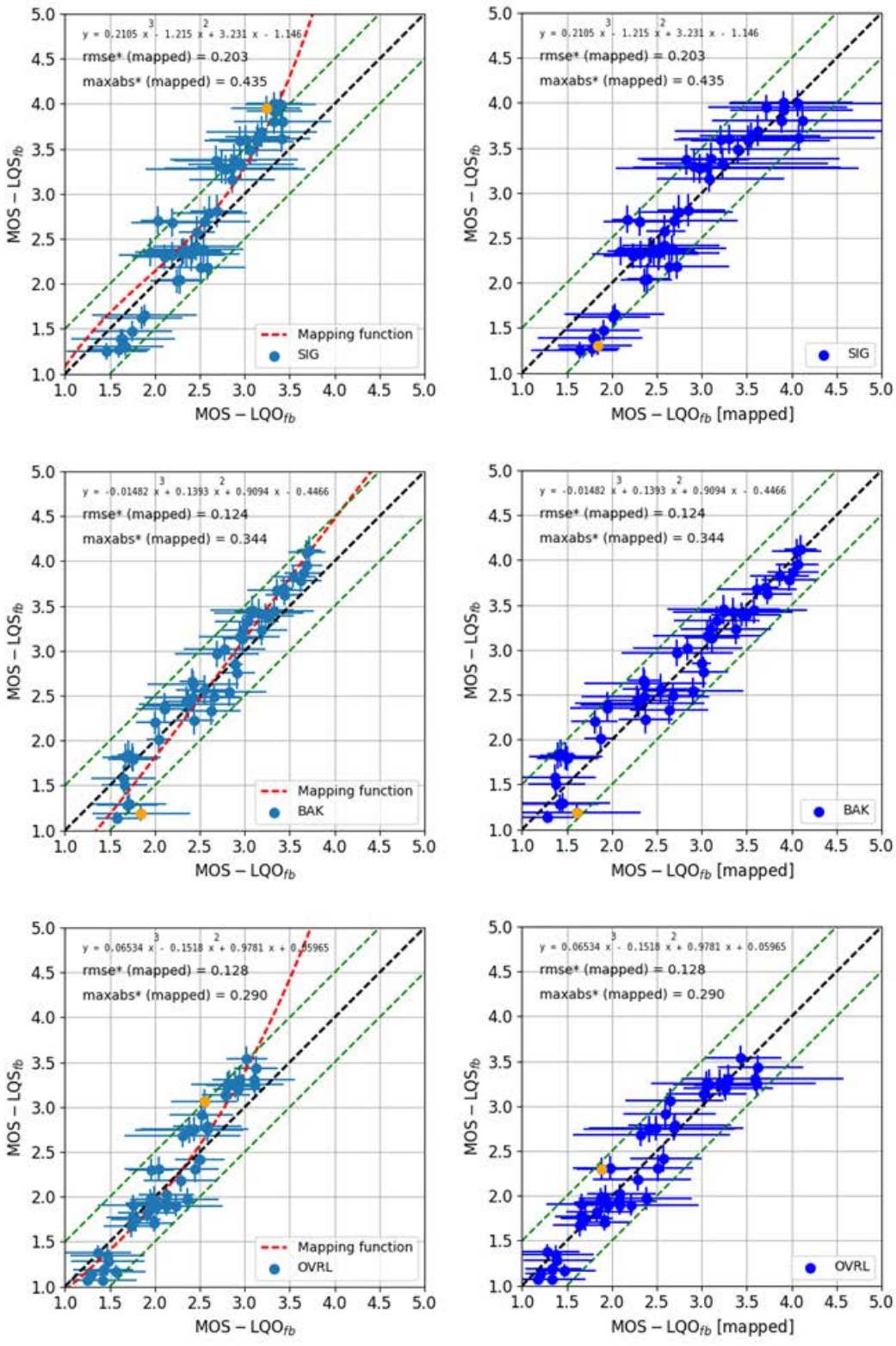


Figure 8.1: Scatter plots from model B for validation database 1

The rmse* and maximum absolute error* (maxabs*) after mapping are shown on all figures, with and orange-colored symbol indicating the condition with the largest overall maximum absolute error. The mapping polynomial is shown in the upper left corner of each panel. The dashed green lines show error of ±0,5 MOS. The error bars indicate the 95 % confidence interval before mapping (left column) and after mapping (right column). The relatively large value of the error bars for the instrumental ratings is due to the relatively small number (6) of speech samples per condition.

Additional performance metrics, including Pearson's ρ correlation coefficient, Spearman's ρ rank order correlation, and Kendall's τ are shown in table 8.2.

Table 8.2: Performance metrics for model B on validation database 1

Dimension	Metric	Raw	Mapped	d*	Mapped & d*
SIG	Rmse	0,396	0,311	0,274	0,203
	Max Abs Error	0,709	0,538	0,573	0,435
	Pearson's ρ	0,956	0,943	0,972	0,973
	Spearman's rank order ρ	0,939	0,939	0,961	0,967
	Kendall's τ	0,816	0,816	0,863	0,881
BAK	Rmse	0,259	0,231	0,157	0,124
	Max Abs Error	0,662	0,429	0,576	0,344
	Pearson's ρ	0,970	0,970	0,987	0,992
	Spearman's rank order ρ	0,973	0,973	0,988	0,991
	Kendall's τ	0,868	0,868	0,922	0,949
OVRL	Rmse	0,279	0,219	0,177	0,128
	Max Abs Error	0,516	0,416	0,382	0,290
	Pearson's ρ	0,964	0,959	0,980	0,985
	Spearman's rank order ρ	0,962	0,962	0,980	0,985
	Kendall's τ	0,841	0,841	0,890	0,917

8.3 Validation database 2 (DES-20)

8.3.1 Database description

Validation database 2 was created using three commercial and pre-commercial handsets. Each was operated in handset and hand-held handsfree mode, set up as described in ETSI TS 126 132 [i.14].

The speech source was full-band Mandarin phonetically-balanced sentences, with two sentences from each of four male and four female talkers, for a total of sixteen samples, with 128 votes collected per condition.

In handset mode, the speech was presented from a properly equalized HATS mouth at -1,7 dB Pa at MRP, while in HHHF mode, the level was +1,3 dB Pa at MRP.

The background noise was reproduced using a system according to ETSI ES 202 396-1[i.1], using the noise types listed in table 8.3. Levels for these noise types can be found in ETSI ES 202 396-1 [i.1].

Table 8.3: Background noises used for validation database 2

Description	Filename according to ETSI ES 202 396-1 [i.1]
Silence	-
Cafeteria	Cafeteria_Noise binaural
Callcenter	Work Office Noise Callcenter binaural
Full-size Car 130 km/h	FullSizeCar_130kmh binaural
Road	Outside Traffic Road binaural
Train Station	Train Station binaural

The handsets were placed in calls using a base station simulator, with EVS super wideband speech encoding [i.4] at 13,2 kbit/sec.

Full-band reference conditions were used, processed according to DESUDAPS-1 [i.17]. Presentation was binaural at 73 dB SPL, using headphones equalized to diffuse field.

8.3.2 Validation database 2: Results for model A

Results are shown as scatter plots, comparing instrumental predicted ratings to subjective ratings. Results from model A on validation database 2 for each of the three ratings, SIG, BAK and OVRL, are shown in figure 8.2. As in figure 8.1, for each rating (rows), two scatter plots are shown, one before a monotonic mapping is applied (right column) and one after a monotonic mapping is applied (left column).

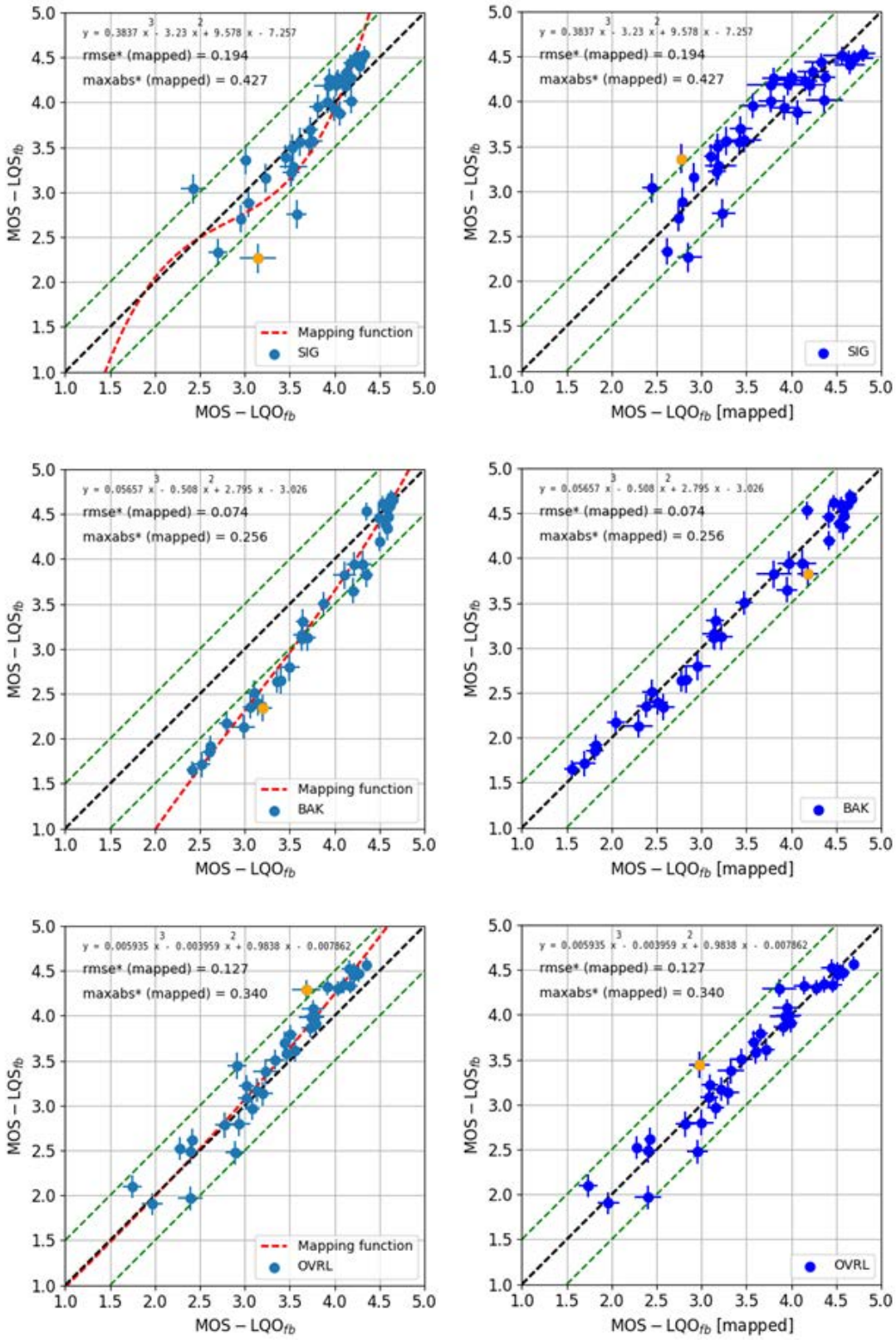


Figure 8.2: Scatter plots from model A for validation database 2

The rmse* and maximum absolute error* (maxabs*) after mapping are shown on all figures, with an orange-colored symbol indicating the condition with the largest overall maximum absolute error. The mapping polynomial is shown in the upper left corner of each panel. The dashed green lines show error of ±0,5 MOS. The error bars indicate the 95 % confidence interval before mapping (left column) and after mapping (right column).

Additional performance metrics, including Pearson's ρ correlation coefficient, Spearman's ρ rank order correlation, and Kendall's τ are shown in table 8.4.

Table 8.4: Performance metrics for model A on validation database 2

Dimension	Metric	Raw	Mapped	d*	Mapped & d*
SIG	Rmse	0,299	0,304	0,205	0,194
	Max Abs Error	0,883	0,590	0,716	0,427
	Pearson's ρ	0,895	0,909	0,935	0,962
	Spearman's rank order ρ	0,937	0,937	0,966	0,972
	Kendall's τ	0,797	0,797	0,860	0,879
BAK	Rmse	0,522	0,157	0,412	0,074
	Max Abs Error	0,858	0,357	0,719	0,256
	Pearson's ρ	0,988	0,990	0,993	0,998
	Spearman's rank order ρ	0,975	0,975	0,985	0,994
	Kendall's τ	0,886	0,886	0,917	0,959
OVRL	Rmse	0,266	0,206	0,177	0,127
	Max Abs Error	0,606	0,473	0,504	0,340
	Pearson's ρ	0,969	0,969	0,982	0,989
	Spearman's rank order ρ	0,974	0,974	0,980	0,989
	Kendall's τ	0,870	0,870	0,898	0,943

8.4 Validation database 3 (DES-25)

8.4.1 Database description

Validation database 3 was created using two commercial handsets. Each was operated in handset, headset, and hand-held handsfree mode, set up as described in ETSI TS 126 132 [i.14].

The speech source was full-band American English phonetically-balanced sentences, with two sentences from each of four male and four female talkers, for a total of sixteen samples, with 128 votes collected per condition.

In handset and headset modes, the speech was presented from a properly equalized HATS mouth at -1,7 dB Pa at MRP, while in HHHF mode, the level was +1,3 dB Pa at MRP.

The background noise was reproduced using a system according to ETSI TS 103 224 [i.19], using the noise types listed in table 8.5. Levels for these noise types can be found in ETSI TS 103 224 [i.19].

Table 8.5: Background noises used for validation database 3

Description	Filename according to ETSI TS 103 224 [i.19]	
	Handset & Headset	Hands-free
Sales counter	SalesCounter_handset	SalesCounter_handsfree
Callcenter	Callcenter2_handset	Callcenter2_handsfree
Cafeteria	Cafeteria_handset	Cafeteria_handsfree
Crossroads	Crossroadnoise_handset	Crossroadnoise_handsfree
Full-size Car 130 km/h	FullSizeCar_130_handset	FullSizeCar_130_handsfree
Pub	Pub_handset	Pub_handsfree
Road	Roadnoise_handset	Roadnoise_handsfree
Train Station	TrainStation_handset	TrainStation_handsfree

The handsets were placed in calls using a base station simulator, with EVS super wideband speech encoding [i.4] at 24,4 kbit/sec.

Full-band reference conditions were used, processed according to DESUDAPS-1 [i.17]. Presentation was binaural at 73 dB SPL, using headphones equalized to diffuse field.

8.4.2 Validation database 3: Results for model A

Results are shown as scatter plots, comparing instrumental predicted ratings to subjective ratings. Results from model A on validation database 3 for each of the three ratings, SIG, BAK, and OVRL, are shown in figure 8.3. For each rating (rows), two scatter plots are shown, one before a monotonic mapping is applied (right column) and one after a monotonic mapping is applied (left column).

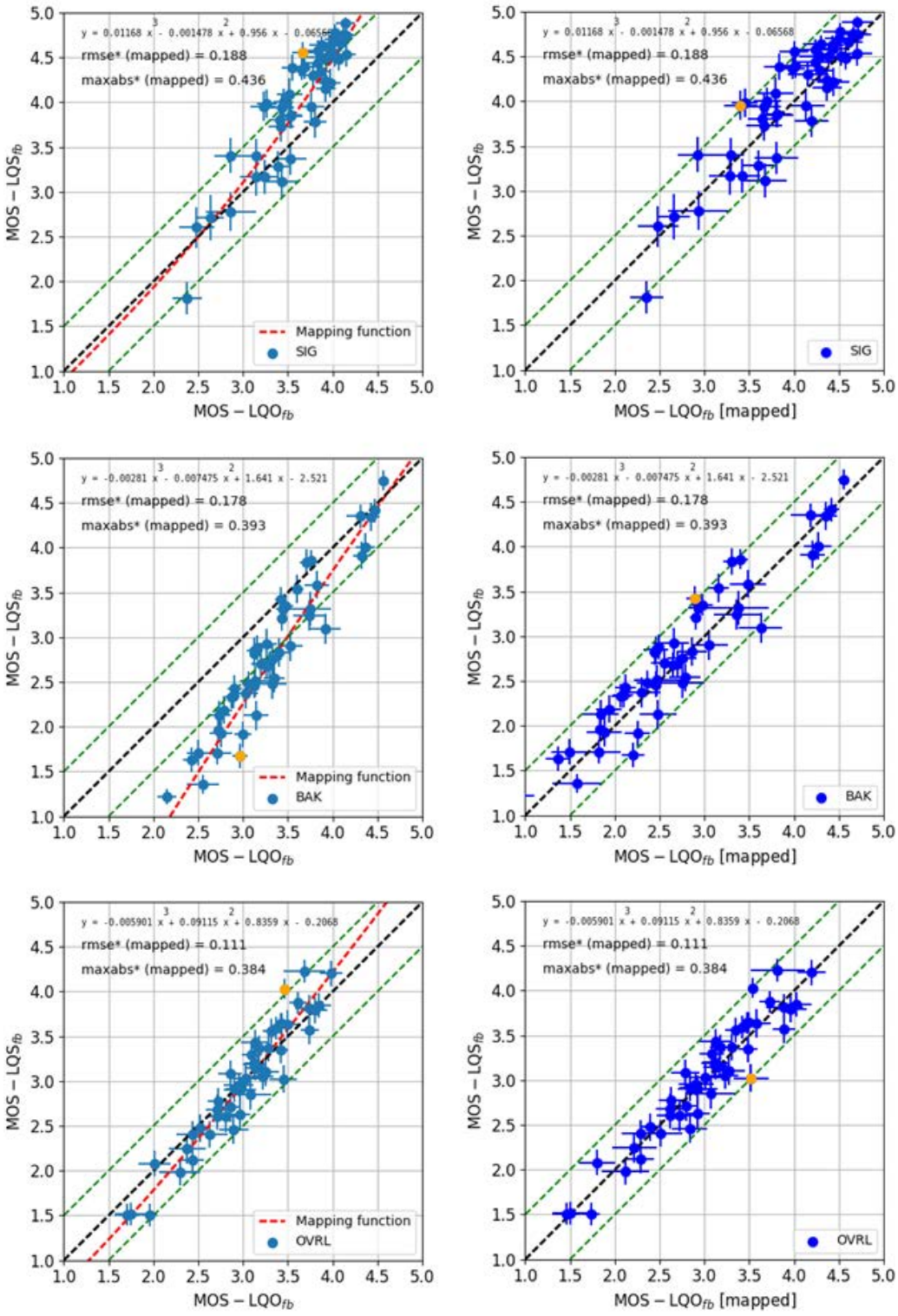


Figure 8.3: Scatter plots from model A for validation database 3

The rmse* and maximum absolute error* (maxabs*) after mapping are shown on all figures, with an orange-colored symbol indicating the condition with the largest overall maximum absolute error. The mapping polynomial is shown in the upper left corner of each panel. The dashed green lines show error of ±0,5 MOS. The error bars indicate the 95 % confidence interval before mapping (left column) and after mapping (right column).

Additional performance metrics, including Pearson's ρ correlation coefficient, Spearman's ρ rank order correlation, and Kendall's τ are shown in table 8.6.

Table 8.6: Performance metrics for model A on validation database 3

Dimension	Metric	Raw	Mapped	d*	Mapped & d*
SIG	Rmse	0,531	0,303	0,420	0,188
	Max Abs Error	0,885	0,555	0,766	0,436
	Pearson's ρ	0,914	0,911	0,937	0,963
	Spearman's rank order ρ	0,880	0,880	0,896	0,955
	Kendall's τ	0,703	0,703	0,730	0,830
BAK	Rmse	0,622	0,292	0,500	0,178
	Max Abs Error	1,292	0,534	1,153	0,393
	Pearson's ρ	0,948	0,949	0,955	0,981
	Spearman's rank order ρ	0,945	0,945	0,949	0,975
	Kendall's τ	0,815	0,815	0,826	0,887
OVRL	Rmse	0,230	0,207	0,135	0,111
	Max Abs Error	0,567	0,493	0,458	0,384
	Pearson's ρ	0,956	0,956	0,980	0,987
	Spearman's rank order ρ	0,943	0,943	0,979	0,983
	Kendall's τ	0,806	0,806	0,902	0,922

8.4.3 Validation database 3: Results for model B

Results are shown as scatter plots, comparing instrumental predicted ratings to subjective ratings. Results from model A on validation database 3 for each of the three ratings, SIG, BAK and OVRL, are shown in figure 8.4. For each rating (rows), two scatter plots are shown, one before a monotonic mapping is applied (right column) and one after a monotonic mapping is applied (left column).

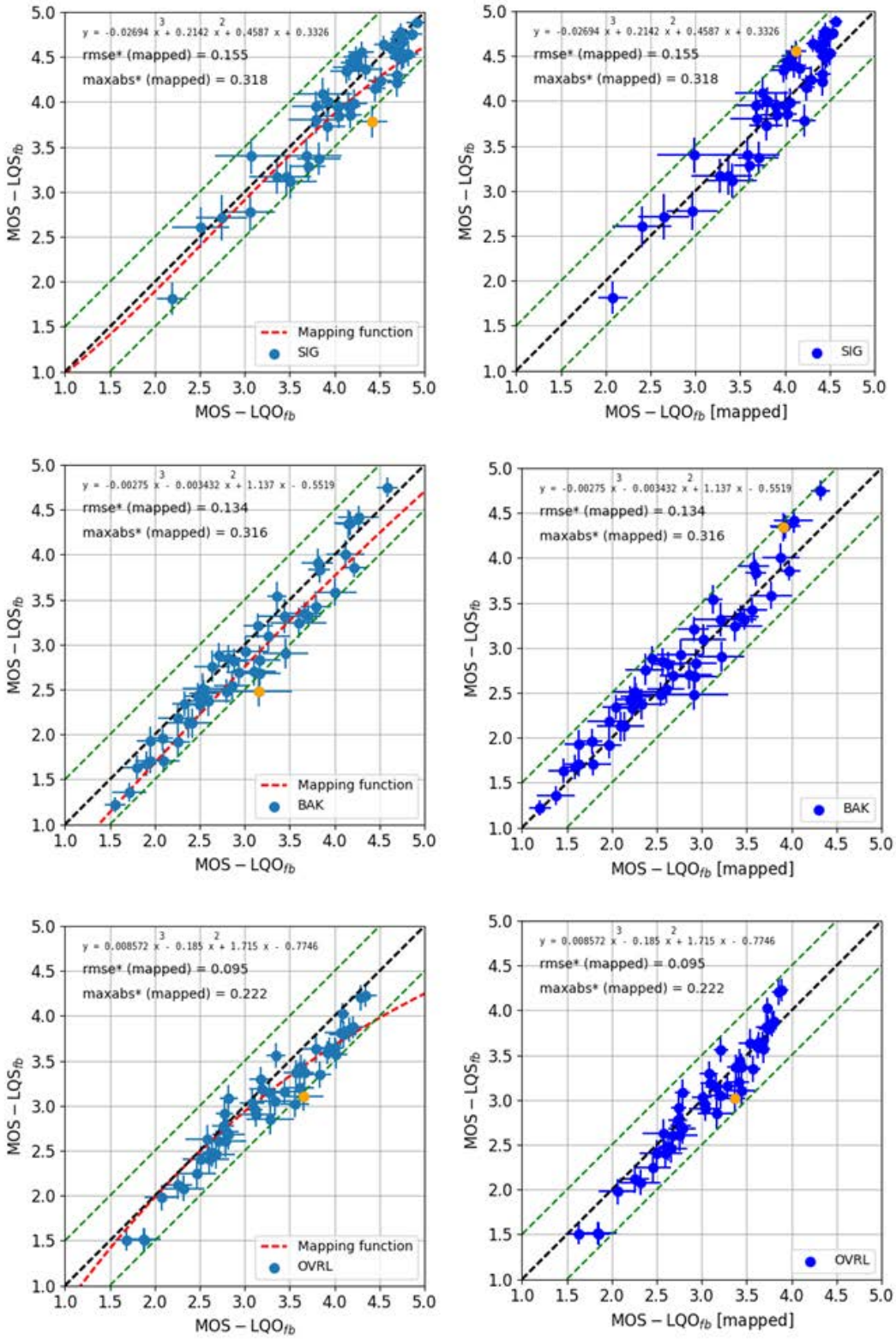


Figure 8.4: Scatter plots from model B for validation database 3

The rmse* and maximum absolute error* (maxabs*) after mapping are shown on all figures, with an orange-colored symbol indicating the condition with the largest overall maximum absolute error. The mapping polynomial is shown in the upper left corner of each panel. The dashed green lines show error of ±0,5 MOS. The error bars indicate the 95 % confidence interval before mapping (left column) and after mapping (right column).

Additional performance metrics, including Pearson's ρ correlation coefficient, Spearman's ρ rank order correlation, and Kendall's τ are shown in table 8.7.

Table 8.7: Performance metrics for model B on validation database 3

Dimension	Metric	Raw	Mapped	d*	Mapped & d*
SIG	Rmse	0,254	0,265	0,134	0,155
	Max Abs Error	0,633	0,434	0,452	0,318
	Pearson's ρ	0,939	0,940	0,981	0,979
	Spearman's rank order ρ	0,895	0,895	0,964	0,956
	Kendall's τ	0,737	0,737	0,871	0,842
BAK	Rmse	0,274	0,245	0,166	0,134
	Max Abs Error	0,668	0,433	0,498	0,316
	Pearson's ρ	0,968	0,967	0,986	0,991
	Spearman's rank order ρ	0,967	0,967	0,988	0,990
	Kendall's τ	0,865	0,865	0,927	0,934
OVRL	Rmse	0,265	0,195	0,125	0,095
	Max Abs Error	0,552	0,350	0,393	0,222
	Pearson's ρ	0,967	0,969	0,986	0,990
	Spearman's rank order ρ	0,968	0,968	0,987	0,993
	Kendall's τ	0,860	0,860	0,922	0,949

8.5 Validation database 4 (DES-26)

8.5.1 Database description

Validation database 4 was created using two commercial handsets, different from those of database 3. Each was operated in handset, headset, and hand-held handsfree mode, set up as described in ETSI TS 126 132 [i.14].

The speech source was full-band American English phonetically-balanced sentences, with two sentences from each of four male and four female talkers, for a total of sixteen samples, with 128 votes collected per condition.

In handset and headset modes, the speech was presented from a properly equalized HATS mouth at -1,7 dB Pa at MRP, while in HHHF mode, the level was +1,3 dB Pa at MRP.

The background noise was reproduced using a system according to ETSI ES 202 396-1 [i.1], using the noise types listed in table 8.8. Levels for these noise types can be found in ETSI ES 202 396-1 [i.1].

Table 8.8: Background noises used for validation database 4

Description	Filename according to ETSI ES 202 396-1 [i.1]
Crossroad	Outside_Traffic Crossroads_binaural
Cafeteria	Mensa binaural
Callcenter	Work Office Noise Callcenter binaural
Full-size Car 130 km/h	FullSizeCar_130kmh binaural
Road	Outside Traffic Road binaural
Train Station	Train Station binaural
Pub	Pub_Noise_binaural_V2
Salescounter	Cafeteria_Noise_binaural

The handsets were placed in calls using a base station simulator, with EVS super wideband speech encoding [i.4] at 13,2 kbit/sec.

Full-band reference conditions were used, processed according to DESUDAPS-1 [i.17]. Presentation was binaural at 73 dB SPL, using headphones equalized to diffuse field.

8.5.2 Validation database 4: Results for model A

Results are shown as scatter plots, comparing instrumental predicted ratings to subjective ratings. Results from model A on validation database 4 for each of the three ratings, SIG, BAK, and OVRL, are shown in figure 8.5. For each rating (rows), two scatter plots are shown, one before a monotonic mapping is applied (right column) and one after a monotonic mapping is applied (left column).

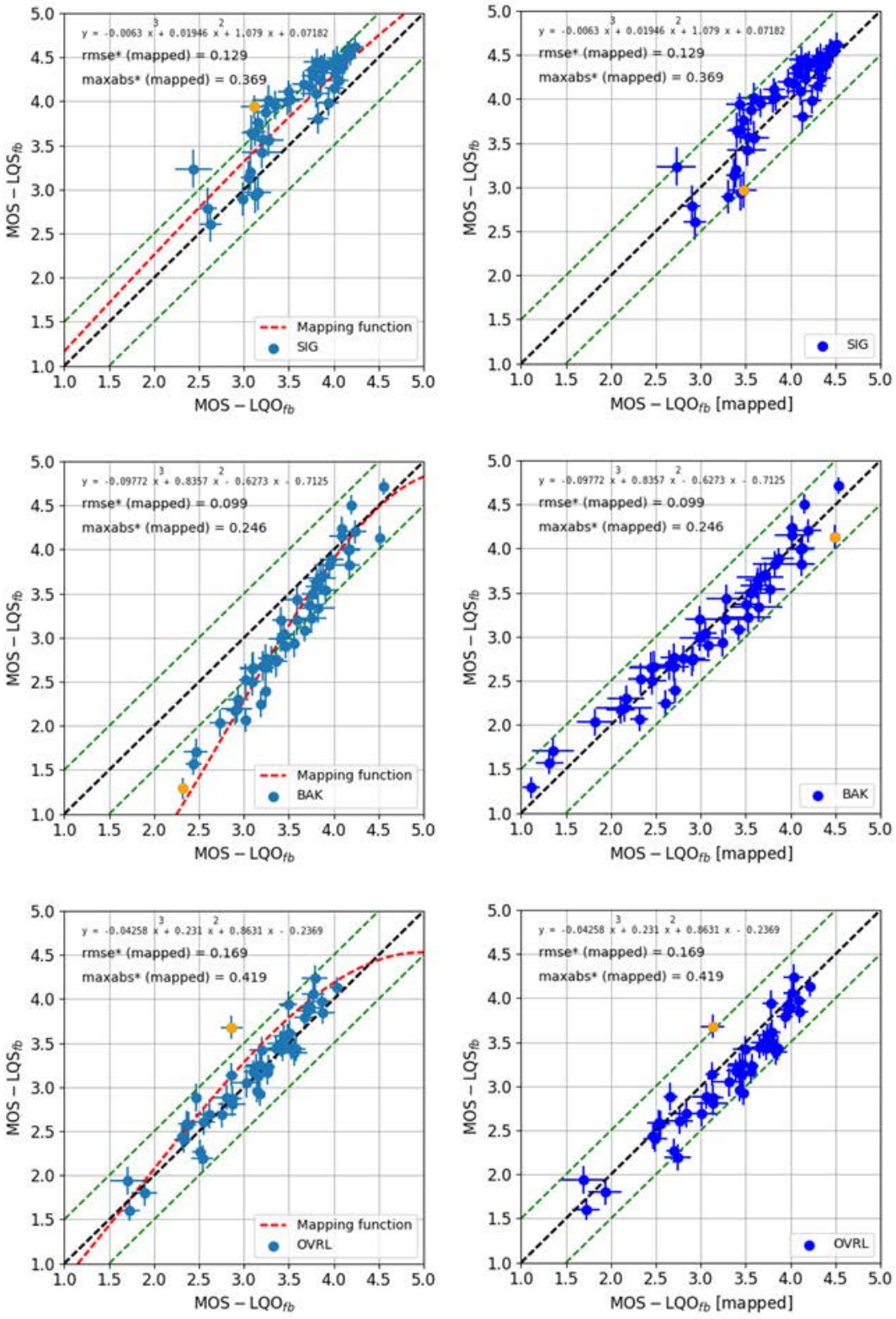


Figure 8.5: Scatter plots from model A for validation database 4

The rmse* and maximum absolute error* (maxabs*) after mapping are shown on all figures, with orange-colored symbol indicating the condition with the largest overall maximum absolute error. The mapping polynomial is shown in the upper left corner of each panel. The dashed green lines show error of ±0,5 MOS. The error bars indicate the 95 % confidence interval before mapping (left column) and after mapping (right column).

Additional performance metrics, including Pearson's ρ correlation coefficient, Spearman's ρ rank order correlation, and Kendall's τ are shown in table 8.9.

Table 8.9: Performance metrics for model A on validation database 4

Dimension	Metric	Raw	Mapped	d*	Mapped & d*
SIG	Rmse	0,456	0,261	0,323	0,129
	Max Abs Error	0,823	0,505	0,687	0,369
	Pearson's ρ	0,896	0,897	0,931	0,966
	Spearman's rank order ρ	0,904	0,904	0,920	0,972
	Kendall's τ	0,758	0,758	0,783	0,881
BAK	Rmse	0,523	0,203	0,398	0,099
	Max Abs Error	1,032	0,359	0,915	0,246
	Pearson's ρ	0,975	0,973	0,979	0,994
	Spearman's rank order ρ	0,980	0,980	0,982	0,995
	Kendall's τ	0,890	0,890	0,901	0,957
OVRL	Rmse	0,218	0,279	0,135	0,169
	Max Abs Error	0,815	0,550	0,684	0,419
	Pearson's ρ	0,948	0,942	0,976	0,974
	Spearman's rank order ρ	0,931	0,931	0,964	0,960
	Kendall's τ	0,801	0,801	0,897	0,854

8.5.3 Validation database 4: Results for model B

Results are shown as scatter plots, comparing instrumental predicted ratings to subjective ratings. Results from model A on validation database 4 for each of the three ratings, SIG, BAK and OVRL, are shown in figure 8.6. For each rating (rows), two scatter plots are shown, one before a monotonic mapping is applied (right column) and one after a monotonic mapping is applied (left column).

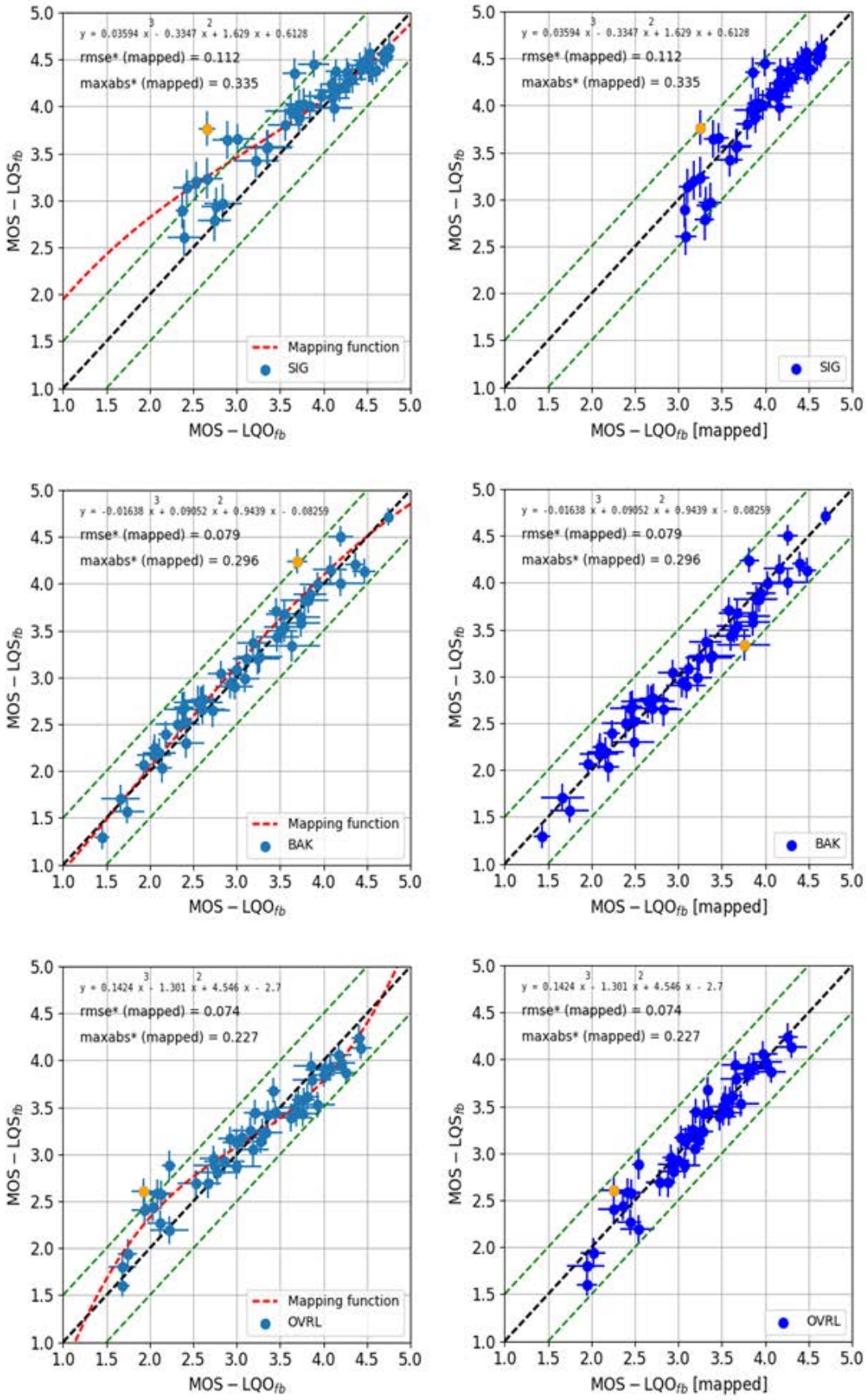


Figure 8.6: Scatter plots from model B for validation database 4

The rmse* and maximum absolute error* (maxabs*) after mapping are shown on all figures, with orange-colored symbol indicating the condition with the largest overall maximum absolute error. The mapping polynomial is shown in the upper left corner of each panel. The dashed green lines show error of ±0,5 MOS. The error bars indicate the 95 % confidence interval before mapping (left column) and after mapping (right column).

Additional performance metrics, including Pearson's ρ correlation coefficient, Spearman's ρ rank order correlation, and Kendall's τ are shown in table 8.10.

Table 8.10: Performance metrics for model B on validation database 4

Dimension	Metric	Raw	Mapped	d*	Mapped & d*
SIG	Rmse	0,346	0,210	0,239	0,112
	Max Abs Error	1,110	0,515	0,921	0,335
	Pearson's ρ	0,934	0,928	0,963	0,977
	Spearman's rank order ρ	0,933	0,933	0,971	0,978
	Kendall's τ	0,797	0,797	0,895	0,927
BAK	Rmse	0,175	0,178	0,084	0,079
	Max Abs Error	0,545	0,425	0,416	0,296
	Pearson's ρ	0,978	0,979	0,995	0,996
	Spearman's rank order ρ	0,977	0,977	0,993	0,995
	Kendall's τ	0,892	0,892	0,416	0,296
OVRL	Rmse	0,257	0,167	0,160	0,074
	Max Abs Error	0,670	0,350	0,528	0,227
	Pearson's ρ	0,962	0,967	0,985	0,993
	Spearman's rank order ρ	0,969	0,969	0,989	0,993
	Kendall's τ	0,861	0,861	0,930	0,956

8.6 Validation database 5 (DES-27)

8.6.1 Database description

Validation database 5 was created using three commercial handsets, different from those of databases 3 and 4. Two were operated in handset, headset, and hand-held handsfree mode, while one was operated in handset and hand-held speakerphone modes. Set up was as described in ETSI TS 126 132 [i.14].

The speech source was full-band American English phonetically-balanced sentences, with two sentences from each of four male and four female talkers, for a total of sixteen samples, with 128 votes collected per condition.

In handset and headset modes, the speech was presented from a properly equalized HATS mouth at -1,7 dB Pa at MRP, while in HHHF mode, the level was +1,3 dB Pa at MRP.

The background noise was reproduced using a system according to ETSI TS 103 224 [i.19], using the noise types listed in table 8.11. Levels for these noise types can be found in ETSI TS 103 224 [i.19].

Table 8.11: Background noises used for validation database 5

Description	Filename according to ETSI TS 103 224 [i.19]	
	Handset & Headset	Hands-free
Sales counter	SalesCounter_handset	SalesCounter_handsfree
Callcenter	Callcenter2_handset	Callcenter2_handsfree
Cafeteria	Cafeteria_handset	Cafeteria_handsfree
Crossroads	Crossroadnoise_handset	Crossroadnoise_handsfree
Full-size Car 130 km/h	FullSizeCar_130_handset	FullSizeCar_130_handsfree
Pub	Pub_handset	Pub_handsfree
Road	Roadnoise_handset	Roadnoise_handsfree
Train Station	TrainStation_handset	TrainStation_handsfree

The handsets were placed in calls using a base station simulator, with EVS super wideband speech encoding [i.4] at 13,2 kbit/sec.

Full-band reference conditions were used, processed according to DESUDAPS-1 [i.17]. Presentation was binaural at 73 dB SPL, using headphones equalized to diffuse field.

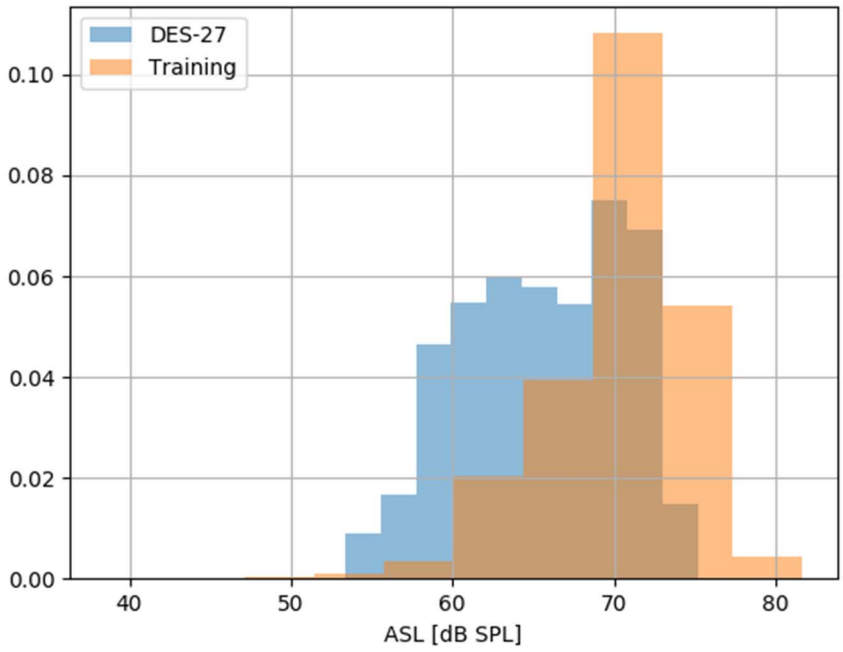


Figure 8.7: Histogram of active speech level distribution of DES-27 compared to level distribution of training databases

Figure 8.7 shows that the level distribution of database DES-27 is significantly lower compared to the level distribution of the training databases. This is due to large SLR differences observed with the terminals available for this database. Since some of them do not even pass the SLR requirements it is believed that these terminals were delivered in a very early stage when super-wideband certification requirements were not yet available. It is expected that future terminals, with respect to the first publication of the present document, will follow the specifications more closely. The performance of the models for this database is somewhat lower for this database, possibly due to the effect of SLR. See also clause 9 for proper application of the models.

8.6.2 Validation database 5: Results for model A

Results are shown as scatter plots, comparing instrumental predicted ratings to subjective ratings. Results from model A on validation database 4 for each of the three ratings, SIG, BAK and OVRL, are shown in figure 8.8. For each rating (rows), two scatter plots are shown, one before a monotonic mapping is applied (right column) and one after a monotonic mapping is applied (left column).

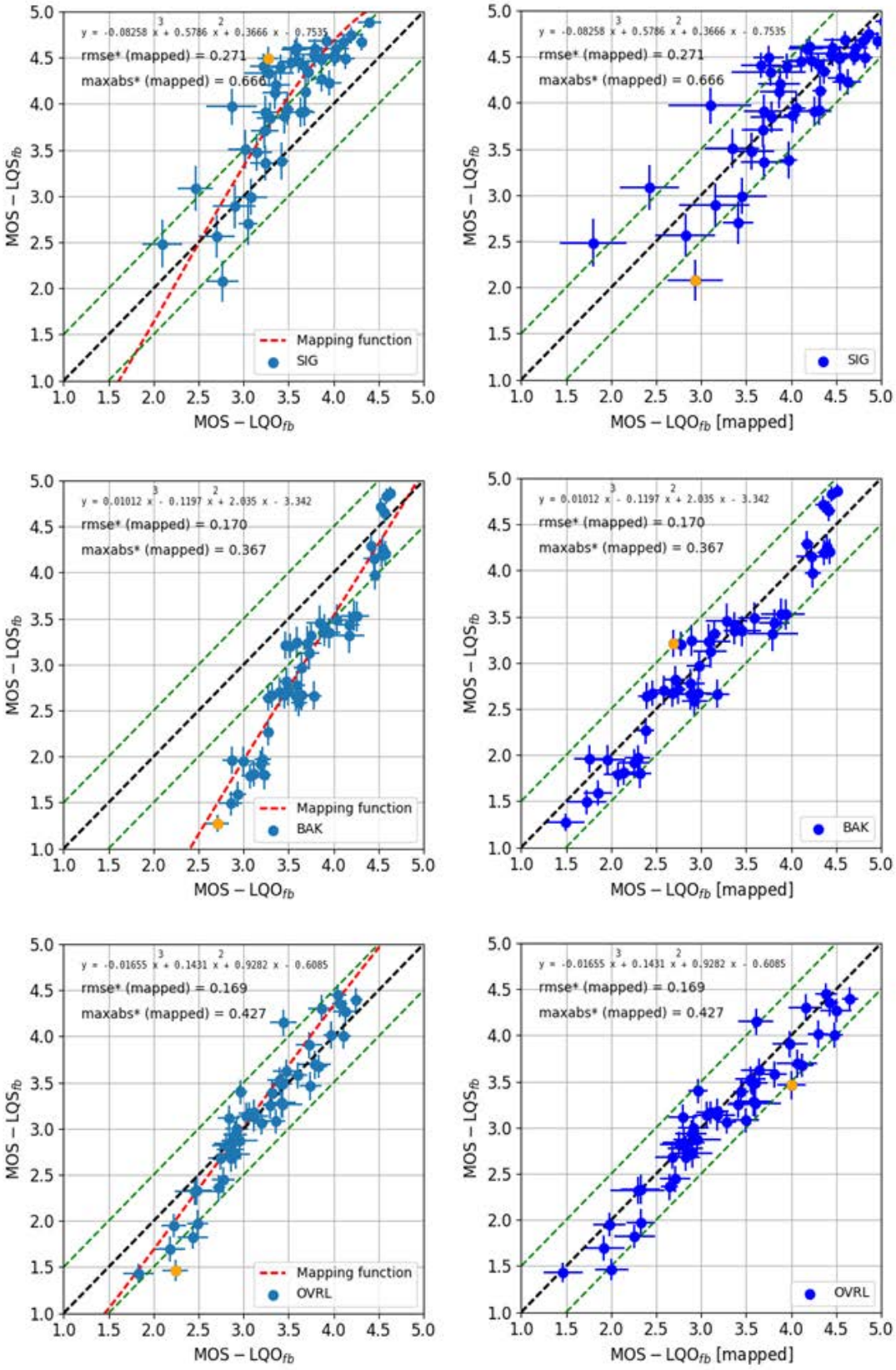


Figure 8.8: Scatter plots from model A for validation database 5

The rmse* and maximum absolute error* (maxabs*) after mapping are shown on all figures, with orange-colored symbol indicating the condition with the largest overall maximum absolute error. The mapping polynomial is shown in the upper left corner of each panel. The dashed green lines show error of ±0,5 MOS. The error bars indicate the 95 % confidence interval before mapping (left column) and after mapping (right column).

Additional performance metrics, including Pearson's ρ correlation coefficient, Spearman's ρ rank order correlation, and Kendall's τ are shown in table 8.12.

Table 8.12: Performance metrics for model A on validation database 5

Dimension	Metric	Raw	Mapped	d*	Mapped & d*
SIG	Rmse	0,662	0,399	0,538	0,271
	Max Abs Error	1,217	0,860	1,093	0,666
	Pearson's ρ	0,827	0,838	0,864	0,924
	Spearman's rank order ρ	0,827	0,827	0,839	0,926
	Kendall's τ	0,651	0,651	0,670	0,796
BAK	Rmse	0,811	0,285	0,682	0,170
	Max Abs Error	1,434	0,519	1,337	0,367
	Pearson's ρ	0,957	0,957	0,960	0,984
	Spearman's rank order ρ	0,952	0,952	0,955	0,982
	Kendall's τ	0,830	0,830	0,839	0,920
OVRL	Rmse	0,282	0,266	0,195	0,169
	Max Abs Error	0,779	0,542	0,664	0,427
	Pearson's ρ	0,955	0,955	0,973	0,980
	Spearman's rank order ρ	0,964	0,964	0,987	0,986
	Kendall's τ	0,843	0,843	0,924	0,917

8.6.3 Validation database 5: Results for model B

Results are shown as scatter plots, comparing instrumental predicted ratings to subjective ratings. Results from model A on validation database 4 for each of the three ratings, SIG, BAK and OVRL, are shown in figure 8.9. For each rating (rows), two scatter plots are shown, one before a monotonic mapping is applied (right column) and one after a monotonic mapping is applied (left column).

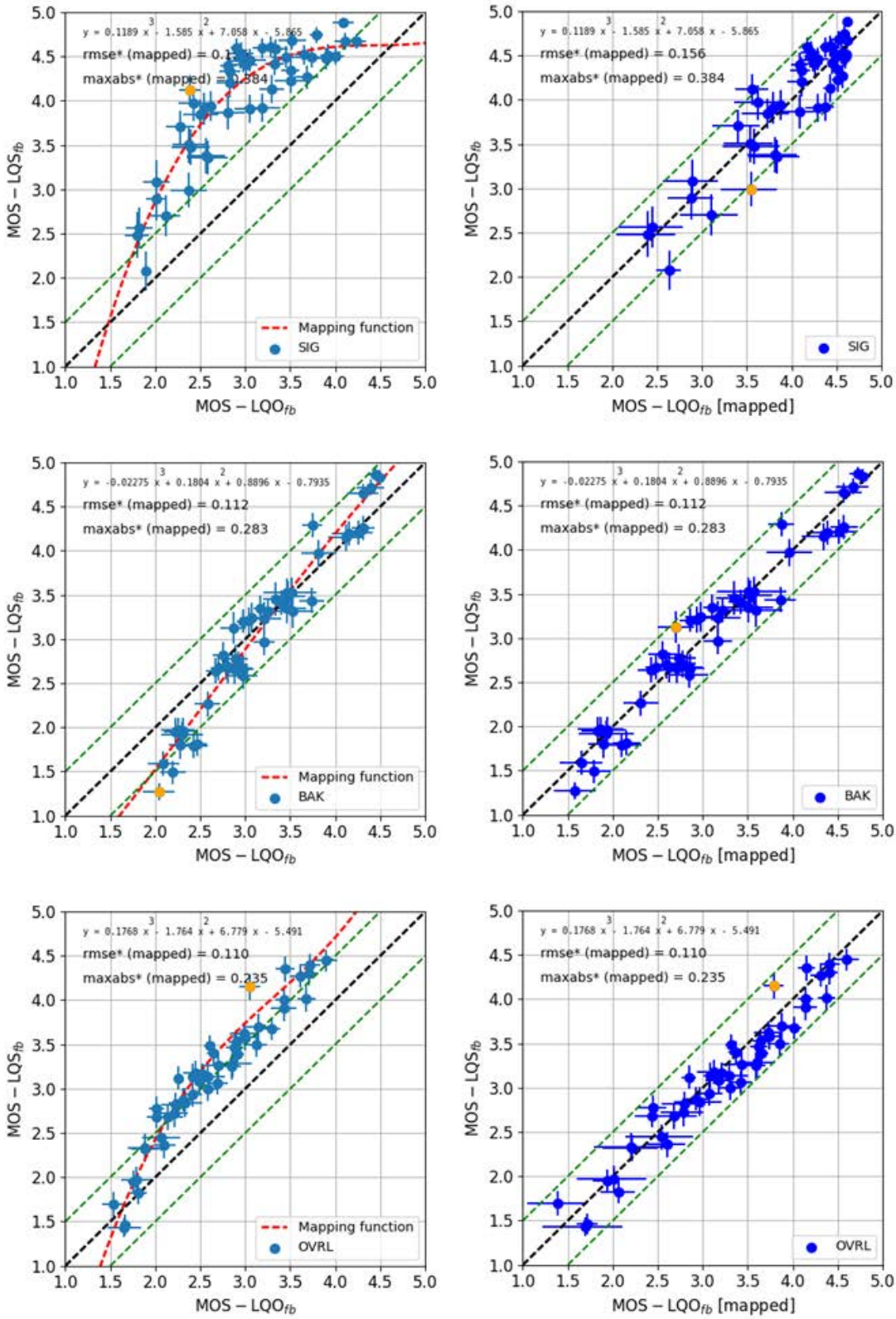


Figure 8.9: Scatter plots from model B for validation database 5

The rmse* and maximum absolute error* (maxabs*) after mapping are shown on all figures, with orange-colored symbol indicating the condition with the largest overall maximum absolute error. The mapping polynomial is shown in the upper left corner of each panel. The dashed green lines show error of ±0,5 MOS. The error bars indicate the 95 % confidence interval before mapping (left column) and after mapping (right column).

Additional performance metrics, including Pearson's ρ correlation coefficient, Spearman's ρ rank order correlation, and Kendall's τ are shown in table 8.13.

Table 8.13: Performance metrics for model B on validation database 5

Dimension	Metric	Raw	Mapped	d*	Mapped & d*
SIG	Rmse	1,118	0,287	0,977	0,156
	Max Abs Error	1,734	0,556	1,580	0,384
	Pearson's ρ	0,835	0,919	0,844	0,973
	Spearman's rank order ρ	0,842	0,842	0,846	0,951
	Kendall's τ	0,660	0,660	0,668	0,830
BAK	Rmse	0,315	0,217	0,220	0,112
	Max Abs Error	0,765	0,424	0,668	0,283
	Pearson's ρ	0,973	0,974	0,985	0,993
	Spearman's rank order ρ	0,964	0,964	0,985	0,994
	Kendall's τ	0,848	0,848	0,918	0,954
OVRL	Rmse	0,571	0,215	0,443	0,110
	Max Abs Error	1,098	0,361	0,957	0,235
	Pearson's ρ	0,956	0,971	0,963	0,992
	Spearman's rank order ρ	0,977	0,977	0,976	0,994
	Kendall's τ	0,884	0,884	0,879	0,949

9 Application of the models

9.1 Introduction

In order to avoid ambiguities in the results, the objective model should be applied in the way it was applied during the training process which also reflects the listening test.

9.2 Speech material

The speech samples used in conjunction with the model should be the ones used in the subjective tests of the training databases. At least 16 sentences of male and female speakers shall be used per measurement run. Each sentence shall be centered in a time window of 4,0 seconds. The minimum duration of an active speech material shall be 1,0 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 3,0 seconds, which correspond to a minimum leading/trailing silence period of 0,5 seconds.

For proper convergence of noise reduction system, the source sequence should contain in addition an initial silence period, as well as at least four different sentences from four different talkers in the beginning.

A set of 16 full-band Chinese, American English and German sentences meeting the requirements described above are provided in annex E. The preferred test sequence is American English because the majority of training and validation material was provided with this vector set.

9.3 Positioning of the device under test

For testing in the handset use case, variations from nominal position defined in clause 8.4 of Recommendation ITU-T P.64 [i.18] representing typical use conditions can also be used.

For testing headsets, positioning recommendations found in Recommendation ITU-T P.380 [i.40] should be followed.

For testing handheld hands-free, positioning recommendations can be found in the corresponding standards, e.g. in ETSI TS 126 132 [i.14].

For testing car hands-free, positioning recommendations can be found in the corresponding standards, e.g. in [i.41].

For each test position, it shall be ensured that the terminal meets the nominal SLR requirements by ±6 dB.

9.4 Background noise playback

The setups according to ETSI ES 202 396-1 [i.1] and ETSI TS 103 224 [i.19] are recommended for the generation of the background noise. The background noises to use in conjunction with the model shall be taken from these two standards.

9.5 Recording and calibration procedure

In general, the signal to be evaluated with the models is recorded at an electrical reference interface. Typically the signal is obtained in a certain physical unit (like Volt) or on a digital scale (e.g. 16-bit integer representation). The level calibration to the acoustical domain (in dB Pa or dB SPL) according to the subjective tests of the training databases is performed as follows.

Prior to the recording of noisy scenarios, a clean speech recording at the nominal position is used to find the correction factor. Note that no Lombard gain (usually +3 dB) for noisy scenarios is applied here. This preparation measurement is carried out with a default active speech level of -4,7 dB Pa in handset / headset mode and -1,7 dB Pa in (handheld) hands-free mode.

The goal of this calibration is to determine a factor, which scales the measured signal in silence to an overall active speech level according to [i.7] of 73 dB SPL, evaluated over all sentences (excluding the convergence sequence). This method preserves individual per-talker levels and leads to a uniformly residual noise level. This correction factor is then used for all test conditions.

9.6 Running the prediction models

The measured sequence as well as the clean speech reference are cropped into segments of 4,0 seconds, corresponding to the sentences described in clause 9.2. Compensating any delay between the two signals is possible, but not required (time alignment is included in the models). Silence/noise parts shall preserve a minimum duration of 0,5 seconds (see also clause 9.2).

The results for S-, N- and G-MOS shall be calculated on a per sentence basis and averaged over all samples. In addition, the standard deviation should be reported to indicate large variance and/or outliers (samples outside of the expected range of values).

9.7 Mapping function for Model A

9.7.1 Derivation of mapping functions

In order to derive proposed mapping functions, the validation databases DES-25 and -26 are investigated. These databases consist of recordings made with real terminals and thus can be regarded as realistic use cases of both models. In order to analyse if one of the models systematically over- or underestimates a certain attribute, figure 9.1 to figure 9.3 show the merged data of the two databases (2x48 conditions = 96 data points) within one scatter plot per attribute.

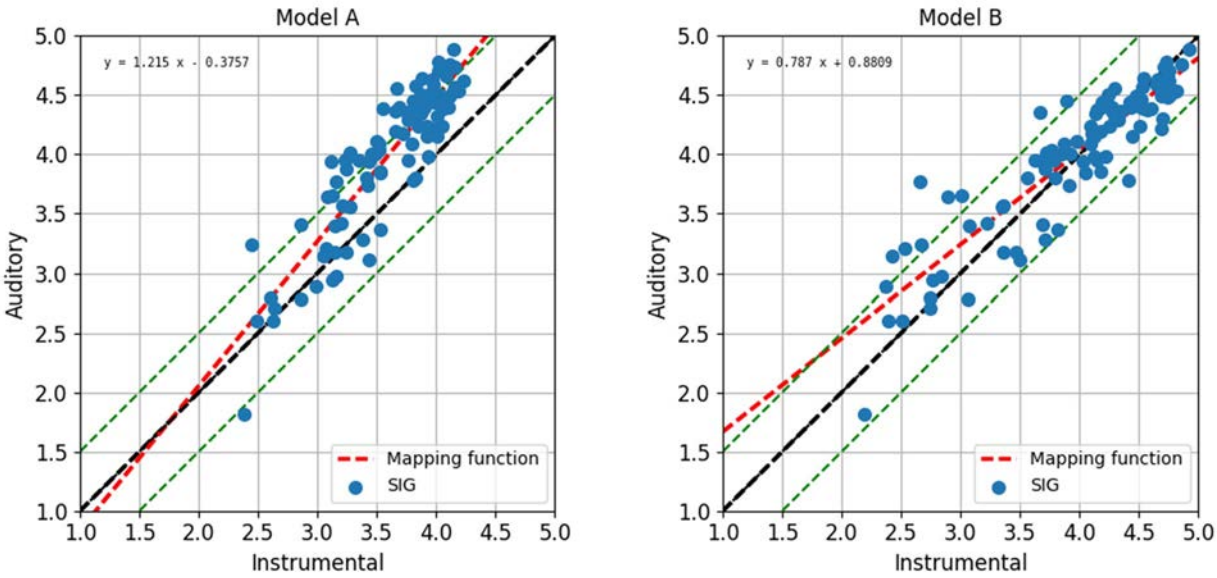


Figure 9.1: Prediction results for databases DES-25/-26 for S-MOS

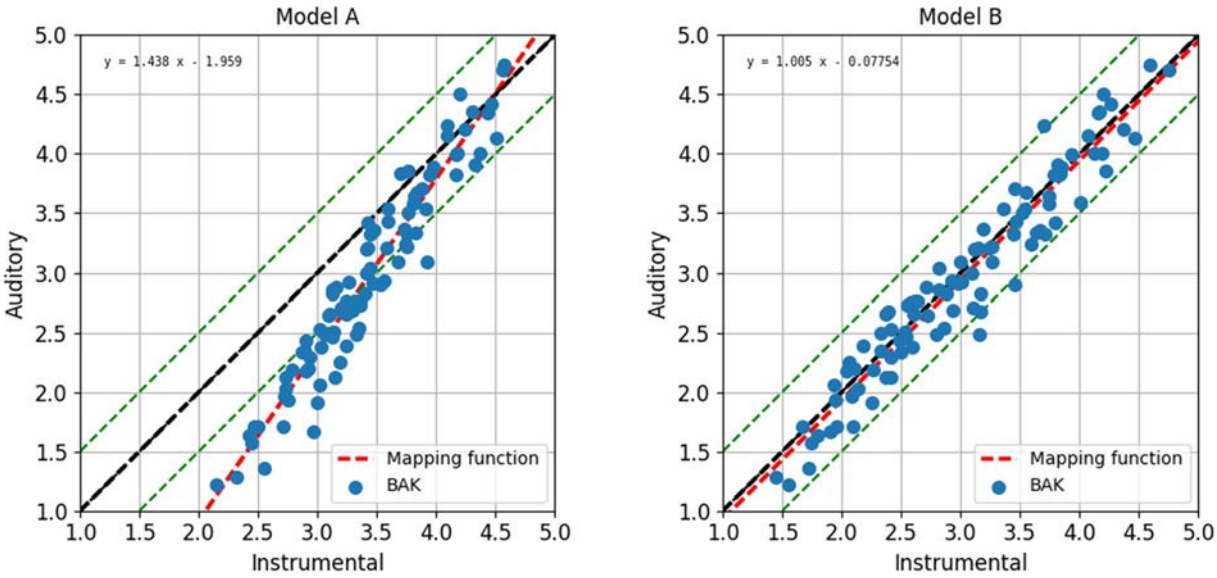


Figure 9.2: Prediction results for databases DES-25/-26 for N-MOS

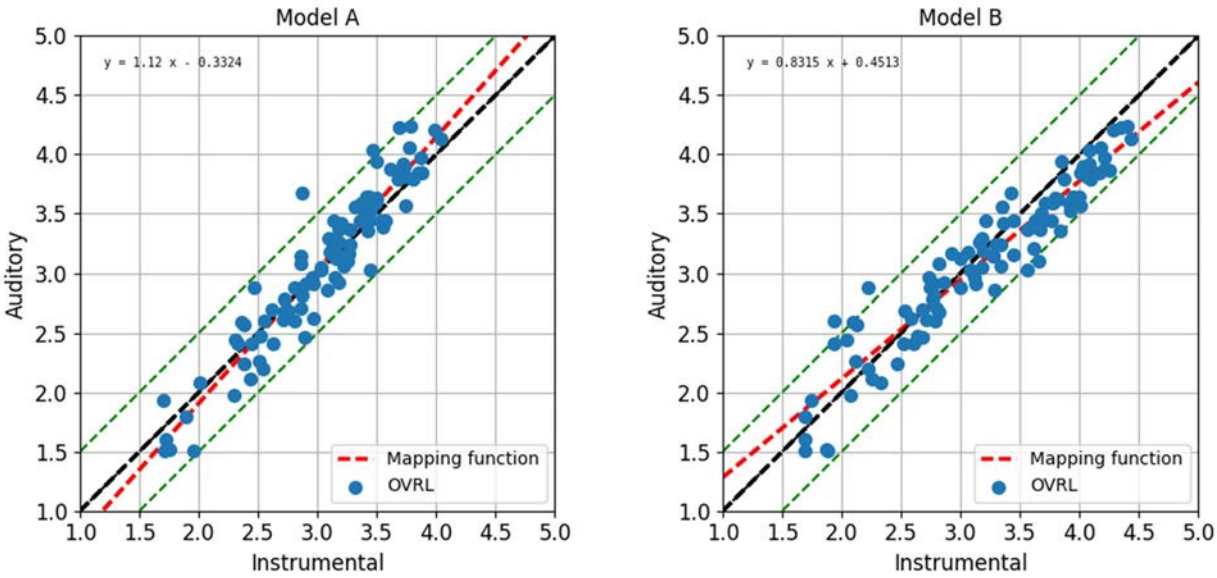


Figure 9.3: Prediction results for databases DES-25/-26 for G-MOS

To avoid mapping function for both models and the three categories (as shown in figure 9.1 to figure 9.3), map only model A is mapped to the scores of model B, since model B provides instrumental results which are closer to the auditory ones. By mapping results of model A, it is possible to define common performance requirements which are valid for both models.

For this purpose, figure 9.4 to figure 9.6 illustrate results of model B vs A for each attribute. In addition, each plot provides a linear mapping function, which can be used to transform results from model A into the range of model B.

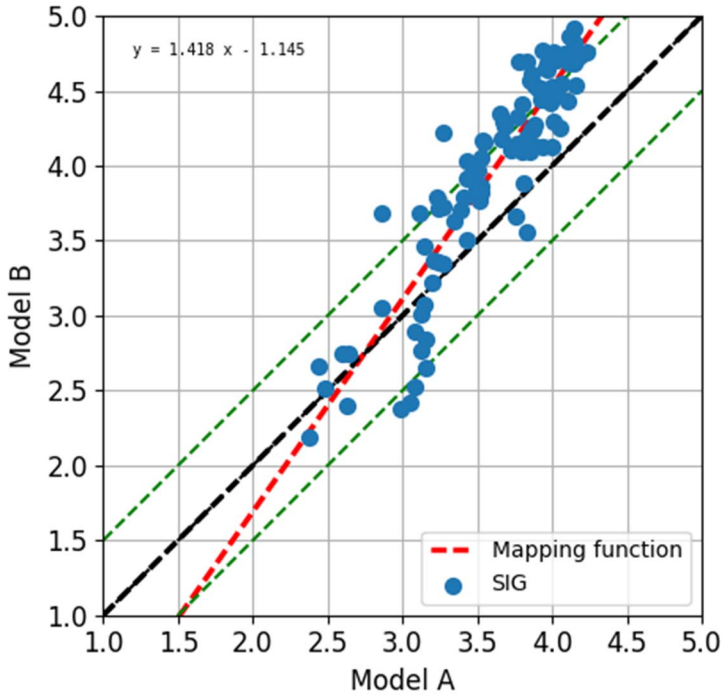


Figure 9.4: Prediction results of model B vs A for databases DES-25/-26 (S-MOS)

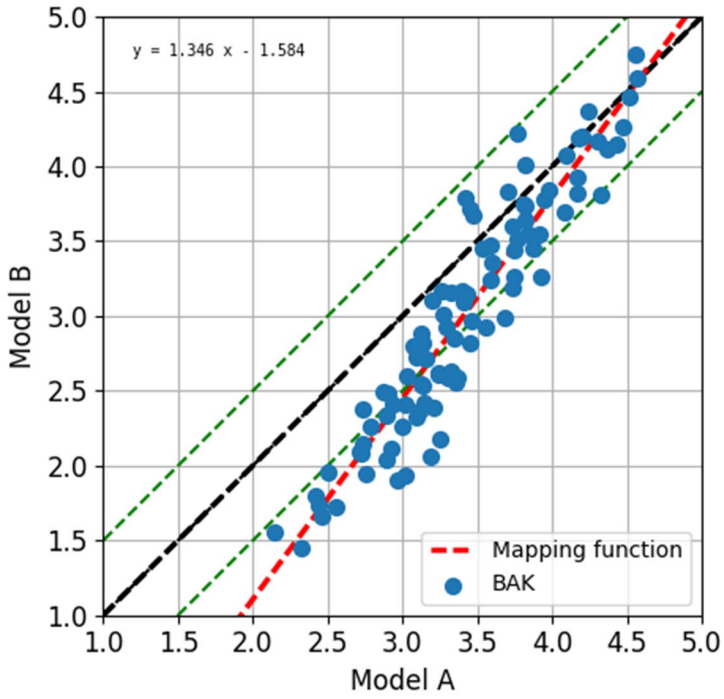


Figure 9.5: Prediction results of model B vs A for databases DES-25/-26 (N-MOS)

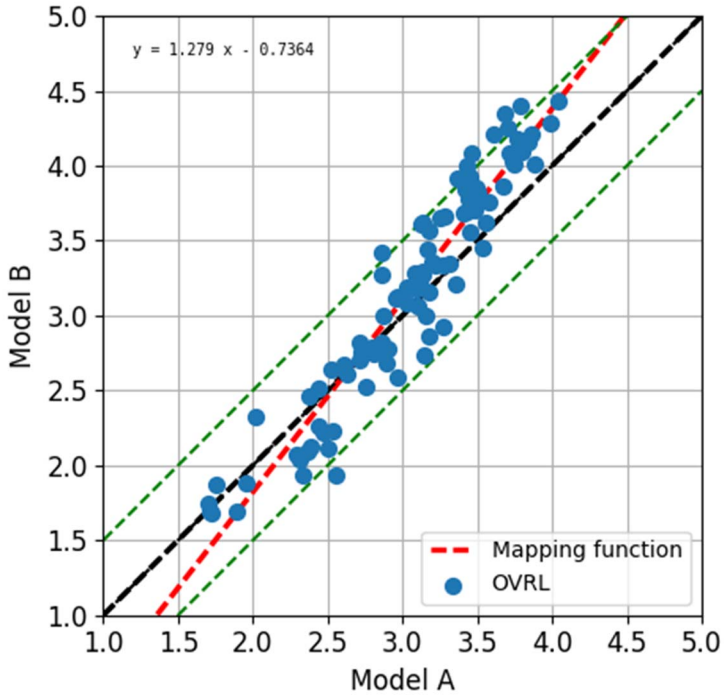


Figure 9.6: Prediction results of model B vs A for databases DES-25/-26 (G-MOS)

9.7.2 Resulting mapping functions

In order to transform results of model A for the evaluation of performance requirements, the following functions may be used:

$$\text{S-MOS}'_{\text{LQOfb}} = 1{,}418 \cdot \text{S-MOS}_{\text{LQOfb}} - 1{,}145$$
$$\text{N-MOS}'_{\text{LQOfb}} = 1{,}346 \cdot \text{N-MOS}_{\text{LQOfb}} - 1{,}584$$
$$\text{G-MOS}'_{\text{LQOfb}} = 1{,}279 \cdot \text{G-MOS}_{\text{LQOfb}} - 0{,}7364$$

Such mapping may be applied if a comparison between results derived from the different models should be compared or unified.

NOTE: Other mapping functions may be used for different purposes. In ETSI TS 126 132 [i.14] the following mapping is used:

S-MOS-LQOfb = S-MOS-LQOfb_modelA

N-MOS-LQOfb = 1,438 × N-MOS-LQOfb_modelA - 1,959

G-MOS-LQOfb = G-MOS-LQOfb_modelA

Annex A (normative): Model configuration files

A.1 Introduction

This annex provides implementation details for models A and B.

A.2 Model A

Random forest regression configuration files are provided in a simple text format. Setup information of the trees and coefficients are denoted line by line in the following format:

Line #1: Number of trees (T)

Line #2: Number of nodes for tree 1 (N_1)

Line #3-6: Description of node 1 of tree 1

- Feature number (1-indexed).
- Feature value to split.
- Following nodes: if split value is lower than feature value, the tree continues with node number mentioned first. Otherwise, the second node number is used.
- Average MOS and standard deviation of this node. If there is no following node (indicated by following nodes 0/0), this is the last one (leaf) and the value determines the predicted MOS output for this single tree.

Line #7-(2 + 4* N_1): node setup is repeated N_1-1 times (for nodes 2.. N_1)

Line #(2 + 4* N_1 + 1): setup for tree 2 (N_2 nodes)

Line #(2 + 4* N_1 + 4* N_2 + 1): setup for tree 3 (N_3 nodes)

…

Line #(2 + 4* N_1 + … + 4* N_T + 1): setup for tree T (N_T nodes)

The feature names used as inputs for the regression are provided in the last line of each configuration file.

NOTE: The configuration files are contained in archive ts_103281v010301p0.zip which accompanies the present document:

- S-MOS model.
- N-MOS model.
- G-MOS model.

A.3 Model B

Deep Neural Network Configuration per Cochlear Prediction model implementation.

NOTE: The *.csv files are contained in archive ts_103281v010301p0.zip which accompanies the present document.

S-MOS Densing layer Parameters:

- SMOS_dense_1_b.csv
- SMOS_dense_1_W.csv
- SMOS_dense_2_b.csv
- SMOS_dense_2_W.csv
- SMOS_dense_3_b.csv
- SMOS_dense_3_W.csv

N-MOS Densing layer Parameters:

- NMOS_dense_1_b.csv
- NMOS_dense_1_W.csv
- NMOS_dense_2_b.csv
- NMOS_dense_2_W.csv
- NMOS_dense_3_b.csv
- NMOS_dense_3_W.csv

G-MOS Densing layer Parameters:

- G-MOS_dense_1_b.csv
- G-MOS_dense_1_W.csv
- G-MOS_dense_2_b.csv
- G-MOS_dense_2_W.csv
- G-MOS_dense_3_b.csv
- G-MOS_dense_3_W.csv

Annex B (normative): Summary of Training Databases

Table B.1

DB#	Source	Bandwidth	Codec(s)	Use case	Noise types	BGN System	Language	Validation for model
DES-01	HEAD acoustics GmbH	FB & SWB	EVS-FB@32kbps, EVS-SWB@24.4kbps	HS	Cafeteria, Road, Car, Pub, Train	ETSI TS 103 224 [i.19]	German	-
DES-02	HEAD acoustics GmbH	FB & SWB	EVS-FB@32kbps, EVS-SWB@24.4kbps	HS	Cafeteria, Road, Car, Pub, Train	ETSI TS 103 224 [i.19]	German	-
DES-03	HEAD acoustics GmbH	FB, SWB, WB, NB	EVS-FB@16.4 kbps, EVS-SWB@13.2kbps, OPUS-SWB@24kbps, OPUS-FB@28kbps, EVS-WB@9.6kbps, EVS-NB@9.6kbps	HS	Cafeteria, Road, Car, Pub, Train	ETSI TS 103 224 [i.19]	German	-
DES-04	Knowles, Inc	SWB	none	HS	Cafeteria, Road, Car, Pub, Train, Crossroad, Sales Counter, Call Center	ETSI TS 103 224 [i.19]	American English	-
DES-05	Knowles, Inc	FB	none	HS	Cafeteria, Road, Car, Pub, Train, Crossroad, Sales Counter, Call Center	ETSI TS 103 224 [i.19]	American English	-
DES-06	Knowles, Inc	FB	none	HHHF	Cafeteria, Road, Car, Pub, Train, Crossroad, Sales Counter, Call Center	ETSI TS 103 224 [i.19]	American English	-
DES-07	Knowles, Inc	SWB, WB, NB	none	HS & HHHF	Train, Road	ETSI TS 103 224 [i.19]	American English	-
DES-08	Knowles, Inc	SWB, WB, NB	EVS-SWB@13.2kbps, AMR-WB@12.65kbps, EVS-NB@9.6kbps	HS & HHHF	Train, Road	ETSI TS 103 224 [i.19]	American English	-
DES-09	Knowles, Inc	SWB, WB, NB	EVS-SWB@13.2kbps, AMR-WB@12.65kbps, EVS-NB@9.6kbps	HS & HHHF	Car, Road, Train, Pub, Sales Counter, Music, Airport departure hall	ETSI TS 103 224 [i.19]	American English	-
DES-10	Qualcomm, Inc.	FB, SWB, WB, NB	EVS-FB@24.4kbps, EVS-SWB@13.2kbps, AMR-WB@12.65kbps, AMR@12.2kbps	HS & HHHF	Car	ETSI ES 202 396-1 [i.1]	American English	-
DES-11	Qualcomm, Inc.	FB, SWB, WB, NB	EVS-FB@24.4kbps, EVS-SWB@13.2kbps, AMR-WB@12.65kbps, AMR@12.2kbps	HS & HHHF	Car	ETSI ES 202 396-1 [i.1]	Mandarin	-

DB#	Source	Bandwidth	Codec(s)	Use case	Noise types	BGN System	Language	Validation for model
DES-12	Knowles, Inc	SWB	EVS-SWB@13.2kbps	HS & HHHF	Car, Road, Train, Pub, Music, Airport departure hall	ETSI TS 103 224 [i.19]	American English	-
DES-13	Knowles, Inc	SWB	EVS-SWB@13.2kbps	HS & HHHF	Car, Road, Train, Pub, Music, Airport departure hall	ETSI TS 103 224 [i.19]	American English	-
DES-14	Knowles, Inc	SWB	EVS-SWB@13.2kbps	HS & HHHF	Car, Road, Train, Pub, Music, Airport departure hall	ETSI TS 103 224 [i.19]	Mandarin	-
DES-15	Knowles, Inc	SWB	EVS-SWB@13.2kbps	HS	Cafeteria, Road, Car, Pub, Train, Crossroad, Sales Counter, Call Center	ETSI TS 103 224 [i.19]	American English	-
DES-16	Knowles, Inc	SWB	EVS-SWB@13.2kbps	HHHF	Cafeteria, Road, Car, Pub, Train, Crossroad, Sales Counter, Call Center	ETSI TS 103 224 [i.19]	American English	-
DES-17	HEAD acoustics GmbH	SWB	EVS-SWB@24.4kbps, EVS-SWB@13.2kbps	HS & HHHF	Cafeteria, Road, Car, Pub, Train, Crossroad, silence	ETSI TS 103 224 [i.19]	German	B
DES-19	Qualcomm, Inc.	FB, SWB, WB, NB	EVS-FB@24.4kbps, EVS-SWB@13.2kbps, AMR-WB@12.65kbps, AMR@12.2kbps	HS & HHHF	Car	ETSI ES 202 396-1 [i.1]	Mandarin	-
DES-20	Qualcomm, Inc.	SWB	EVS-SWB@13.2kbps	HS & HHHF	Silence, Car, Café, Train, Road, Callctr	ETSI ES 202 396-1 [i.1]	Mandarin	A
DES-22	Qualcomm, Inc.	SWB	EVS-SWB@24.4kbps	HS & HHHF	Silence, Pub, Mensa, Car	ETSI ES 202 396-1 [i.1]	American English	-
DES-23	HEAD acoustics GmbH	SWB	EVS-SWB@13.2kbps	HS & HHHF	Car, Road, Train, Pub, Music, Airport departure hall	ETSI TS 103 224 [i.19]	German	-
DES-24	Knowles, Inc	SWB	EVS-SWB@13.2kbps	HS & HHHF	Car, Road, Train, Pub, Music, Airport departure hall	ETSI TS 103 224 [i.19]	American English	-
DES-25	Knowles, Inc	SWB	EVS-SWB@24.4kbps	HS & HE & HHHF	Cafeteria, Road, Car, Pub, Train, Crossroad, Sales Counter, Call Center	ETSI TS 103 224 [i.19]	American English	A, B
DES-26	Knowles, Inc	SWB	EVS-SWB@13.2kbps	HS & HE & HHHF	Cafeteria, Road, Car, Pub, Train, Crossroad, Sales Counter, Call Center	ETSI ES 202 396-1 [i.1]	American English	A, B
DES-27	Knowles, Inc	SWB	EVS-SWB@13.2kbps	HS & HE & HHHF	Cafeteria, Road, Car, Pub, Train, Crossroad, Sales Counter, Call Center	ETSI TS 103 224 [i.19]	American English	A, B

Annex C (normative): Test vectors for model verification

The test vectors for verification of an objective model implementation are given in this annex. A model implemented according to clause 6 and claiming to be compliant to the present document shall achieve all scores with an accuracy of ±0,1 MOS. For calibration purposes, the RMS level (unweighted) is specified as well in table C.1 (calculation carried out on the entire sample). The speech samples can be downloaded here: https://docbox.etsi.org/stq/Open/TS%20103%20281%20Wave%20files/Annex_C%20test%20vectors.

Table C.1: Test vectors and instrumental results for the models

				Model A			Model B		
Condition	**Talker**	**Sample**	**RMS [dBPa]**	**S-MOS**	**N-MOS**	**G-MOS**	**S-MOS**	**N-MOS**	**G-MOS**
C01	f1	s1	-24,51	3,43	3,14	2,90	3,86	2,79	2,70
C02	f1	s1	-22,60	2,16	1,91	1,28	2,64	1,49	1,55
C03	f1	s1	-28,46	3,33	2,83	2,64	3,96	2,21	2,42
C04	f1	s1	-23,59	2,82	2,31	1,93	3,99	1,51	2,30
C05	f1	s1	-24,29	3,32	2,66	2,65	4,51	2,15	2,70
C06	f1	s1	-26,10	3,77	2,92	2,99	4,24	2,91	3,01
C07	f1	s1	-23,24	3,88	3,39	3,25	4,29	3,49	3,27
C08	f1	s1	-25,73	4,00	3,00	3,14	4,42	2,57	2,91
C09	f1	s1	-24,50	3,71	3,57	3,40	4,15	3,81	3,40
C10	f1	s1	-33,46	2,73	2,56	2,01	2,19	1,91	1,66
C11	f1	s1	-25,86	3,40	3,19	2,89	3,31	2,42	2,65
C12	f1	s1	-19,15	2,55	2,32	1,76	3,28	1,53	1,79
C13	f1	s1	-22,92	3,23	2,87	2,36	3,47	2,09	2,24
C14	f1	s1	-25,58	3,19	3,16	2,58	3,86	2,73	2,71
C15	f1	s1	-23,42	4,15	3,75	3,54	4,70	3,32	3,47
C16	f1	s1	-23,86	3,90	3,36	3,28	4,75	2,96	3,40
C17	f1	s1	-21,12	4,28	4,18	4,01	5,17	4,44	4,38
C18	f1	s1	-22,14	3,99	3,75	3,48	4,90	4,41	4,31
C19	f1	s1	-24,84	3,79	2,38	2,90	4,46	1,91	2,10
C20	f1	s1	-22,91	4,23	3,30	3,40	4,75	3,55	3,66
C21	f1	s1	-22,91	4,00	2,95	3,17	5,02	2,33	3,09
C22	f1	s1	-22,22	4,17	3,44	3,37	5,00	3,93	4,03
C23	f1	s1	-22,13	4,07	3,22	3,23	4,99	2,79	3,52
C24	f1	s1	-21,76	4,20	3,47	3,40	4,98	3,80	3,99
C25	f1	s1	-22,83	4,29	4,46	4,03	5,10	4,82	4,56
C26	f1	s1	-23,16	4,01	4,54	4,07	4,99	4,88	4,65
C27	f1	s1	-24,26	3,20	3,53	2,68	3,34	3,23	3,00
C28	f1	s1	-23,26	3,82	4,45	3,69	4,61	4,29	4,24
C29	f1	s1	-23,25	3,97	3,88	3,53	4,84	2,96	3,48
C30	f1	s1	-23,19	3,91	4,57	3,89	4,74	4,34	4,34
C31	f1	s1	-22,86	3,98	4,16	3,73	4,88	3,25	3,83
C32	f1	s1	-22,83	4,13	4,41	3,87	4,91	4,48	4,43
C33	f1	s1	-22,21	4,01	3,43	3,39	4,08	3,20	3,26
C34	f1	s1	-26,82	3,11	2,21	2,01	3,77	1,70	1,85
C35	f1	s1	-24,86	4,05	3,04	3,23	4,12	2,73	3,02
C36	f1	s1	-24,59	3,64	2,68	2,75	4,48	2,15	2,70
C37	f1	s1	-22,39	4,05	3,33	3,30	4,53	2,97	3,24
C38	f1	s1	-22,26	3,60	3,01	2,97	4,23	2,40	2,85
C39	f1	s1	-21,62	3,95	3,33	3,24	4,84	3,37	3,68
C40	f1	s1	-19,67	4,20	4,13	3,88	4,81	3,94	3,90
C41	f1	s1	-23,00	3,69	3,67	3,24	4,34	4,03	3,85
C42	f1	s1	-24,61	2,14	3,06	1,89	2,47	2,20	1,80
C43	f1	s1	-23,07	3,85	3,33	3,27	4,41	3,16	3,42
C44	f1	s1	-21,29	3,12	2,45	2,34	4,19	1,86	2,36
C45	f1	s1	-22,99	3,59	3,28	3,04	4,28	3,04	3,15
C46	f1	s1	-22,21	3,42	2,72	2,70	4,26	2,22	2,65
C47	f1	s1	-22,75	3,92	3,58	3,42	4,75	3,20	3,68
C48	f1	s1	-22,73	4,16	3,92	3,61	4,74	3,81	3,99

Annex D (informative): Subjective testing framework

D.1 Introduction

This annex is an excerpt of [i.17] and describes the framework for conducting subjective testing used for the training and validation of the model described in the present document. Such a framework is seen as necessary in order to minimize variations between subjective tests performed in different listening laboratories. The framework can be used for conducting further subjective experiments which are intended to be compatible with the prediction model (validation, verification, further development).

D.2 Subjective test plan

D.2.1 Traceability

The subjective test method is described in Recommendation ITU-T P.835 [i.5] and the ITU-T Handbook of subjective testing practical procedures [i.15]. With the additional observations given in the following clauses D.2.2 to D.2.9, traceability can be improved.

D.2.2 Speech database requirements

The source speech database (near end signal) to be used for data collection and listening tests needs to consist of at least 8 samples (2 male and 2 female talkers, 2 samples per talker).

The speech material needs to conform to the guidelines specified in the ITU-T handbook of subjective testing practical procedures, clause 5, and clause B.3 of Recommendation ITU-T P.501 [i.10]. Each sample needs to be constructed according to the guidelines described in Recommendation ITU-T P.835 [i.5] clause 5.1.4 (including 1 second of leading and 1 second of trailing silence) and normalized to an active speech level [i.7] of -26 dBov. It is recommended that the source speech material be 16 bit/48 kHz.

D.2.3 Reference Conditions

Reference conditions need to follow the proposal in [i.16], which incorporates a spectral subtraction based distortion instead of the MNRU-based distortion typically used in subjective tests. The conditions used for the new SIG reference system and specification for NS Levels are listed in table D.1, the flow chart of the generation process is shown in figure D.1. Further details as well as an example implementation can be found in [i.17].

D.2.4 Test Conditions

Test conditions need to be recorded from real handset devices or from mock-up terminals for offline processing as described in clause D.3. Table D.2 lists the recommended test conditions used for the recordings and listening tests. At least 6 out of the 8 noise types described should be included in the test to provide similarity of context between different labs. 2 of the 8 noise types can be replaced by either a clean speech transmission scenario (i.e. the background noise reproduction is disabled) or other noise types taken from the ETSI ES 202 396-1 [i.1] database (except for the Male Single Voice Distractor noise type, see note).

NOTE: As speech and music carry contextual information, they can be viewed as a separate class of distractors and more study was felt necessary for their inclusion.

Either handset, headset or handheld hands-free usage modes are acceptable. The inclusion of hands-free test and headset cases is optional and intended to span a larger range of degradations for the purposes of re-training of the objective predictor model.

The preferred size equals 48 test conditions per database. This procedure provides a reasonable balance between reference (20 %) and test samples (80 %). As borderline cases, the amount of test conditions per database should be between 12 (50 %) minimum and 80 (85 %).

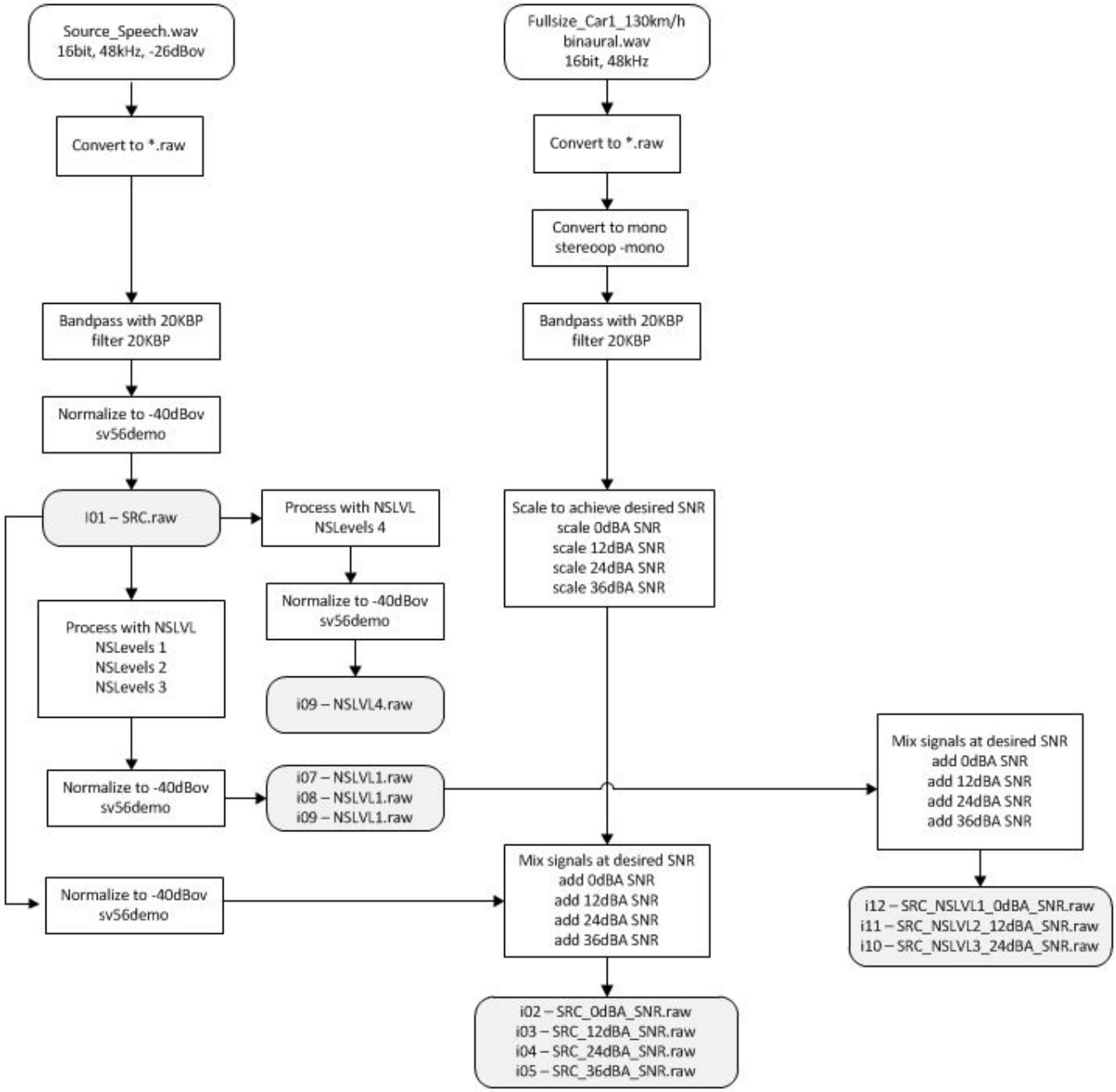


Figure D.1: Generation of reference conditions

D.2.5 Post-processing of test conditions

The uplink recordings of processed speech materials are normalized for use in the subjective tests. For the test conditions, the normalization gain is the gain necessary to obtain a recorded active speech level of -26 dBov with a clean speech condition (no noise applied in the room). As a result, this normalization gain needs to be applied to all other test conditions for the same device (noise suppressed speech signals). In this way, the effect of level changes introduced by terminals in the presence of noise needs to be part of the quality measurement.

Table D.1: Reference conditions for super-wideband and fullband subjective evaluation of noise reduction

Reference Conditions				
Condition	**Speech Distortion**	**SNR (A)**	**Noise Type**	**Description**
i01	Source	No Noise	-	Best anchor for SIG, BAK, OVRL
i02	Source	0 dB	Fullsize_Car1_130Kmh_binaural	Lowest anchor for BAK
i03	Source	12 dB	Fullsize_Car1_130Kmh_binaural	[…]
i04	Source	24 dB	Fullsize_Car1_130Kmh_binaural	[…]
i05	Source	36 dB	Fullsize_Car1_130Kmh_binaural	Second-best anchor for BAK
i06	NS Level 1	No Noise	-	Lowest anchor for SIG
i07	NS Level 2	No Noise	-	[…]
i08	NS Level 3	No Noise	-	[…]
i09	NS Level 4	No Noise	-	Second-best anchor for SIG
i10	NS Level 3	24 dB	Fullsize_Car1_130Kmh_binaural	Second-best anchor for OVRL
i11	NS Level 2	12 dB	Fullsize_Car1_130Kmh_binaural	[…]
i12	NS Level 1	[0 dB]	Fullsize_Car1_130Kmh_binaural	Lowest anchor for OVRL

Table D.2: Test conditions per device for super-wideband and fullband subjective evaluation of noise reduction

Test Conditions			
Speech level @ MRP Handset/ handsfree	**Noise level @ HATS ear simulators with ID correction**	**Noise Type**	**Description of Noise from ETSI ES 202 396-1 [i.1]**
-1,7/+1,3 dBPa	L: 75,0 dB(A) / R: 73,0 dB(A)	Pub_Noise_binaural_V2	Recording in a pub
-1,7/+1,3 dBPa	L: 74,9 dB(A) / R: 73,9 dB(A)	Outside_Traffic_Road_binaural	Recording at pavement
-1,7/+1,3 dBPa	L: 69,1 dB(A) / R: 69,6 dB(A)	Outside_Traffic_Crossroads_binaural	Recording at pavement
-1,7/+1,3 dBPa	L: 68,2 dB(A) / R:69,8 dB(A)	Train_Station_binaural	Recording at departure platform
-1,7/+1,3 dBPa	L: 69,1 dB(A) / R: 68,1 dB(A)	Fullsize_Car1_130Kmh_binaural	Recording in passenger cabin
-1,7/+1,3 dBPa	L: 68,4 dB(A) / R: 67,3 dB(A)	Cafeteria_Noise_binaural	Recording at sales counter
-1,7/+1,3 dBPa	L: 63,4 dB(A) / R: 61,9 dB(A)	Mensa_binaural	Recording in a cafeteria
-1,7/+1,3 dBPa	L: 56,6 dB(A) / R: 57,8 dB(A)	Work_Noise_Office_Callcenter_binaural	Recording in a business office

D.2.6 Calibration and equalization of headphones for presentation

Headphones used for presentation of the test material to the listening panel should be calibrated and equalized using a HATS conforming to Recommendation ITU-T P.58 [i.11] and an artificial ear type 3.3 according to Recommendation ITU-T P.57 [i.12]. The HATS is diffuse field equalized. The resulting one-third octave frequency response characteristic of the headphones used in the subjective experiments should be within the mask given in Recommendation ITU-R BS.708 [i.42], annex 1, figure 1.

Alternatively, equalization can be made using a subjective method as in IEC 60268-7:2010 [i.43], ensuring that all frequencies for full-band listening are satisfactorily reproduced.

The presentation of the test and reference conditions to listeners should be diotic. The system gain is adjusted so that a speech segment of -26 dBov corresponds to a presentation level of 73 dB SPL measured at the DRP with diffuse-field equalization.

D.2.7 Requirements on the listening laboratory

Listening laboratory facilities need to comply with the recommendations provided in Recommendation ITU-T P.800 [i.13].

D.2.8 Experimental design

The use of the Balanced Blocks experimental design described in [i.15], clause 3.3.2 is recommended. The experimental design needs to include the 12 reference conditions and 8 test conditions per device under test, described in table D.1. A minimum of two and a maximum of six devices needs to be included in any one test.

The test and reference conditions should be reported for a total of 32 naïve listeners. The listeners need to be native speakers of the language used for the test.

An example of subjective test presentation sequence (i.e. randomizations) is provided in annex A of [i.17] for a test with 32 listeners, 4 talkers, 4 samples per talker, 12 reference conditions and 48 test conditions (6 devices and 8 noise types). Each of the 4 presentation sequences in annex A are presented to 8 of the 32 listeners.

128 votes per condition should be obtained, in order to achieve adequately low variance per condition. The number of votes per sample will depend on the number of samples per talker chosen (see clause D.2.2). A minimum of 2 samples per talker and 12 votes per sample should be obtained in order to achieve adequately low variance per sample for model training purposes.

D.2.9 Training session

Prior to administration of the test, subjects need to be provided with written instructions on the test procedures. The use of training materials (e.g. videos, presentations) is encouraged to ensure the participants fully understand the task being requested. The training session needs to be followed by a practice session containing 16 trials. The practice session needs to include conditions representative of those presented in the test. An example is provided in table D.3.

Table D.3: Example sequence for training session

Trial	Sample	Condition
1	m1s3.r01	Reference - Source/No noise
2	f2s1.x06	Test - Cafeteria
3	m2s4.r11	Reference - NS-L2/12 dB SNR
4	f1s1.r02	Reference - Source/0 dB SNR
5	m2s3.x03	Test - Traffic-crossroads
6	f1s1.x05	Test - Fullsize car
7	m2s1.r07	Reference - NS-L2/No noise
8	f2s2.x02	Traffic-road
9	m2s2.r03	Reference - Source/12 dB SNR
10	f2s2.r06	Reference - NSL1/No noise
11	m2s4.x01	Pub
12	f2s3.x08	Test - Call-center
13	m2s4.r04	Reference - Source/24 dB SNR
14	f2s1.x04	Test - Train station
15	m2s3.r12	Reference - NS-L1/0 dB SNR
16	f2s3.x07	Test - Mensa babble
NOTE: x is a device outside the set of DUTs.		

D.3 Set-up for acquisition of test conditions

D.3.1 Terminal positioning and HATS calibration

For reproduction of the near-end signal, a HATS conforming to Recommendation ITU-T P.58 [i.11] is used. The mouth simulator needs to be equalized to achieve the reproduction accuracy described in ETSI TS 126 132 [i.14], clause 5.3.

For handset and headset mode testing, the mouth sensitivity gain needs to be adjusted to produce an active speech level of -1,7 dBPa at MRP for a -26 dBov input speech signal.

The handset terminals or mock-ups under test need to be set-up on HATS and the handset mounting position documented as described in ETSI TS 126 132 [i.14], clause 5.1.1.

Headsets need to be set-up on HATS as described in ETSI TS 126 132 [i.14], clause 5.1.2.

For handheld hands-free mode the device is set-up using HATS as described in ETSI TS 126 132 [i.14], clause 5.1.3.3.

For handheld hands-free mode testing, the mouth sensitivity gain needs to be adjusted to produce an active speech level of +1,3 dBPa at MRP for a -26 dBov input speech signal.

D.3.2 Background Noise reproduction

The background noise reproduction system should be setup and equalized according to ETSI ES 202 396-1 [i.1] and/or ETSI TS 103 224 [i.19]. Noise types should be reproduced at their realistic levels according to ETSI ES 202 396-1 [i.1], clause 8 and ETSI TS 103 224 [i.19], clause 7.2. The test conditions and noise files are specified in table D.1.

D.3.3 Noise and speech playback synchronization

The noise and speech playback needs to be time aligned and synchronized. This is generally the case when playing the noise and speech files out of multiple channels of a same hardware interface but appropriate synchronization needs to be ensured when using separate hardware for noise and speech playback.

D.3.4 Convergence sequence

For proper convergence of terminal noise suppression the following time sequencing should be applied:

1) the terminal is set-up and a call is established in noise free conditions;

2) 2 seconds of noise only is applied in the test room with a linear amplitude fade-in from 0 to 2 seconds (noise ramp-up period), immediately followed by;

3) 6 seconds of noise only, immediately followed by;

4) 16 seconds (4 samples) of simultaneous speech and noise, immediately followed by;

5) actual test material to be used for listening panel presentation.

The speech sequence provided in annex C already includes the points addressed above.

D.3.5 Example of noise and speech playback sequence including convergence period

Figure D.1 illustrates an example of a playback time history for speech and one particular noise signal (Fullsize car 1 at 130 km/h, binaural). The following applies to the example in figure D.1:

1) The speech signal is constructed by concatenating 8 seconds of silence with 36 speech samples of 4 seconds each. The total length is therefore 152 seconds. The first 24 seconds according to clause D.3.4 are used for convergence of the noise suppression algorithm and not used for the purposes of listening panel presentation.

2) The noise signal is constructed by concatenating 6 repetitions of a noise sample and the first 8 seconds of the 7^{th} repetition. The noise sample is cut out, or generated from, the original noise file in ETSI ES 202 396-1 [i.1] database to be 24 seconds in length, and fade-in and fade-out processing is applied to the first and last 50 samples (assuming noise at 48 kHz sampling rate) to ensure zero-crossing of the signal amplitude at beginning and end of the sample. A linear fade-in is applied to the first 2 seconds of the concatenated noise signal, as this was found necessary for proper convergence of some terminals.

It is noted that by looping the noise every 24 seconds (a multiple of the speech sample length of 4 seconds) the sharp transitions in the noise amplitude at the looping point coincide with the location of sample cutting for listening panel presentation. This avoids audible sharp transitions to fall during a speech segment.

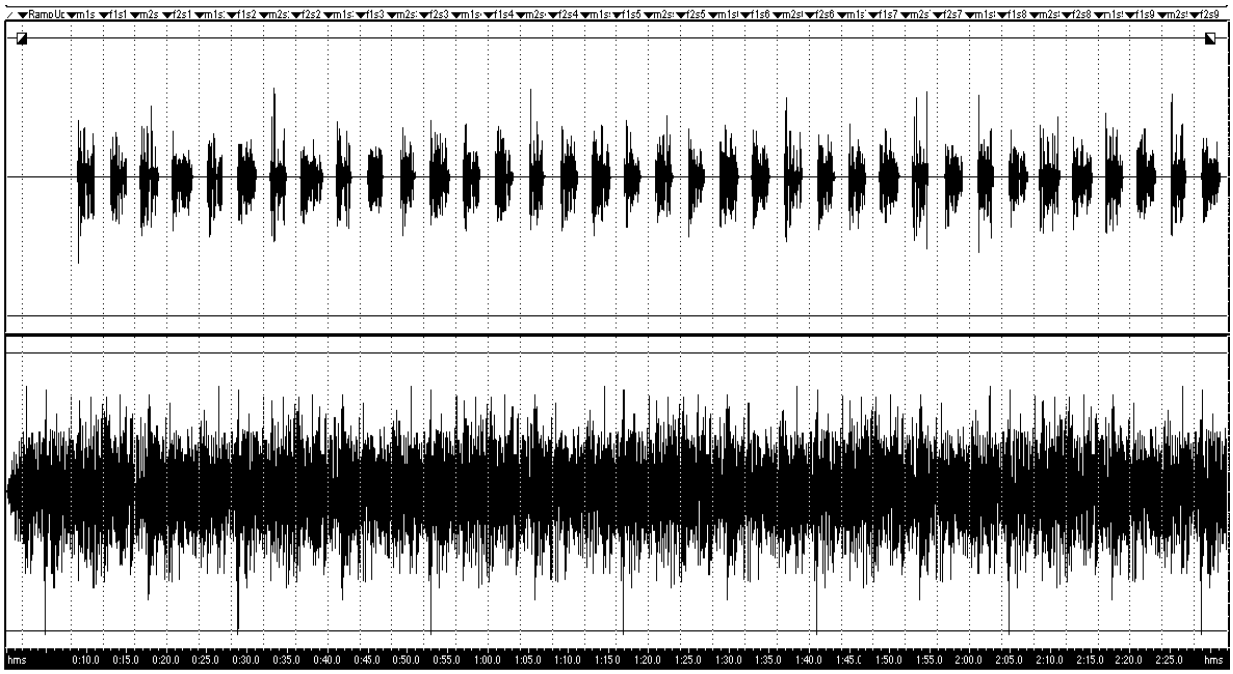


Figure D.2: Noise and speech playback sequence, including convergence period

D.3.6 Recordings at the network simulator electrical reference point

The network simulator needs to be configured for a call setup with EVS-SWB speech codec bitrate of 24,4 kbps in super-wideband mode [i.4]. In order to obtain a lower quality range as well, a bitrate of 13,2 kbps can be utilized in addition. The send signal is recorded at the electrical reference point of a network simulator to generate the test conditions (noise suppressed speech) for the subjective test.

D.3.7 Recordings at the MRP and terminal's primary microphone location

In addition to the recordings at the electrical reference point of a network simulator, the acoustic signals at MRP and primary microphone position can be recorded for further reference and analysis.

Annex E (normative): Speech material to be used for objective testing

The following speech samples provided in tables E.1 to E.3 are used in conjunction with the model. Preferred test sequence for the application of the models is American English.

The first 4 sentences in each sequence are used during the adaptation period of the noise canceller under test, the remaining 16 samples are used for calculating the objective scores.

The speech samples can be downloaded here:
https://docbox.etsi.org/stq/Open/TS%20103%20281%20Wave%20files/Annex_E%20speech%20data.

Table E.1: American English test sequence

Seq	Sample	Harvard Sentences	
1	m1s8	We tried to replace the coin but failed.	Preliminary (convergence)
2	f1s8	A rod is used to catch pink salmon.	
3	m2s8	Corn cobs can be used to kindle a fire.	
4	f2s8	The crooked maze failed to fool the mouse.	
5	m1s1	The empty flask stood on the tin tray.	
6	f1s1	It is easy to tell the depth of a well.	
7	m2s1	Acid burns holes in wool cloth.	
8	f2s1	Note closely the size of the gas tank.	
9	m1s2	He broke a new shoelace that day.	
10	f1s2	The box was thrown beside the parked truck.	
11	m2s2	Eight miles of woodland burned to waste.	
12	f2s2	Mend the coat before you go out.	
13	m1s3	The urge to write short stories is rare.	
14	f1s3	Four hours of steady work faced us.	
15	m2s3	A young child should not suffer fright.	
16	f2s3	The stray cat gave birth to kittens.	
17	m1s4	The pirates seized the crew of the lost ship.	
18	f1s4	The boy was there when the sun rose.	
19	m2s4	The fruit of a fig tree is apple shaped.	
20	f2s4	The frosty air passed through the coat.	

Table E.2: German test sequence

Seq	Sample	German Sentences	
1	m3s2	Ich hole den Mantel lieber gleich.	Preliminary (convergence)
2	f3s2	Die Firma setzt Maßstäbe.	
3	m4s2	Die Katze schleicht langsam heran.	
4	f4s2	Faulheit ist auch erholsam.	
5	f1s1	Arbeit im Garten ist besinnlich.	
6	f1s2	Blumen muss man häufig gießen.	
7	f2s1	Kühl und klar ist die Luft.	
8	f2s2	Wir müssen das Licht anschalten.	
9	f3s1	Hier richten Zimmerleute ein Dach.	
10	f3s2	Die Firma setzt Maßstäbe.	
11	f4s1	Bitte verlier doch keine Zeit!	
12	f4s2	Faulheit ist auch erholsam.	
13	m1s1	Der Hammer trifft den Nagel	
14	m1s2	Strohhalme brechen leicht.	
15	m2s1	Im Hof wartet man schon auf uns.	
16	m2s2	Der Spatz frisst am liebsten Körner.	
17	m3s1	Man zahl Eintritt an der Kasse.	
18	m3s2	Ich hole den Mantel lieber gleich.	
19	m4s1	Er erklärt die Dinge schlecht.	
20	m4s2	Die Katze schleicht langsam heran.	

Table E.3: Mandarin test sequence

Sample	Gender	Mandarin	Pin-yin	Translation
1	male1	北京近来很寒冷		
2	male1	短裙长度正合适		
3	female1	外孙出生在农村		
4	female1	星期二别打篮球		
5	male1	北京近来很寒冷	Běijīng jìnlái hěn hánlěng	Beijing is very cold recently
6	male1	短裙长度正合适	Duǎn qún chángdù zhèng héshì	The length of the skirt is just fine
7	female1	外孙出生在农村	Wàisūn chūshēng zài nóngcūn	The grandson was born in the countryside
8	female1	星期二别打篮球	Xīngqí'èr bié dǎ lánqiú	Don't play basketball on Tuesday
9	male2	我确实没接到信	Wǒ quèshí méi jiē dào xìn	I did not receive the email
10	male2	顾客有较多意见	Gùkè yǒu jiào duō yìjiàn	The customers have quite many complaints
11	female2	讨论很快结束了	Tǎolùn hěn kuài jiéshùle	The discussion ended quickly
12	female2	成绩面前不骄傲	Chéngjī miànqián bù jiāo'ào	Don't be arrogant with good achievement
13	male3	明早乘船去上海	Míngzǎo chéng chuán qù shànghǎi	Travel to Shanghai by ship tomorrow morning
14	male3	旁边桂花树真多	Pángbiān guìhuā shù zhēn duō	There are so many Osmanthus trees close-by
15	female3	那块绿窗帘很贵	Nà kuài lǜ chuānglián hěn guì	That green curtain is very expensive
16	female3	严禁随处丢垃圾	Yánjìn suíchù diū lèsè	Do not litter
17	male4	彩虹有七种颜色	Cǎihóng yǒu qī zhǒng yánsè	There are seven colors in a rainbow
18	male4	躺着看书会近视	Tǎngzhe kànshū huì jìnshì	You will get near-sighted if reading lying down
19	female4	平时要注意卫生	Píngshí yào zhùyì wèishēng	Pay attention to hygiene everyday
20	female4	谦虚会使人进步	Qiānxū huì shǐ rén jìnbù	Being modest makes a person improve

History

Document history		
V1.1.1	April 2017	Publication
V1.2.1	January 2018	Publication
V1.3.1	May 2019	Publication