ETSI TS 103 106 V1.4.1 (2016-11)

Speech and multimedia Transmission Quality (STQ);
Speech quality performance in
the presence of background noise:
Background noise transmission for
mobile terminals-objective test methods

Reference
RTS/STQ-224

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 5
Foreword 5
Modal verbs terminology 5
1 Scope 6
2 References 6
2.1 Normative references 6
2.2 Informative references 7
3 Abbreviations 8
4 Introduction 9
5 Underlying speech databases and preparations 9
6 Modifications to the model described in ETSI EG 202 396-3 10
6.1 Prefiltering in Narrowband Mode (NB) 10
6.2 Detection of the speech parts 10
6.3 Speech level adjustment in wideband 11
6.4 Replacement of parameter regression for S-MOS 11
6.5 Retraining of parameter regression for N-MOS and G-MOS 14
7 Comparison of objective and subjective results after the training process 15
7.0 General 15
7.1 Results in wideband mode 15
7.1.0 General 15
7.1.1 Results for database "Audience - Test 3" 15
7.1.2 Results for database "Audience - Test 3L" (excluded during retraining) 16
7.1.3 Results for database "Audience - Test 4" 16
7.1.4 Results for database "Audience - Test 4L" 17
7.1.5 Results for database "Nokia - Test 1" 18
7.1.6 Results for database "Nokia - Test 2" (excluded during retraining) 18
7.1.7 Results for database "Orange" 19
7.1.8 Results for database "Qualcomm - Test 3" 20
7.1.9 Results for database "Qualcomm - Test 4" 20
7.2 Results in narrowband mode 21
7.2.0 General 21
7.2.1 Results for database "Audience - Test 1" 21
7.2.2 Results for database "Audience - Test 1L" 22
7.2.3 Results for database "Audience - Test 2" 22
7.2.4 Results for database "Audience - Test 2L" 23
7.2.5 Results for database "Qualcomm- Test 1" 24
7.2.6 Results for database "Qualcomm- Test 2" 24
8 Validation results 25
8.0 Preamble 25
8.1 Audience validation data 25
8.1.1 Description of tests 25
8.1.2 Description of validation results 27
8.1.2.1 General explanation 27
8.1.2.1 Experiment 5: Narrowband 27
8.1.2.2 Experiment 6: Narrowband 29
8.1.2.3 Experiment 7: Wideband 31
8.1.2.4 Experiment 8: Wideband 33
8.2 Orange validation data 35
8.2.1 Description of tests 35
8.2.2 Description of validation results 36
8.3 Qualcomm validation data 37
8.3.1 Description of tests 37

8.3.2 Description of validation results 40
8.4 Validation data for additional use cases 44
8.4.1 Tests 1 & 2: Description 44
8.4.2 Tests 1 & 2: Results 45
8.4.3 Tests 3, 4, 5 & 6: Description 54
8.4.4 Tests 3 & 4: Results, Narrowband 55
8.4.5 Tests 5 & 6: Results, Wideband 63
9 Application of the retrained model 70
Annex A (normative): Summary of Retraining Databases 71
Annex B (normative): Test vectors for model verification 72
B.0 Test vectors 72
B.1 Audience test vectors 72
Annex C (normative): Speech material to be used for objective testing 75
Annex D (informative): Subjective testing framework used for the present document 76
D.1 Introduction 76
D.2 Subjective test plan 76
D.2.1 Traceability 76
D.2.2 Speech database requirements 76
D.2.3 Reference Conditions 76
D.2.4 Test Conditions 76
D.2.5 Pre-processing of reference conditions 78
D.2.6 Post-processing of test conditions 78
D.2.7 Calibration and equalization of headphones for presentation 78
D.2.8 Requirements on the listening laboratory 78
D.2.9 Experimental design 78
D.2.10 Training session 79
D.3 Set-up for acquisition of test conditions 79
D.3.1 Terminal positioning and HATS calibration 79
D.3.2 Background Noise reproduction 79
D.3.3 Noise and speech playback synchronization 80
D.3.4 Convergence sequence 80
D.3.5 Example of noise and speech playback sequence including convergence period 80
D.3.6 Recordings at the network simulator electrical reference point 81
D.3.7 Recordings at the MRP and terminal's primary microphone location 81
D.4 Processing test plan block diagram 82
History 84

Intellectual Property Rights

IPRs essential or potentially essential to the present document 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.

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 the ETSI standard series EG/S 202 396 [i.2] to [i.4]:

ETSI ES 202 396 1: "Background noise simulation technique and background noise database";

ETSI EG 202 396 2: "Background noise transmission - Network simulation - Subjective test database and results";

ETSI EG 202 396-3: "Background noise transmission - Objective test methods".

The present document is based on the objective test method described in ETSI EG 202 396-3 [i.4] and contains modifications of the model required in order to provide a good prediction of the uplink speech quality in the presence of background noise of modern mobile terminals.

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 narrowband and wideband 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 both, terminals and networks. The present document provides information about the testing methods applicable to objectively evaluate the speech quality of mobile terminals with AMR and AMR-WB codecs 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 description of the adaptation of the test method described in ETSI ES 202 396-1 [i.2].
- The model results in comparison with the underlying subjective tests used for the retraining of the objective model.
- The model validation results:
  - Additional validation results are provided for cases which include some conditions outside the scope of ETSI ES 202 396-1 [i.2]. These include music as background noise, and user holding a handset in other than nominal position, as defined in Recommendation ITU-T P.64 [i.24]. In addition, validation results are provided for Chinese language.

The present document is to be used in conjunction with:

- ETSI ES 202 396-1 [i.2] 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 EG 202 396-2 [i.3] which describes the simulation of network impairments and how to simulate realistic transmission network scenarios and which contains the methodology and results of the subjective scoring for the data forming the basis of the present document.
- ETSI EG 202 396-3 [i.4] which describes the basic objective model underlying to the Model described in the present document.
- 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] 3GPP S4-120542: "Common subjective testing framework for training of P.835 test predictors".

[i.2] 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.3] ETSI EG 202 396-2: "Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality performance in the presence of background noise; Part 2: Background Noise Transmission - Network Simulation - Subjective Test Database and Results".

[i.4] 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.5] ETSI TS 126 073: "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; ANSI C code for the Adaptive Multi Rate (AMR) speech codec (3GPP TS 26.073)".

[i.6] Recommendation ITU-T P.835: "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm".

[i.7] Recommendation ITU-T G.722.2: "Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)".

[i.8] Recommendation ITU-T P.56: "Objective measurement of active speech level".

[i.9] Recommendation ITU-T P.1401: "Methods, metrics and procedures for statistical evaluation, qualifying and comparison of objective quality prediction models".

[i.10] 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.11] Recommendation ITU-T G.191: "Software tools for speech and audio coding standardization".

[i.12] Hastie, T.; Tibshirani, R.; Friedman, J.: "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", New York: Springer-Verlag, 2001.

[i.13] Recommendation ITU-T P.501: "Test Signals for Use in Telephonometry".

[i.14] Recommendation ITU-T P.58: "Head and Torso simulator for telephonometry".

[i.15] Recommendation ITU-T P.57: "Artificial ears".

[i.16] ETSI TS 126 131: "Universal Mobile Telecommunications System (UMTS); LTE; Terminal acoustic characteristics for telephony; Requirements (3GPP TS 26.131 version 10.2.0 Release 10)".

[i.17] Recommendation ITU-T P.800: "Methods for subjective determination of transmission quality".

[i.18] ETSI TS 126 132: "Universal Mobile Telecommunications System (UMTS); LTE; Speech and video telephony terminal acoustic test specification (3GPP TS 26.132)".

[i.19] Void.

[i.20] Recommendation ITU-T TD 477 (GEN/12): "Handbook of subjective test practical procedures" (temporary document) - Geneva, 18-27 January 2011.

[i.21] 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.22] 3GPP, Tdoc S4(12)0621, Ext-ATS Permanent document (EATS-3): "Common subjective testing framework for validation of P.835 test predictors".

[i.23] Recommendation ITU-T P.50: "Artificial voices".

[i.24] Recommendation ITU-T P.64: "Determination of sensitivity/frequency characteristics of local telephone systems".

3 Abbreviations

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

AMR	Adaptive MultiRate
AMR-NB	Adaptive Multirate Codec - Narrow Band
AMR-WB	Adaptive Multi-Rate Wideband Speech Codec
BAK	Background Noise Component
dB SPL	Sound Pressure Level re 20 µPa in dB
DRP	Drum Reference Point
DTX	Discontinuous Transmission
G-MOS	Global MOS

NOTE: MOS related to the overall sample.

HATS	Head and Torso Simulator
HHHF	Hand-Held Hands-Free
IRS	Intermediate Reference System
NTT	Nippon Telegraph and Telephone
ITU	International Telecommunication Union
ITU-T	Telecommunication Standardization Sector of ITU
MOS	Mean Opinion Score
MRP	Mouth Reference Point
MSIN	Mobile Station Input Filter
NB	NarrowBand
N-MOS	Noise MOS

NOTE: MOS related to the noise transmission only.

NS	Noise Suppression
OVRL	Overall (speech + noise) Component
RCV	ReCeiVe
RMSE	Root Mean Square Error
RMSE*	epsilon insensitive Root Mean Square Error
SIG	SIGnal component
S-MOS	Speech MOS

NOTE: MOS related to the speech signal only.

SND	Sending Direction
SNR	Signal to Noise Ratio
SPL	Sound Pressure Level
WB	WideBand
WCDMA	Wideband Code Division Multiple Access

4 Introduction

The present document describes the modifications of the ETSI EG 202 396-3 [i.4] model which were necessary to adapt to the training databases provided by the 3GPP contributors listed in annex A. The core model itself retains mainly unmodified except the points given in the clauses below. Modifications affect the narrow- and wideband mode in different ways.

The adapted objective method described in the present document is intended to be used for all types of modern mobile terminals using different bitrates of AMR [i.5] and AMR-WB [i.7] coding.

5 Underlying speech databases and preparations

The base for each mode of the objective model (wideband/narrowband) as described in ETSI EG 202 396-3 [i.4] are listening tests conducted according to Recommendation ITU-T P.835 [i.6]. From the beginning of the development, these listening test databases were designed to be a training set for predicting Recommendation ITU-T P.835 [i.6] scores. They included a huge amount of conditions (> 170) and a wide range of speech and noise quality. Besides real terminals also terminal simulations and transmission impairments were included. However, the data and processing included were based on technologies actual at the time when the standard and its updates were created.

The underlying databases for the retraining as described in the present document were created using real state-of-the-art mobile devices and thus the quality ranges yielded may not be normally distributed over all MOS scales. The context between the databases can also differ (e.g. pure handset recordings vs. mixed handset/hands-free databases). Furthermore new reference conditions extensively discussed in different standards groups and described in [i.1] were included in the tests.

Table 1: Set of reference conditions

File	SIG.	SNR	Noise Type
i01	Source (filtered)	No Noise	-
i02	Source (filtered)	0 dB	Fullsize_Car1_130Kmh_binaural
i03	Source (filtered)	12 dB	Fullsize_Car1_130Kmh_binaural
i04	Source (filtered)	24 dB	Fullsize_Car1_130Kmh_binaural
i05	Source (filtered)	36 dB	Fullsize_Car1_130Kmh_binaural
i06	NS Level 1	No Noise	-
i07	NS Level 2	No Noise	-
i08	NS Level 3	No Noise	-
i09	NS Level 4	No Noise	-
i10	NS Level 3	24 dB	Fullsize_Car1_130Kmh_binaural
i11	NS Level 2	12 dB	Fullsize_Car1_130Kmh_binaural
i12	NS Level 1	[0 dB]	Fullsize_Car1_130Kmh_binaural

Each training database was provided together with 12 reference conditions, mainly created according to the annex of [i.1], table 1 shows one possible arrangement. Although it was observed that not all reference sets included exactly the same speech material, used background noise, SNR ranges and speech distortion configuration, this data indicates which range of speech and noise degradations can be expected in the databases.

For transforming the different databases (to achieve at least approximately on a common base for the retraining of the model), thus the 12 x 3 values of the reference conditions (averaged over all samples) were used to linearly transform the subjective MOS data. In a first step, the reference conditions of all databases included in the retraining process were weighted together to an average reference condition set. The weight per database depends on the number of samples it provides for the training.

For each database, a mapping between the reference conditions and the average reference condition set is calculated. To catch also inter-relations between speech, noise and global ratings, a matrix transformation instead a per-scale regression was chosen. To compensate biases, a constant column was added to the reference set. Then a transformation Tj is calculated for each database j with reference set Rj which minimizes the distance to the average reference set A:

$$\underbrace{\begin{pmatrix} 1 & S_{i01} & N_{i01} & G_{i12} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & S_{i12} & N_{i12} & G_{i12} \end{pmatrix}}_{R_j(Ref.\ set\ j)} \times T_j = \underbrace{\begin{pmatrix} \overline{S_{i01}} & \overline{N_{i01}} & \overline{G_{i12}} \\ \vdots & \vdots & \vdots \\ \overline{S_{i12}} & \overline{N_{i12}} & \overline{G_{i12}} \end{pmatrix}}_{A\ (Avg.\ ref.\ set)} \tag{1}$$

The transformation matrix Tj (size 4 x 3) can easily be determined to:

$$T_j = \left({R_j}^T \times R_j\right)^{-1} \times {R_j}^T \times A \tag{2}$$

If the three scales (S-MOS/N-MOS/G-MOS) are independent from each other for any database, the matrix transformation T_j equals a linear per-scale transformation. Before the retraining of the model, the transformation is applied to the whole test data on a per-sample base:

$$\underbrace{\begin{pmatrix} 1 & S_1 & N_1 & G_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & S_N & N_N & G_N \end{pmatrix}}_{S_j\ (scores\ of\ samples\ of\ database\ j)} \times T_j = \underbrace{\begin{pmatrix} \tilde{S}_1 & \tilde{N}_1 & \tilde{G}_1 \\ \vdots & \vdots & \vdots \\ \tilde{S}_N & \tilde{N}_N & \tilde{G}_N \end{pmatrix}}_{\tilde{S}_j\ (transformed\ scores\ of\ samples\ of\ database\ j)} \tag{3}$$

6 Modifications to the model described in ETSI EG 202 396-3

6.1 Prefiltering in Narrowband Mode (NB)

In the narrowband mode described in ETSI EG 202 396-3 [i.4], the listening test audio files included a far-end handset simulation, realized with an IRS RCV filter. In the requirements described in ETSI EG 202 396-3 [i.4], neither for narrow- nor for wideband such a listening filter was described or used in the databases.

The narrowband mode internally filters the unprocessed and clean reference with IRS SND and IRS RCV to simulate a transmission over high-quality listening devices and network. The principle of IRS seems to be outdated, modern state-of-the-art mobiles do not have this frequency characteristic. Even more when using these newly created NB databases, where the used devices have almost flat frequency responses in sending direction.

Thus the filtering with IRS SND and RCV of the two reference signals was replaced by filtering with the MSIN [i.11] filter, which is mainly a band pass. Also no listening filter was applied to the processed signals.

6.2 Detection of the speech parts

The detection of signal parts belonging to either speech or noise was updated. Now the clean speech signal is segmented into frames and classified according to Recommendation ITU-T G.160 [i.10]. The signal parts classified as silence are assumed as background noise sections, all other frames are assumed as speech.

6.3 Speech level adjustment in wideband

The current ETSI EG 202 396-3 [i.4] implementation assumes 79 dB SPL/-15 dB Pa active speech level due to the underlying listening test based on the underlying subjective databases in the wideband model of ETSI EG 202 396-3 [i.4].

For the objective model as described in the present document the level adjustment of the recordings of the training databases was applied in such a way, that the active speech level over the full sequence test should be about 73 dB SPL/-21 dB Pa (for the listening test) as described in ETSI EG 202 396-3 [i.4].

6.4 Replacement of parameter regression for S-MOS

The model described in ETSI EG 202 396-3 [i.4] calculates several parameters out of the psycho-acoustically motivated inner representation for the estimation of S- and N-MOS. The parameters are shown in tables 2 and 3. A detailed description of the calculation for the parameters can be found in [i.4].

Table 2: Extracted parameters for N-MOS

P_1	$N_{BGN,\,P}$	P_4	$\mu(RA_{BGN,\,U})$
P_2	$\mu(RA_{BGN,\,P})$	P_5	$\sigma^2(RA_{BGN,\,U})$
P_3	$\sigma^2(RA_{BGN,\,P})$	P_6	$\sigma^2(\Delta RA_{BGN,\,P\text{-}U})$

Table 3: Extracted Parameters for S-MOS

P_1	ΔSNR	P_4	$\mu(\Delta RA_{Sp,\,P\text{-}C})$
P_2	$\mu(RA_{Sp,\,P})$	P_5	$\sigma^2(\Delta RA_{Sp,\,P\text{-}C})$
P_3	$\mu(\Delta RA_{Sp,\,P\text{-}U})$	P_6	$\sigma^2(\Delta RA_{Sp,\,P\text{-}U})$

The calculation of the objective S-MOS in clause 6.5.2 of [i.4] is performed with a linear quadratic regression of the parameters mentioned above. In addition, the regression coefficients are switched with regard to the N-MOS calculated before which models the expectation to speech [i.4] quality of the listener.

The applied modification is the replacement of the linear quadratic regression with a feed forward neural network. In consequence, the switching of the regression coefficients depending on the N-MOS is removed. Only one network is trained with input (6 parameters of table 3) and output (S-MOS) data by a simple back-propagation algorithm [i.12].

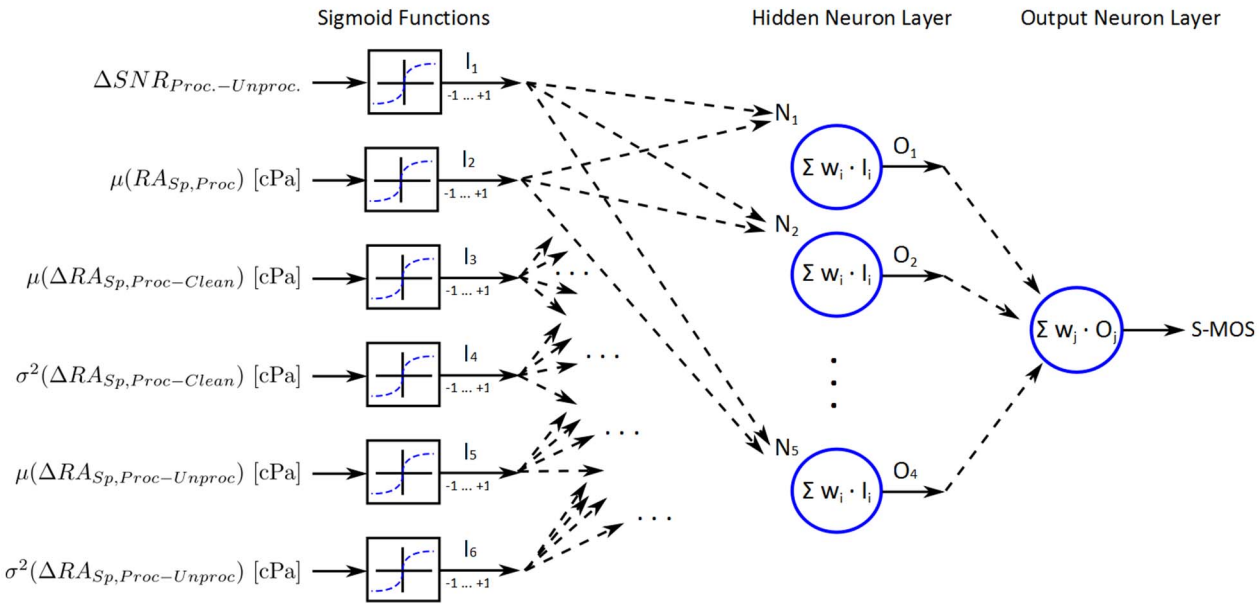


Figure 1: Structure of neural network for S-MOS

The setup of the neural network is shown in figure 1. It consists of 5 units in one hidden layer; each unit N_j includes a connection from each transformed input parameter I_i. The output O_j of each unit is calculated as the weighted sum of each input I_i using the weights w_{ij}. The outputs O_j are then weighted by w_j and summed up to the output S-MOS. Both, w_{ij} and w_j are the result of the training of the network.

The parameters according to table 3 are composed to a vector **P** including a bias as the first element:

$$\mathbf{P} = (1 \quad P_1 \quad P_2 \quad P_3 \quad P_4 \quad P_5 \quad P_6) \tag{4}$$

The output calculation of the neural network shown in figure 1 can be described as concatenated matrix operations:

$$SMOS_{objective,raw} = f_{sigmoid}\left(f_{sigmoid}\left(\frac{\mathbf{P} - \mathbf{M_{in}}}{\mathbf{S_{in}}}\right) \times \mathbf{H}\right) \times \mathbf{O} \tag{5}$$

First the parameter vector **P** is normalized to mean 0,0 and standard deviation 1,0. This is done by subtracting the average of all training data for each parameter from each item of the input parameter vector. The averages for each parameter P_i can be described as a vector, which is different for narrow- and wideband mode:

$$\mathbf{M_{in,WB}} = (0{,}0 \quad 12{,}7309 \quad 4{,}2076 \quad -1{,}2456 \quad 0{,}8834 \quad 12{,}2522 \quad 7{,}0541)$$

$$\mathbf{M_{in,NB}} = (0{,}0 \quad 13{,}7519 \quad 2{,}0884 \quad -0{,}3124 \quad 0{,}2511 \quad 6{,}7091 \quad 5{,}2951) \tag{6}$$

NOTE 1: The first element is set to zero to be compatible with the bias element in **P**.

A similar approach can be made for the standard deviation for each parameter P_i, also separated for wide- and narrowband:

$$\mathbf{S_{in,WB}} = (1{,}0 \quad 11{,}8503 \quad 1{,}2824 \quad 1{,}1981 \quad 0{,}9572 \quad 6{,}7848 \quad 4{,}8380)$$

$$\mathbf{S_{in,NB}} = (1{,}0 \quad 11{,}4341 \quad 0{,}4047 \quad 0{,}3877 \quad 0{,}3309 \quad 3{,}1189 \quad 2{,}5976) \tag{7}$$

NOTE 2: The first element is set to one to be compatible with the bias element in **P**.

After normalizing the input data, the sigmoid function $f_{sigmoid}(x)$ is applied to the each normalized parameter P_i. This ensures that each input of each neuron of the hidden layer is soft-limited to the range ±1,0 and guarantees that parameters out of the training range cannot produce an overflow which results in eventually unreasonable scores. For the current model, the hyperbolic tangent was chosen to a sigmoid function:

$$f_{sigmoid}(x) = \tanh(x) \tag{8}$$

Thus the input of the hidden neuron layers can also be given as a transformed parameter vector $\widetilde{\mathbf{P}}$:

$$\widetilde{\mathbf{P}} = f_{sigmoid}\left(\frac{\mathbf{P} - \mathbf{M_{in}}}{\mathbf{S_{in}}}\right) = (1 \quad \widetilde{P_1} \quad \widetilde{P_2} \quad \widetilde{P_3} \quad \widetilde{P_4} \quad \widetilde{P_5} \quad \widetilde{P_6}) \tag{9}$$

NOTE 3: The sigmoid function is not applied to the bias component.

The output of the hidden layer is calculated with a matrix multiplication of $\widetilde{\mathbf{P}}$ and **H**. **H** describes all weights from each input parameter to each neuron in the hidden layer. These weights are the results of the training with the back-propagation algorithm. In consequence, **H** is different for each bandwidth mode:

$$\mathbf{H_{WB}} = \begin{pmatrix} -0{,}4336 & -0{,}9873 & 0{,}0091 & -0{,}0845 & 0{,}0203 \\ 0{,}1141 & -0{,}0004 & -0{,}7133 & -0{,}2798 & -1{,}8189 \\ 1{,}0265 & 0{,}5001 & 0{,}5120 & 0{,}0537 & 0{,}1265 \\ -0{,}8627 & -1{,}7518 & -0{,}0374 & -0{,}2908 & 0{,}3064 \\ 2{,}1381 & 0{,}4190 & 1{,}0715 & -1{,}6716 & 0{,}4973 \\ -1{,}3933 & 0{,}5972 & 0{,}0852 & 0{,}1977 & 0{,}2222 \\ -0{,}3793 & -1{,}7785 & -0{,}5306 & -1{,}7538 & -2{,}9630 \end{pmatrix} \tag{10}$$

$$\mathbf{H_{NB}} = \begin{pmatrix} -0{,}3608 & -0{,}3805 & 0{,}5359 & -1{,}1131 & -0{,}1322 \\ 0{,}7348 & -4{,}4639 & -1{,}2552 & 0{,}3338 & 0{,}5452 \\ 0{,}9117 & 2{,}7177 & 0{,}8876 & 0{,}1712 & -2{,}1279 \\ -0{,}2383 & 1{,}7228 & -0{,}0354 & -1{,}0284 & 1{,}0483 \\ 1{,}4511 & 2{,}1467 & 1{,}0010 & 0{,}7356 & 0{,}1154 \\ -0{,}5573 & -0{,}6137 & -0{,}2648 & 1{,}6202 & 0{,}5966 \\ -3{,}2194 & -7{,}9575 & -0{,}7736 & -0{,}8676 & 0{,}1663 \end{pmatrix} \tag{11}$$

The outputs of the hidden layer are then again soft-limited with the same sigmoid function to assure a valid range (±1,0) for the output neuron layer. The five transformed output values of the hidden layer are then given to the output layer. Here the output of the neural network is calculated with another matrix multiplication with the matrix **O**, which weights the outputs of the hidden layers to an output score $S\text{-}MOS_{objective,\,raw}$. This output layer matrix **O** is also given for wide and narrowband mode independently:

$$\mathbf{O_{WB}} = (0{,}1777 \quad -0{,}2835 \quad -0{,}3147 \quad 0{,}1837 \quad -0{,}3237)$$

$$\mathbf{O_{NB}} = (0{,}3832 \quad -0{,}5250 \quad -0{,}1878 \quad -0{,}2674 \quad -0{,}1548) \tag{12}$$

Another part of the back-propagation algorithm is also to normalize the output data to mean 0,0 and standard deviation 1,0. To revise this step and transform the output of the neural network back to the MOS scale, the objective S-MOS is calculated from the raw score:

$$SMOS_{objective} = \max(1{,}0, \min(\mathbf{S_{out}} \cdot (SMOS_{objective\ ,raw} + \mathbf{M_{out}}), 5{,}0) \tag{13}$$

The objective S-MOS is calculated with M_{out} = (3,0), S_{out} = (2,0) and a hard limiter [1,0; 5,0].

6.5 Retraining of parameter regression for N-MOS and G-MOS

The objective N-MOS is the result of a linear, quadratic regression algorithm applied to the six parameters of table 2 according to equation (14):

$$NMOS = c_0 + \sum_{j=1}^{2}\sum_{i=1}^{6} c_{ji} \cdot P_i^{\,j} \quad (1)$$

(14)

The overall or global quality G-MOS is calculated by using the previously calculated N-MOS and S-MOS as input parameters for a linear quadratic regression according to equation (15):

$$GMOS = c_0 + \sum_{j=1}^{2} c_{Sj} \cdot SMOS^{\,j} + \sum_{j=1}^{2} c_{Nj} \cdot NMOS^{\,j} \quad (1)$$

(15)

The calculation steps for N-MOS and G-MOS are not modified, only the coefficients for the linear regressions according to equations (14) and (15) are adapted to the new training material. The new coefficients are given in tables 4 to 7:

Table 4: N-MOS coefficients for narrowband; Parameters P_i according to table 2

	Bias	P_1	P_2	P_3	P_4	P_5	P_6
Order j = 1	2,2231	-0,0395	-0,0359	0,2825	0,0023	-0,3959	-2,6965
Order j = 2	-	-	0,0021	-0,0239	-0,0003	0,0542	0,8684

Table 5: N-MOS coefficients for wideband; Parameters P_i according to table 2

	Bias	P_1	P_2	P_3	P_4	P_5	P_6
Order j = 1	1,4279	-0,0484	0,0994	0,2189	-0,0732	-0,3346	-1,3108
Order j = 2	-	-	-0,0018	-0,0079	0,0011	0,0891	0,2566

Table 6: G-MOS coefficients for narrowband

	Bias	S-MOS	N-MOS
Order j = 1	-0,4879	0,2647	0,8274
Order j = 2	-	0,0696	-0,0737

Table 7: G-MOS coefficients for wideband

	Bias	S-MOS	N-MOS
Order j = 1	-0,2141	0,2735	0,4542
Order j = 2	-	0,0708	-0,0065

7 Comparison of objective and subjective results after the training process

7.0 General

The comparison between the results of the subjective tests and the objective prediction of the conditions used in the training process are given in this clause. The metrics used in the statistical evaluation process are derived from Recommendation ITU-T P.1401 [i.9]. Besides the RMSE or RMSE* values, the different metrics and scatterplots are given in this clause.

A summary of the databases and the conditions used for retraining is given in annex A.

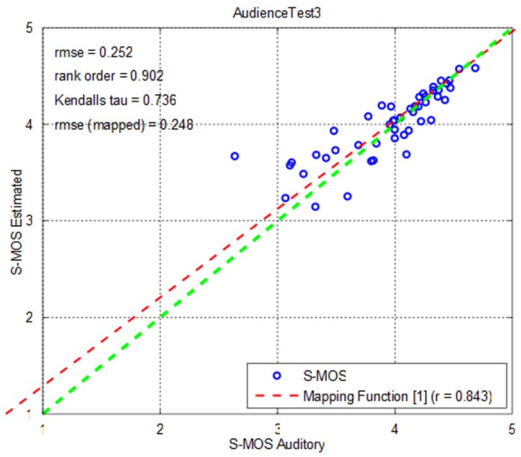
7.1 Results in wideband mode

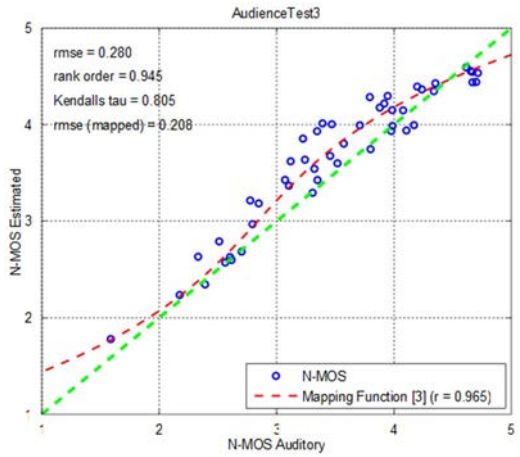
7.1.0 General

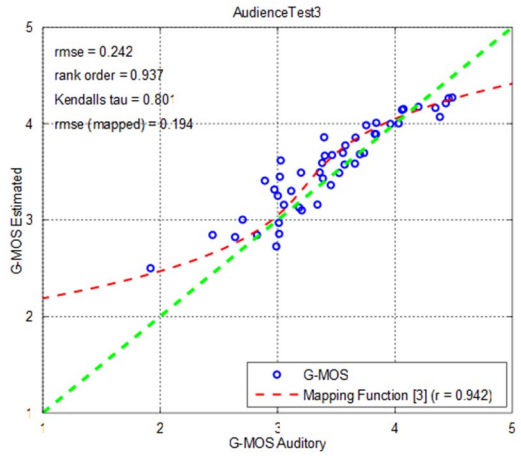
For the wideband retraining procedure two databases were not included within the training for several reasons. Removal of these databases significantly increases the performance. Further analysis is required why these databases seem to be "incompatible" with the remaining training set.

Overall, 7 databases with 387 conditions and 5 544 samples were used.

7.1.1 Results for database "Audience - Test 3"



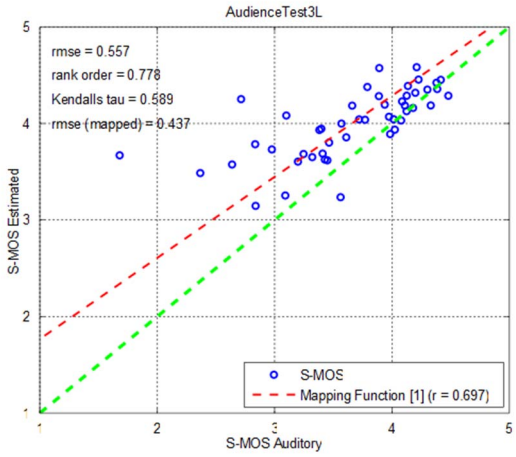


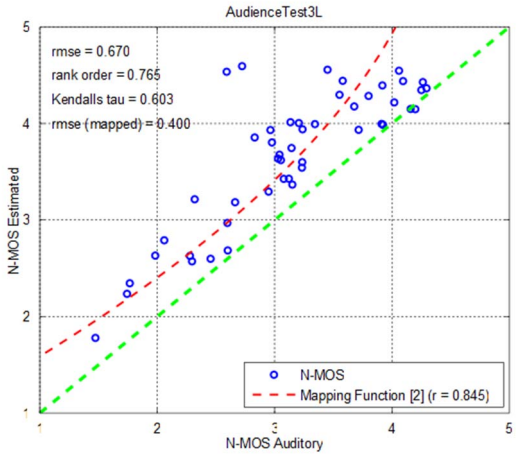


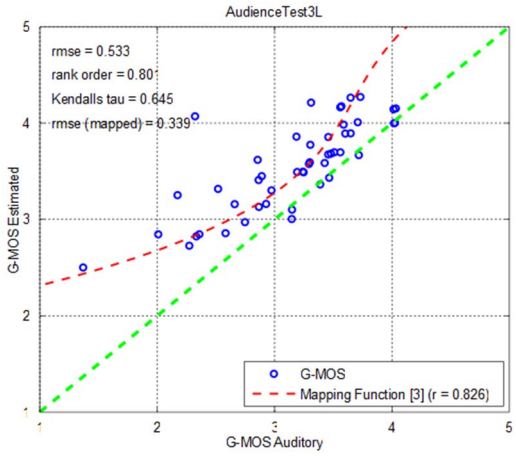
RMSE:	no Mapping	0,25	0,28	0,24
	1st Ord. Mapping	0,25	0,23	0,22
	3rd Ord. Mapping	**0,24**	**0,21**	**0,19**

RMSE*:	no Mapping	0,17	0,18	0,15
	1st Ord. Mapping	0,16	0,13	0,12
	3rd Ord. Mapping	**0,15**	**0,11**	**0,10**

7.1.2 Results for database "Audience - Test 3L" (excluded during retraining)



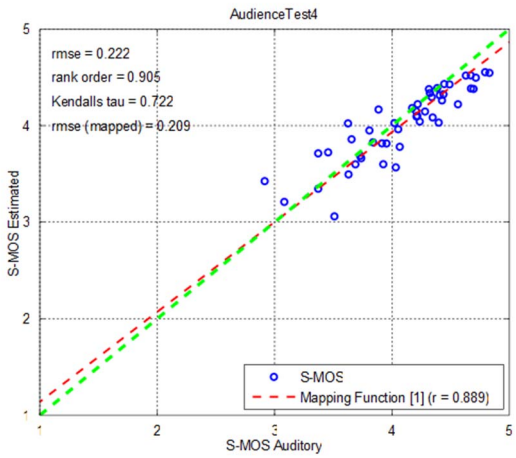


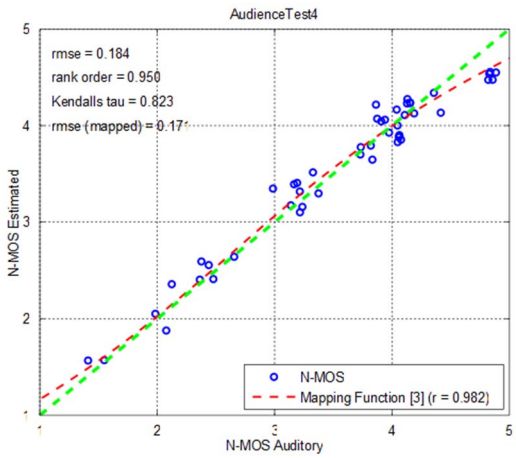


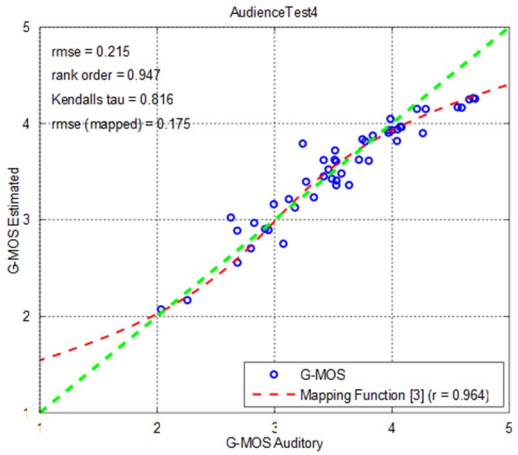
		SIG	BAK	OVRL
RMSE:	no Mapping	0,56	0,67	0,53
	1st Ord. Mapping	0,44	0,40	0,36
	3rd Ord. Mapping	**0,42**	**0,39**	**0,34**

RMSE*:	no Mapping	0,45	0,56	0,43
	1st Ord. Mapping	0,34	0,30	0,26
	3rd Ord. Mapping	**0,31**	**0,28**	**0,25**

7.1.3 Results for database "Audience - Test 4"



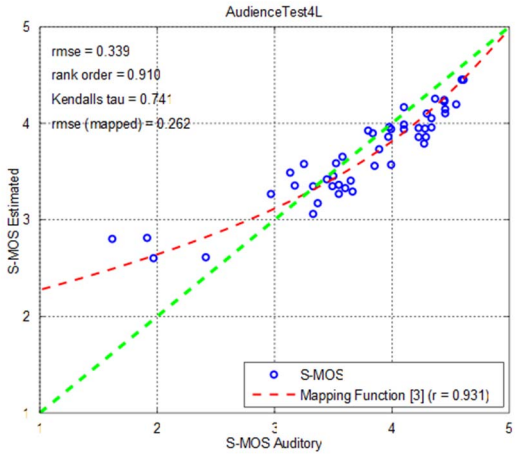


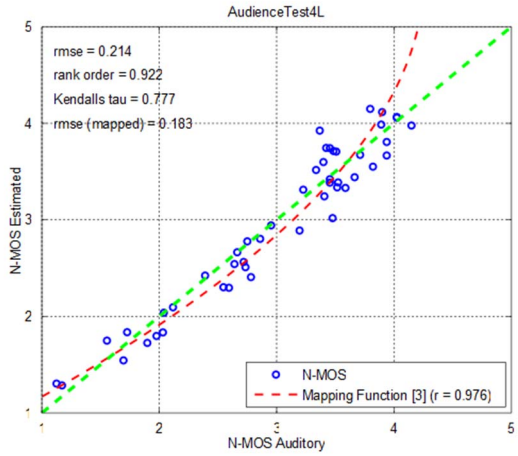


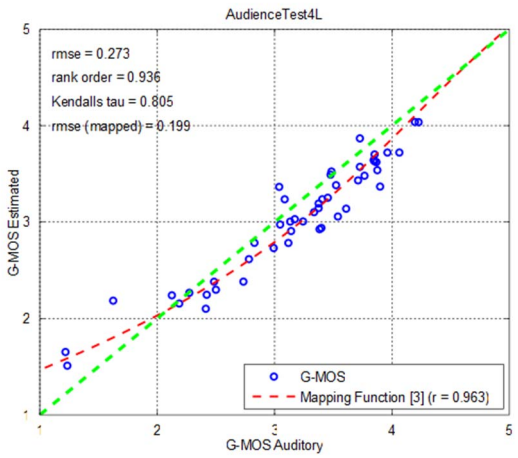
		SIG	BAK	OVRL
RMSE:	no Mapping	0,22	0,18	0,21
	1st Ord. Mapping	0,21	0,18	0,20
	3rd Ord. Mapping	**0,21**	**0,17**	**0,18**

RMSE*:	no Mapping	0,14	0,11	0,14
	1st Ord. Mapping	0,12	0,10	0,12
	3rd Ord. Mapping	**0,12**	**0,08**	**0,10**

7.1.4 Results for database "Audience - Test 4L"



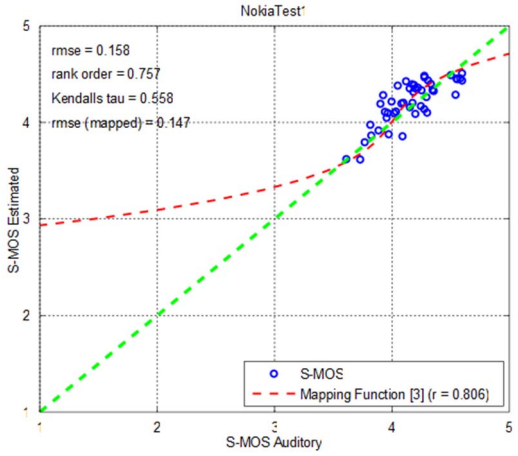


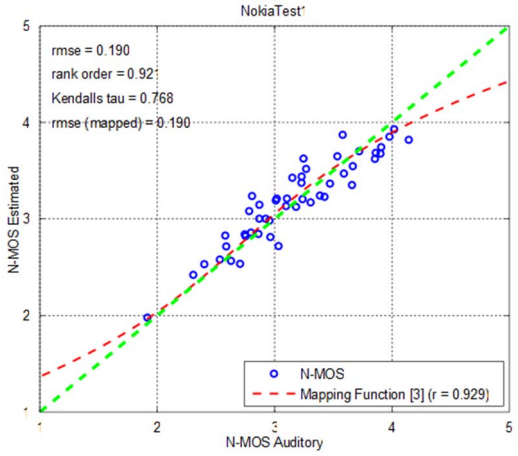


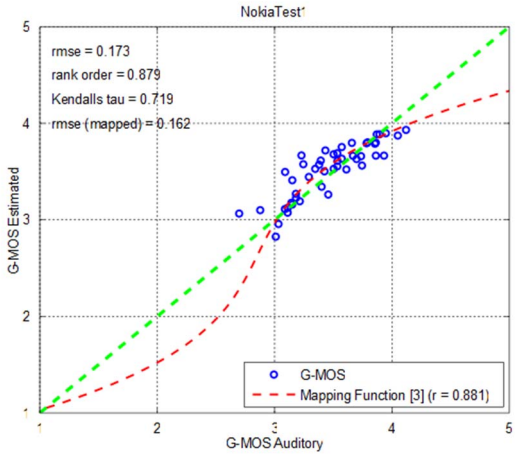
		SIG	BAK	OVRL
RMSE:	no Mapping	0,34	0,21	0,27
	1st Ord. Mapping	0,28	0,21	0,22
	3rd Ord. Mapping	**0,26**	**0,18**	**0,20**

RMSE*:	no Mapping	0,23	0,11	0,17
	1st Ord. Mapping	0,17	0,11	0,14
	3rd Ord. Mapping	**0,15**	**0,08**	**0,12**

7.1.5 Results for database "Nokia - Test 1"



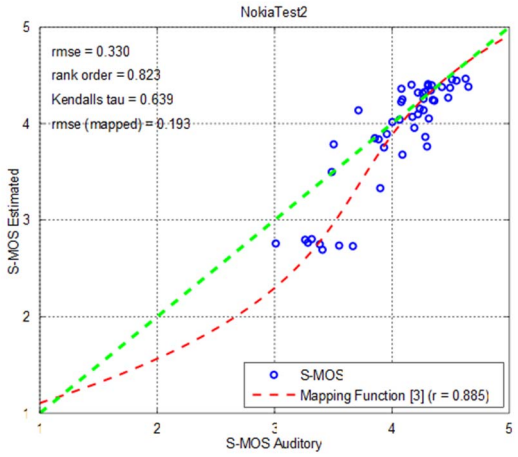


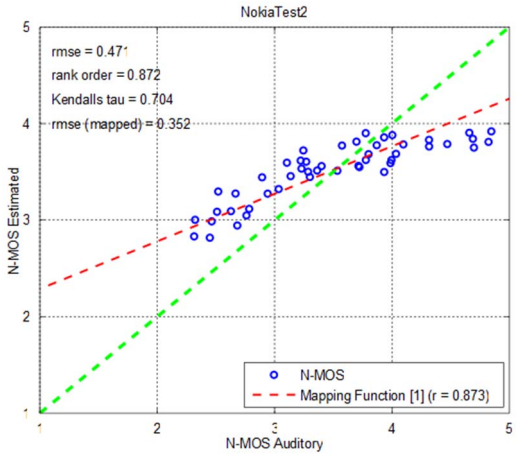


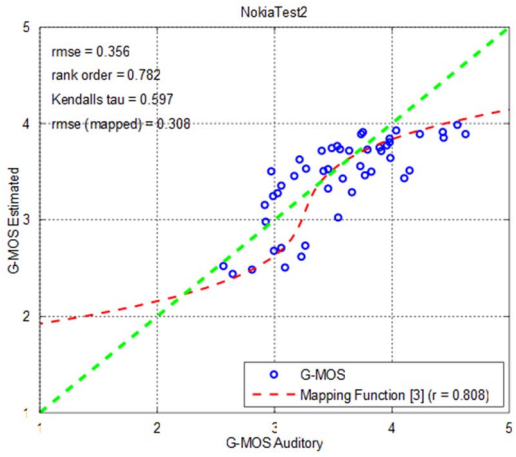
		SIG	BAK	OVRL
RMSE:	no Mapping	0,16	0,19	0,17
	1st Ord. Mapping	0,16	0,19	0,18
	3rd Ord. Mapping	**0,16**	**0,20**	**0,18**

RMSE*:	no Mapping	0,06	0,08	0,09
	1st Ord. Mapping	0,07	0,08	0,09
	3rd Ord. Mapping	**0,07**	**0,09**	**0,09**

7.1.6 Results for database "Nokia - Test 2" (excluded during retraining)



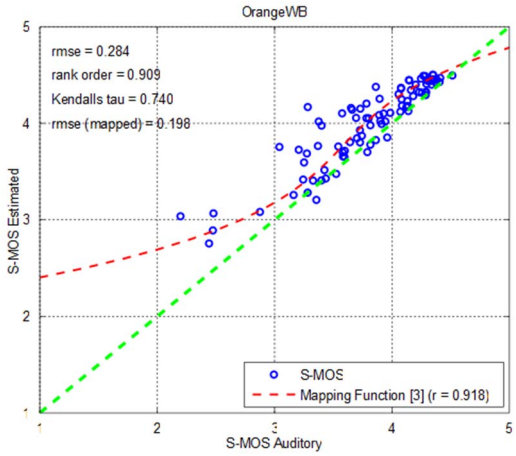


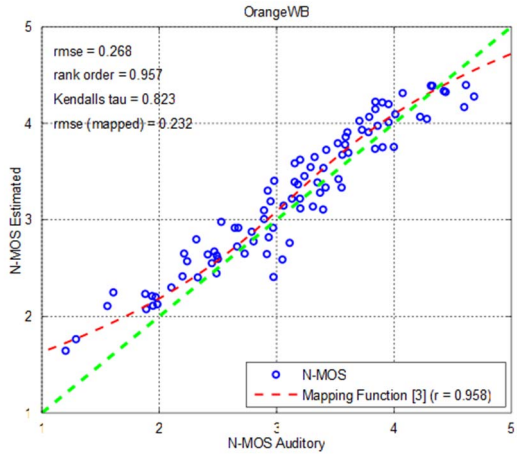


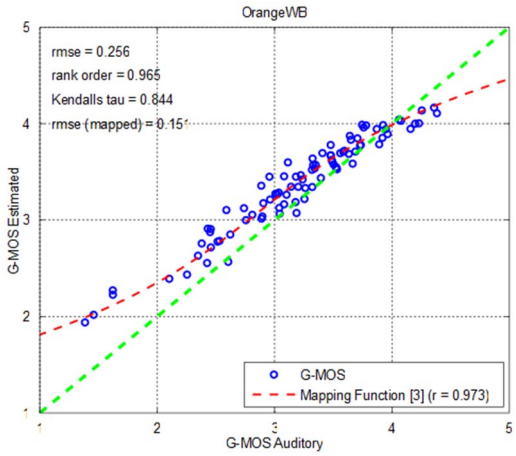
		SIG	BAK	OVRL
RMSE:	no Mapping	0,33	0,47	0,36
	1st Ord. Mapping	0,33	0,48	0,36
	3rd Ord. Mapping	**0,34**	**0,49**	**0,37**

RMSE*:	no Mapping	0,23	0,37	0,26
	1st Ord. Mapping	0,23	0,38	0,26
	3rd Ord. Mapping	**0,24**	**0,38**	**0,26**

7.1.7 Results for database "Orange"



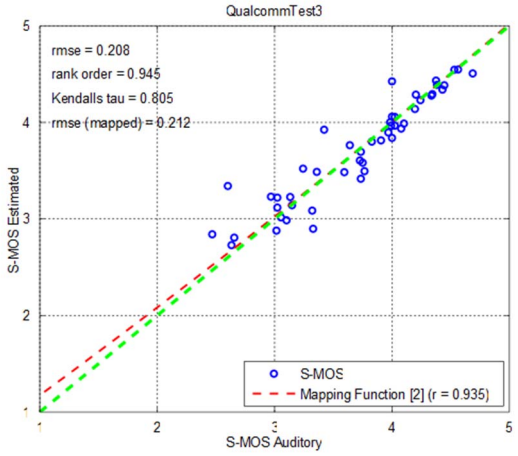


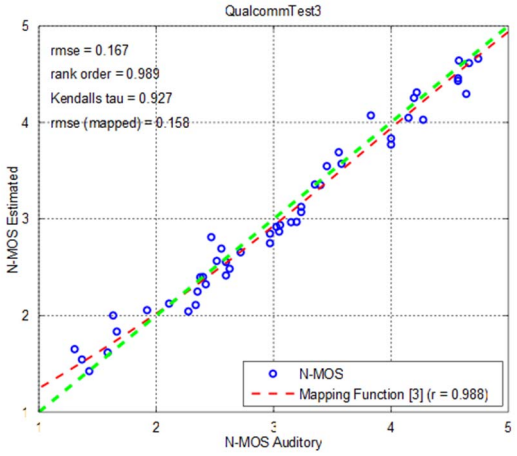


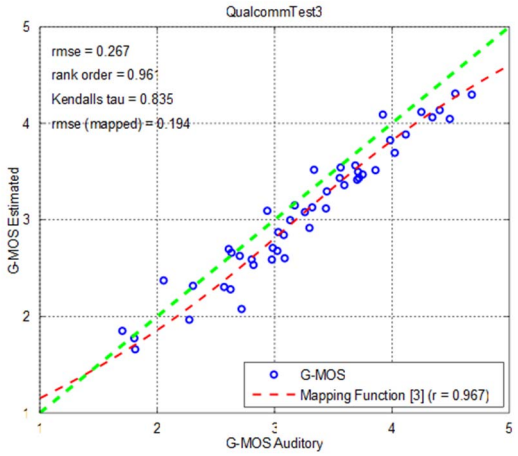
		SIG	BAK	OVRL
RMSE:	no Mapping	0,28	0,27	0,26
	1st Ord. Mapping	0,20	0,24	0,16
	3rd Ord. Mapping	**0,20**	**0,23**	**0,15**

RMSE*:	no Mapping	0,22	0,21	0,20
	1st Ord. Mapping	0,13	0,19	0,10
	3rd Ord. Mapping	**0,13**	**0,18**	**0,10**

7.1.8 Results for database "Qualcomm - Test 3"



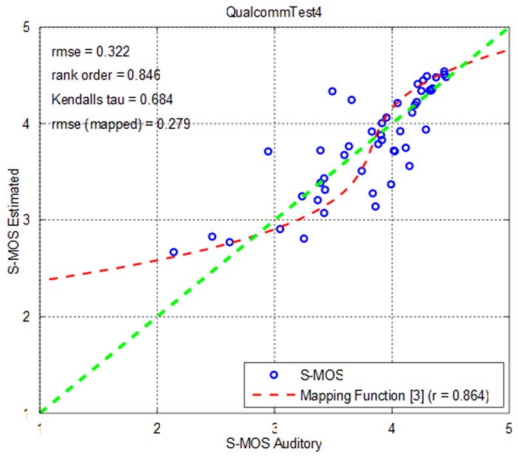


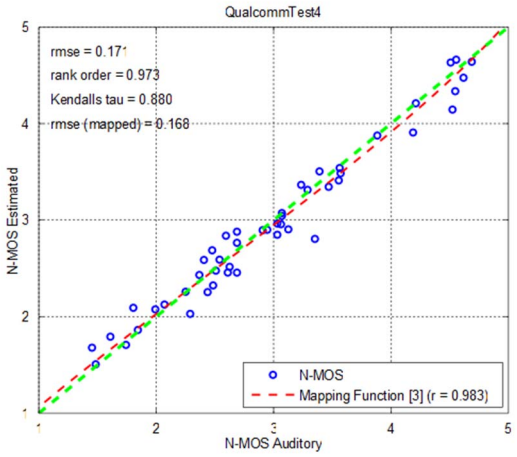


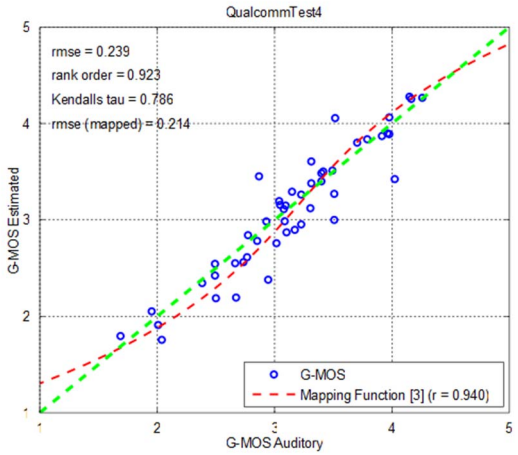
		SIG	BAK	OVRL
RMSE:	no Mapping	0,21	0,17	0,27
	1st Ord. Mapping	0,21	0,17	0,28
	3rd Ord. Mapping	**0,21**	**0,16**	**0,19**

RMSE*:	no Mapping	0,11	0,08	0,16
	1st Ord. Mapping	0,11	0,08	0,18
	3rd Ord. Mapping	**0,11**	**0,07**	**0,10**

7.1.9 Results for database "Qualcomm - Test 4"







		SIG	BAK	OVRL
RMSE:	no Mapping	0,32	0,17	0,24
	1st Ord. Mapping	0,31	0,28	0,24
	3rd Ord. Mapping	**0,28**	**0,17**	**0,21**

RMSE*:	no Mapping	0,22	0,09	0,16
	1st Ord. Mapping	0,21	0,18	0,14
	3rd Ord. Mapping	**0,17**	**0,08**	**0,12**

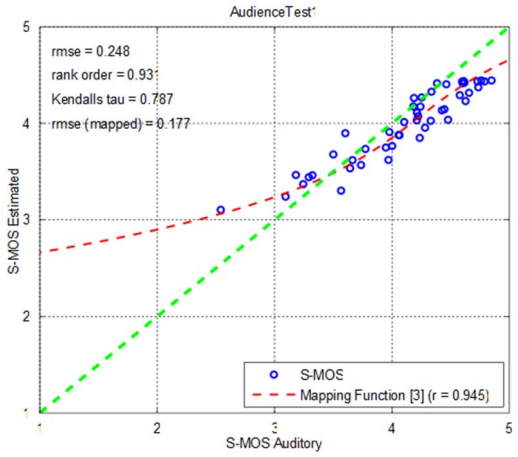
7.2 Results in narrowband mode

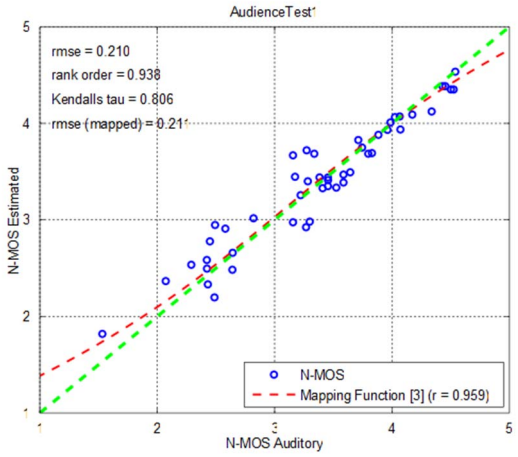
7.2.0 General

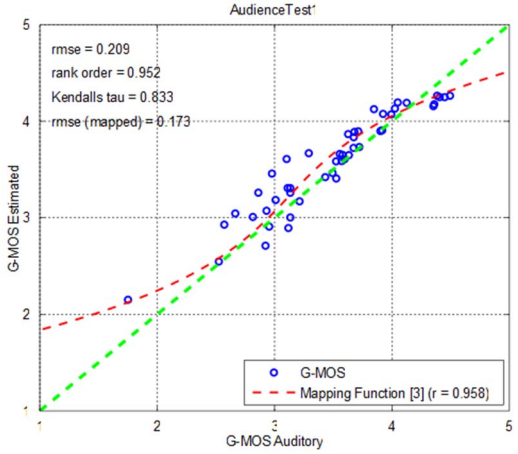
For the narrowband retraining procedure, no database was excluded.

Overall, 6 databases with 288 conditions and 3 840 samples were used.

7.2.1 Results for database "Audience - Test 1"



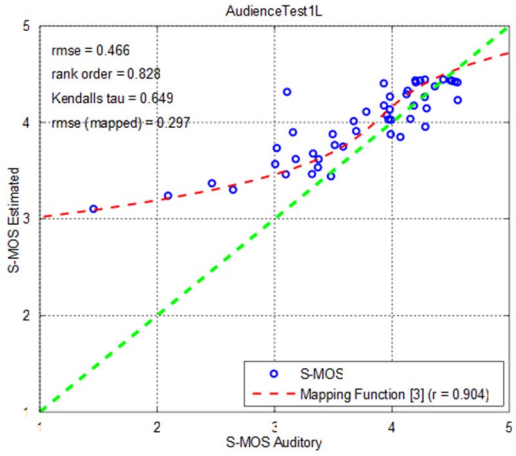


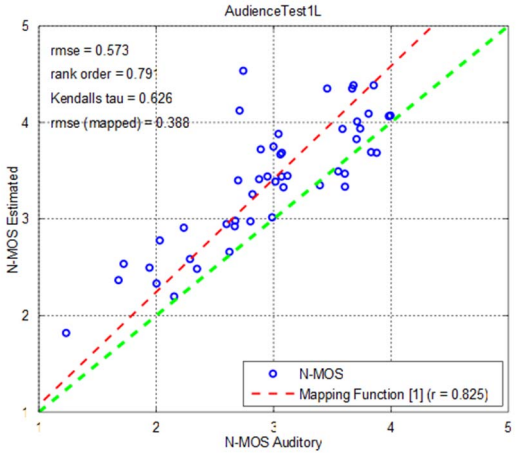


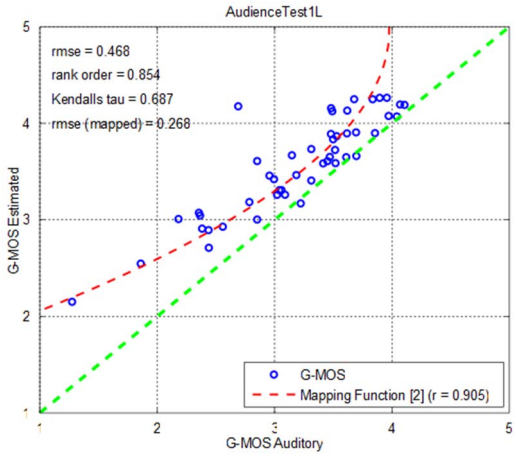
		SIG	BAK	OVRL
RMSE:	no Mapping	0,25	0,21	0,21
	1st Ord. Mapping	0,18	0,21	0,19
	3rd Ord. Mapping	**0,18**	**0,21**	**0,17**

RMSE*:	no Mapping	0,15	0,12	0,12
	1st Ord. Mapping	0,08	0,11	0,10
	3rd Ord. Mapping	**0,08**	**0,11**	**0,07**

7.2.2 Results for database "Audience - Test 1L"



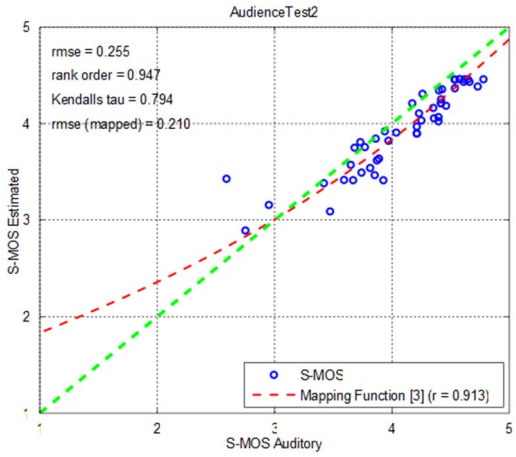


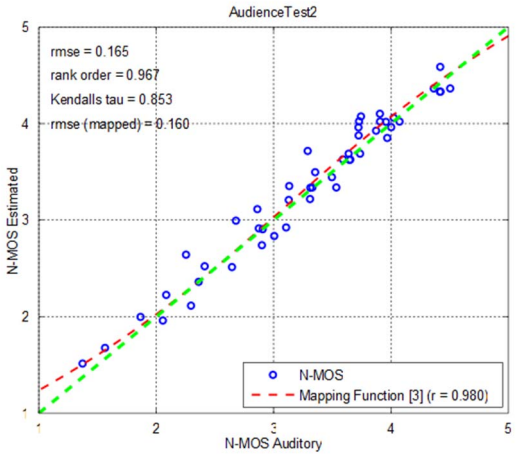


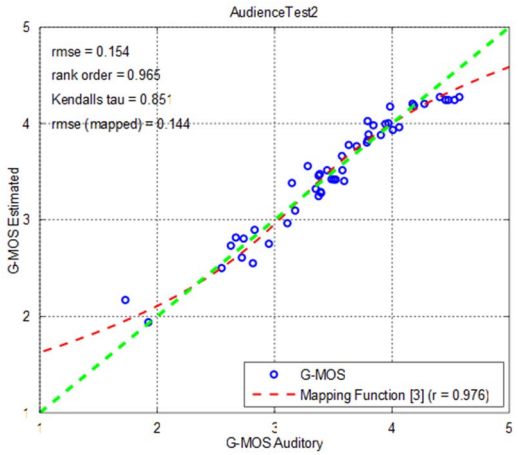
		SIG	BAK	OVRL
RMSE:	no Mapping	0,47	0,57	0,47
	1st Ord. Mapping	0,34	0,39	0,29
	3rd Ord. Mapping	**0,30**	**0,35**	**0,27**

RMSE*:	no Mapping	0,38	0,47	0,37
	1st Ord. Mapping	0,25	0,29	0,20
	3rd Ord. Mapping	**0,21**	**0,24**	**0,19**

7.2.3 Results for database "Audience - Test 2"



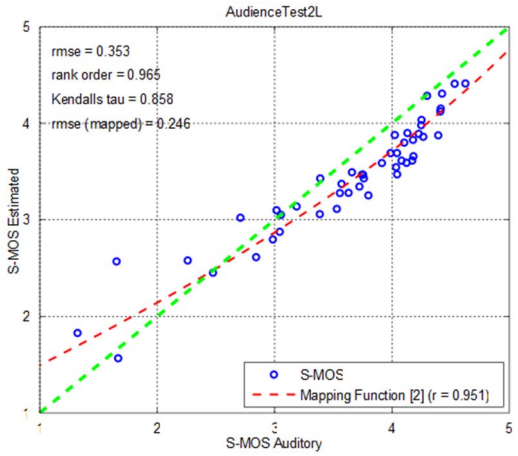


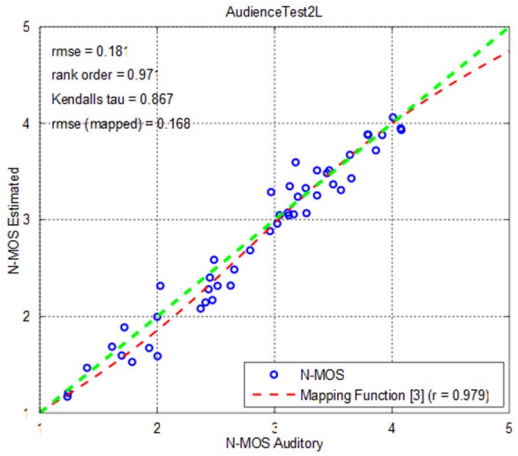


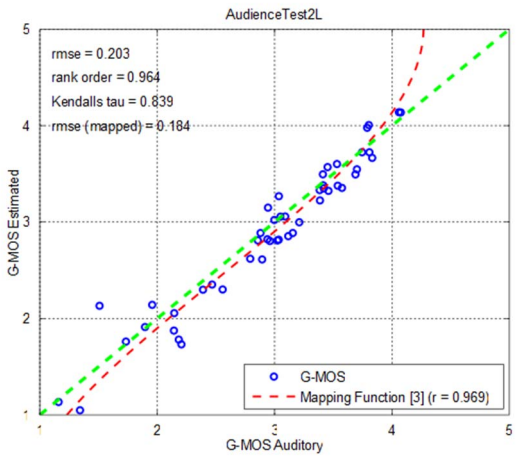
		SIG	BAK	OVRL
RMSE:	no Mapping	0,47	0,57	0,47
	1st Ord. Mapping	0,34	0,39	0,29
	3rd Ord. Mapping	**0,30**	**0,35**	**0,27**

RMSE*:	no Mapping	0,38	0,47	0,37
	1st Ord. Mapping	0,25	0,29	0,20
	3rd Ord. Mapping	**0,21**	**0,24**	**0,19**

7.2.4 Results for database "Audience - Test 2L"



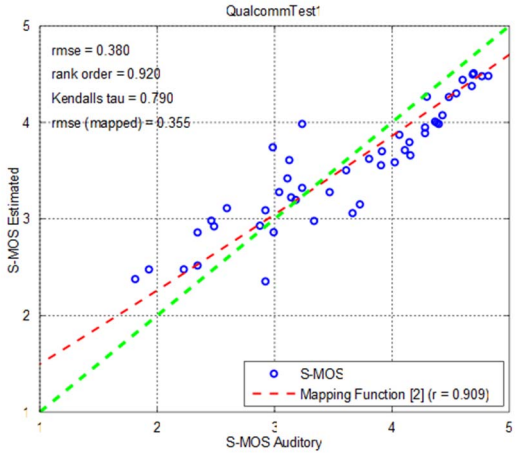


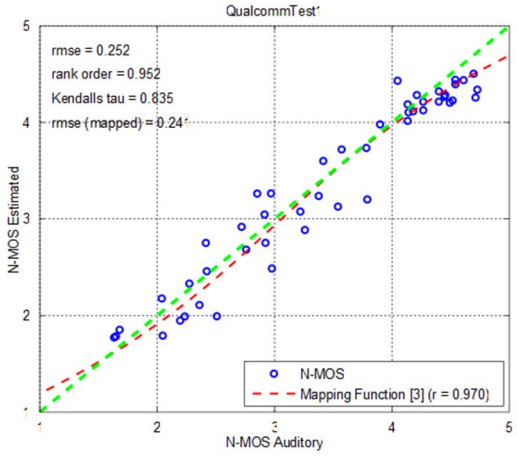


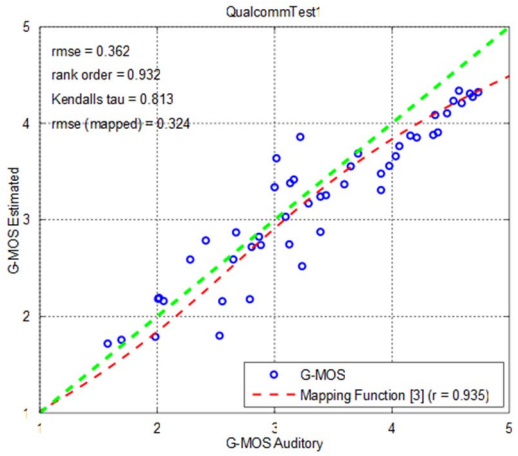
		SIG	BAK	OVRL
RMSE:	no Mapping	0,35	0,18	0,20
	1st Ord. Mapping	0,25	0,17	0,18
	3rd Ord. Mapping	**0,21**	**0,17**	**0,18**

RMSE*:	no Mapping	0,23	0,08	0,11
	1st Ord. Mapping	0,15	0,07	0,11
	3rd Ord. Mapping	**0,11**	**0,08**	**0,11**

7.2.5 Results for database "Qualcomm- Test 1"



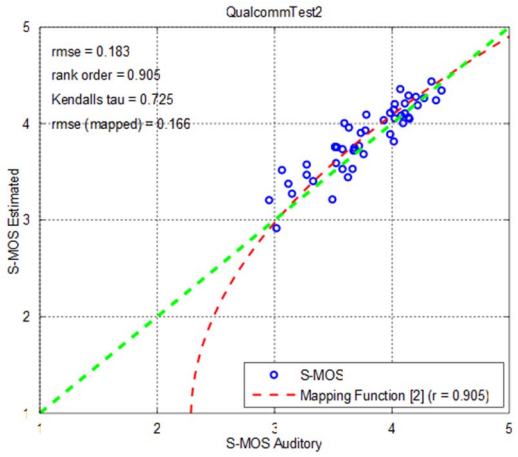


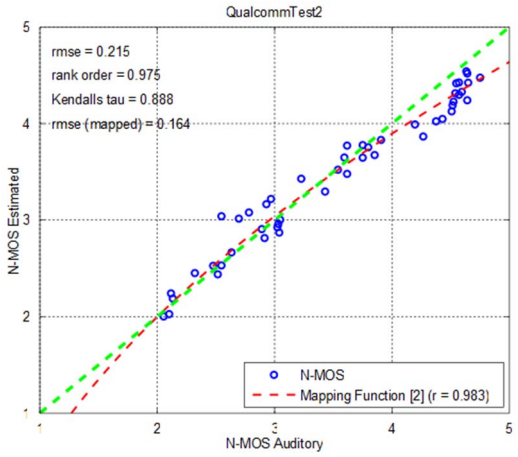


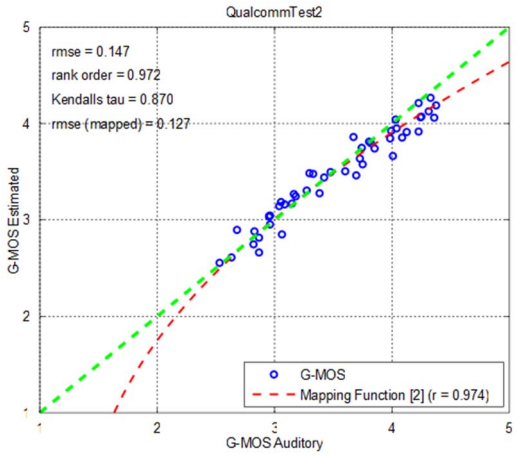
		SIG	BAK	OVRL
RMSE:	no Mapping	0,38	0,26	0,37
	1st Ord. Mapping	0,35	0,33	0,41
	3rd Ord. Mapping	**0,36**	**0,24**	**0,33**

RMSE*:	no Mapping	0,26	0,16	0,27
	1st Ord. Mapping	0,24	0,23	0,31
	3rd Ord. Mapping	**0,24**	**0,13**	**0,22**

7.2.6 Results for database "Qualcomm- Test 2"







		SIG	BAK	OVRL
RMSE:	no Mapping	0,32	0,31	0,37
	1st Ord. Mapping	0,18	0,53	0,22
	3rd Ord. Mapping	**0,18**	**0,17**	**0,16**

RMSE*:	no Mapping	0,21	0,19	0,26
	1st Ord. Mapping	0,08	0,41	0,12
	3rd Ord. Mapping	**0,08**	**0,06**	**0,08**

8 Validation results

8.0 Preamble

For the validation of the model different databases were provided. The databases included different types of conditions and different terminals and simulations. The details of the validation databases are described separately for each set of databases provided by the validation labs.

8.1 Audience validation data

8.1.1 Description of tests

Four tests were conducted, two narrowband (5 and 6) and two wideband (7 and 8). In each test, the noise types listed in [i.1] were used, but the noise levels were increased by 6 dB as in five of the training databases. Six different devices, new to this sequence of validation tests, were used, again a mix of commercial and simulated handsets. All devices were tested in both handset and handheld speakerphone use cases, counterbalanced between the pair of tests at a given bandwidth.

Devices

In each experiment, six devices were evaluated, the maximum number allowed in the EATS-3 [i.1] test plan. In each experiment at one bandwidth, half of the devices were tested in handset mode and half tested in handheld speakerphone mode, in order to provide a consistent and wide range of listening conditions, so that all six devices were tested in both handset and handheld speakerphone modes across the two tests at each bandwidth. The devices included a mix of real and simulated devices with both 1- and 2-microphone noise suppression systems.

The reference conditions and noise types are as defined in table 1 of [i.1] (see table 7a).

Table 7a

Reference Conditions				
File	**SIGNAL**	**SNR**	**Noise Type**	
i01	Source (filtered)	No Noise	-	
i02	Source (filtered)	0 dB	Fullsize_Car1_130Kmh_binaural	
i03	Source (filtered)	12 dB	Fullsize_Car1_130Kmh_binaural	
i04	Source (filtered)	24 dB	Fullsize_Car1_130Kmh_binaural	
i05	Source (filtered)	36 dB	Fullsize_Car1_130Kmh_binaural	
i06	NS Level 1	No Noise	-	
i07	NS Level 2	No Noise	-	
i08	NS Level 3	No Noise	-	
i09	NS Level 4	No Noise	-	
i10	NS Level 3	24 dB	Fullsize_Car1_130Kmh_binaural	
i11	NS Level 2	12 dB	Fullsize_Car1_130Kmh_binaural	
i12	NS Level 1	[0 dB]	Fullsize_Car1_130Kmh_binaural	
Test Conditions				
File	**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.2]**
i13	-1,7/+1,3 dBPa	L: 75,0 dB(A)/R: 73,0 dB(A)	Pub_Noise_binaural_V2	Recording in a pub
i14	-1,7/+1,3 dBPa	L: 74,9 dB(A)/R: 73,9 dB(A)	Outside_Traffic_Road_binaural	Recording at pavement
i15	-1,7/+1,3 dBPa	L: 69,1 dB(A)/R: 69,6 dB(A)	Outside_Traffic_Crossroads_binaural	Recording at pavement
i16	-1,7/+1,3 dBPa	L: 68.2 dB(A)/R:69,8 dB(A)	Train_Station_binaural	Recording at departure platform
i17	-1,7/+1,3 dBPa	L: 69,1 dB(A)/R: 68,1 dB(A)	Fullsize_Car1_130Kmh_binaural	Recording in passenger cabin
i18	-1,7/+1,3 dBPa	L: 68,4 dB(A)/R: 67,3 dB(A)	Cafeteria_Noise_binaural	Recording at sales counter
i19	-1,7/+1,3 dBPa	L: 63,4 dB(A)/R: 61,9 dB(A)	Mensa_binaural	Recording in a cafeteria
i20	-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

However, as noted above for these tests, the noise levels were increased by 6 dB as was done in five of the training databases.

8.1.2 Description of validation results

8.1.2.0 General explanation

For each test, three scatter plots are shown, plotting the results of the predictions versus the subjective data. In each plot, three sets of data are shown, one for no mapping, one for a first-order remapping, and one for a third-order remapping. Tables of correlation, RMSE, and RMSE* [i.9] follow each set of scatter plots. The 1st and 3rd order remappings were derived for each experiment from the 48 test conditions, according to the procedure defined in [i.9]. The intention behind showing scatter plots for the three mapping cases is to demonstrate visually that there is only a small impact of the remapping procedure for these data.

8.1.2.1 Experiment 5: Narrowband

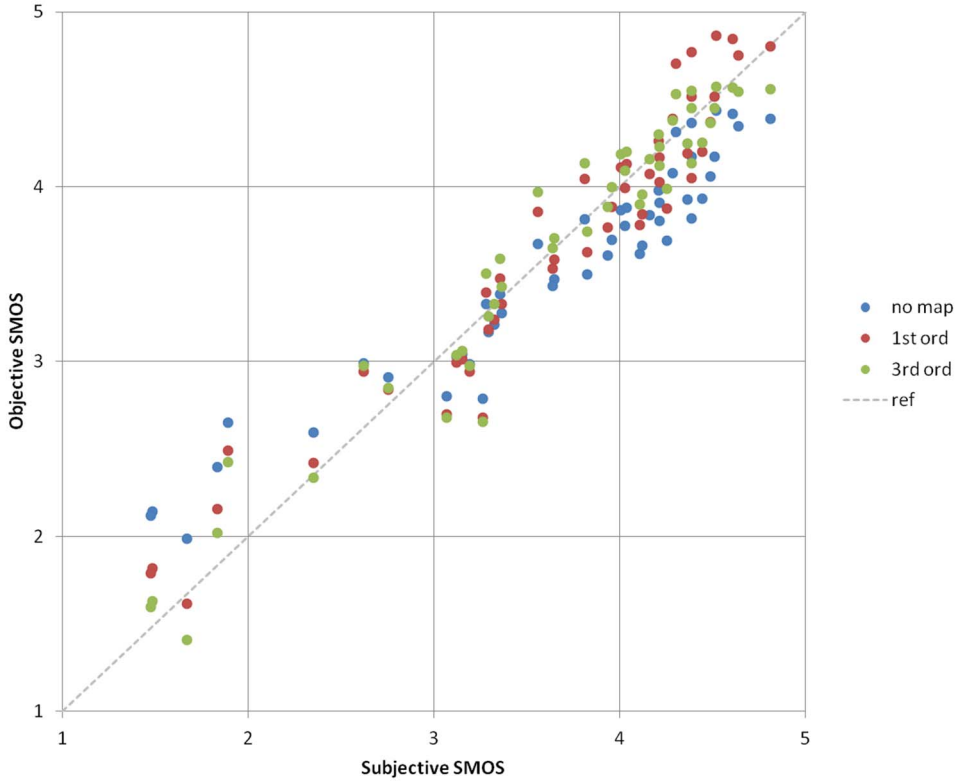


Figure 2: Experiment 5 S-MOS scatter plot

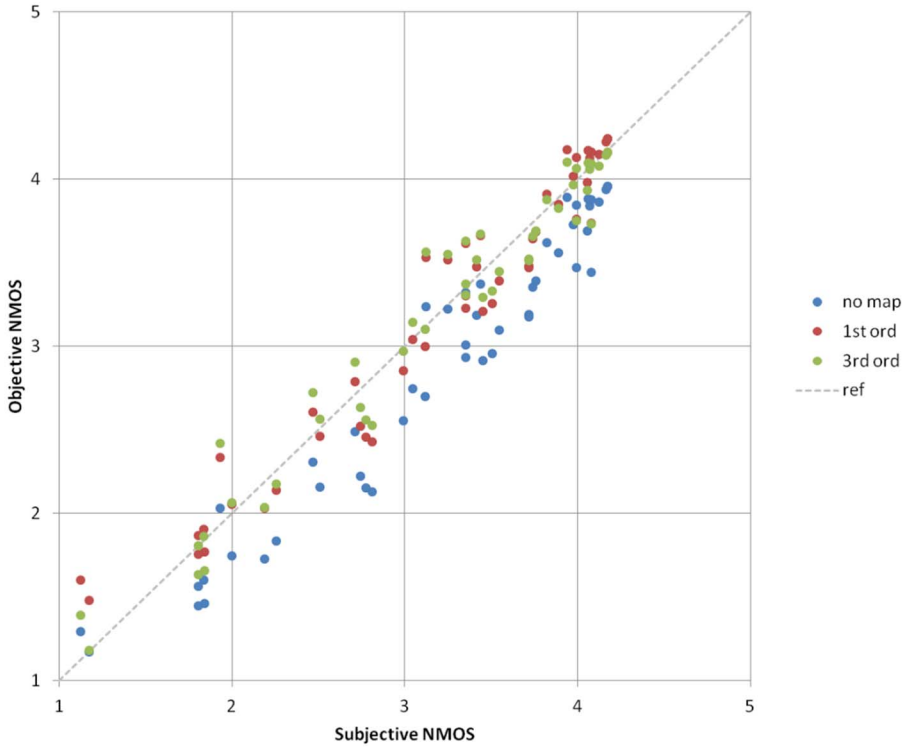


Figure 3: Experiment 5 N-MOS scatter plot

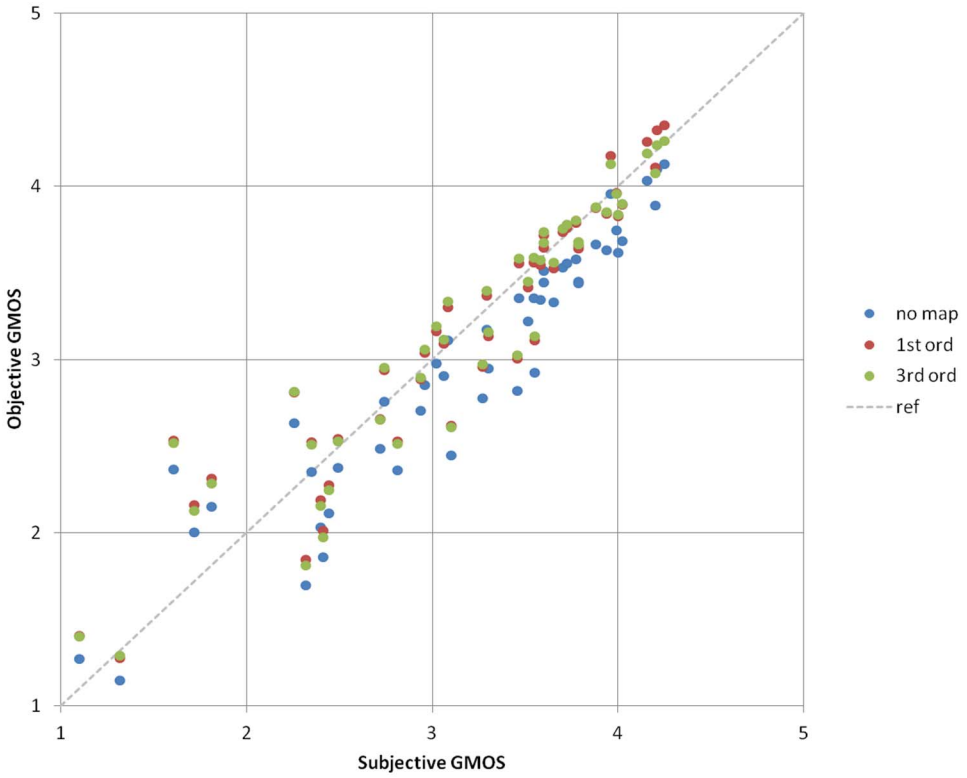


Figure 4: Experiment 5 G-MOS scatter plot

Table 8: Correlation, RMSE, and RMSE* for experiment 5

Condition	S-MOS	N-MOS	G-MOS
Correlation	0,96	0,97	0,94
RMSE, no mapping	0,35	0,36	0,33
RMSE, 1st order mapping	0,25	0,20	0,27
RMSE, 3rd order mapping	**0,22**	**0,18**	**0,28**
RMSE*, no mapping	0,24	0,25	0,23
RMSE*, 1st order mapping	0,14	0,12	0,20
RMSE*, 3rd order mapping	**0,12**	**0,10**	**0,20**

8.1.2.2 Experiment 6: Narrowband

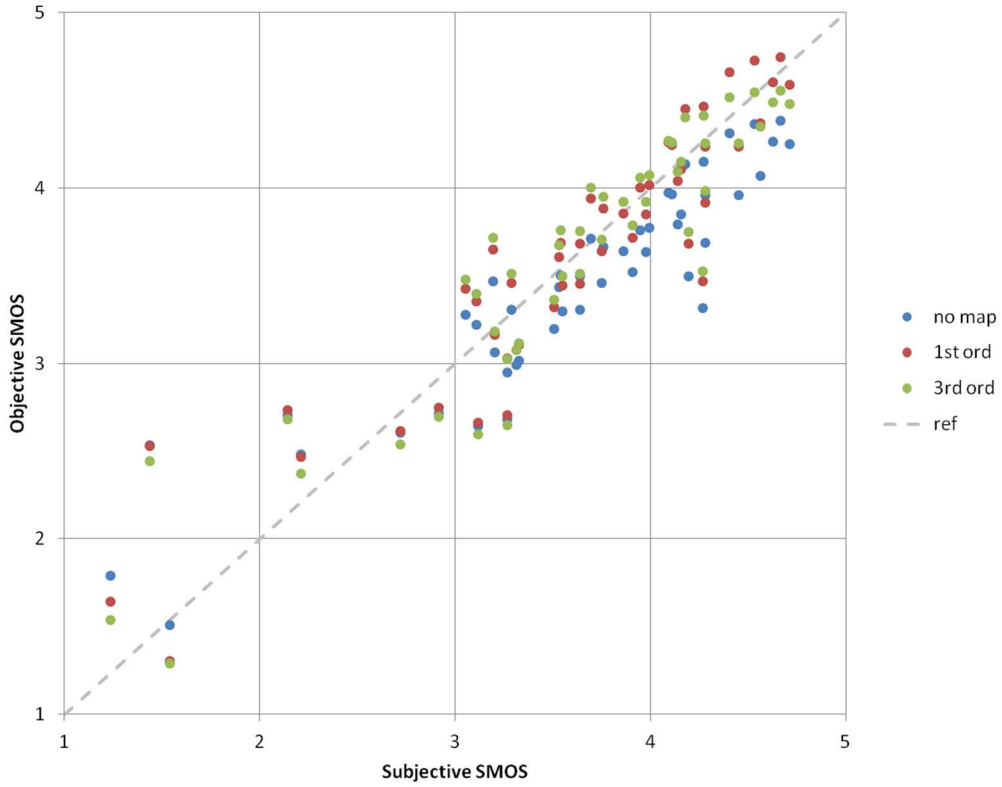


Figure 5: Experiment 6 G-MOS scatter plot

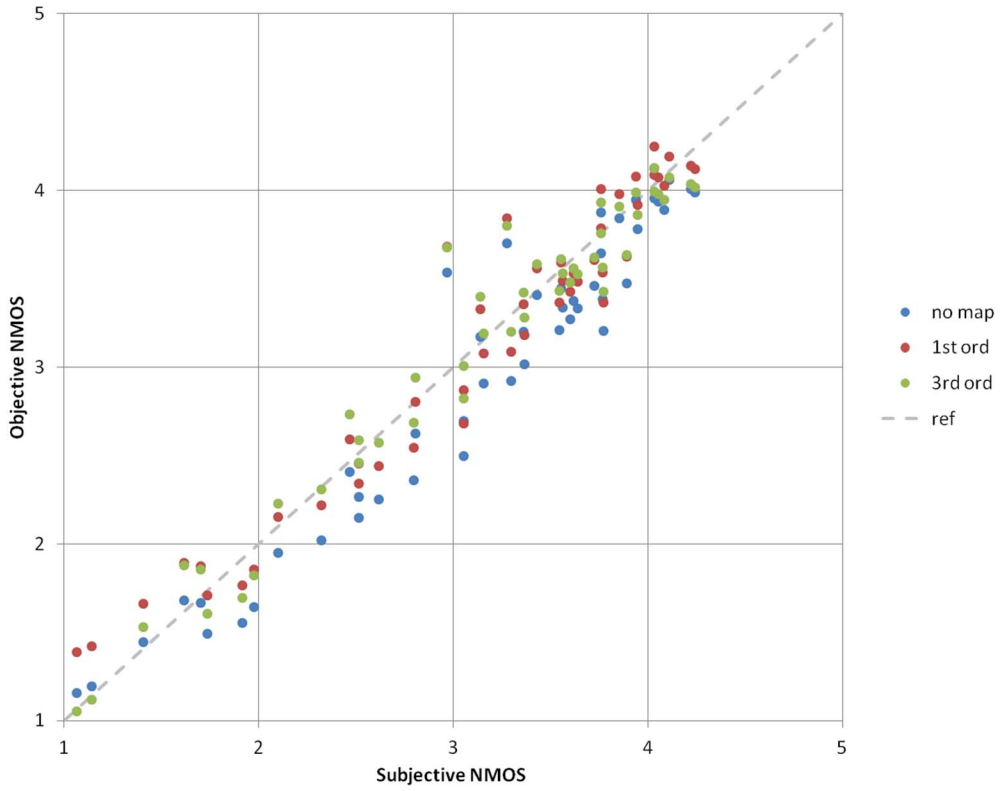


Figure 6: Experiment 6 N-MOS scatter plot

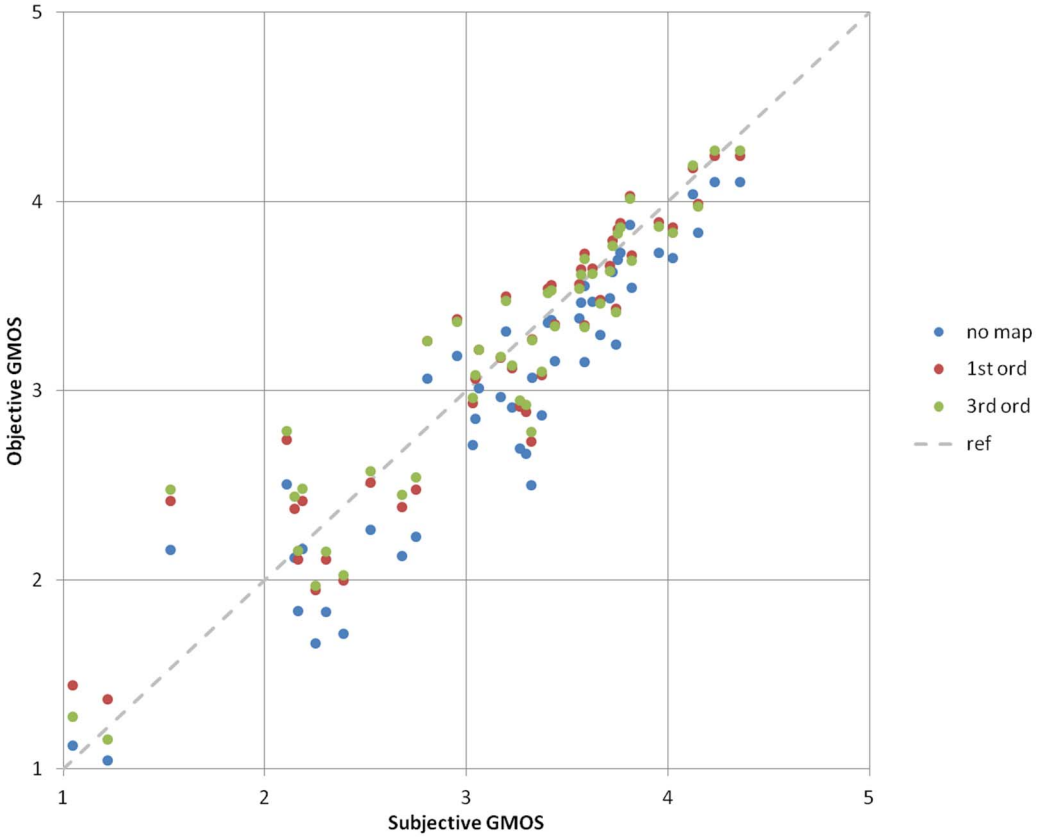


Figure 7: Experiment 6 G-MOS scatter plot

Table 9: Correlation, RMSE, and RMSE* for Experiment 6

Condition	S-MOS	N-MOS	G-MOS
Correlation	0,93	0,97	0,93
RMSE, no mapping	0,38	0,28	0,35
RMSE, 1st order mapping	0,32	0,22	0,28
RMSE, 3rd order mapping	**0,32**	**0,20**	**0,28**
RMSE*, no mapping	0,28	0,18	0,25
RMSE*, 1st order mapping	0,22	0,14	0,19
RMSE*, 3rd order mapping	**0,22**	**0,12**	**0,20**

8.1.2.3 Experiment 7: Wideband

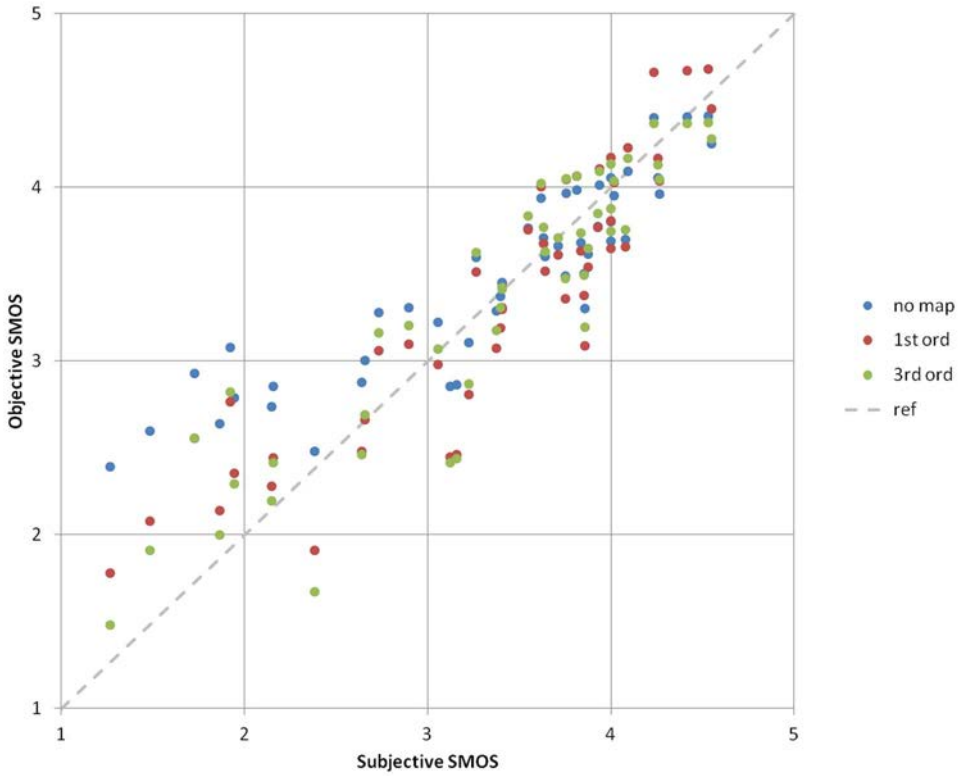


Figure 8: Experiment 7 S-MOS scatter plot

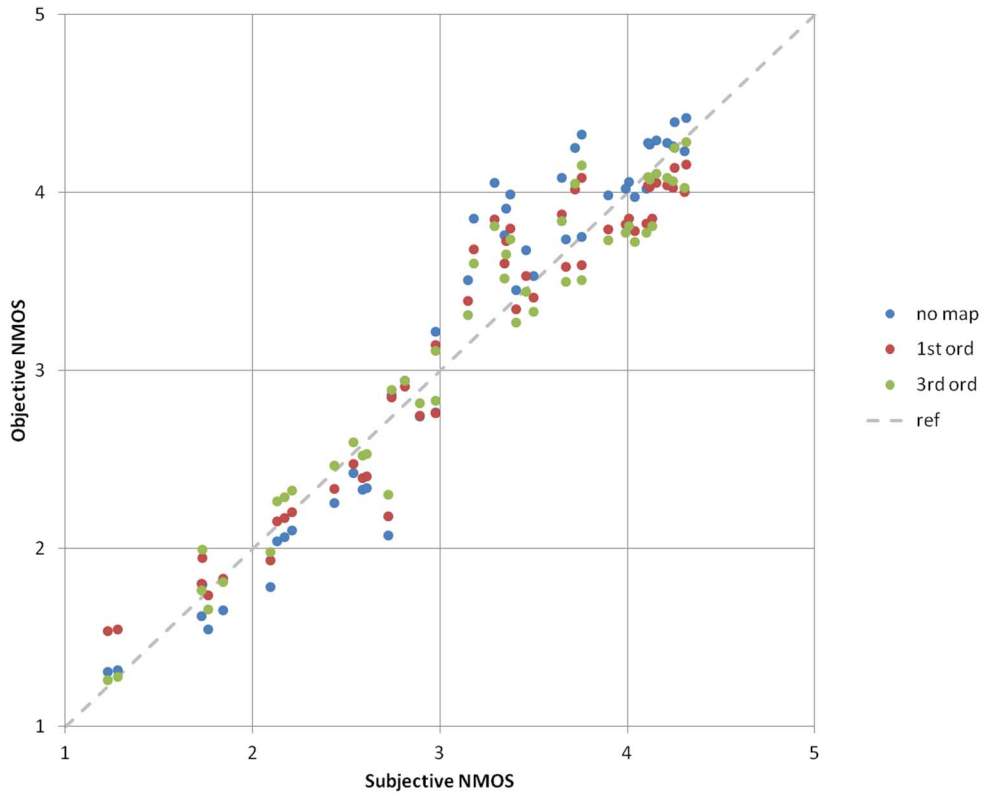


Figure 9: Experiment 7 N-MOS scatter plot

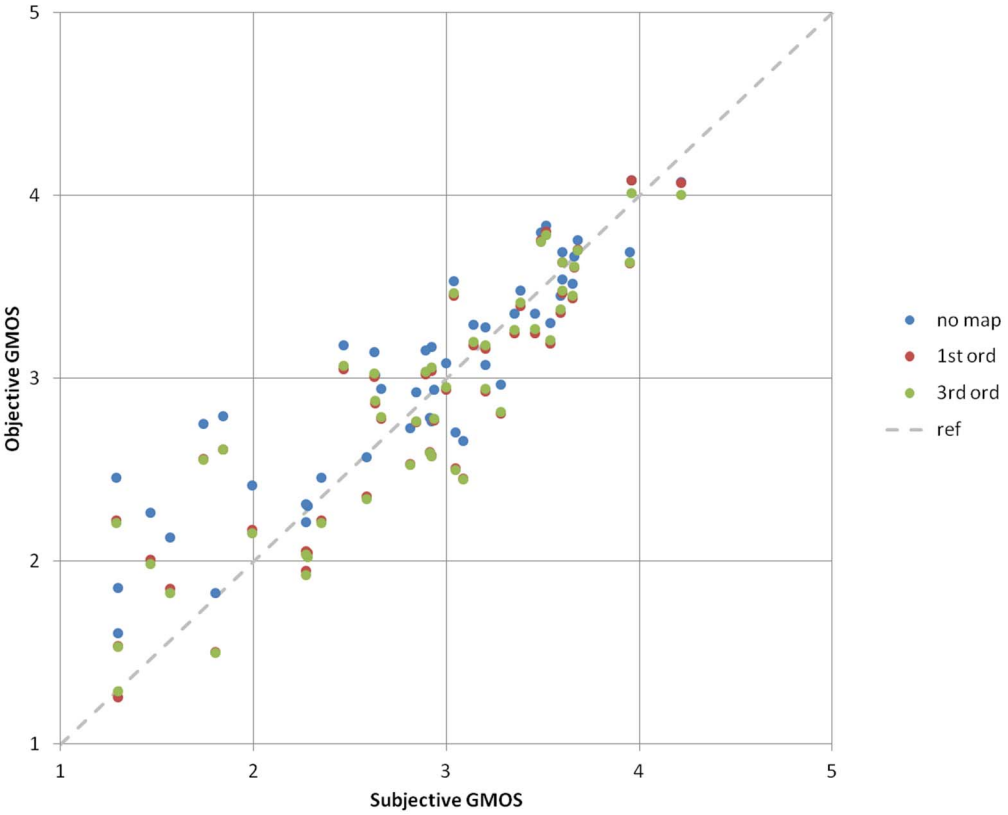


Figure 10: Experiment 7 G-MOS scatter plot

Table 10: Correlation, RMSE, and RMSE* for Experiment 7

Condition	S-MOS	N-MOS	G-MOS
Correlation	0,90	0,96	0,89
RMSE, no mapping	0,46	0,29	0,39
RMSE, 1st order mapping	0,37	0,24	0,35
RMSE, 3rd order mapping	**0,36**	**0,22**	**0,36**
RMSE*, no mapping	0,36	0,20	0,32
RMSE*, 1st order mapping	0,26	0,13	0,26
RMSE*, 3rd order mapping	**0,25**	**0,12**	**0,27**

8.1.2.4 Experiment 8: Wideband

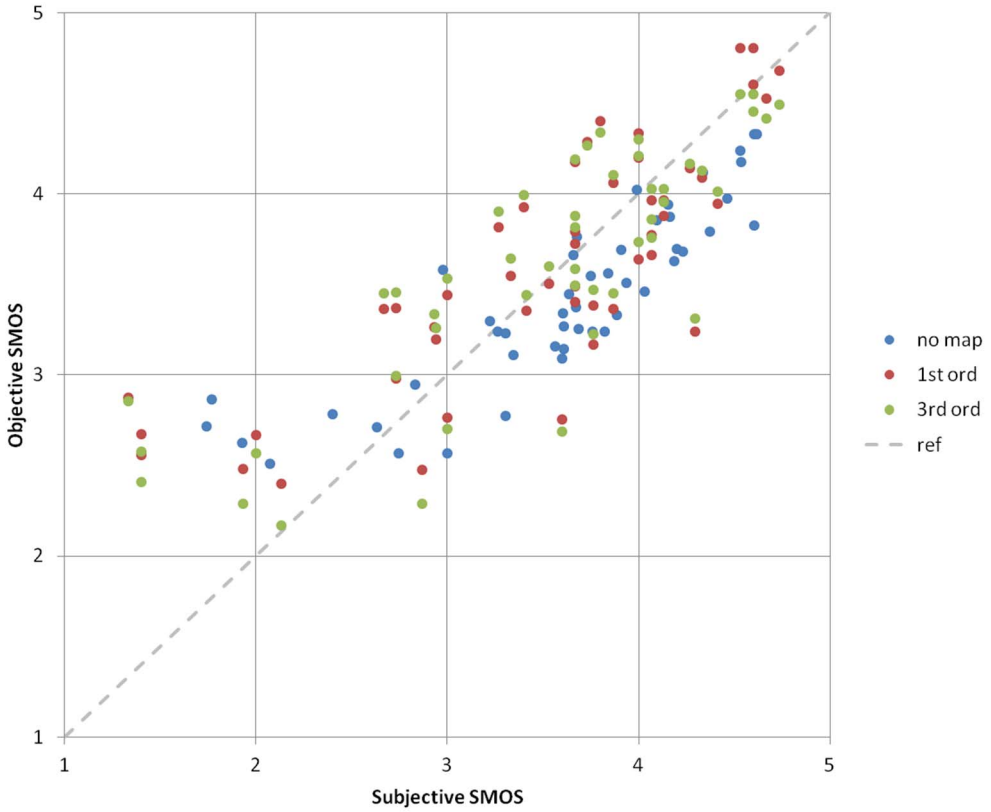


Figure 11: Experiment 8 S-MOS scatter plot

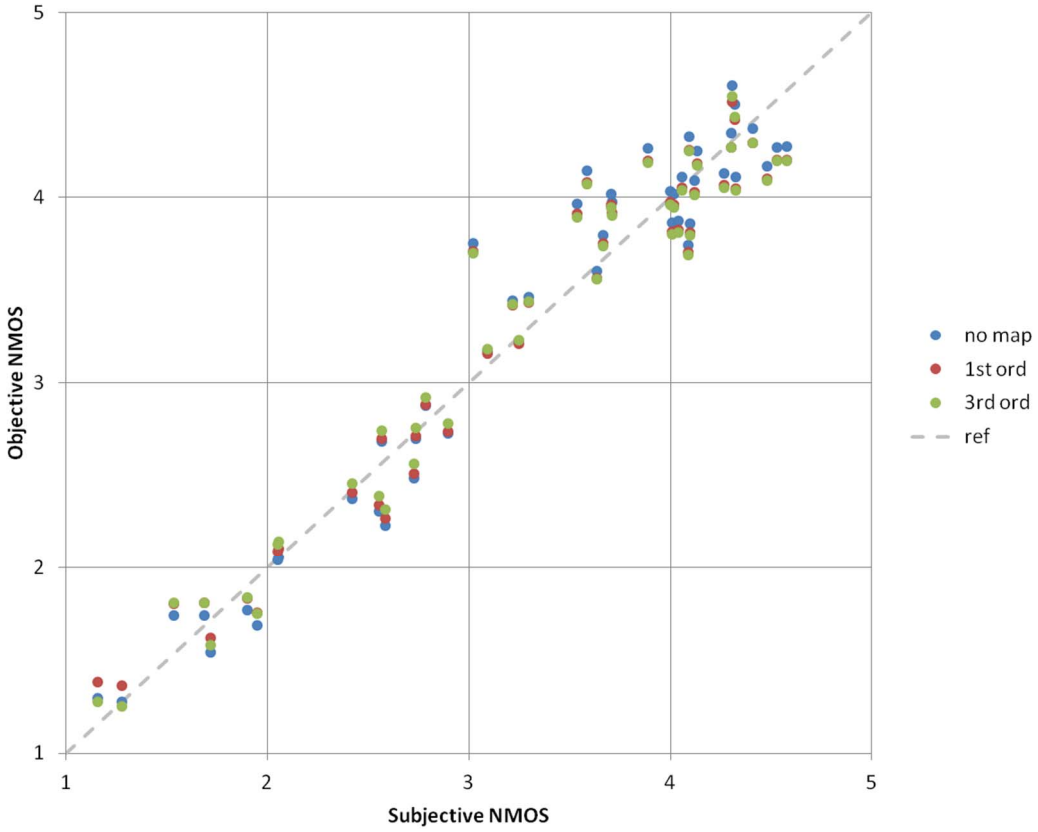


Figure 12: Experiment 8 N-MOS scatter plot

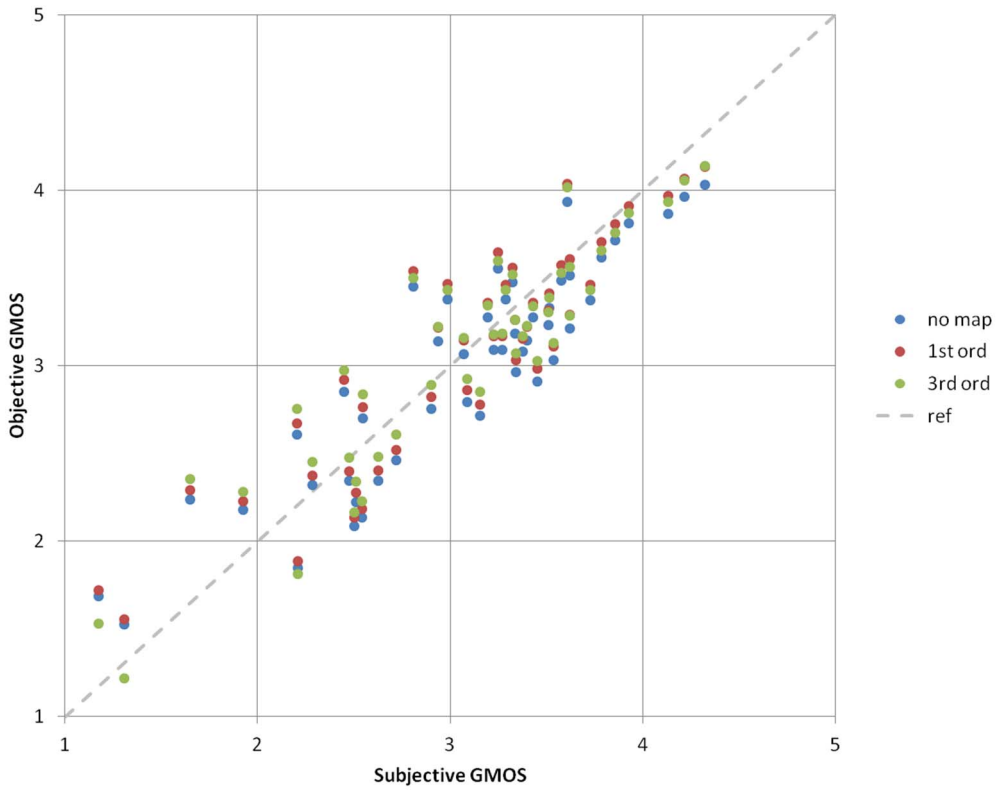


Figure 13: Experiment 8 G-MOS scatter plot

Table 11: Correlation, RMSE, and RMSE* for Experiment 8

Condition	S-MOS	N-MOS	G-MOS
Correlation	**0,87**	**0,97**	**0,90**
RMSE, no mapping	0,45	0,24	0,31
RMSE, 1st order mapping	0,38	0,23	0,31
RMSE, 3rd order mapping	**0,37**	**0,24**	**0,30**
RMSE*, no mapping	0,32	0,14	0,20
RMSE*, 1st order mapping	0,26	0,14	0,20
RMSE*, 3rd order mapping	**0,26**	**0,14**	**0,20**

8.2 Orange validation data

8.2.1 Description of tests

The Orange validation database includes six wideband mobile devices, and three noises from ES 202-396-1 [i.2] at nominal level are used (see table 12). As for speech samples, four talkers are used: two males and two females, with two sentences for each talker. The resulting tests conditions are summarized in table 13. Except for f3, all talkers come from Recommendation ITU-T P.501 [i.13].

Table 12: Noise names and descriptions for Orange validation database

Noise type	Description	ETSI ES 202 396-1 [i.2] filename
Crossroad	Recording at pavement	Outside_Traffic_Crossroads_binaural
Mensa	Recording in a cafeteria	Mensa_binaural
Pub	Recording in a Pub	Pub_Noise_binaural_V2

Table 13: Definition of tests conditions parameters for Orange WB validation test

Test conditions	Number	Designation
Noises	3	N1, N2, N3
SNR	1	Nominal level
Devices	6	D1, …, D6
Talkers	4	m1, m2, f2, f3
Sentences per talker	2	s1, s2

All test conditions were processed with the 4 talkers and 2 sentences. Level adjustment was performed as described in EATS-3.

Reference conditions which incorporate a spectral subtraction based distortion were included in the test and are listed in table 14. These reference conditions are exactly the same as the one provided in EATS-3, table 2 of [i.1].

Table 14: Reference set conditions for wideband testing

Reference Conditions			
File	**SIG.**	**SNR**	**Noise Type**
i01	Source (filtered)	No Noise	-
i02	Source (filtered)	10 dB	Outside_Traffic_Crossroads_binaural
i03	Source (filtered)	20 dB	Outside_Traffic_Crossroads_binaural
i04	Source (filtered)	30 dB	Outside_Traffic_Crossroads_binaural
i05	Source (filtered)	40 dB	Outside_Traffic_Crossroads_binaural
i06	NS Level 1, 2nd set of parameters	No Noise	-
i07	NS Level 2, 2nd set of parameters	No Noise	-
i08	NS Level 3, 2nd set of parameters	No Noise	-
i09	NS Level 4, 2nd set of parameters	No Noise	-
i10	NS Level 3, 2nd set of parameters	30 dB	Outside_Traffic_Crossroads_binaural
i11	NS Level 2, 2nd set of parameters	20 dB	Outside_Traffic_Crossroads_binaural
i12	NS Level 1, 2nd set of parameters	10 dB	Outside_Traffic_Crossroads_binaural

8.2.2 Description of validation results

Scatter plots on a per condition basis are provided in figures 14 to 16: they show the distribution over the quality range for the three dimensions (Speech, Noise, Overall quality).

The RMSE and RMSE* performance parameters specified in [i.9] were computed. Results before mapping and after monotonic 3rd order mapping are presented in tables 15 and 16 respectively. The Pearson correlation is also reported in table 17. These results are meeting the performance requirements specified for RMSE and RMSE* on the 3rd order remapping, as given in [i.9].

Table 15: Statistical analysis results before mapping

	S-MOS	N-MOS	G-MOS
RMSE	0,68	0,29	0,62
RMSE*	0,58	0,23	0,53

Table 16: Statistical analysis results after monotonic 3rd order mapping

	S-MOS	N-MOS	G-MOS
RMSE	0,38	0,23	0,29
RMSE*	0,30	0,16	0,21

Table 17: Pearson correlation (after monotonic 3rd order mapping)

	S-MOS	N-MOS	G-MOS
before mapping	0,90	0,97	0,90
after monotonic 3rd order mapping	0,91	0,98	0,93

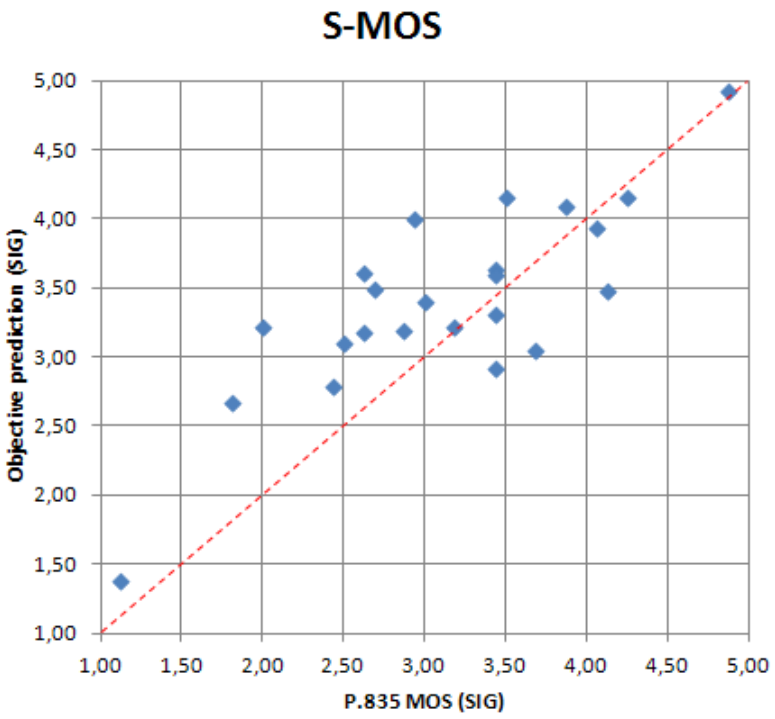


Figure 14: S-MOS scatter plot

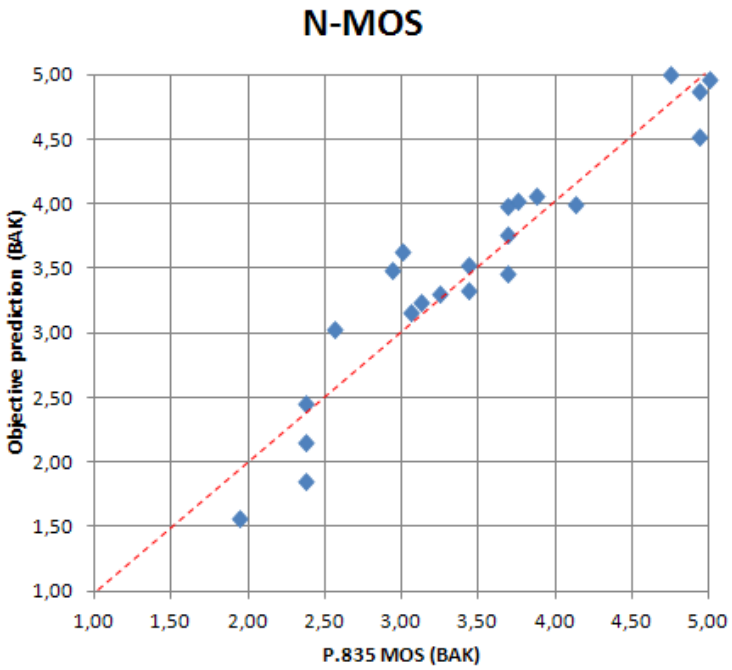


Figure 15: N-MOS scatter plot

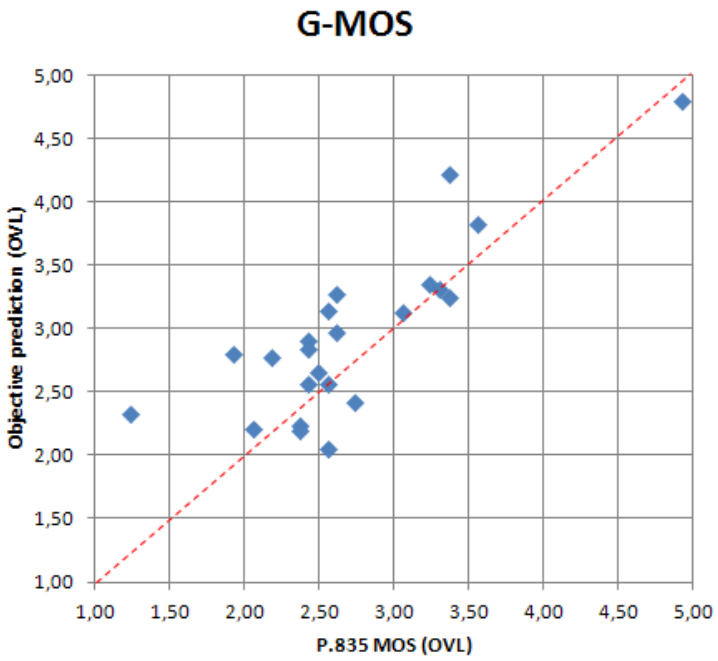


Figure 16: G-MOS scatter plot

8.3 Qualcomm validation data

8.3.1 Description of tests

Two narrowband experiments following the EATS-3 subjective test plan [i.1] were conducted. The test set-up, background noise reproduction calibration and levels, noise types and convergence sequencing are according to the EATS-3 subjective test plan [i.1], except where noted. The reference conditions are according to [i.1], table 1.

In the first validation experiment (Exp 6), 2 devices were tested with 7 noise types and a clean condition (no noise added). The devices were tested in the following modes:

- handset with AMR 12,2 kbps;
- handset with AMR 5,9 kbps;
- handheld Hands-free with AMR 5,9 kbps;

resulting in a total of 48 test conditions. The inclusion of AMR 5,9 kbps was used in order to increase the range of degradations for the validation tests. Commercial devices in a call with a CMU200 network simulator were used.

In the second validation experiment (Exp 7), 1 device was tested with 7 noise types and a clean condition (no noise added). The device was tested in the following modes:

- handset with AMR 12,2 kbps;
- handset with AMR 5,9 kbps;
- handheld Hands-free with AMR 5,9 kbps;
- handset with AMR 12,2 kbps (Noise levels increased by 6 dB);
- handset with AMR 5,9 kbps (Noise levels increased by 6 dB);
- handheld Hands-free with AMR 5,9 kbps (Noise levels increased by 6 dB);

resulting in a total of 48 test conditions. A commercial device in a call with the CMU200 network simulator was used.

The same reference set (exact same signals) was used in the narrowband experiments reported in previous contributions in order to keep consistency and facilitate any necessary mapping or normalization of the data.

Tables 18 and 19 detail the conditions for both experiments.

Table 18: Summary of experimental conditions for EXP 6 (NB)

Experiment	6
Number of devices	2 (HS AMR 12.2; HS AMR 5.9; HHHF AMR 5.9)
Number of noise conditions per device	8 noise conditions
Number of reference conditions	12
Number of test conditions	48
Number of talkers	4
Number of samples per talker	4
Number of votes per condition	128
Method of presentation	Diotic
Presentation level (for -26 dBov)	73dBSPL
Headphones	HD280 PRO
Reference set	According to table 1 and batch processing script in clause 8.3 of [i.1].
Noise conditions	Pub_Noise_binaural_V2
	Outside_Traffic_Road_binaural
	Outside_Traffic_Crossroads_binaural
	Clean (no noise)
	Fullsize_Car1_130Kmh_binaural
	Cafeteria_Noise_binaural
	Mensa_binaural
	Work_Noise_Office_Callcenter_binaural

Table 19: Summary of experimental conditions for EXP 7 (NB)

Experiment	7
Number of devices	1 (HS AMR12.2; HS AMR5.9, HHHF AMR12.2, HHHF AMR5.9)
Number of noise conditions per device	16 noise conditions
Number of reference conditions	12
Number of test conditions	48
Number of talkers	4
Number of samples per talker	4
Number of votes per condition	128
Method of presentation	Diotic
Presentation level (for -26 dBov)	73dBSPL
Headphones	HD280 PRO
Reference set	According to table 1 and batch processing script in clause 8.3 of [i.1].
Noise conditions	Pub_Noise_binaural_V2 (nominal and +6 dB)
	Outside_Traffic_Road_binaural (nominal and +6 dB)
	Outside_Traffic_Crossroads_binaural (nominal and +6 dB)
	Clean (no noise, two different recordings)
	Fullsize_Car1_130Kmh_binaural (nominal and +6 dB)
	Cafeteria_Noise_binaural (nominal and +6 dB)
	Mensa_binaural (nominal and +6 dB)
	Work_Noise_Office_Callcenter_binaural (nominal and +6 dB)

The results for Experiments 6 and 7 are summarized in figures 17 and 18. The results for S-MOS (SIG), N-MOS (BAK) and G-MOS (OVRL) of 60 conditions (being 48 test and 12 reference conditions) are reported for each experiment. Results are sorted by OVRL.

It can be seen that both experiments exercised the entire range of degradations for the SIG, BAK and OVRL scales. About 67 % of the scores for OVRL are > 3,0 in both tests. This is in contrast with previous experiments conducted by the source where 3,0 represented the median of the scores for OVRL. This effect is observed despite an attempt to increase the range of degradations by including raised noise levels and AMR 5,9 kbps speech coding.

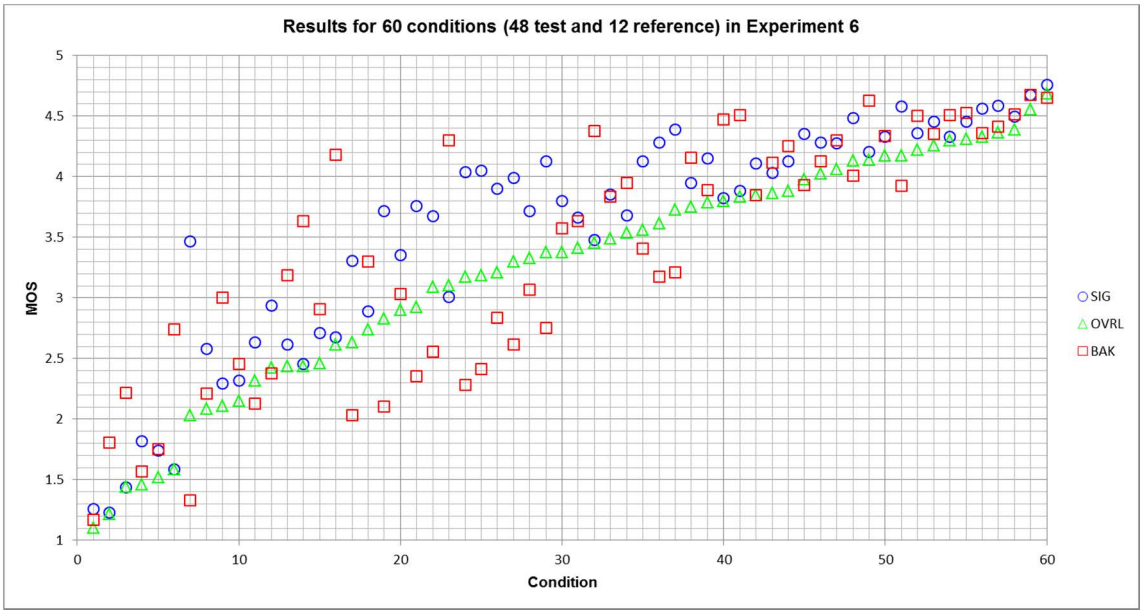


Figure 17: Results of Experiment 6

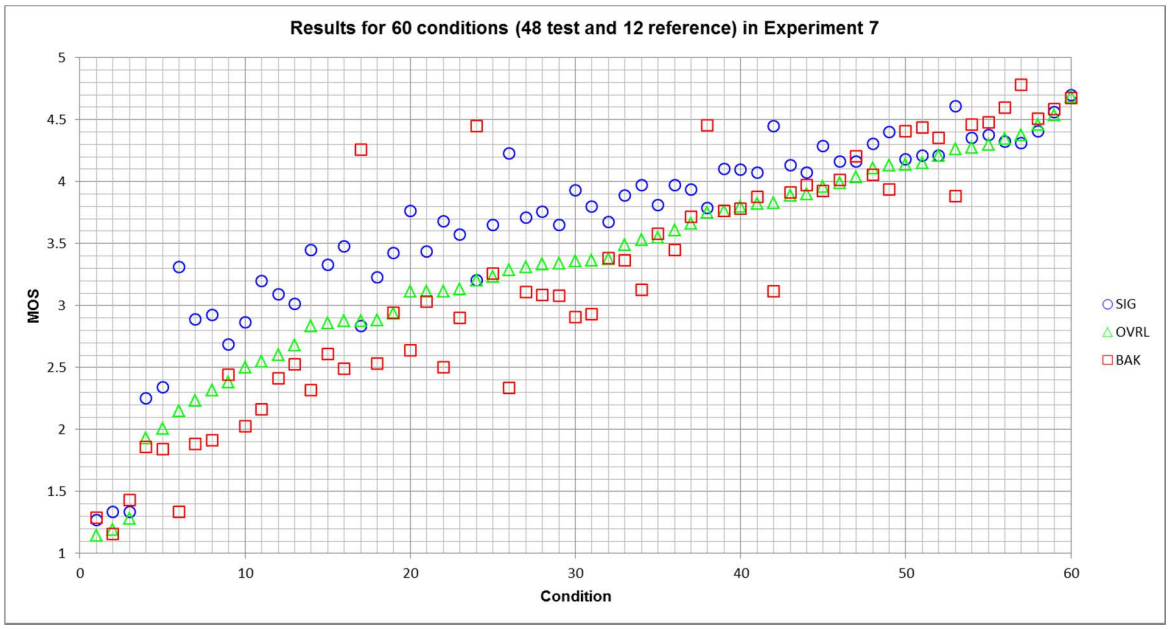


Figure 18: Results of Experiment 7

8.3.2 Description of validation results

Each individual sample used in Experiments 6 and 7 was processed by HEAD Acoustics GmbH using the re-trained P.835 objective predictor model. An average of the objective scores per condition (average of the scores of 16 samples), as well as the 95 % confidence interval was computed and plotted against the results of the subjective test. Scatter plots for N-MOS, S-MOS and G-MOS are shown in figures 19 to 24.

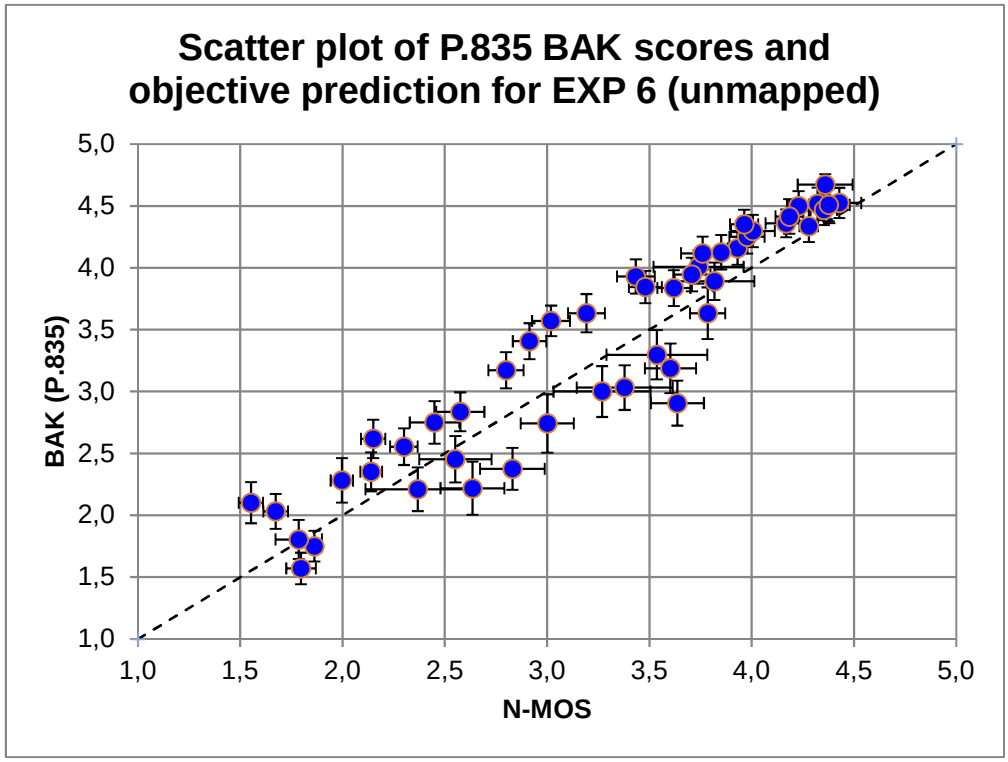


Figure 19: Experiment 6 N-MOS scatter plot

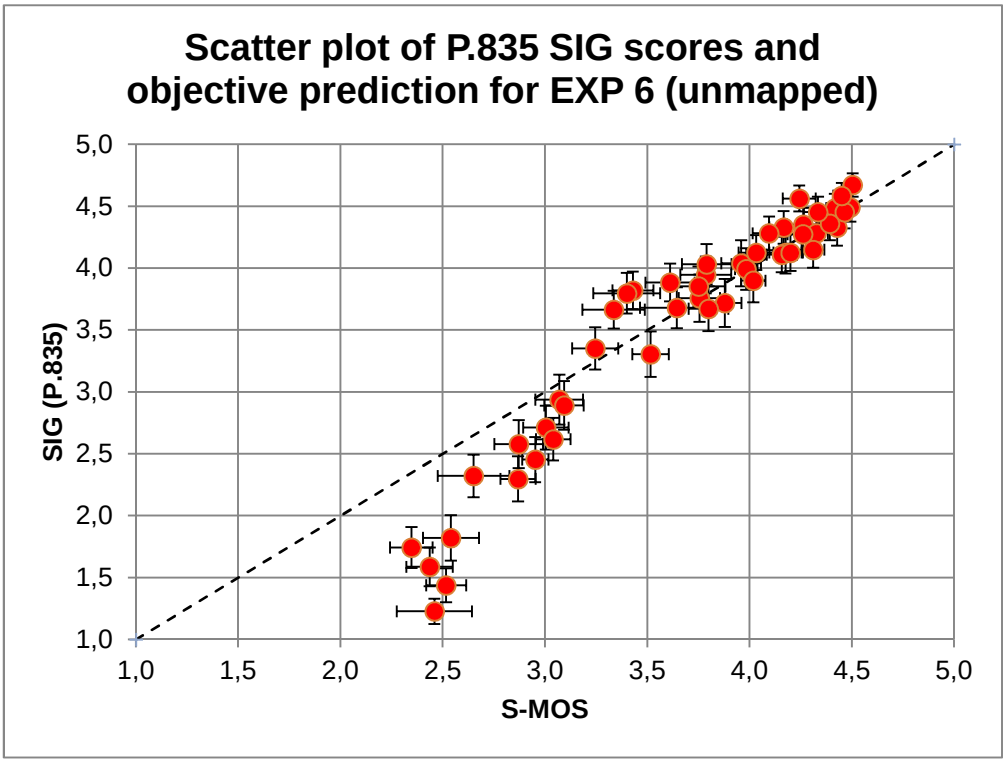


Figure 20: Experiment 6 S-MOS scatter plot

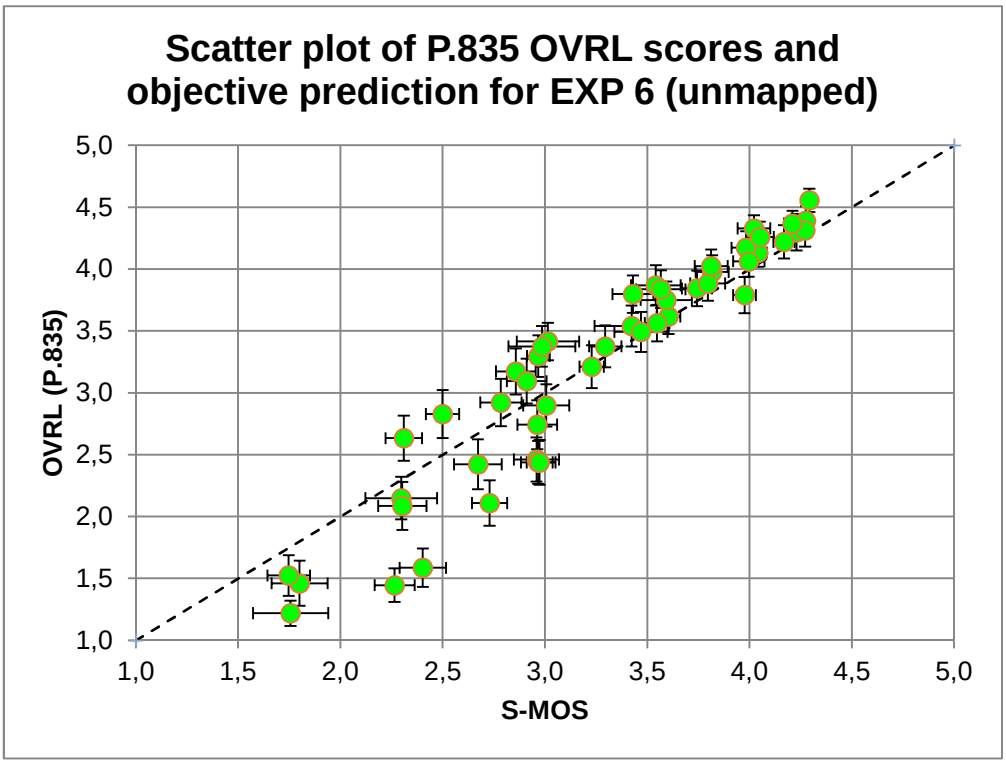


Figure 21: Experiment 6 G-MOS scatter plot

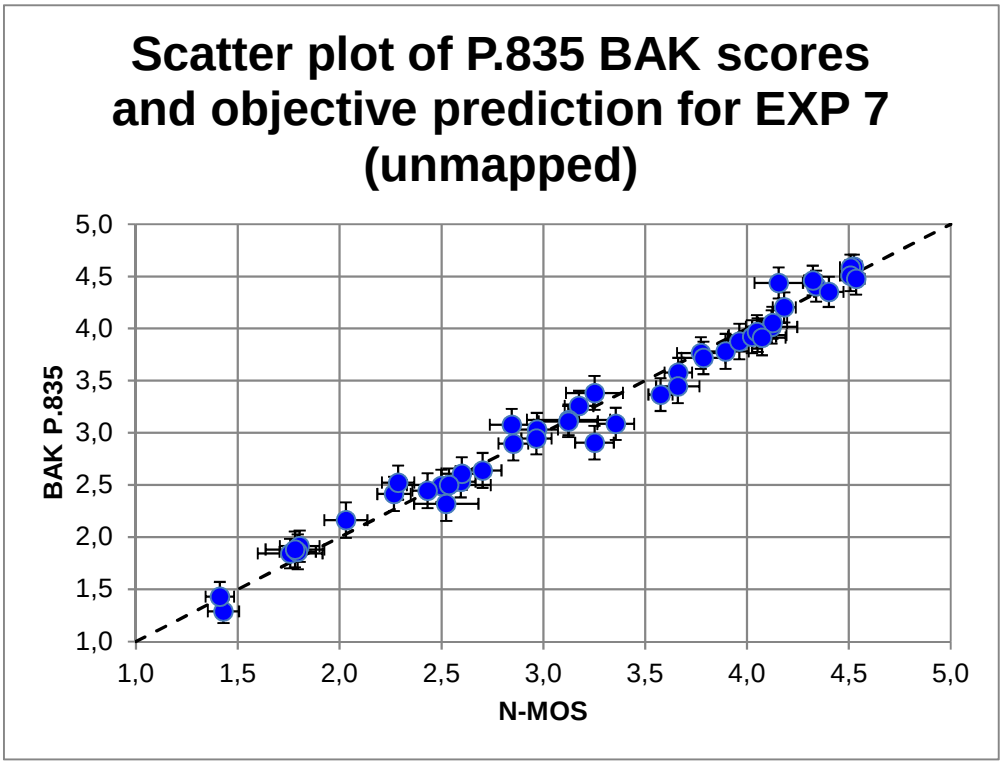


Figure 22: Experiment 7 N-MOS scatter plot

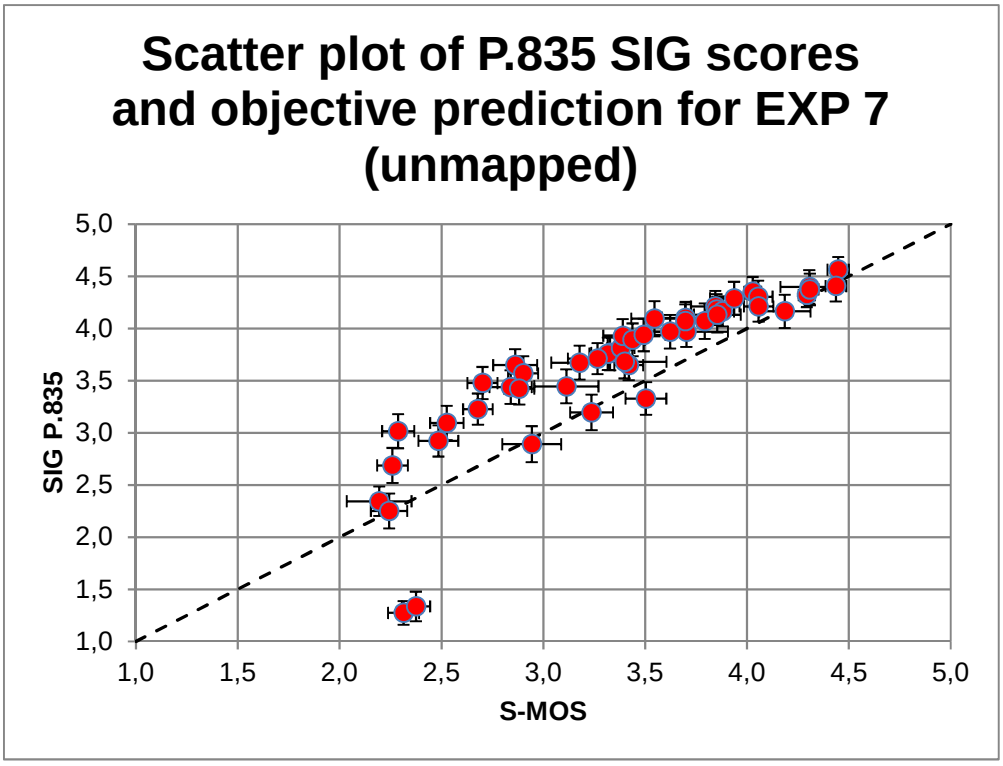


Figure 23: Experiment 7 S-MOS scatter plot

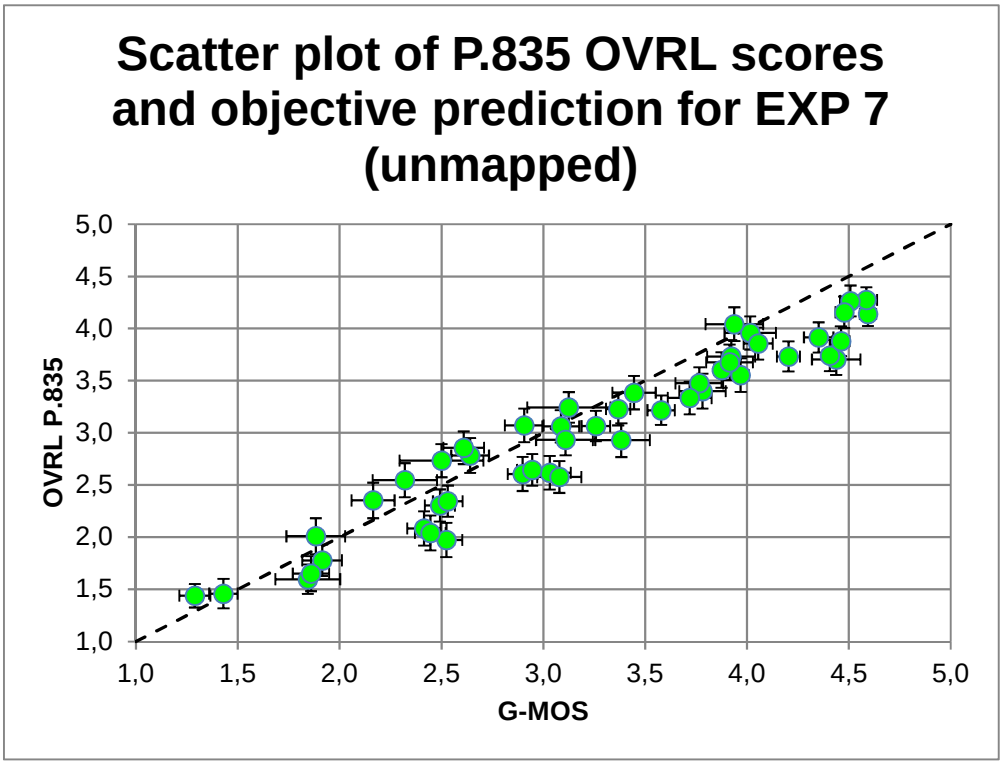


Figure 24: Experiment 7 G-MOS scatter plot

The Pearson correlation coefficient, RMSE and RMSE* performance parameters specified in [i.9] were computed for both validation databases and reported in tables 20 and 21 along with results before and after 1st and 3rd order mapping.

Table 20: Performance of the objective predictor on NB validation database from EXP6

	Condition	S-MOS	N-MOS	G-MOS
	Correlation	0,96	0,95	0,95
RMSE:	no Mapping	0,37	0,32	0,32
	1st Ord. Map.	0,26	0,30	0,28
	3rd Ord. Map	0,19	0,30	0,28
RMSE*:	no Mapping	0,28	0,20	0,22
	1st Ord. Map.	0,17	0,18	0,18
	3rd Ord. Map	0,09	0,17	0,18

Table 21: Performance of the objective predictor on NB validation database from EXP7

	Condition	S-MOS	N-MOS	G-MOS
	Correlation	0,87	0,99	0,97
RMSE:	no Mapping	0,45	0,13	0,36
	1st Ord. Map.	0,36	0,13	0,19
	3rd Ord. Map	0,33	0,12	0,16
RMSE*:	no Mapping	0,33	0,04	0,23
	1st Ord. Map.	0,28	0,04	0,12
	3rd Ord. Map	0,25	0,04	0,07

8.4 Validation data for additional use cases

8.4.1 Tests 1 & 2: Description

Two listening tests according to Recommendation ITU-T P.835 [i.6] were conducted, one in American English, and one in Mandarin. Apart from the source speech, the processing was identical for both tests, based on a simulated handset.

The noise types included four from ETSI ES 202 396-1 [i.2], Car, Pub, Road, and Single-Voice, and one new type, Music.

Noise types

The following noise samples from ETSI ES 202 396-1 [i.2] were used:

Table 22: Noise types for Tests 1& 2 taken from ETSI ES 202 396-1 [i.2]

Description	File name	Duration	Type
Recording in pub	Pub_Noise_binaural_V2	30 s	Binaural
Recording at pavement	Outside_Traffic_Road_binaural	30 s	Binaural
Recording at the drivers position	Fullsize_Car1_130Kmh_binaural	30 s	Binaural
Single voice, male and female	na (see below)	na	single source

The voice distractor consisting of full-band, alternating male and female talkers, anechoically recorded, was produced by a single equalized artificial mouth positioned directly in front of the HATS in the setup of ETSI ES 202 396-1 [i.2], clause 6.5. American English sentences were used for the American English listening panel. Chinese sentences were used for the Chinese listening panel. The distance from HATS MRP to distractor artificial mouth lip ring was 1 m.

In addition, an additional noise type was used:

Description	File name	Duration	Type
Music, with guitar and drums	na	na	binaural

The music distractor contains electric guitar and drums, with short pauses containing near silence, so the dynamic range is quite large A binaural recording, following ETSI ES 202 396-1 [i.2] was made of stereo reproduction in a room compliant to ETSI ES 202 396-1 [i.2].

Source speech

For American English, the 16 FB American English sentences included in the present document were used.

For Mandarin, four sentences from each of two male and two female talkers from the NTT Speech Database for Telephonometry, also included in the speech databases in Recommendation ITU-T P.50 [i.23], were selected for pronunciation and quality by native Mandarin expert listeners, resulting in 16 WB Mandarin sentences.

All speech was reproduced at a nominal level of -4,7 dBPa at MRP through an appropriately equalized HATS in ETSI ES 202 396-1 [i.2] set up.

Simulation processing

The processing was a single simulated handset using two microphones. An acoustic mock-up containing two microphones was built. Recordings were made by placing the mock-up on the HATS in ETSI ES 202 396-1 [i.2] setup.

Recordings were made separately for target speech and distracters, allowing for mixing prior to simulations.

Mixtures were produced at SNRs of 0 dB, 6 dB and 12 dB, using Recommendation ITU-T P.56 [i.8] Active Speech Level for the speech signals (target and distractor) and A-weighting for all distracters. Additionally, a clean-speech (no noise) condition was also produced.

Simulation consisted of one- and two-microphone noise suppression for all signals at the SNRs noted above. For the two-microphone simulation, the degree of suppression was set at 3 levels, low, medium, high. Identical processing was used for both English and Mandarin.

P.835 test

Two panels were recruited, one consisting of 32 native speakers of American English, and one consisting of 32 native speakers of Mandarin. Mandarin speakers were screened for reading skills in English to understand the common English-language material describing rating scales, and for training on the task.

For each P.835 test and for each listening condition, 128 votes were obtained as basis for the mean opinion scores. A single sentence sample was used for each rating. The sixteen sentences were counter-balanced across conditions within each P.835 test, so that a given listener heard sentences from two male and two female talkers.

Presentation was monaural, at 79 dBSPL, using diffuse-field equalized closed-back headphones.

In addition to the test conditions listed above, twelve reference conditions using the noise-suppression simulation described in annex D of the present document were used.

A common presentation describing the task and rating scales of the task was given to all listeners on both panels before listening. Each listening conducted a practice block of 16 trials, containing the reference conditions and additional conditions not part of the test, prior to collecting data.

8.4.2 Tests 1 & 2: Results

Below, results for prediction by the present document are shown. Figure 25 shows a scatter plot of the subjective SIG ratings versus the S-MOS predictions. Only the transformation described in clause 5, remapping based on scores for reference conditions, has been applied to the subjective scores in figure 25. Figure 26 shows results after applying a 3rd order mapping to the transformed subjective scores

Figures 25 and 26, the '+' symbols indicate results for the single-voice, and the 'x' symbols indicate results for the music. Filled dots are used for the other noises from ETSI ES 202 396-1 [i.2].

Table 23 provides correlation, RMSE, and RMSE* according to Recommendation ITU-T P.1401 [i.9].

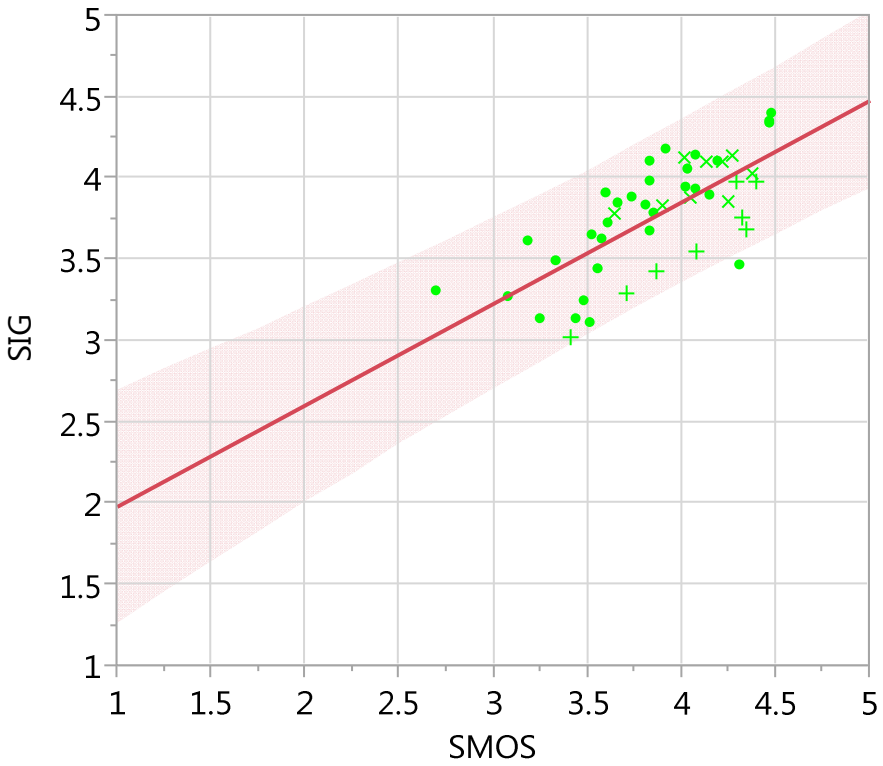


Figure 25: S-MOS fit to SIG, Chinese results, shaded area is 95 % confidence interval

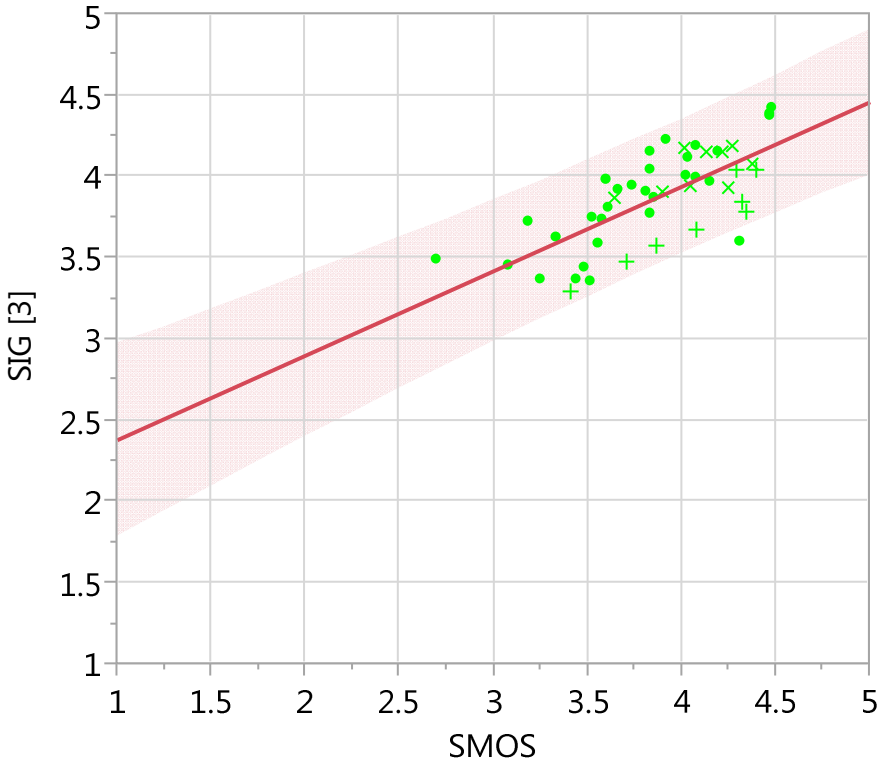


Figure 26: S-MOS fit to SIG after 3rd order mapping, Chinese results, shaded area is 95 % confidence interval

Table 23: Correlation, RMSE, and RMSE* for Chinese SIG

Metric	SIG	SIG [3]
Correlation	0,719	0,721
RMSE	0,301	0,278
RMSE*	0,294	0,273

NOTE 1: If the reference transformation described in clause 5, equations (1), (2), and (3) is not applied to the subjective results, then the resulting RMSE* values after 3rd order mapping is 0,183.

Figures 27 and 28 show results for English SIG, after reference transformation, and then 3rd order remapping, respectively. Table 24 provides correlation, RMSE, and RMSE* according to Recommendation ITU-T P.1401 [i.9].

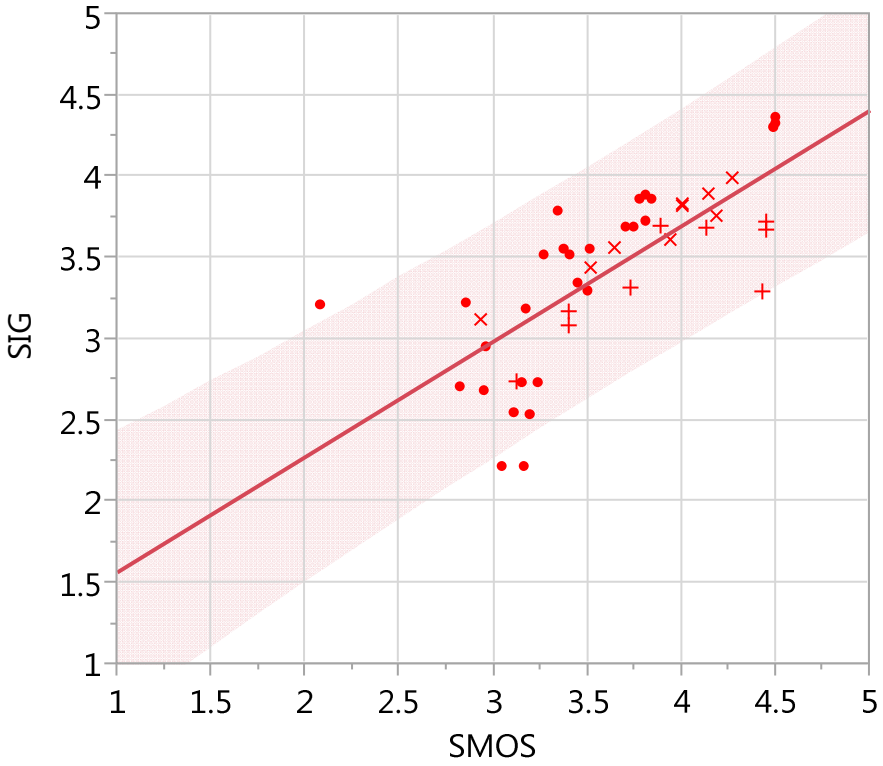


Figure 27: S-MOS fit to SIG, English results, shaded area is 95 % confidence interval

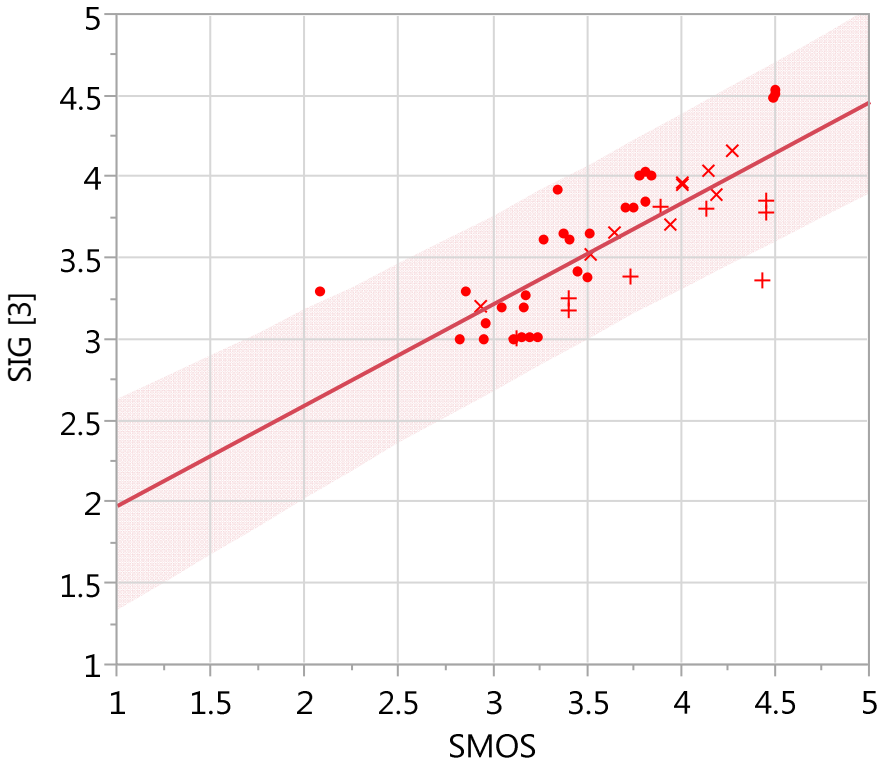


Figure 28: S-MOS fit to SIG after 3rd order mapping, English results, shaded area is 95 % confidence interval

Table 24: Correlation, RMSE, and RMSE* for English SIG

Metric	SIG	SIG [3]
Correlation	0,739	0,788
RMSE	0,418	0,326
RMSE*	0,415	0,316

NOTE 2: If the reference transformation described in clause 5, equations (1), (2), and (3) is not applied to the subjective results, then the resulting RMSE* values after 3rd order mapping is 0,287.

Figures 29 and 30 show results for Chinese BAK, after reference transformation, and then 3rd order remapping, respectively. Table 25 provides correlation, RMSE, and RMSE* according to Recommendation ITU-T P.1401 [i.9].

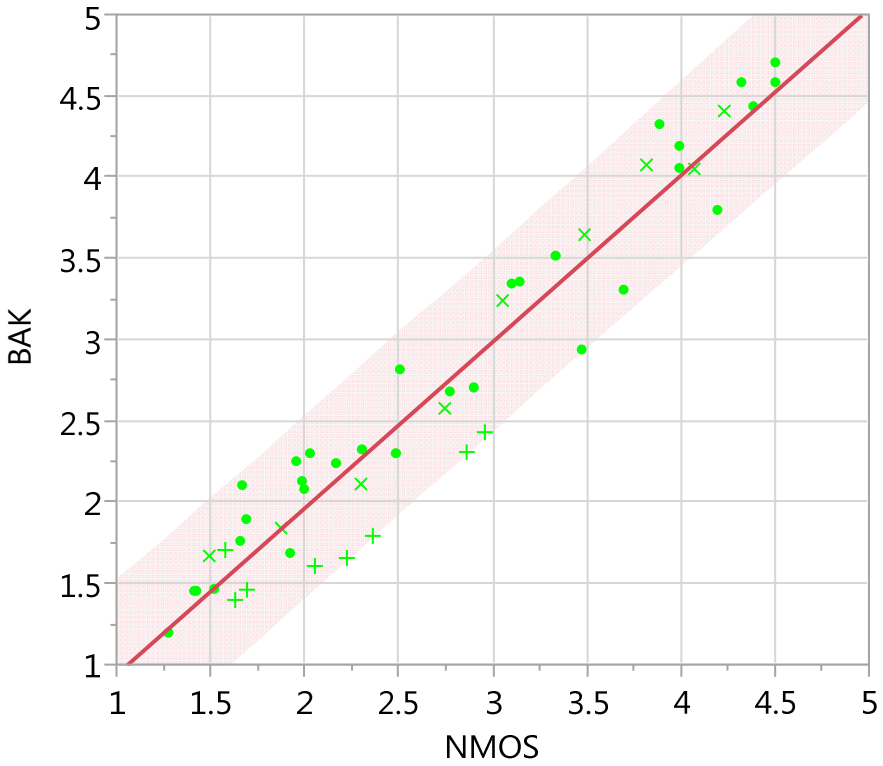


Figure 29: N-MOS fit to BAK, Chinese results, shaded area is 95 % confidence interval

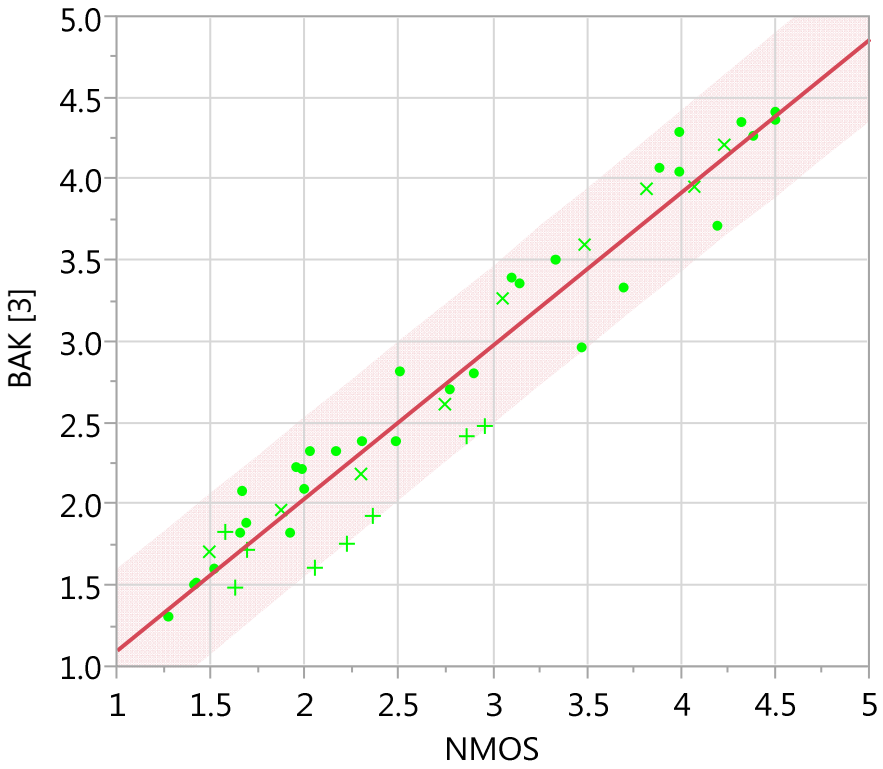


Figure 30: N-MOS fit to BAK after 3rd order mapping, Chinese results, shaded area is 95 % confidence interval

Table 25: Correlation, RMSE, and RMSE* for Chinese BAK

Metric	BAK	BAK [3]
Correlation	0,966	0,971
RMSE	0,271	0,243
RMSE*	0,268	0,236

NOTE 3: If the reference transformation described in clause 5, equations (1), (2), and (3) is not applied to the subjective results, then the resulting RMSE* values after 3rd order mapping is 0,215.

Figures 31 and 32 show results for English BAK, after reference transformation, and then 3rd order remapping, respectively. Table 26 provides correlation, RMSE, and RMSE* according to Recommendation ITU-T P.1401 [i.9].

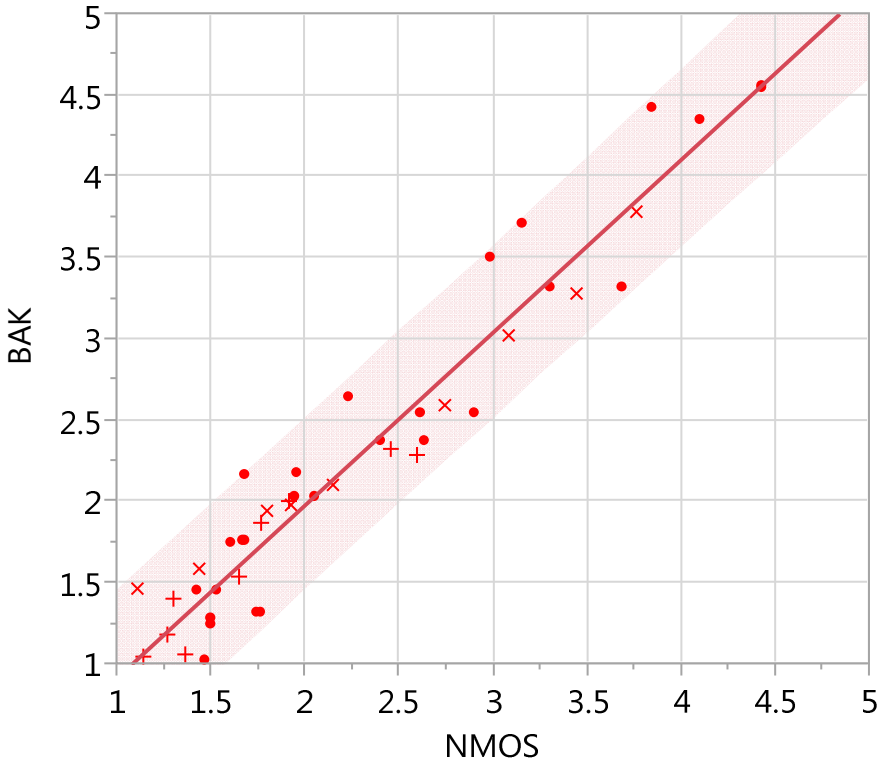


Figure 31: N-MOS fit to BAK, English results, shaded area is 95 % confidence interval

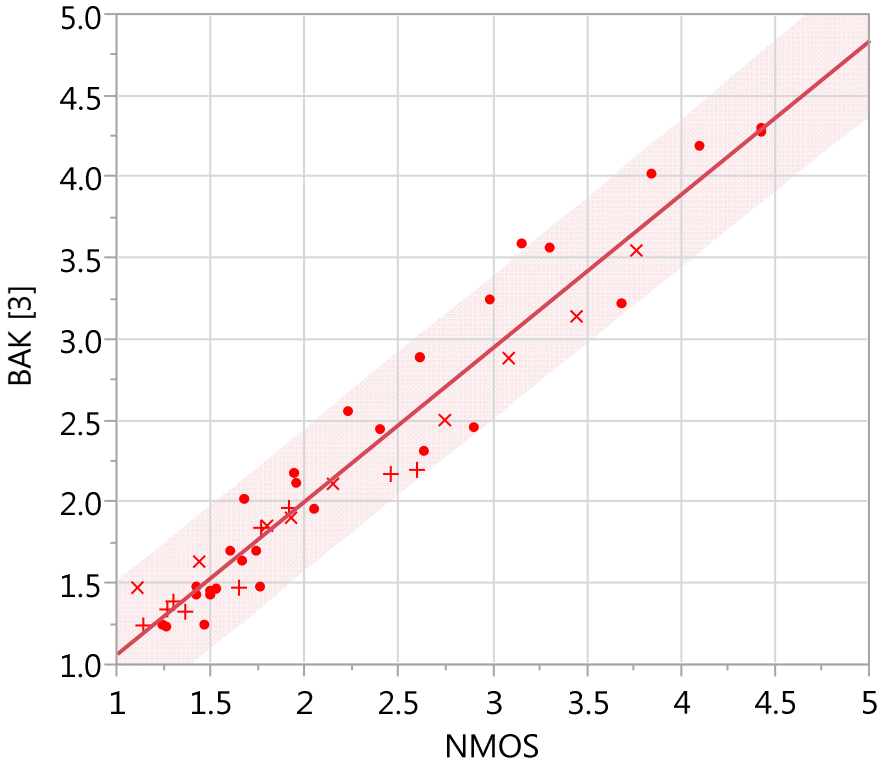


Figure 32: N-MOS fit to BAK after 3rd order mapping, English results, shaded area is 95 % confidence interval

Table 26: Correlation, RMSE, and RMSE* for English BAK

Metric	BAK	BAK [3]
Correlation	0,967	0,971
RMSE	0,261	0,217
RMSE*	0,253	0,212

NOTE 4: If the reference transformation described in clause 5, equations (1), (2), and (3) is not applied to the subjective results, then the resulting RMSE* values after 3rd order mapping is 0,239.

Figures 33 and 34 show results for Chinese OVRL, after reference transformation, and then 3rd order remapping, respectively. Table 27 provides correlation, RMSE, and RMSE* according to Recommendation ITU-T P.1401 [i.9].

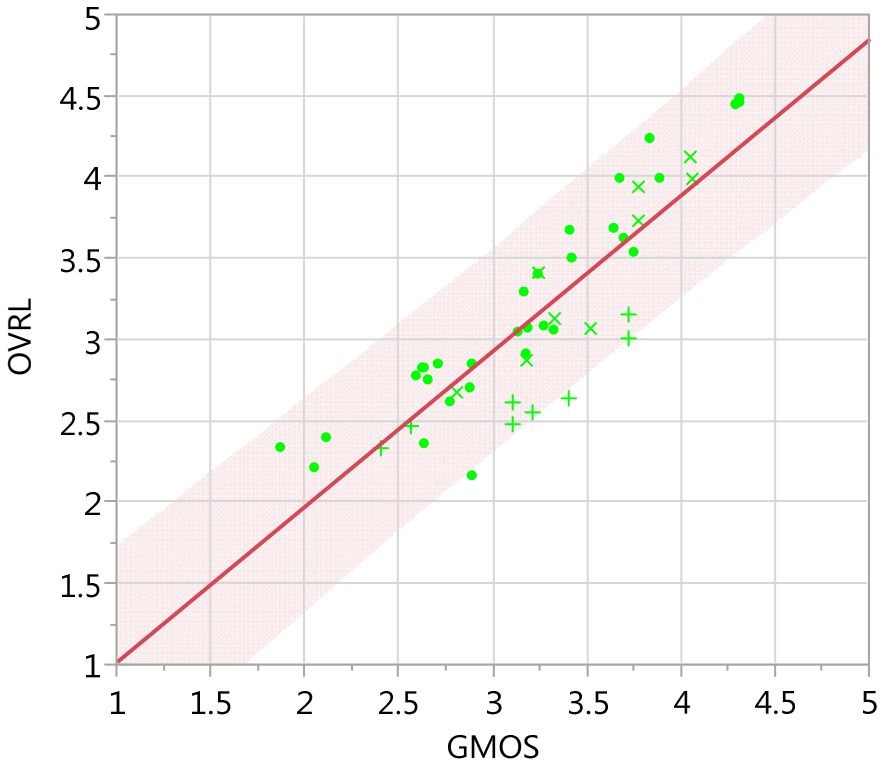


Figure 33: G-MOS fit to OVRL, Chinese results, shaded area is 95 % confidence interval

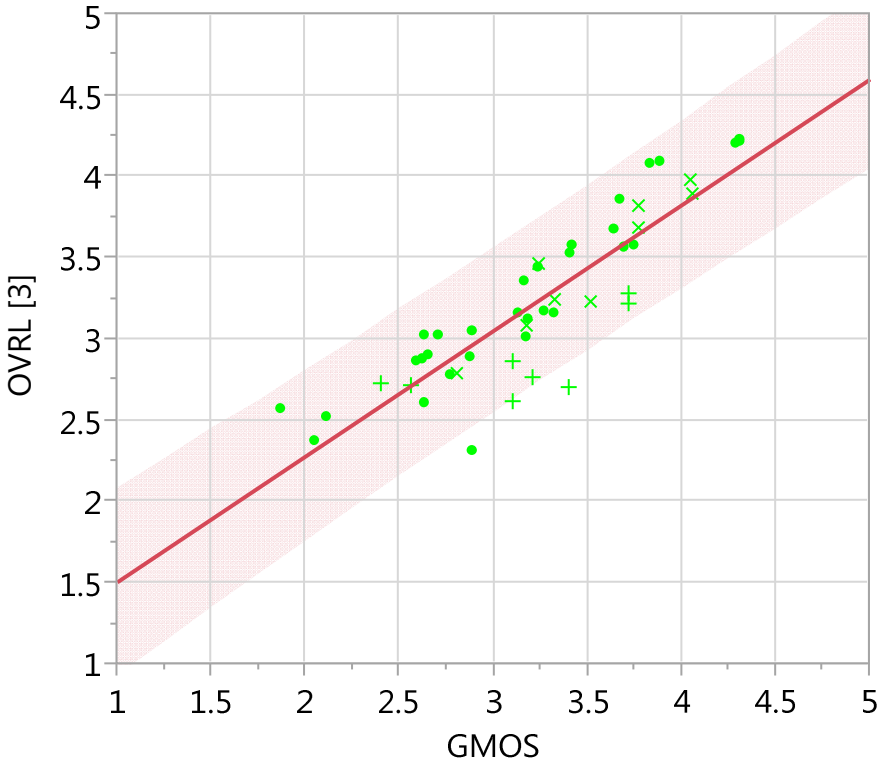


Figure 34: G-MOS fit to OVRL BAK after 3rd order mapping, Chinese results, shaded area is 95 % confidence interval

Table 27: Correlation, RMSE, and RMSE* for Chinese OVRL

Metric	OVRL	OVRL [3]
Correlation	0,878	0,879
RMSE	0,313	0,277
RMSE*	0,309	0,274

NOTE 5: If the reference transformation described in clause 5, equations (1), (2), and (3) is not applied to the subjective results, then the resulting RMSE* values after 3rd order mapping is 0,217.

Figures 35 and 36 show results for English OVRL, after reference transformation, and then 3rd order remapping, respectively. Table 28 provides correlation, RMSE, and RMSE* according to Recommendation ITU-T P.1401 [i.9].

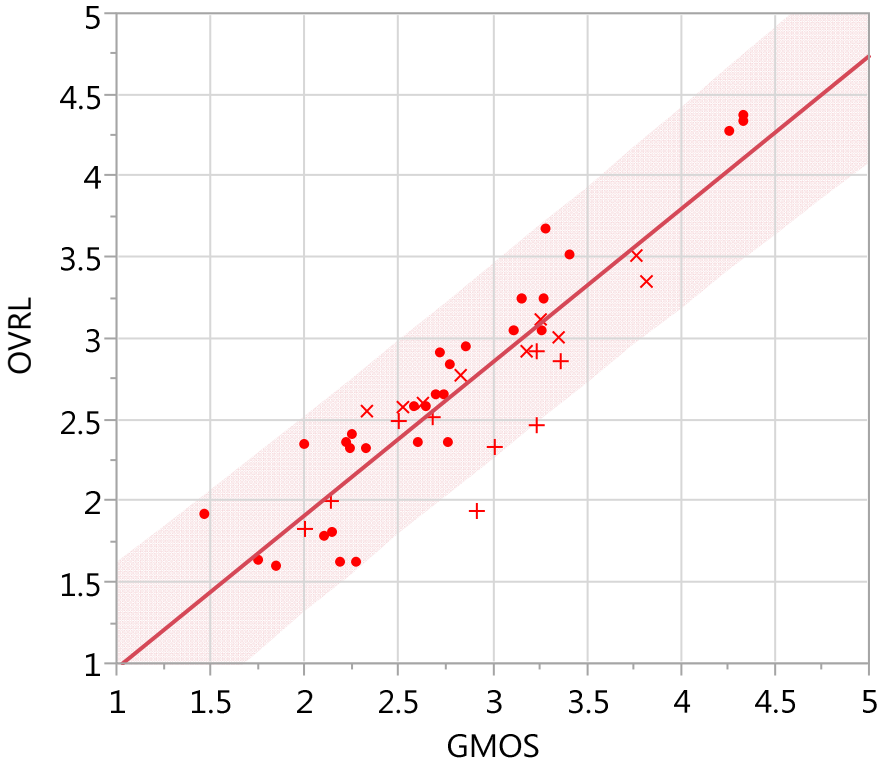


Figure 35: G-MOS fit to OVRL, English results, shaded area is 95 % confidence interval

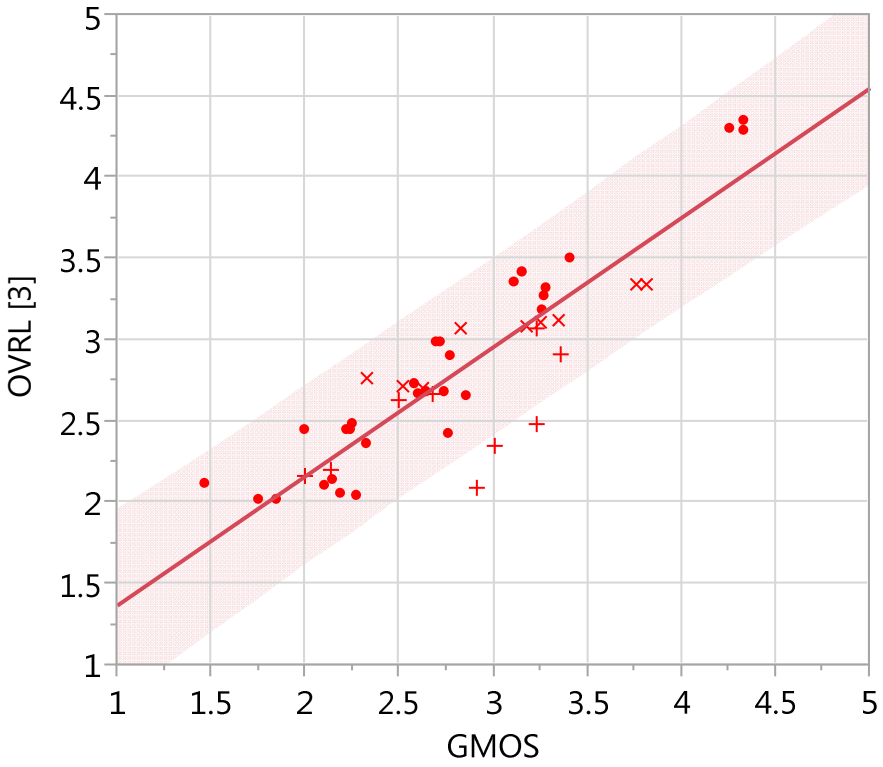


Figure 36: G-MOS fit to OVRL BAK after 3rd order mapping, English results, shaded area is 95 % confidence interval

Table 28: Correlation, RMSE, and RMSE* for English OVRL

Metric	OVRL	OVRL [3]
Correlation	0,905	0,892
RMSE	0,310	0,293
RMSE*	0,306	0,289

NOTE 6: If the reference transformation described in clause 5, equations (1), (2), and (3) is not applied to the subjective results, then the resulting RMSE* values after 3rd order mapping is 0,221.

In general, the results for Chinese and English are quite similar.

8.4.3 Tests 3, 4, 5 & 6: Description

Four additional listening tests according to Recommendation ITU-T P.835 [i.6] were conducted, two in Narrowband and two in Wideband, all in American English, using nine commercially available handsets and one development handset. The noise types includes two from ETSI ES 202 396-1 [i.2] (Road, Call Center), with a hybrid condition consisting of Call Center with a single talker added. The material and conditions for the single talker were as in Test 1. A fourth test condition was no additional noise ("clean").

Noise types

The following noise samples from ETSI ES 202 396-1 [i.2] were used:

Description	File name	Duration	Level	Type
Recording at pavement	Outside_Traffic_Road_binaural	30 s	L: 74,9 dB(A) R: 73,9 dB(A)	Binaural
Recording in a business office	Work_Noise_Office_callcenter_binaural	30 s	L: 56,6 dB(A) R: 57,8 dB(A)	Binaural

A third noise condition was created by combining the Callcenter with the following:

Description	File name	Duration	Level	Type
Single voice, male and female	na	na	-4,7 dB Pa at MRP	single source

The voice distractor consisting of full-band, alternating male and female talkers, anechoically recorded, was produced by a single equalized artificial mouth positioned directly in front of the HATS in the setup of ETSI ES 202 396-1 [i.2], clause 6.5. American English sentences were used for the American English listening panel. The distance from HATS MRP to distractor artificial mouth lip ring was 1 m, and the level at the MRP of the second HATS was -4,7 dBPa, active speech. This was added to the binaural reproduction of the Call Center noise ("Call center + Voice").

A fourth condition was no background noise ("Clean" conditions).

Source speech

The 16 FB American English sentences included in the present document were used, and were reproduced at a nominal level of -4,7 dBPa at MRP through an appropriately equalized HATS in ETSI ES 202 396-1 [i.2] set up.

Device setup

All devices were mounted on the test HATS in ETSI ES 202 396-1 [i.2] set up. For the Road and Clean conditions, three positions of the handsets were tested: standard test position according to Recommendation ITU-T P.64 [i.24] ("Nominal" position); a second position with elevation increased by 40º and azimuth increased by 10º ("Up" position); a third position with elevation decreased by 25º and azimuth increased by 10º ("Down" position). These alternative positions were selected based on information indicating that these were representative of relatively common user behaviour.

Calls were placed via a radiated connection to a base station simulator. For Narrowband, AMR-NB 12,2 kbps was used, and for Wideband, AMR-WB 12,65 kbps was used.

P.835 test

Four panels were recruited, each consisting of 32 native speakers of American English.

For each P.835 test and for each listening condition, 128 votes were obtained as basis for the mean opinion scores. A single sentence sample was used for each rating. The sixteen sentences were counter-balanced across conditions within each P.835 test, so that a given listener heard sentences from two male and two female talkers.

For Narrowband, presentation was monaural, at 79 dBSPL, while for Wideband, presentation was diotic, at 73 dBSPL, in both cases using diffuse-field equalized closed-back headphones.

Due to the relatively large number of conditions, devices were assigned to the panels according to the following plan:

Table 29: Assignment of devices to panels/tests

Device	Test 3 NB	Test 4 NB	Test 5 WB	Test 6 WB
A	X	X	X	X
B	X	X	X	X
C	X		X	
D	X		X	
E	X		X	
F	X		X	
G		X		X
H		X		X
I		X		X
J		X		X

The two devices in common allow for analysis of consistency between panels, to support the possibility of combining data across each of the two panels listening to common bandwidth.

In addition to the test conditions listed above, twelve reference conditions using the noise-suppression simulation described in annex D of the present document were used. Based on preliminary listening, the parameters of the noise-suppression system were selected to be appropriate for Wideband testing.

A common presentation describing the task and rating scales of the task was given to all listeners on all panels before listening. Each listening conducted a practice block of 16 trials, containing the reference conditions and additional conditions not part of the test, prior to collecting data.

8.4.4 Tests 3 & 4: Results, Narrowband

Figure 37 shows a scatter plot of BAK versus SIG for the ratings from Tests 3 & 4. For the two devices in common, the numeral 1 (e.g. "A 1") is used to indicate results from Test 3 and numeral 2 (e.g. "B 2") is used to indicate results from Test 4. The '+' symbol represents reference conditions for Test 3 and 'x' represents reference conditions for Test 4.

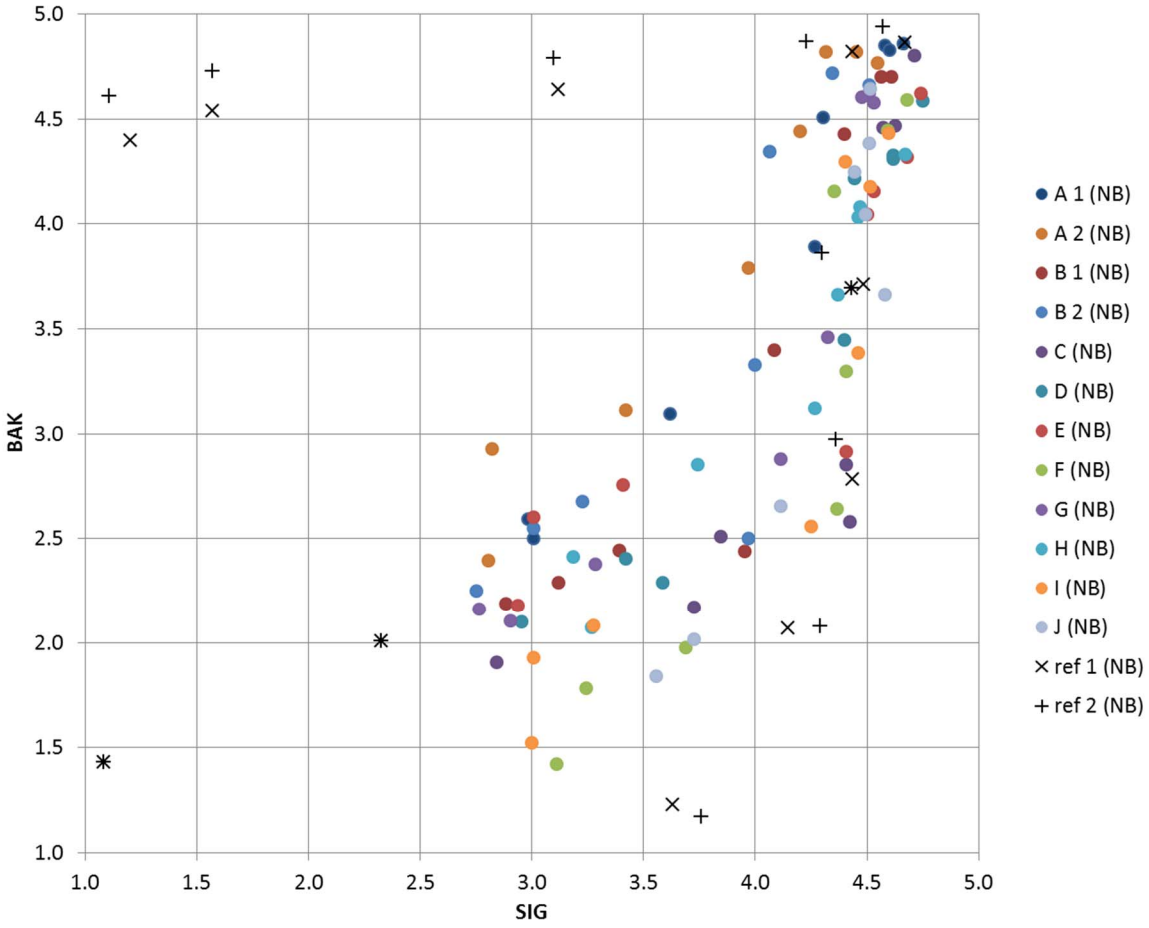


Figure 37: BAK vs SIG results, Tests 3 and 4

In Figure 37, it can be seen that the scores for the reference conditions span the space, and generally bound the scores from the test conditions. The range of the BAK scores for test conditions spans nearly the same range as that of the BAK scores for reference conditions. As was seen in the validation results for the present document, the range of the SIG scores is limited compared to that of the SIG scores for the reference conditions, but still spans the range from 4,7 to 2,7.

Figure 38 contains a scatter plot comparing the SIG and BAK scores for the reference conditions between Tests 3 and 4.

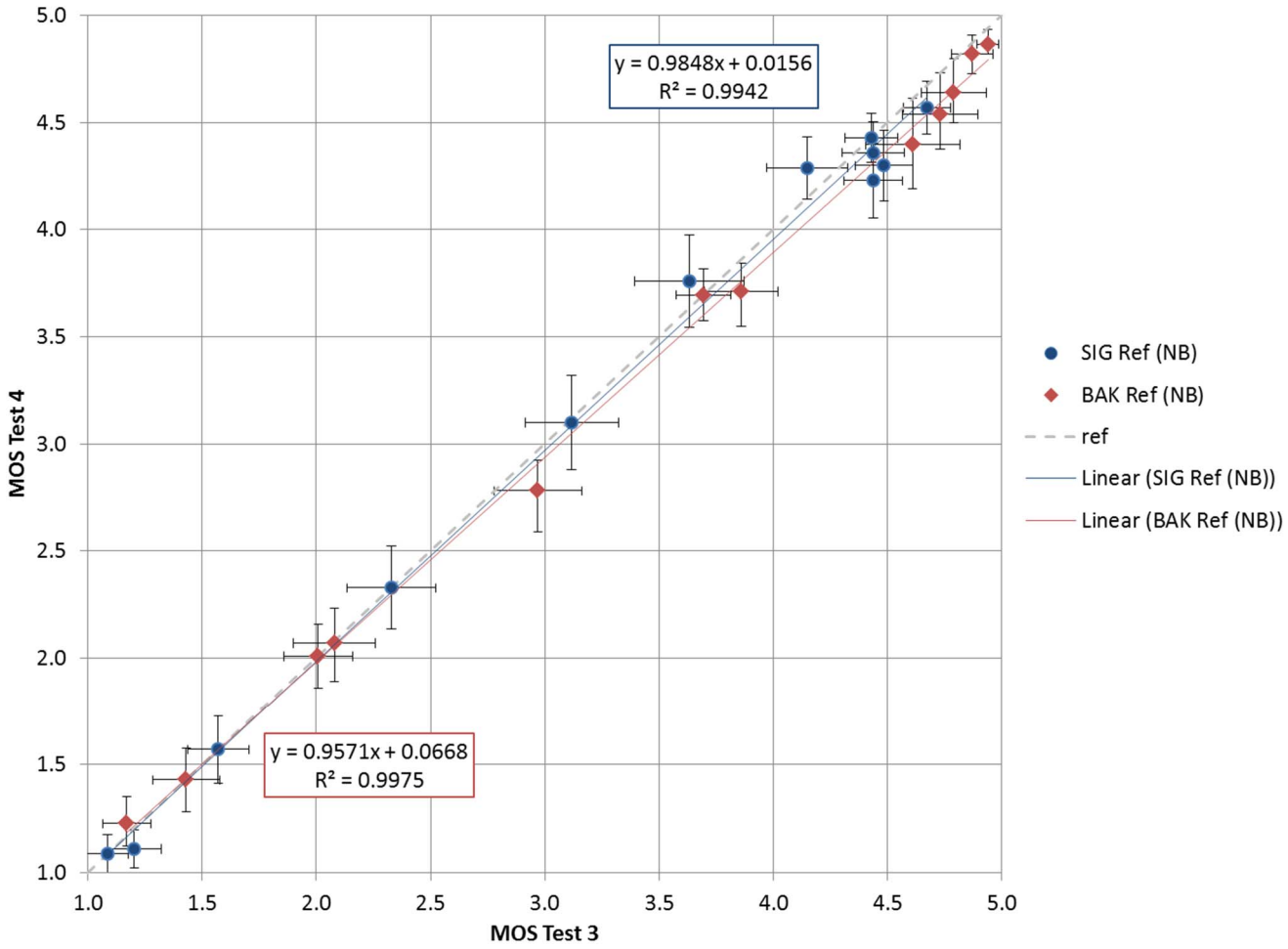


Figure 38: Scatterplot of SIG and BAK scores for reference conditions, Tests 3 & 4

The error bars show the 95 % confidence intervals of the scores. The scores lie very close to the positive diagonal and have very high correlation (> 0,997), indicating very high consistency between panels on the ratings of the reference conditions.

Figure 39 shows a scatter plot comparing SIG and BAK scores between Tests 3 and 4 for devices A and B, the two devices that were common to both tests. As in figure 38, the results lie close to the positive diagonal, and also have high correlation (> 0,988), again indicating high consistency between panels in Test 3 and Test 4 for the ratings of the test conditions for the two devices that are in common to both tests.

Based on these consistency checks, for subsequent evaluation of test conditions, results from Tests 3 and Test 4 will be combined, for a total of 10 unique devices.

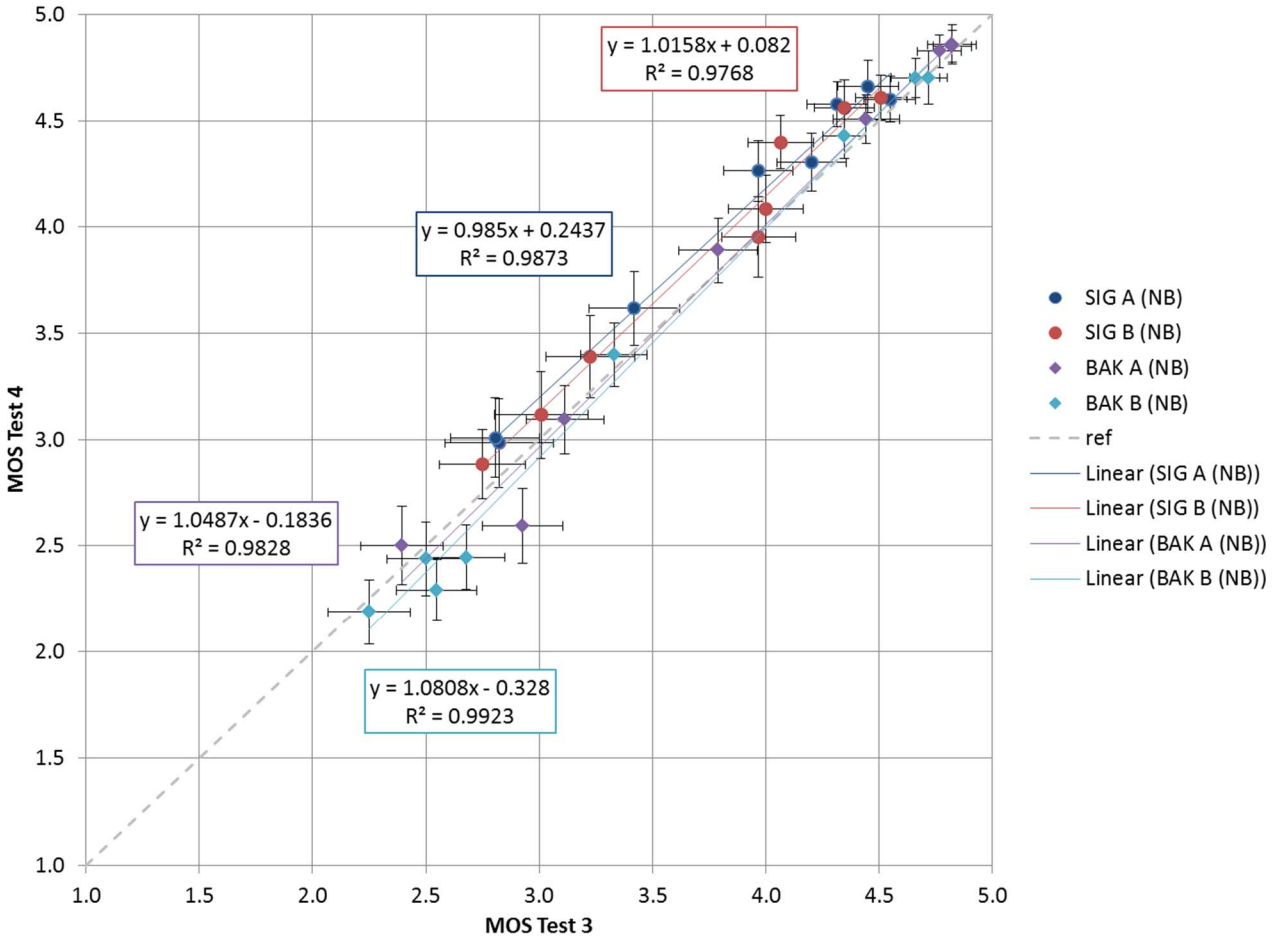


Figure 39: Scatterplot of SIG and BAK scores for devices A and B, Tests 3 & 4

Figure 40 shows a scatter plot of the subjective SIG ratings versus the S-MOS predictions. Figure 41 shows results after the transformation described in clause 5, remapping based on scores for reference conditions, has been applied to the subjective scores in figure 40.

In figures 40 and 41 the Green symbols indicate results for the Road, the Blue symbols for Callcenter, the Red symbols for Callcenter+Voice, and the Black symbols for the Clean (no noise) conditions.

Table 30 provides correlation, RMSE, and RMSE* according to Recommendation ITU-T P.1401 [i.9].

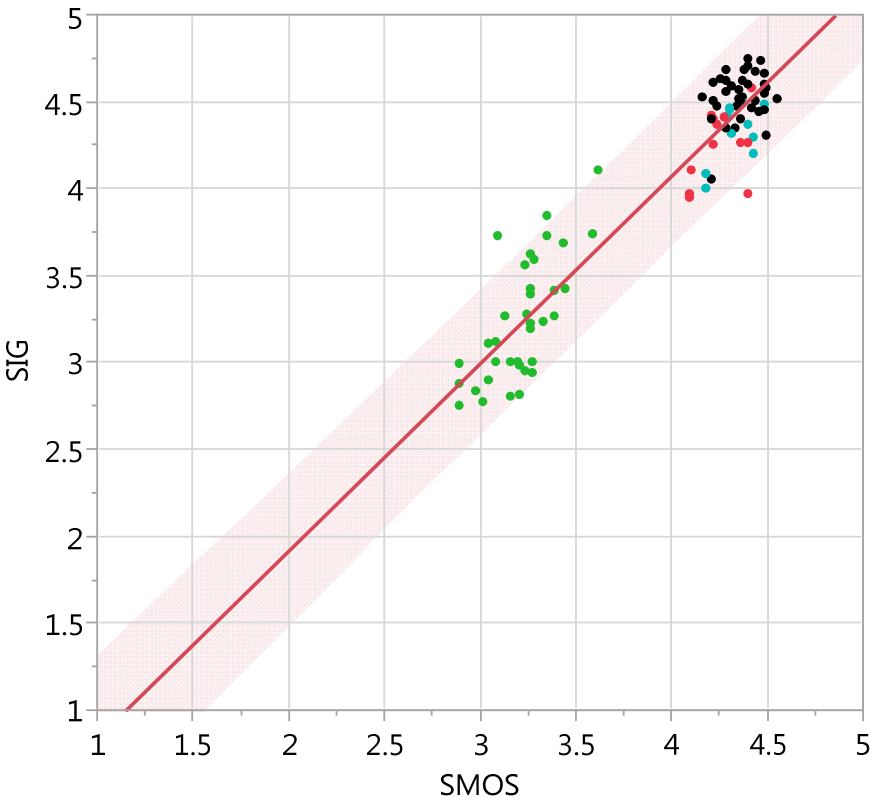


Figure 40: S-MOS fit to SIG, Tests 3 & 4, shaded area is 95 % confidence interval

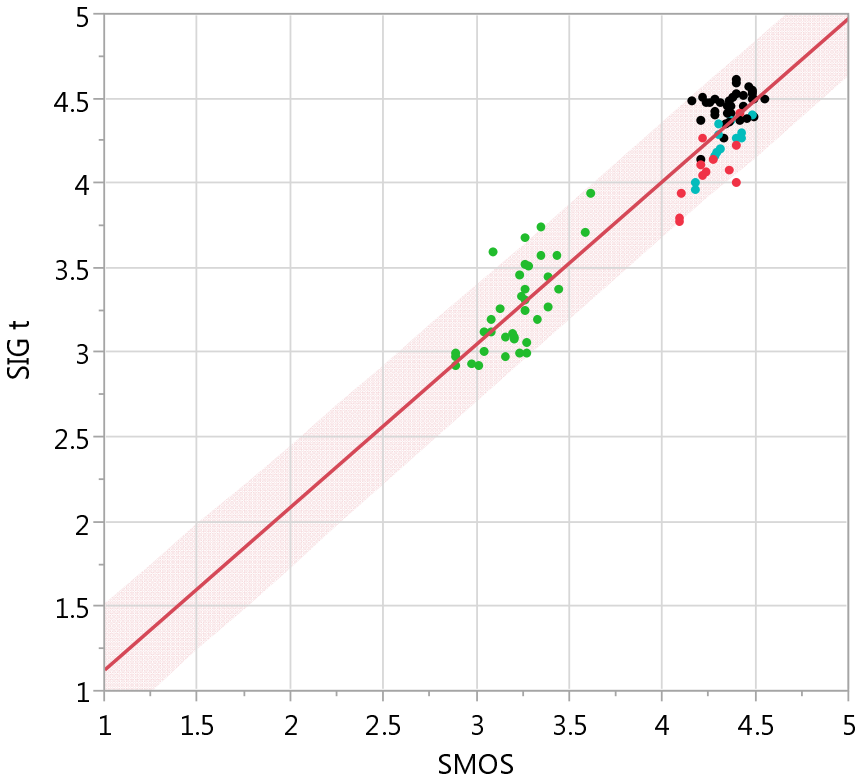


Figure 41: S-MOS fit to SIG-transformed, Tests 3 & 4, shaded area is 95 % confidence interval

Table 30: Correlation, RMSE, and RMSE* for SIG

Metric	SIG	SIG_t
Correlation	0,922	0,954
RMSE	0,223	0,173
RMSE*	0,221	0,156

NOTE 1: If the reference transformation described in clause 5, equations (1), (2), and (3) is not applied to the subjective results, then the resulting RMSE* values after 3rd order mapping is 0,121.

Figures 42 and 43 plot results for N-MOS fit to BAK, and table 31 provides correlation, RMSE, and RMSE*.

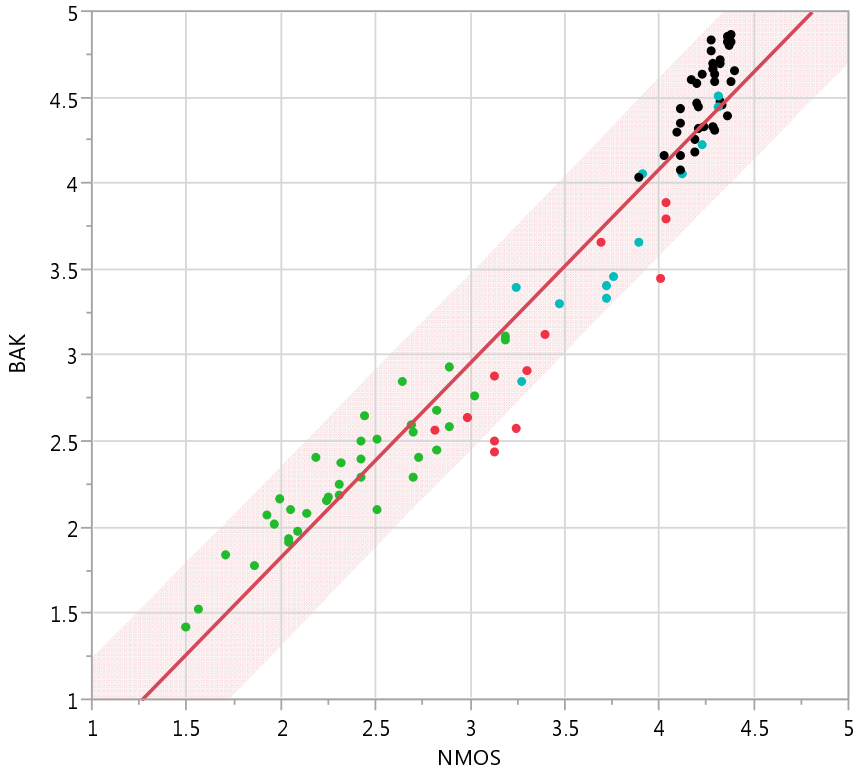


Figure 42: N-MOS fit to BAK, Tests 3 & 4, shaded area is 95 % confidence interval

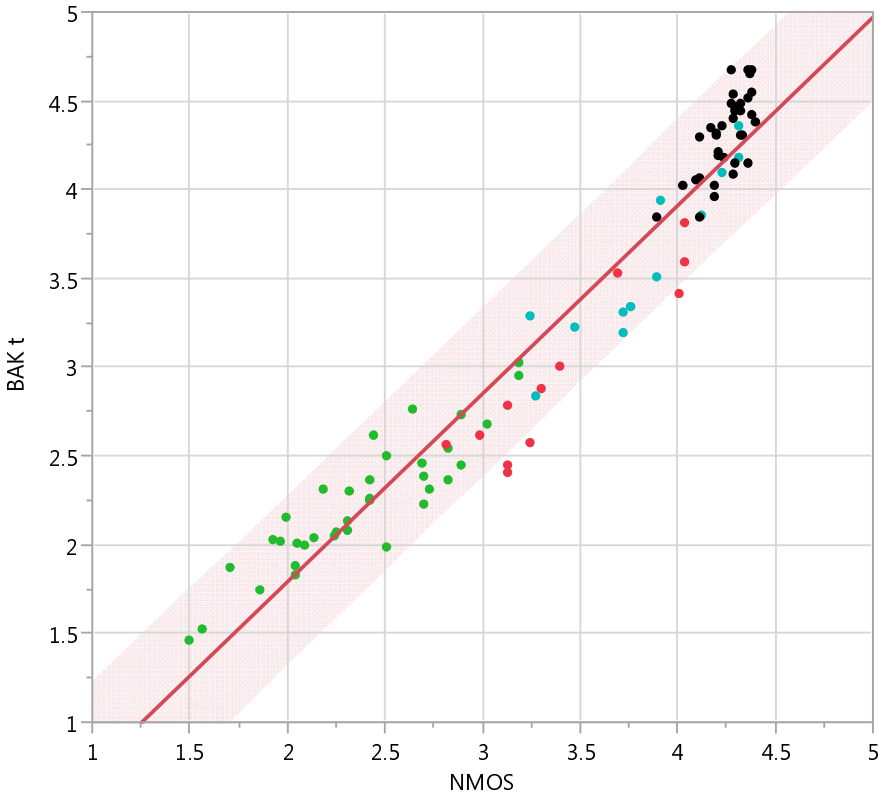


Figure 43: N-MOS fit to BAK-transformed, Tests 3 & 4, shaded area is 95 % confidence interval

Table 31: Correlation, RMSE, and RMSE* for BAK

Metric	BAK	BAK_t
Correlation	0,969	0,970
RMSE	0,280	0,267
RMSE*	0,274	0,262

NOTE 2: If the reference transformation described in clause 5, equations (1), (2), and (3) is not applied to the subjective results, then the resulting RMSE* values after 3rd order mapping is 0,187.

Figures 44 and 45 plot results for G-MOS fit to OVRL, and table 32 provides correlation, RMSE and RMSE*.

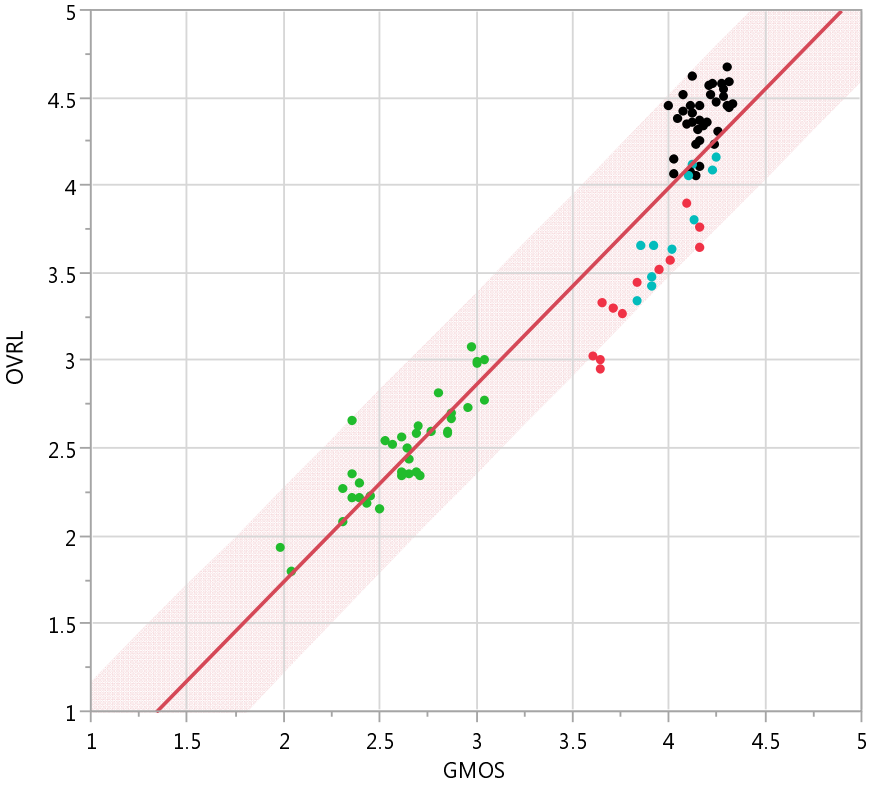


Figure 44: G-MOS fit to OVRL, Tests 3 & 4, shaded area is 95 % confidence interval

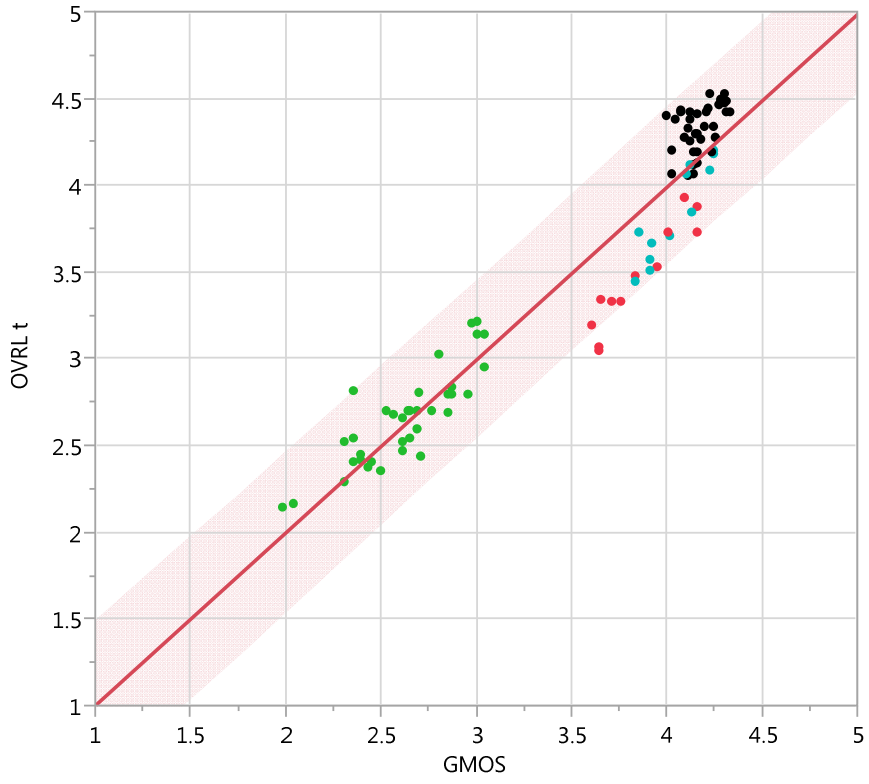


Figure 45: G-MOS fit to OVRL-transformed, Tests 3 & 4, shaded area is 95 % confidence interval

Table 32: Correlation, RMSE, and RMSE* for OVRL

Metric	OVRL	OVRL_t
Correlation	0,955	0,956
RMSE	0,127	0,225
RMSE*	0,108	0,225

NOTE 3: If the reference transformation described in clause 5, equations (1), (2), and (3) is not applied to the subjective results, then the resulting RMSE* values after 3rd order mapping is 0,179.

8.4.5 Tests 5 & 6: Results, Wideband

Figure 46 shows a scatter plot of BAK versus SIG for the ratings from Tests 5 & 6. For the two devices in common, the numeral 1 (e.g. "A 1") is used to indicate results from Test 5 and numeral 2 (e.g. "B 2") is used to indicate results from Test 6. The '+' symbol represents reference conditions for Test 5 and 'x' represents reference conditions for Test 6.

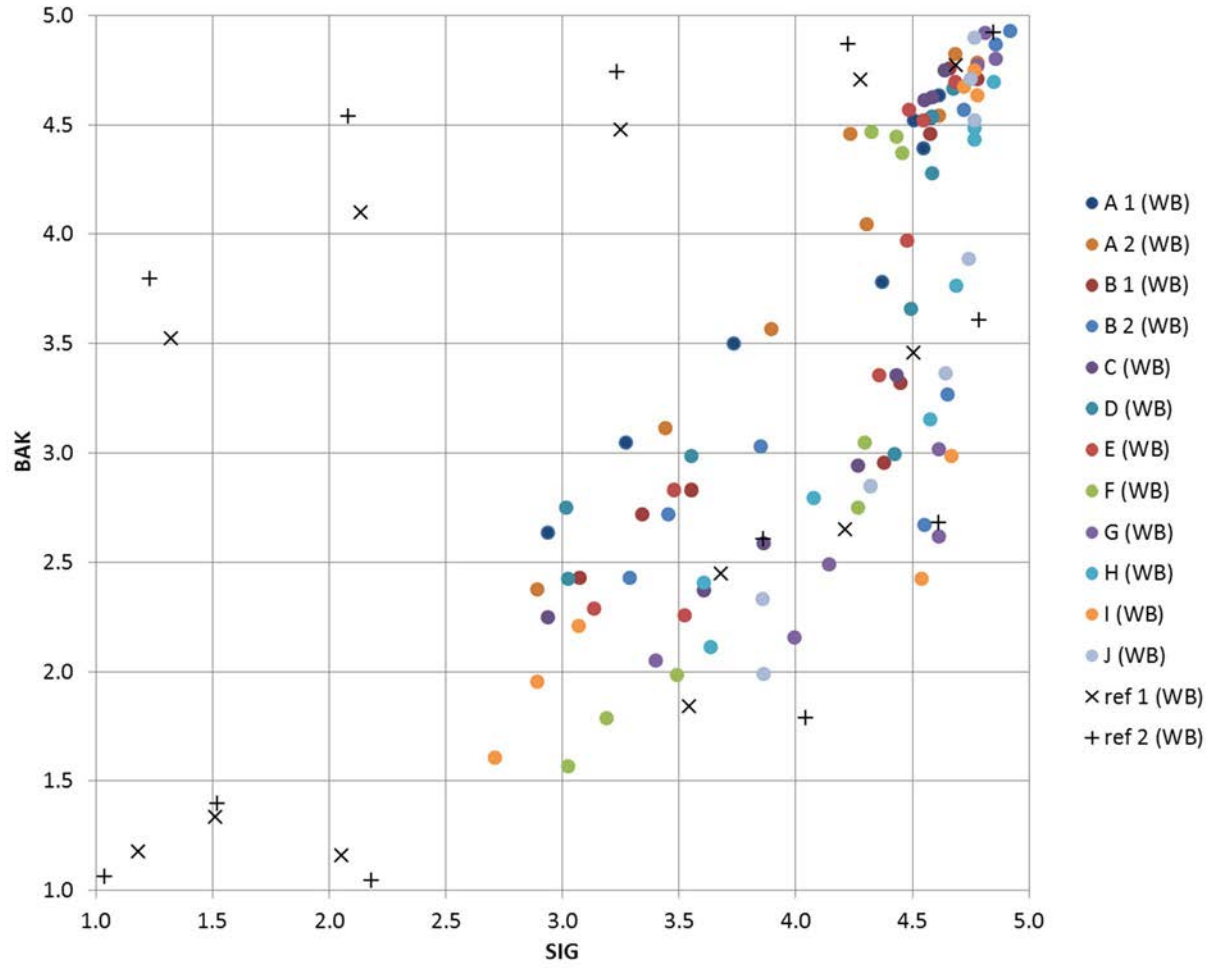


Figure 46: BAK vs SIG results, Tests 5 and 6

In Figure 46, it can be seen that the scores for the reference conditions span the space, and generally bound the scores from the test conditions. The range of the BAK scores for test conditions spans nearly the same range as that of the BAK scores for reference conditions. As was seen in the validation results for the present document, the range of the SIG scores is limited compared to that of the SIG scores for the reference conditions, but still spans the range from 4,7 to 2,7. In contrast to results for Tests 3 and 4 shown in figure 37, there seems to be larger separation for some of the reference scores between the two tests, in particular the noise-suppression distortion references, at the upper portion of the plot.

Figure 47 contains a scatter plot comparing the SIG and BAK scores for the reference conditions between Tests 5 and 6. The separation in scores for reference conditions noted in figure 46 appears in figure 47 as a slight offset at the higher values of SIG and BAK. However, the scores are still quite similar between the two tests.

The error bars show the 95 % confidence intervals of the scores. The scores lie very close to the positive diagonal and have very high correlation (> 0,994), indicating very high consistency between panels on the ratings of the reference conditions.

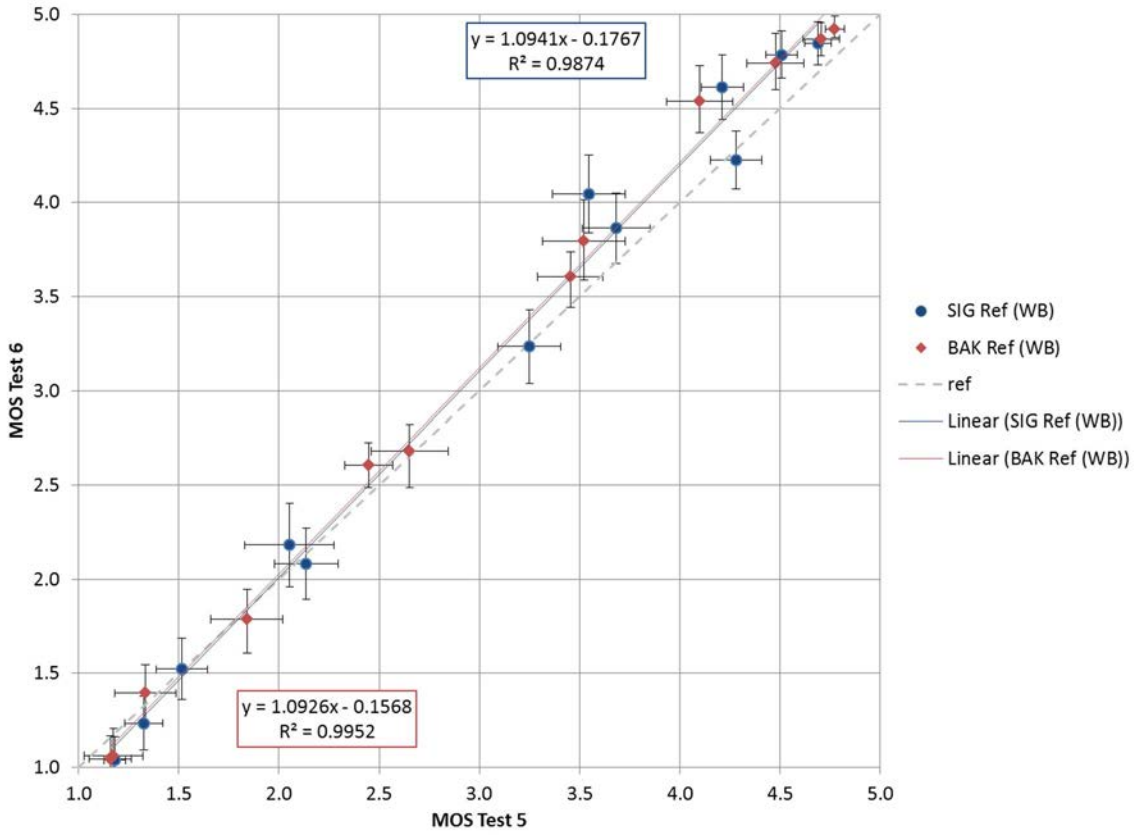


Figure 47: Scatterplot of SIG and BAK scores for reference conditions, Tests 5 & 6

Figure 48 shows a scatter plot comparing SIG and BAK scores between Tests 5 and 6 for devices A and B, the two devices that were common to both tests. As in figure 47, the results lie close to the positive diagonal, and also have high correlation (> 0,962), again indicating very good consistency between panels in Test 5 and Test 6 for the ratings of the test conditions for the two devices that are in common to both tests.

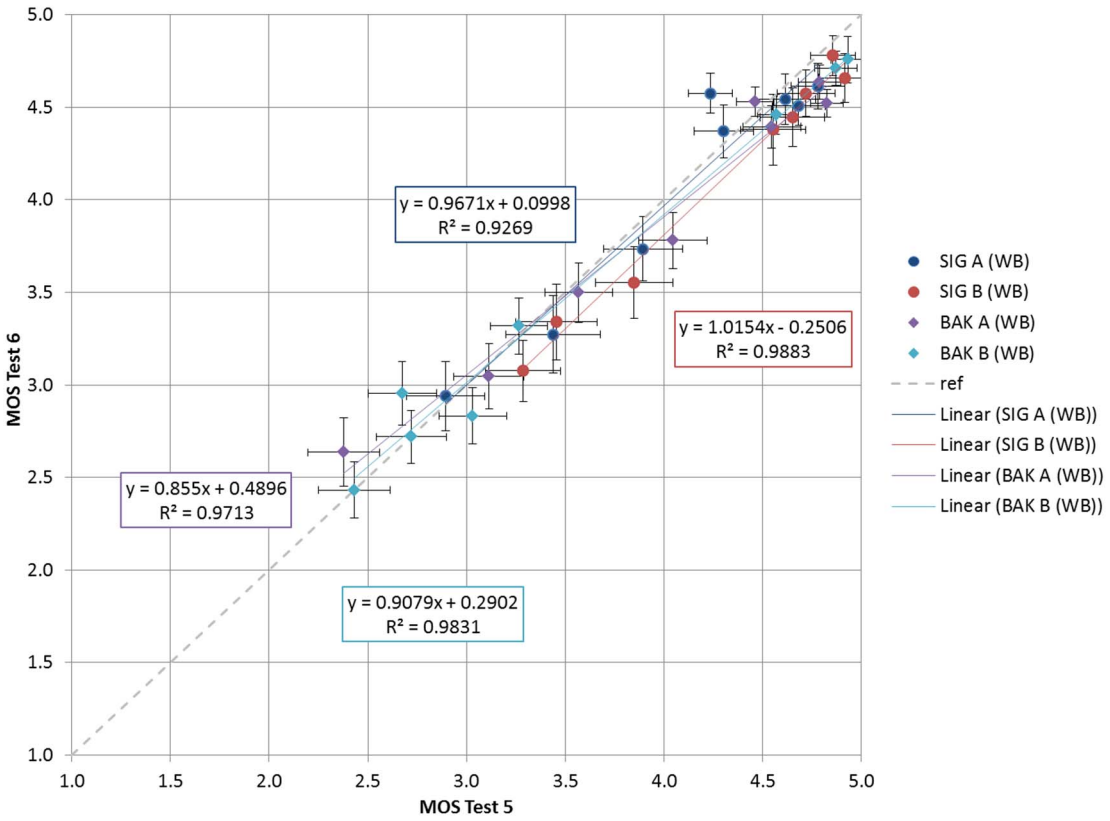


Figure 48: Scatterplot of SIG and BAK scores for devices A and B, Tests 5 & 6

Based on these consistency checks, for subsequent evaluation of test conditions, results from Tests 5 and Test 6 will be combined, for a total of 10 unique devices.

Figure 49 shows a scatter plot of the subjective SIG ratings versus the S-MOS predictions. Figure 50 shows results after the transformation described in clause 5, remapping based on scores for reference conditions, has been applied to the subjective scores in figure 49. Table 33 provides correlation, RMSE and RMSE* according to Recommendation ITU-T P.1401 [i.9].

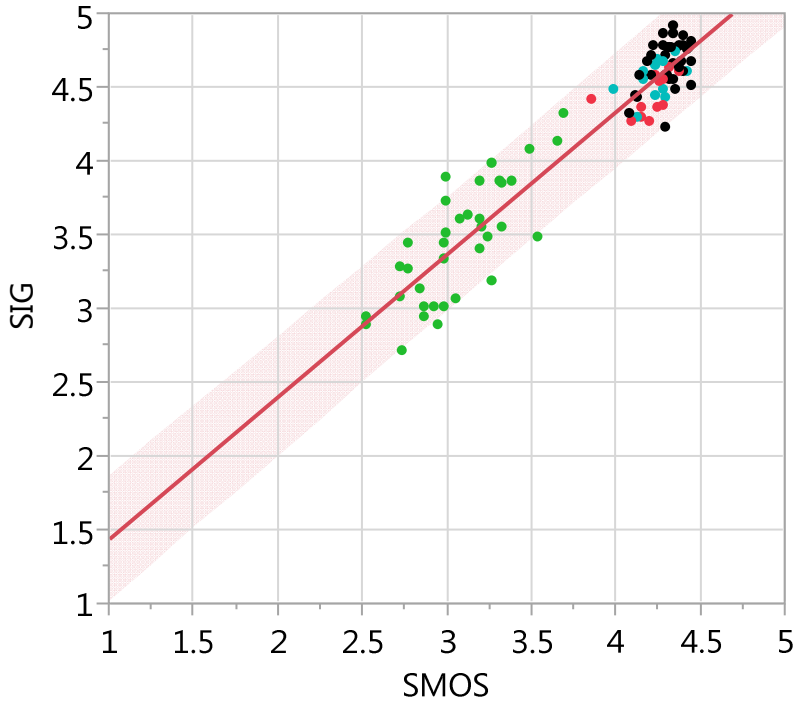


Figure 49: S-MOS fit to SIG, Tests 5 & 6, shaded area is 95 % confidence interval

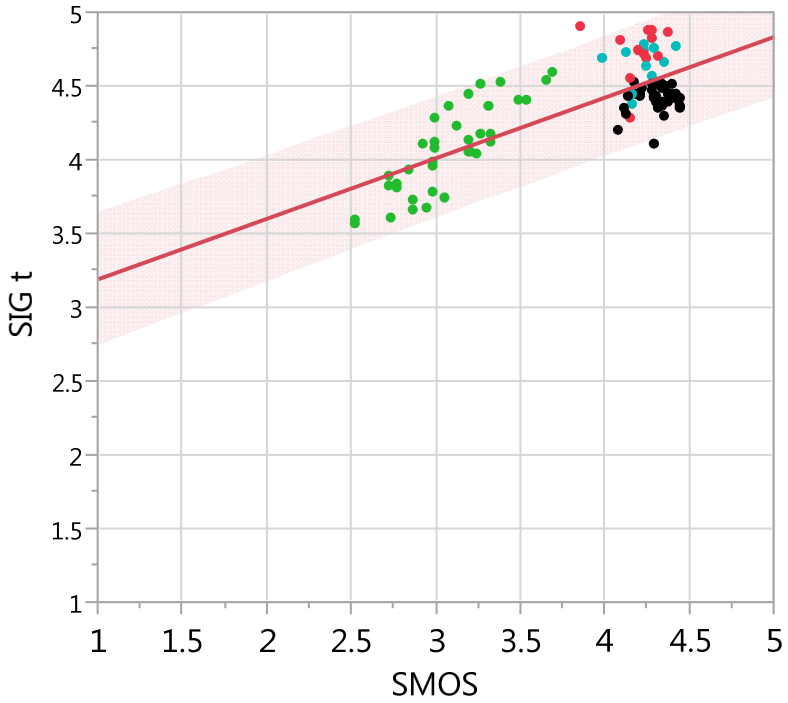


Figure 50: S-MOS fit to SIG-transformed, Tests 5 & 6, shaded area is 95 % confidence interval

Table 33: Correlation, RMSE, and RMSE* for SIG

Metric	SIG	SIG_t
Correlation	0,951	0,779
RMSE	0,395	0,675
RMSE*	0,393	0,674

NOTE 1: If the reference transformation described in clause 5, equations (1), (2), and (3) is not applied to the subjective results, then the resulting RMSE* values after 3rd order mapping is 0,240.

Figures 51 and 52 plot results for N-MOS fit to BAK, and table 34 provides correlation, RMSE and RMSE*.

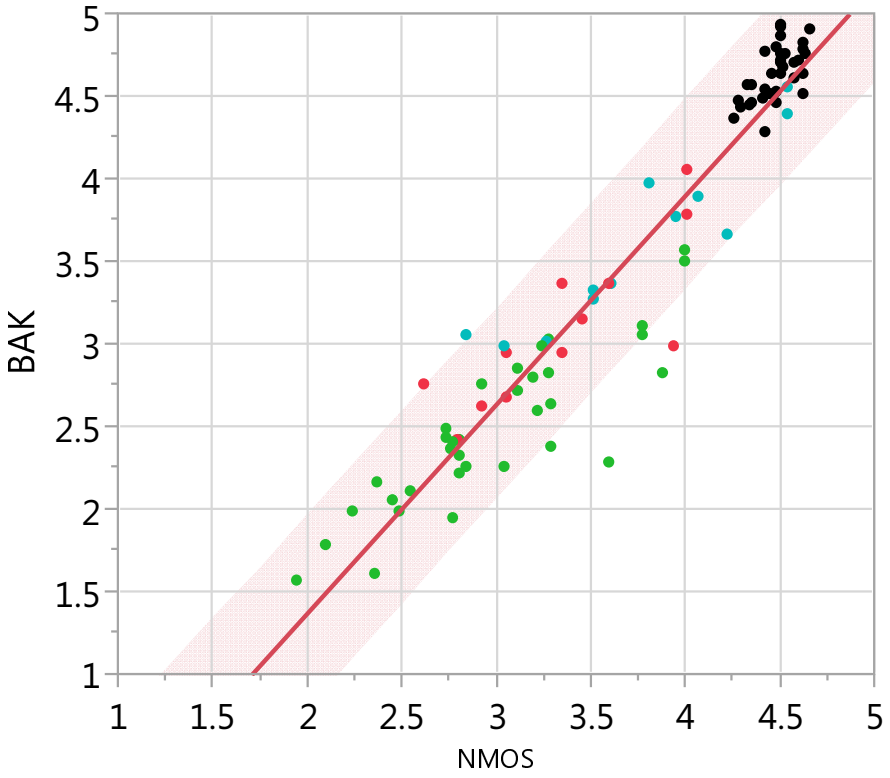


Figure 51: N-MOS fit to BAK, Tests 5 & 6, shaded area is 95 % confidence interval

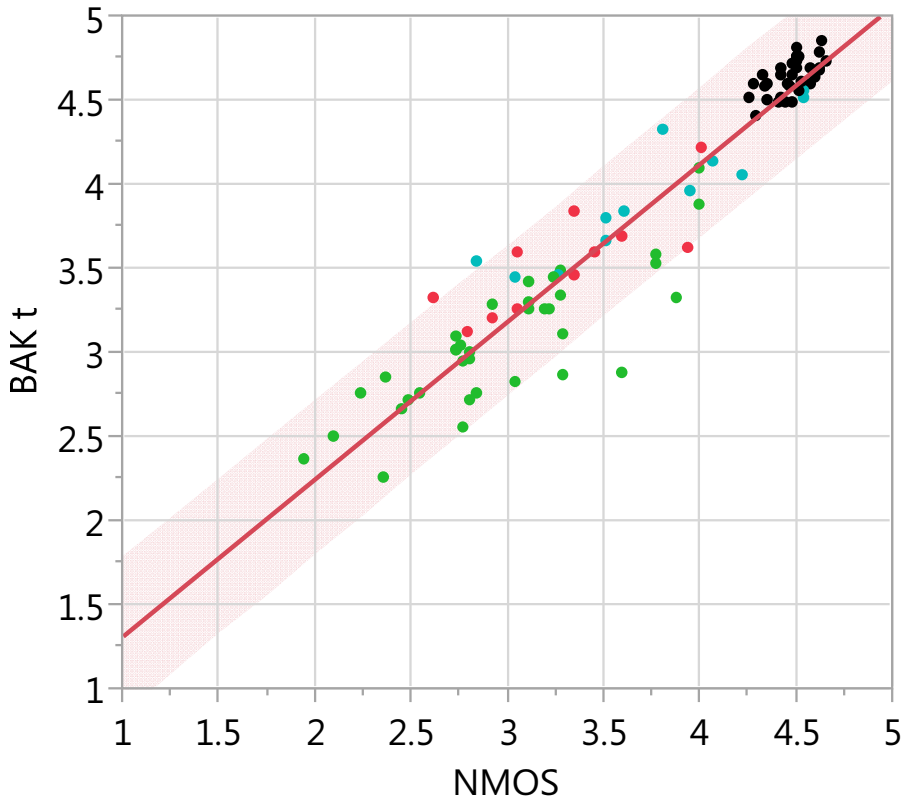


Figure 52: N-MOS fit to BAK-transformed Tests 5 & 6, shaded area is 95 % confidence interval

Table 34: Correlation, RMSE, and RMSE* for BAK

Metric	BAK	BAK_t
Correlation	0,960	0,957
RMSE	0,394	0,270
RMSE*	0,392	0,265

NOTE 2: If the reference transformation described in clause 5, equations (1), (2), and (3) is not applied to the subjective results, then the resulting RMSE* values after 3rd order mapping is 0,253.

Figures 53 and 54 plot results for G-MOS fit to OVRL, and table 35 provides correlation, RMSE and RMSE*.

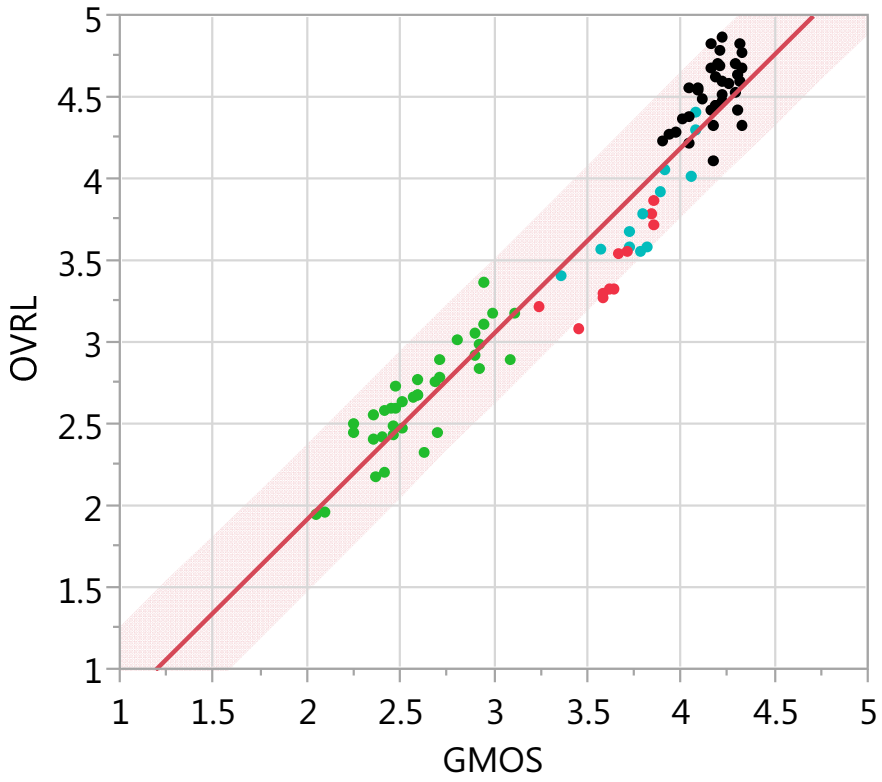


Figure 53: G-MOS fit to OVRL, Tests 5 & 6, shaded area is 95 % confidence interval

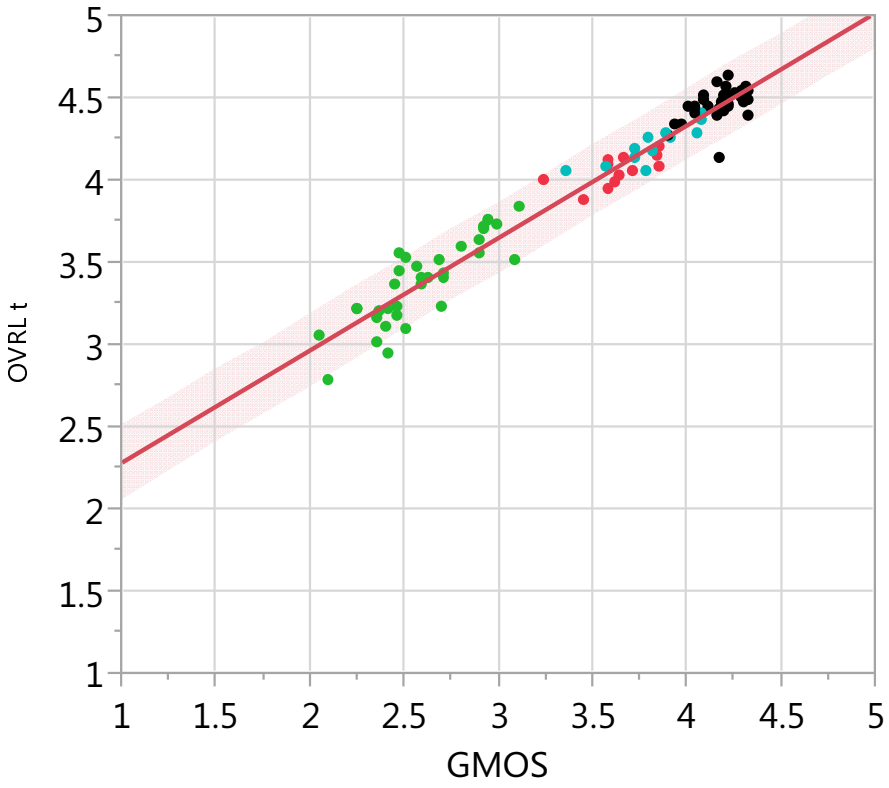


Figure 54: G-MOS fit to OVRL-transformed, Tests 5 & 6, shaded area is 95 % confidence interval

Table 35: Correlation, RMSE, and RMSE* for OVRL

Metric	OVRL	OVRL_t
Correlation	0,966	0,978
RMSE	0,270	0,175
RMSE*	0,262	0,160

NOTE 3: If the reference transformation described in clause 5, equations (1), (2), and (3) is not applied to the subjective results, then the resulting RMSE* values after 3rd order mapping is 0,179.

The RMSE* values in this clause in general are similar to the ones in clauses 8.1 to 8.3.

9 Application of the retrained model

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:

1) The speech samples used in conjunction with the model should be the ones used in the subjective tests: 16 sentences of male and female speakers, American English. A set of 16 full-band Chinese sentences meeting requirements of Recommendation ITU- T P.800 [i.17] can also be used, for testing in narrowband.

2) The results should be calculated on a per sentence basis and averaged over all 16 samples.

3) The background noises to be used in conjunction with the model shall be taken from ETSI ES 202 396-1 [i.2]. In addition, music can be used as a background noise.

4) The setup is according to ETSI ES 202 396-1 [i.2]. For testing in the handset use case, variations from nominal position defined in Recommendation ITU-T P.64 [i.24], and described in clause 8.4 can also be used.

Annex A (normative): Summary of Retraining Databases

					References															
Database	Lab	Test #	BW	#Noise Types	NSLevel	Ref SNR [dB]	# Reference Conditions	#Test Conditions	# Devices	Use case	Listening Instrument	Listening Mode	Presentation Level [dBSPL]	#talkers	# samples per talker	# of listeners	# votes per sample	# votes per condition	Signals available	Contribution
1	Audience	1	NB	8 (replace Crossroads with clean speech)	Table 1	0, 12, 24, 36	12	48	6	HS & HHHF	HD280 PRO	diotic	73	4 (2M, 2F)	2	32	16	128	CMU_OUT, PRI_MIC_IN	S4-120322
2	Audience	2	NB	8 (replace Crossroads with clean speech)	Table 1	0, 12, 24, 36	12	48	6	HS & HHHF	HD280 PRO	diotic	73	4 (2M, 2F)	2	32	16	128	CMU_OUT, PRI_MIC_IN	S4-120322
3	Audience	3	WB	8 (replace Crossroads with clean speech)	Table 1	10, 20, 30, 40	12	48	6	HS & HHHF	HD280 PRO	diotic	73	4 (2M, 2F)	2	32	16	128	CMU_OUT, PRI_MIC_IN	S4-120322
4	Audience	4	WB	8 (replace Crossroads with clean speech)	Table 1	10, 20, 30, 40	12	48	6	HS & HHHF	HD280 PRO	diotic	73	4 (2M, 2F)	2	32	16	128	CMU_OUT, PRI_MIC_IN	S4-120322
5	Qualcomm/Dynastat	1	NB	6 (Pub, Road, Train, Car, Mensa, clean speech)	Table 1	0, 12, 24, 36	12	48	8	HS	HD25	diotic	73	4 (2M, 2F)	8	32	4	128	CMU_OUT, PRI_MIC_IN, MRP	S4-120375
6	Qualcomm	2	NB	8 (replace Train with clean speech)	Table 1	0, 12, 24, 36	12	48	3	HS & HHHF	HD280 PRO	diotic	73	4 (2M, 2F)	4	32	8	128	CMU_OUT, PRI_MIC_IN, MRP	S4-120375
7	Qualcomm	3	NB	8 (replace Train with clean speech)	Table 1	0, 12, 24, 36	12	48	6	HS	HD280 PRO	diotic	73	4 (2M, 2F)	4	32	8	128	CMU_OUT, PRI_MIC_IN, MRP	S4-120375
8	Orange SA	1	WB	5 (Car, Road, Train, Cafeteria, Office)	Table 2	10, 20, 30, 40	12	90	6	HS	HD25	monaural	79	6 (3M, 3F)	2	24	24	288	CMU_OUT, PRI_MIC_IN	SA-120348
9	Qualcomm	4	WB	8 (replace Train with clean speech)	Table 2	10, 20, 30, 40	12	48	6	HS & HHHF	HD280 PRO	diotic	73	4 (2M, 2F)	4	32	8	128	CMU_OUT, PRI_MIC_IN, MRP	S4-120467
10	Qualcomm	5	WB	8 (replace Train with clean speech)	Table 1	10, 20, 30, 40	12	48	6	HS	HD280 PRO	diotic	73	4 (2M, 2F)	4	32	8	128	CMU_OUT, PRI_MIC_IN, MRP	S4-120619
11	Audience	1A	NB	8 (noise level +6dB)	Table 1	0, 12, 24, 36	12	48	6	HS & HHHF	HD280 PRO	diotic	73	4 (2M, 2F)	2	32	16	128	CMU_OUT, PRI_MIC_IN	S4-120655
12	Audience	2A	NB	8 (noise level +6dB)	Table 1	0, 12, 24, 36	12	48	6	HS & HHHF	HD280 PRO	diotic	73	4 (2M, 2F)	2	32	16	128	CMU_OUT, PRI_MIC_IN	S4-120655
13	Audience	3A	WB	8 (noise level +6dB)	Table 1	10, 20, 30, 40	12	48	6	HS & HHHF	HD280 PRO	diotic	73	4 (2M, 2F)	2	32	16	128	CMU_OUT, PRI_MIC_IN	S4-120655
14	Audience	4A	WB	8 (noise level +6dB)	Table 1	10, 20, 30, 40	12	48	6	HS & HHHF	HD280 PRO	diotic	73	4 (2M, 2F)	2	32	16	128	CMU_OUT, PRI_MIC_IN	S4-120655
15	NOKIA Corp/Dynastat	1	WB	8	Table 1	0, 12, 24, 36	12	48	6	HS	HD25	diotic	73	4 (2M, 2F)	4	32	8	128	CMU_OUT, PRI_MIC_IN	S4-120813
16	NOKIA Corp/Dynastat	2	WB	8	Table 1	0, 12, 24, 36	12	48	6	HS	HD25	diotic	73	4 (2M, 2F)	4	32	8	128	CMU_OUT, PRI_MIC_IN	S4-120813

Annex B (normative): Test vectors for model verification

B.0 Test vectors

The test vectors for verification of an objective model implementation are given in this annex. A model claiming to be compatible with the present document shall achieve all scores with an accuracy of ±0,1 MOS.

B.1 Audience test vectors

The validation results below are for signals used in the validation Experiments 6 (NB) and 8 (WB) as reported in clause 9. The reference conditions and noise types are as described in [i.1], but with levels of noise increased by 6 dB. The [six] devices have been tested in a mix of handset and handheld speakerphone use cases. Predictions from the model are presented at both sample and condition level. For each Experiment, 50 sample files and value sets are provided for validation of implementations of this model.

The test vectors can be downloaded here:
http://docbox.etsi.org/stq/Open/TS%20103%20106%20Wave%20files/Annex_B1_1_2_Audience%20Verification%20Data/

Table B.1: Audience experiment 6 test vectors and objective scores to be achieved by an objective model implementation

				Per sample		
Noise	**Device**	**Talker**	**Sample**	**S-MOS**	**N-MOS**	**G-MOS**
cafeteria	A	m1	s2	3,87	3,29	3,50
car1	A	m1	s2	3,51	3,42	3,27
crossroads	A	f1	s4	3,54	3,42	3,29
crossroads	A	f2	s3	3,18	3,34	3,00
mensa	A	f1	s5	4,24	3,25	3,80
mensa	A	m2	s4	3,62	3,45	3,36
office	A	f2	s6	4,35	3,59	4,00
pub	A	f1	s6	2,31	2,84	2,25
pub	A	m1	s6	2,31	3,07	2,34
trafficRoad	A	f1	s7	2,82	2,85	2,57
cafeteria	B	f1	s1	3,42	3,21	3,13
cafeteria	B	f2	s2	3,22	3,20	2,98
car1	B	f2	s3	3,33	3,59	3,19
car1	B	m1	s3	3,45	3,59	3,27
crossroads	B	m1	s3	3,30	3,56	3,15
mensa	B	f1	s5	4,18	3,33	3,77
office	B	f2	s6	4,33	3,58	3,98
office	B	m2	s6	4,06	3,57	3,75
pub	B	m1	s6	2,30	2,45	2,08
pub	B	m2	s7	2,32	2,69	2,19
trafficRoad	B	m1	s8	2,68	2,31	2,24
trafficRoad	B	m2	s8	2,89	2,26	2,35
train	B	m1	s1	2,22	3,07	2,29
train	B	m2	s8	2,62	3,17	2,57
cafeteria	D	m2	s2	3,64	2,87	3,17
car1	D	m2	s3	3,20	4,07	3,22
crossroads	D	m2	s4	3,78	3,88	3,61
mensa	D	m2	s5	3,94	4,16	3,80
office	D	m2	s6	4,28	3,75	3,99
pub	D	m2	s7	3,42	3,90	3,34
trafficRoad	D	m2	s8	3,11	2,71	2,71
train	D	m2	s8	3,54	3,85	3,42
cafeteria	E	m2	s2	2,67	1,93	2,04

				Per sample		
Noise	**Device**	**Talker**	**Sample**	**S-MOS**	**N-MOS**	**G-MOS**
car	E	m2	s3	1,85	1,92	1,56
crossroads	E	m2	s4	2,98	1,65	2,08
mensa	E	m2	s5	3,52	2,23	2,79
office	E	m2	s6	3,48	2,41	2,84
pub	E	m2	s7	1,13	1,27	1,00
trafficRoad	E	m2	s8	1,22	1,23	1,00
train	E	m2	s8	2,82	1,34	1,79
cafeteria	F	m2	s2	3,41	2,63	2,89
car	F	m2	s3	3,72	2,57	3,10
crossroads	F	m2	s4	3,86	2,55	3,20
mensa	F	m2	s5	4,14	2,95	3,60
office	F	m2	s6	4,18	3,63	3,87
pub	F	m2	s7	2,95	1,65	2,06
trafficRoad	F	m2	s8	2,71	1,43	1,78
train	F	m2	s8	3,87	1,96	2,92

Table B.2: Audience experiment 8 test vectors and objective scores to be achieved by an objective model implementation

				Per sample		
Noise	**Device**	**Talker**	**Sample**	**S-MOS**	**N-MOS**	**G-MOS**
cafeteria	A	m2	s1	3,64	3,50	3,23
car1	A	m1	s2	3,69	4,17	3,54
crossroads	A	m2	s3	3,57	3,65	3,23
mensa	A	f1	s4	3,60	3,76	3,30
mensa	A	m1	s5	3,91	4,21	3,73
office	A	f1	s5	4,07	4,37	3,94
office	A	f2	s6	4,20	4,31	4,02
pub	A	m2	s6	2,44	4,02	2,60
trafficRoad	A	m2	s7	2,03	3,40	2,10
train	A	m1	s1	3,04	4,06	3,01
cafeteria	B	f1	s2	3,37	3,47	3,01
car1	B	m2	s3	3,80	3,91	3,53
crossroads	B	m2	s3	3,52	3,64	3,19
mensa	B	f1	s5	3,79	3,73	3,44
mensa	B	m2	s4	3,82	3,95	3,56
office	B	f2	s5	4,01	4,04	3,75
office	B	m1	s5	4,17	4,26	3,98
pub	B	f2	s7	2,91	2,49	2,27
trafficRoad	B	m2	s7	2,02	3,30	2,06
train	B	f1	s1	2,99	4,26	3,05
cafeteria	D	f2	s1	3,63	4,43	3,60
car1	D	f2	s3	3,52	4,66	3,60
mensa	D	f2	s4	3,79	3,85	3,50
office	D	f1	s6	4,21	4,84	4,24
pub	D	f1	s7	2,85	4,03	2,87
trafficRoad	D	f2	s7	2,40	3,82	2,49
train	D	m1	s8	3,85	4,53	3,81
cafeteria	E	m2	s2	3,20	2,58	2,52
car	E	f2	s3	3,31	2,48	2,55
mensa	E	f1	s5	4,06	2,04	2,96
office	E	f1	s6	3,84	2,65	3,03
pub	E	m2	s6	2,17	1,56	1,41
trafficRoad	E	f2	s8	2,44	1,96	1,74
train	E	m2	s1	2,73	1,92	1,90
cafeteria	F	f1	s1	3,59	2,13	2,62
car	F	f2	s2	3,66	2,89	2,99
mensa	F	f2	s5	3,83	3,01	3,18
office	F	m1	s6	4,09	3,36	3,54
office	F	m2	s5	4,14	3,37	3,59
pub	F	m2	s7	3,13	2,45	2,41
trafficRoad	F	f1	s8	3,14	2,43	2,41

				Per sample		
Noise	Device	Talker	Sample	S-MOS	N-MOS	G-MOS
train	F	f1	s8	4,12	3,23	3,51

B.2 Orange test vectors

A subset of Orange validation database, comprised of the three scores [S-MOS, N-MOS, G-MOS] and the associated audio material [Clean, Noisy Input, Noise-reduced output] for each sample is provided for purposes of validation. This subset covers as much as possible the entire quality range and includes samples of conditions 2, 10, 19, 23, 26 and 30 as detailed in table B.3.

The test vectors can be downloaded here:
http://docbox.etsi.org/stq/Open/TS%20103%20106%20Wave%20files/Annex_B2_Orange%20Verification%20Data/

Table B.3: Orange test vectors and objective scores to be achieved by an objective model implementation

					Per sample			Per condition		
File name	Noise	Device	Talker	Sample	S-MOS	N-MOS	G-MOS	S-MOS	N-MOS	G-MOS
1bff2s01.c02	Mensa	D1	f2	s1	4,04	3,79	3,68			
1bff2s02.c02	Mensa	D1	f2	s2	3,70	4,39	3,64			
1bfm1s01.c02	Mensa	D1	m1	s1	3,97	2,89	3,25			
1bfm1s02.c02	Mensa	D1	m1	s2	4,13	4,08	3,87			
1bfm2s01.c02	Mensa	D1	m2	s1	4,42	4,43	4,26			
1bfm2s02.c02	Mensa	D1	m2	s2	4,09	3,92	3,77	4,06	3,92	3,74
1bff2s01.c10	Crossroads	D4	f2	s1	3,59	3,15	3,05			
1bff2s02.c10	Crossroads	D4	f2	s2	3,87	3,68	3,49			
1bfm1s01.c10	Crossroads	D4	m1	s1	3,67	2,98	3,04			
1bfm1s02.c10	Crossroads	D4	m1	s2	3,71	3,45	3,26			
1bfm2s01.c10	Crossroads	D4	m2	s1	3,80	2,82	3,07			
1bfm2s02.c10	Crossroads	D4	m2	s2	3,77	3,02	3,14	3,74	3,18	3,18
1bff2s01.c19	No noise	Source	f2	s1	4,75	4,64	4,40			
1bff2s02.c19	No noise	Source	f2	s2	4,74	4,39	4,40			
1bfm1s01.c19	No noise	Source	m1	s1	4,76	3,47	4,19			
1bfm1s02.c19	No noise	Source	m1	s2	4,76	4,43	4,40			
1bfm2s01.c19	No noise	Source	m2	s1	4,72	4,63	4,40			
1bfm2s02.c19	No noise	Source	m2	s2	4,75	4,26	4,40	4,75	4,30	4,37
1bff2s01.c23	No noise	NS Level 1	f2	s1	2,32	4,67	2,78			
1bff2s02.c23	No noise	NS Level 1	f2	s2	2,58	4,75	2,98			
1bfm1s01.c23	No noise	NS Level 1	m1	s1	2,65	4,55	2,94			
1bfm1s02.c23	No noise	NS Level 1	m1	s2	2,33	4,70	2,80			
1bfm2s01.c23	No noise	NS Level 1	m2	s1	2,30	4,63	2,75			
1bfm2s02.c23	No noise	NS Level 1	m2	s2	2,69	4,73	3,04	2,48	4,67	2,88
1bff2s01.c26	Crossroads	Source	f2	s1	4,76	2,48	3,77			
1bff2s02.c26	Crossroads	Source	f2	s2	4,74	2,40	3,73			
1bfm1s01.c26	Crossroads	Source	m1	s1	4,72	2,35	3,69			
1bfm1s02.c26	Crossroads	Source	m1	s2	4,76	2,36	3,73			
1bfm2s01.c26	Crossroads	Source	m2	s1	4,67	2,45	3,68			
1bfm2s02.c26	Crossroads	Source	m2	s2	4,72	2,48	3,74	4,73	2,42	3,72
1bff2s01.c30	Crossroads	NS Level 1	f2	s1	3,36	1,95	2,37			
1bff2s02.c30	Crossroads	NS Level 1	f2	s2	3,20	2,02	2,28			
1bfm1s01.c30	Crossroads	NS Level 1	m1	s1	2,79	1,97	1,97			
1bfm1s02.c30	Crossroads	NS Level 1	m1	s2	3,29	2,01	2,34			
1bfm2s01.c30	Crossroads	NS Level 1	m2	s1	2,08	1,94	1,52			
1bfm2s02.c30	Crossroads	NS Level 1	m2	s2	2,94	2,08	2,12	2,94	2,00	2,10

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

The following speech samples are used in conjunction with the model: 4 talkers (2 Males/2 Females), 8 Harvard sentences per talker, each sample is 4 s duration.

The first 4 sentences 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:
http://docbox.etsi.org/stq/Open/TS%20103%20106%20Wave%20files/Annex_C_Dynastat%20Speech%20Data/

Table C.1

Seq	Sample	Harvard Sentence	
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.	

Annex D (informative): Subjective testing framework used for the present document

D.1 Introduction

This annex describes the framework for conducting subjective testing used for the 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.

D.2 Subjective test plan

D.2.1 Traceability

The subjective test method is described in Recommendation ITU-T P.835 [i.6] and the ITU-T Handbook of subjective testing practical procedures [i.20], with the following observations:

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, section 5, and section B.3 of Recommendation ITU-T P.501 [i.13]. Each sample needs to be constructed according to the guidelines described in Recommendation ITU-T P.835 [i.6], section 5.1.4 (including 1 s of leading and 1 s of trailing silence) and normalized to an active speech level [i.8] of -26 dBov. It is recommended that the source speech material be 16 bit/48kHz.

D.2.3 Reference Conditions

Reference conditions need to follow the proposal in [i.21], 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 tables D.1 and D.2.

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 3. Table D.1 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.2] database (except for the Male Single Voice Distractor noise type, see note 1).

NOTE 1: 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.

NOTE 2: During the derivation of the training databases it was found that relatively few conditions had SIG scores below 3.0. For the validation stage it is desired that the tests also include conditions that cover a broader range of the SIG scale, while also achieving a good distribution of BAK and OVRL scores. This can be accomplished by one (or a combination of) the following means:

- A pre-selection of UEs with known poor speech quality for use in the test (preferred).
- A higher proportion of UEs operating in hand-held hands-free mode.
- Use of lower speech codec bit-rates (as these were not part of the training databases, the lower speech codec bit-rates will be only considered in the validation, in situations in which the model is capable of handling these types of impairments).

Table D.1: Test and Reference conditions for narrowband subjective evaluation of noise reduction

Reference Conditions				
File	**SIG.**	**SNR**	**Noise Type**	
i01	Source (filtered)	No Noise	-	
i02	Source (filtered)	0 dB	Fullsize_Car1_130Kmh_binaural	
i03	Source (filtered)	12 dB	Fullsize_Car1_130Kmh_binaural	
i04	Source (filtered)	24 dB	Fullsize_Car1_130Kmh_binaural	
i05	Source (filtered)	36 dB	Fullsize_Car1_130Kmh_binaural	
i06	NS Level 1	No Noise	-	
i07	NS Level 2	No Noise	-	
i08	NS Level 3	No Noise	-	
i09	NS Level 4	No Noise	-	
i10	NS Level 3	24 dB	Fullsize_Car1_130Kmh_binaural	
i11	NS Level 2	12 dB	Fullsize_Car1_130Kmh_binaural	
i12	NS Level 1	[0 dB]	Fullsize_Car1_130Kmh_binaural	
Test Conditions				
File	**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.2]**
i13	-1,7/+1,3 dBPa	L: 75,0 dB(A)/R: 73,0 dB(A)	Pub_Noise_binaural_V2	Recording in a pub
i14	-1,7/+1,3 dBPa	L: 74,9 dB(A)/R: 73,9 dB(A)	Outside_Traffic_Road_binaural	Recording at pavement
i15	-1,7/+1,3 dBPa	L: 69,1 dB(A)/R: 69,6 dB(A)	Outside_Traffic_Crossroads_binaural	Recording at pavement
i16	-1,7/+1,3 dBPa	L: 68,2 dB(A)/R:69,8 dB(A)	Train_Station_binaural	Recording at departure platform
i17	-1,7/+1,3 dBPa	L: 69,1 dB(A)/R: 68,1 dB(A)	Fullsize_Car1_130Kmh_binaural	Recording in passenger cabin
i18	-1,7/+1,3 dBPa	L: 68,4 dB(A)/R: 67,3 dB(A)	Cafeteria_Noise_binaural	Recording at sales counter
i19	-1,7/+1,3 dBPa	L: 63,4 dB(A)/R: 61,9 dB(A)	Mensa_binaural	Recording in a cafeteria
i20	-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

Table D.2: Test and Reference conditions for wideband subjective evaluation of noise reduction

Reference Conditions			
File	**SIG.**	**SNR**	**Noise Type**
i01	Source (filtered)	No Noise	-
i02	Source (filtered)	10 dB	Outside_Traffic_Crossroads_binaural
i03	Source (filtered)	20 dB	Outside_Traffic_Crossroads_binaural
i04	Source (filtered)	30 dB	Outside_Traffic_Crossroads_binaural
i05	Source (filtered)	40 dB	Outside_Traffic_Crossroads_binaural
i06	NS Level 1, 2nd set of parameters	No Noise	-
i07	NS Level 2, 2nd set of parameters	No Noise	-
i08	NS Level 3, 2nd set of parameters	No Noise	-
i09	NS Level 4, 2nd set of parameters	No Noise	-
i10	NS Level 3, 2nd set of parameters	30 dB	Outside_Traffic_Crossroads_binaural
i11	NS Level 2, 2nd set of parameters	20 dB	Outside_Traffic_Crossroads_binaural
i12	NS Level 1, 2nd set of parameters	10 dB	Outside_Traffic_Crossroads_binaural

D.2.5 Pre-processing of reference conditions

For the reference conditions, the clean speech and noise signals need to be filtered with the LP35 and MSIN (for narrowband) and 78 KBP (for wideband) filters available with a modified "filter" demo program from Recommendation ITU-T G.191 [i.11] (available from ORANGE SA upon request). Appropriate resampling needs to be used prior to application of the filters. The necessary upsampling/downsampling are also performed through the use of Recommendation ITU-T G.191 [i.11] "filter" demo program. The clean speech is then processed with the spectral subtraction algorithm in annex A at the appropriate settings and, prior to mixing, normalized to an active speech level of -26 dBov. The mixing needs to be performed with the appropriate Recommendation ITU-T G.191 [i.11] tool to obtain the SNRs described in table D.1. The SNR is defined as the ratio between active speech levels to A-weighted noise level, for more details on NS Levels, see [i.22]. Clause D.4 shows a block diagram of the necessary processing steps.

D.2.6 Post-processing of test conditions

The uplink recordings of processed speech materials be 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.

D.2.7 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.14] and an artificial ear type 3.3 according to Recommendation ITU-T P.57 [i.15]. The HATS is diffuse field equalized. The resulting frequency response characteristic of the headphones used in the subjective experiments needs to be within the mask given in ETSI TS 126 131 [i.16], clause 6.4.2.

The presentation of the test and reference conditions to listeners needs to 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.2.8 Requirements on the listening laboratory

Listening laboratory facilities need to comply with the recommendations provided in Recommendation ITU-T P.800 [i.17].

D.2.9 Experimental design

The use of the Balanced Blocks experimental design described in [i.20], section 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 (alternatively, for wideband, table D.2 can be used). 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.

128 votes per condition need to be obtained. The number of votes per sample will depend on the number of samples per talker chosen. A minimum of 2 samples per talker and 8 votes per sample needs to be used.

D.2.10 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 below:

Table D.3

Trial	Sample	Condition
1	m1s3.r01	Reference - Source/No noise
2	f2s1.x06	Test - Cafeteria
3	m2s4.r11	Reference - NS-L2/12dB SNR
4	f1s1.r02	Reference - Source/0dB 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/12dB SNR
10	f2s2.r06	Reference - NSL1/No noise
11	m2s4.x01	Pub
12	f2s3.x08	Test - Call-center
13	m2s4.r04	Reference - Source/24dB SNR
14	f2s1.x04	Test - Train station
15	m2s3.r12	Reference - NS-L1/0dB 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.14] is used. The mouth simulator needs to be equalized to achieve the reproduction accuracy described in ETSI TS 126 132 [i.18], 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.18] clause 5.1.1.

Headsets need to be set-up on HATS as described in ETSI TS 126 132 [i.18], clause 5.1.2.

For handheld hands-free mode the device is set-up using HATS as described in ETSI TS 126 132 [i.18], 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 needs to be set-up and equalized according to ETSI ES 202 396-1 [i.2]. Noise types need to be reproduced at their realistic levels according to ETSI ES 202 396-1 [i.2], clause 8. 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.

To allow for proper convergence of noise suppressors, an additional four sentences need to be placed immediately prior to the 32 sentences used in subjective testing.

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 1 130 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 s each. The total length is therefore 152 seconds. The first 24 seconds 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.2] database to be 24 s 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 s) 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.

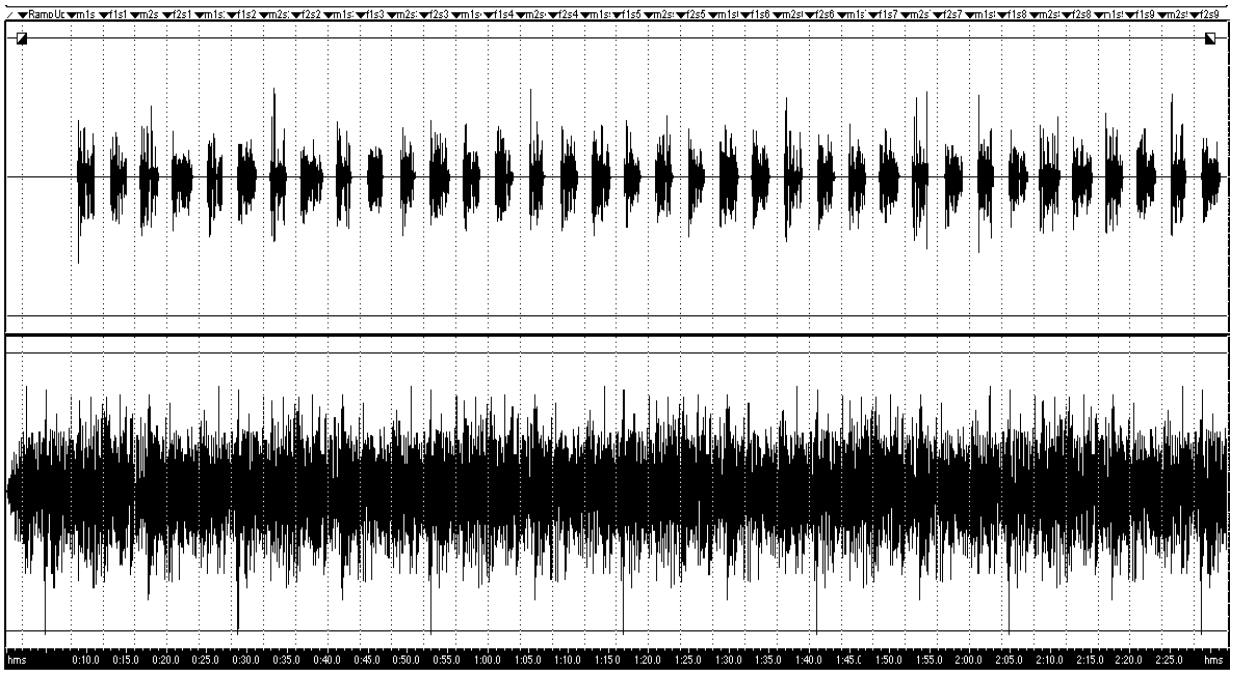


Figure D.1: Noise and speech playback sequence, including convergence period

D.3.6 Recordings at the network simulator electrical reference point

The network simulator needs to be set to a WCDMA voice call with the AMR 12,2 kbps speech codec in narrowband tests and AMR-WB 12,65 kbps speech codec in wideband tests. DTX needs to be enabled. 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. The send frequency response of the terminals needs to be measured according to ETSI TS 126 132 [i.18], clauses 7.4.1, 7.4.3, 8.4.1 or 8.4.3 (depending on the testing condition being performed) and the results documented.

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 use on objective predictor retraining.

D.4 Processing test plan block diagram

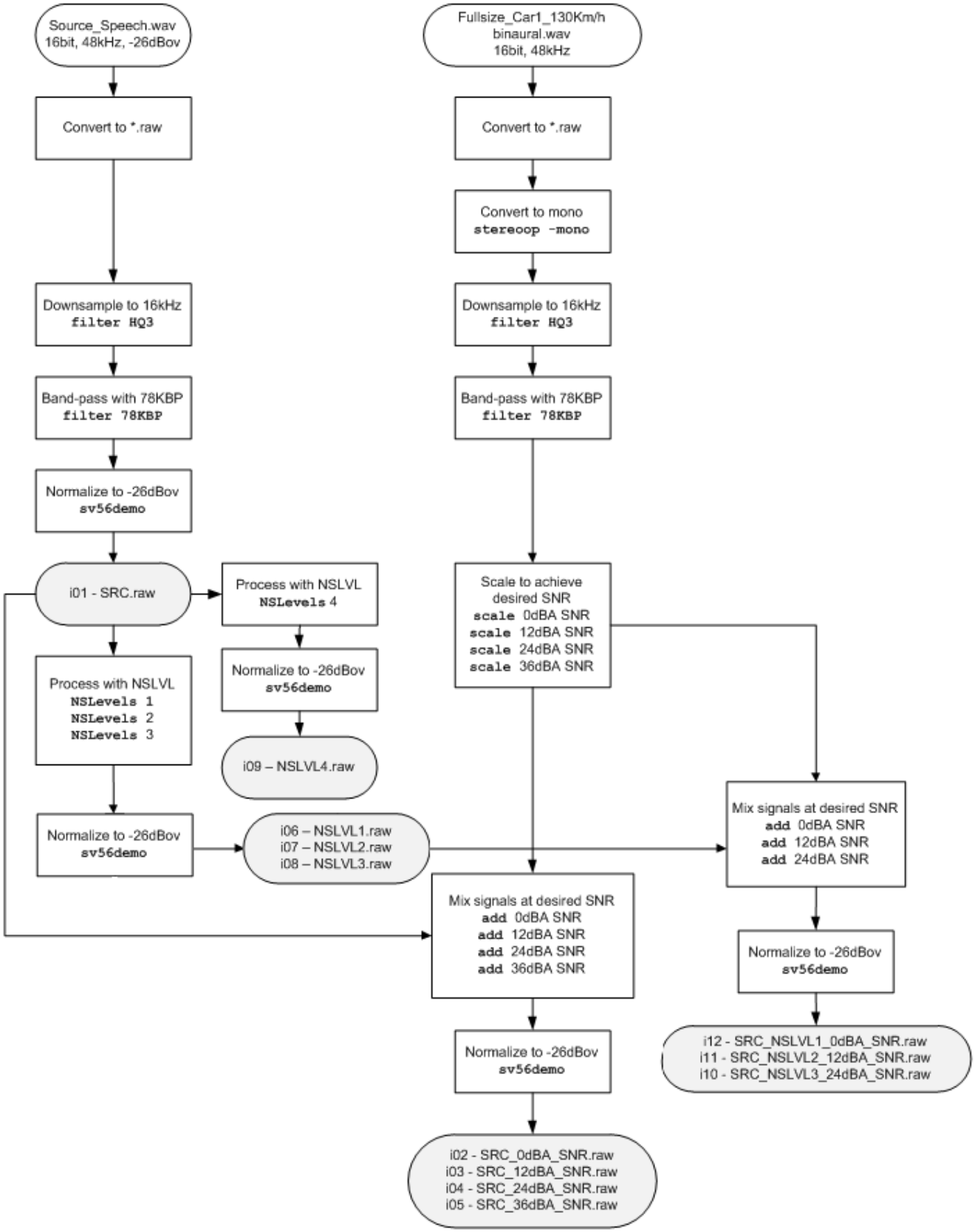


Figure D.2: Wideband processing for reference set

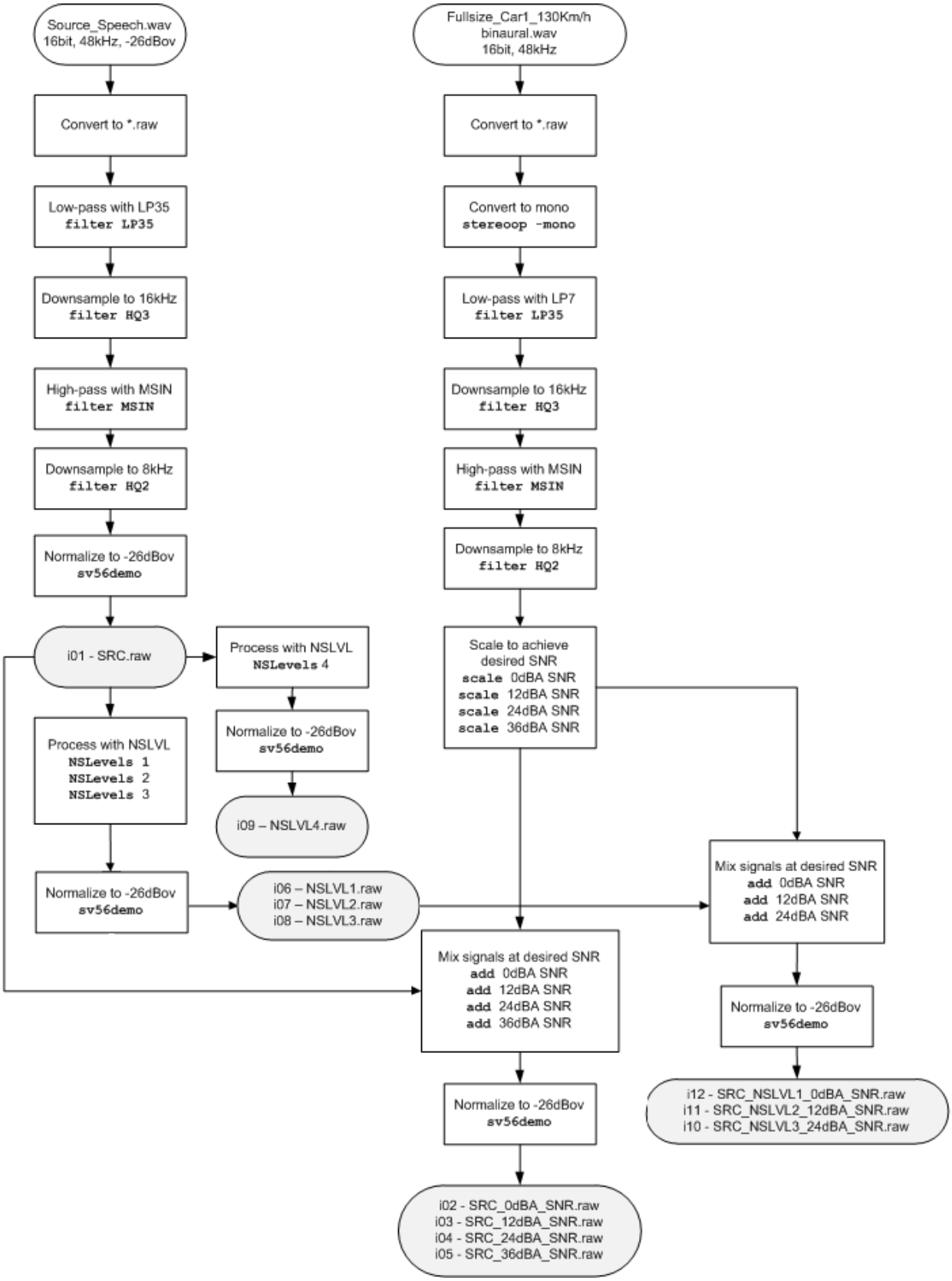


Figure D.3: Narrowband processing for reference set

History

Document history		
V1.1.1	August 2012	Publication
V1.2.1	March 2013	Publication
V1.3.1	April 2014	Publication
V1.4.1	November 2016	Publication