3QUEST:

Comparison of EG 202 396-3 to TS 103 106

HEAD acoustics
Application Note

HEAD acoustics GmbH
Ebertstraße 30a
D-52134 Herzogenrath
Tel: +49 (0) 2407-577-0
Fax: +49 (0) 2407-577-99
E-mail: telecom@head-acoustics.de
WEB: www.head-acoustics.de

3QUEST: Comparison of EG 202 396-3 to TS 103 106

Contents

1. Introduction

The objective model according to ETSI EG 202 396-3 [1] was developed to predict speech, noise and global quality of noisy speech signals for wide- and narrowband terminals according to ITU-T recommendation P.835 [2]. Especially for narrowband applications, signal processing capabilities of modern terminals (mobiles, smartphones) have rapidly progressed in the last years so that even 2-channel-microphone noise reduction solutions are currently state of the art. As a result, a much higher speech and noise quality can be achieved with these devices than without this technique.

Furthermore, it was discussed whether the underlying subjective narrowband data used in Annex H of EG 202 396-3 [1] from 2007 do not represent the latest technologies used for noise cancellation and it was suspected that there may be impairments potentially not covered correctly by the ETSI model.

In addition, some difficulties were reported [3] in achieving a reasonable correlation between a series of subjective tests and objective measurements following ETSI EG 202 396-3. It was found that this applies specifically to conditions which were not in scope of the ETSI databases, mainly conditions with different “SNR-sweeps” as described in Amendment 1 of ITU-T P.835 [2].

While revising the measurement standards TS 26.131 and TS 26.132 in 3GPP, it was agreed to include a method for objectively predicting speech, noise and global quality for handset terminals in sending direction and to possibly replace the ANR measurement by this new method. Due to the reported issues above, it was planned to create a modified version of the algorithm originally defined in ETSI EG 202 396-3 which should be adapted to a completely new set of subjective databases. After several parties contributed to this project with a huge effort of subjective testing, the model was retrained to this new material.

Due to the success of this process and the good prediction performance of the validation databases, the work of 3GPP was directly transferred into a new ETSI standard TS 103 106 [5].

This application note gives an overview about this new method and the consequences for the HEAD acoustics measurement system HEAD Analyzer ACQUA. Finally, several comparison measurements between the EG 202 396-3 and the TS 103 106 method are presented to provide an informative basis for the daily usage of both algorithms.

2. Development and Implementation

In this chapter, a brief description of the process of development and the algorithm modification is given. The following sections are intended to provide summary of the work which led to the new ETSI TS 103 106 standard, for further details refer to [5].

2.1. Status of EG 202 396-3 and Discussion in 3GPP

In the beginning of 2012, it was discussed to include a psycho-acoustically motivated test procedure for speech quality in background noise scenarios in the established measurement standards TS 26.131 and TS 26.132 for mobile phones. Prior to these discussions it became obvious that neither the existing ANR method nor other methodologies such as SNRI [7] were suitable to adequately predict the performance of advanced noise cancelling techniques used in modern mobile phones. It was evident that only a method which can predict the speech quality as perceived by the user could be a solution to this problem.

Since the only standard describing an objective method for this purpose was EG 202 396-3, the first proposal was to use this method for the next revision of the standard. But as already mentioned, some problems with the correlation of objective results of this method and subjective listening tests conducted by parties in 3GPP were reported [3]. The contexts of these listening tests always were state-of-the-art mobiles including multi-channel noise reduction systems.

As the need for an objective predictor of speech, noise and global quality found mutual consent, a new approach was chosen: several partners in 3GPP created new subjective listening tests and provided these databases to HEAD acoustics. The idea was to retrain the existing ETSI model to these new databases with as little changes as possible.

2.2. New Databases

The new subjective databases and the corresponding listening tests were created and conducted according to ITU-T recommendation P.835. This recommendationbest describes the basic procedure for conducting subjective tests and deriving the different quality dimensions S-MOS (speech mean opinion score), N-MOS (noise mean opinion score) and G-MOS (overall mean opinion score). Because of the rather general descriptions in this standard, a more detailed test plan was developed [6].

Several parties contributed with different types of databases which were divided into training and validation parts. Between 30 and 102 conditions per database were provided (including 12 reference conditions described in [6]), which led to an enormous amount of subjectively rated samples (8–24 samples per condition, 8–24 votes per sample).

An overview of the contributors and the extents of the databases is given in table 1. In particular, it is worth mentioning that well over 5000 single samples were made available for retraining in each mode (wide- and narrowband).

	Samples per condition	Narrowband				Wideband			
		Training		Validation		Training		Validation	
		(S)	(C)	(S)	(C)	(S)	(C)	(S)	(C)
Audience	8	1920	240	768	96	1440	180	768	96
Nokia	16	1920	120	0	0	960	60	0	0
Orange France	12	0	0	0	0	1080	90	360	30
Qualcomm	16	1920	120	1536	96	1920	120	0	0
Total		5760	480	2304	192	5400	450	1128	126

Table 1: Amount of conditions (C) and samples (S) provided by several parties

2.3. Modifications of the ETSI EG model

A short overview of the changes of the EG 202 396-3 algorithm is presented in the following. For detailed descriptions of the modifications, please refer to [5].

2.4. Data Transformation

One of the most important differences between the new retrained model and the ETSI EG 202 396-3 method is the usage of multiple databases for the training. Due to the different spread of conditions amongst the quality scale, each listening test usually has its own relative quality dimensions because all the conditions are judged subjectively in the context of this individual listening test.

In order to better align different tests, a set of reference conditions as defined in [6] was used for these series of subjective tests so that the scales for S-, N- and G-MOS may eventually be set into a common context. This anchoring was applied with the reference conditions, which were available for each listening test database.

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 of a per-scale regression was chosen. To compensate for biases, a constant column was added to the reference set. Then, a transformation T_j is calculated for each database j with reference set R_j which minimizes the distance to the average reference set A:

$$\underbrace{\begin{pmatrix} 1 & S_{i01} & N_{i01} & G_{i12} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & S_{i12} & N_{i12} & G_{i12} \end{pmatrix}}_{R_j(\text{Ref. set } j)} \times T_j = \underbrace{\begin{pmatrix} \overline{S_{\imath 01}} & \overline{N_{\imath 01}} & \overline{G_{\imath 12}} \\ \vdots & \vdots & \vdots \\ \overline{S_{\imath 12}} & \overline{N_{\imath 12}} & \overline{G_{\imath 12}} \end{pmatrix}}_{A\ (\text{Avg. ref. set})}$$

The transformation matrix T_j (size 4 × 3) can easily be determined to be

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

If the three scales (S-MOS / N-MOS / G-MOS) are independent from each other for each 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:

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

After transforming all scores individually per database, they could then be used for the retraining of the ETSI model. In consequence, all MOS values predicted by the new algorithm are also derived from this "average" context which is a mix of several databases.

2.5. Pre-Processing

Pre-Filtering (NB)

In the narrowband extension of [1], the listening test audio files included a far-end handset simulation, realized with an *intermediate reference system* (IRS) RCV filter according to annex D of [8]. In the requirements described in [6], such a listening filter was described or used in the databases neither for narrow- nor for wideband.

The narrowband mode of ETSI EG 202 396-3 internally filters the unprocessed and clean reference with IRS SND and IRS RCV to simulate a transmission over high-quality listening devices and networks. The principle of IRS seems to be outdated, modern state-of-the-art mobiles do not have this frequency characteristic. In these newly created NB databases, the used devices show 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 *mobile station in* (MSIN) filter [10], which is mainly a band pass. A listening filter is not applied to the processed signals in the retrained model.

Speech part detection (NB & WB)

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 G.160 [7]. The signal parts classified as silence are assumed as background noise sections, all other frames are assumed as speech.

Speech level adjustment (WB)

According to [6], 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 must be set to 73 dB SPL (-21 dB Pa) for the listening test.

The EG 202 396-3 implementation assumes 79 dB SPL (-15 dB Pa) active speech level due to the underlying listening test. Thus, this constant was adapted to the new databases.

2.6. Algorithm Changes

The ETSI model calculates several parameters out of the psycho-acoustically motivated inner representation for the estimation of S- and N-MOS. The parameters for S-MOS are presented in table 2. A detailed description of the calculation for the parameters can be found in [1].

The calculation of the objective S-MOS in chapter 6.5.2 of [1] 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 listeners' expectation to speech quality.

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

Table 2: Extracted Parameters for S-MOS

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 (six parameters of table 2) and output (S-MOS) data by a simple back-propagation algorithm [9].

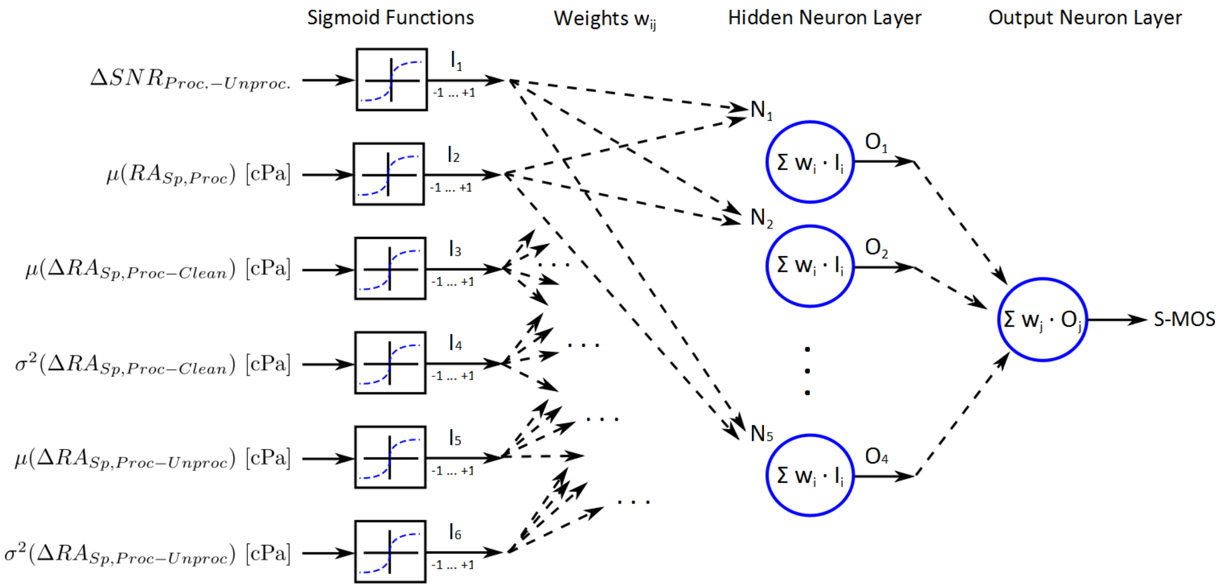


Figure 1: Structure of neural network for S-MOS

The setup of the neural network is shown in figure 1. It consists of 5 hidden layers; each layer N_j includes a connection from each transformed input parameter I_i. The output O_j of each layer 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 2 are composed to a vector $\mathbf{P}$ including a bias as the first element:

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

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

$$S\text{-}MOS_{\text{objective,raw}} = f_{\text{sigmoid}}\left(f_{\text{sigmoid}}\left(\frac{\mathbf{P}-\mathbf{M}_{\text{in}}}{\mathbf{S}_{\text{in}}}\right) \times \mathbf{H}\right) \times \mathbf{O}$$

First the parameter vector $\mathbf{P}$ 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. The numeric content of these vectors described can be found in [5].

After normalizing the input data, the sigmoid function $f_{\text{sigmoid}}(x)$ is applied to 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:

$$f_{\text{sigmoid}}(x) = \tanh(x)$$

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

$$\widetilde{\mathbf{P}} = f_{\mathrm{sigmoid}}\left(\frac{\mathbf{P}-\mathbf{M}_{\mathrm{in}}}{\mathbf{S}_{\mathrm{in}}}\right) = (1 \quad \widetilde{P_1} \quad \widetilde{P_2} \quad \widetilde{P_3} \quad \widetilde{P_4} \quad \widetilde{P_5} \quad \widetilde{P_6})$$

(Note: 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 $\mathbf{H}$. The matrix $\mathbf{H}$ 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, $\mathbf{H}$ is different for each bandwidth mode. The numeric values for $\mathbf{H}$ are given in [5].

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 $\mathbf{O}$, which weights the outputs of the hidden layers to an output score $S\text{-}MOS_{\mathrm{objective,raw}}$. This output layer matrix $\mathbf{O}$ is also given for wide- and narrowband mode independently.

Another part of the back-propagation algorithm is to normalize also 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:

$$S\text{-}MOS_{\mathrm{objective}} = \max\left(1.0, \min\left(\mathbf{S}_{\mathrm{out}} \cdot \left(SMOS_{\mathrm{objective,raw}} + \mathbf{M}_{\mathrm{out}}\right), 5.0\right)\right).$$

The objective S-MOS is calculated with $\mathbf{M}_{\mathrm{out}} = (3.0)$, $\mathbf{S}_{\mathrm{out}} = \ (2.0)$ in addition with the hard limiter [1.0; 5.0].

This retraining procedure was already successfully applied with other listening test databases [4].

The estimation for N- and G-MOS remains identical with the calculations presented in [1]. Only the coefficients for both regressions are updated and can be found in [5].

2.7. Prediction Results of Training Data

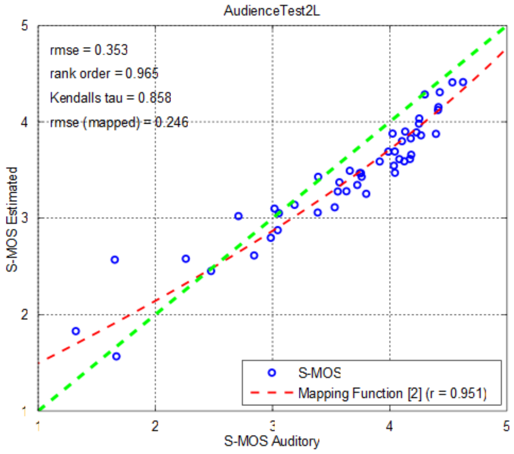
As difference metrics, the Pearson correlation coefficients, RMSE and RMSE* are calculated for the comparison between subjective and objective MOS data. For further details on these metrics, please refer to [11]. Prediction scores are compared without any further processing, with 1st order mapping and 3rd order mapping (see [11]).

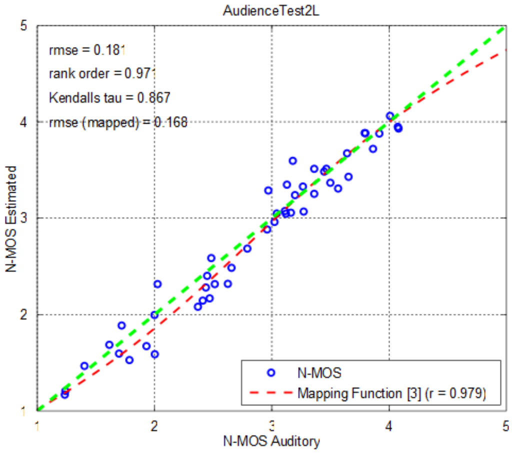
As already described in section 2.4, this is necessary because the raw / unmapped output of the new retrained model refers to an average reference system. When comparing this output to subjective data, the rank order correlation of the prediction is of interest. If the rank correlation is high, possible shifts and offsets can be compensated with mapping functions.

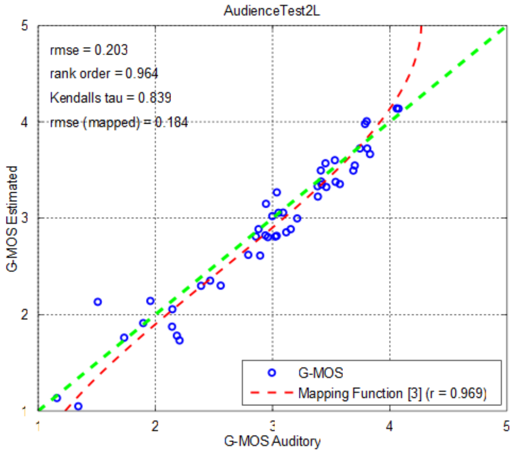
Although the prediction results of the training databases do not have to yield certain RMSE or RMSE* values, the difference metrics and scatterplots are given in this section.

2.8. Narrowband Training

For the new narrowband mode, a total of six databases with 288 conditions and 3840 samples were used. Table 3 exemplarily shows one training database and its prediction performance results.







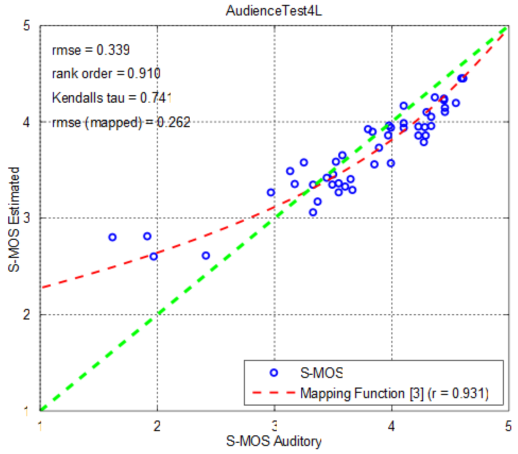
		S-MOS	N-MOS	G-MOS
RMSE:	no Mapping	0,35	0,18	0,20
	1st Ord. Mapping	0,25	0,17	0,18
	3rd Ord. Mapping	**0,21**	**0,17**	**0,18**

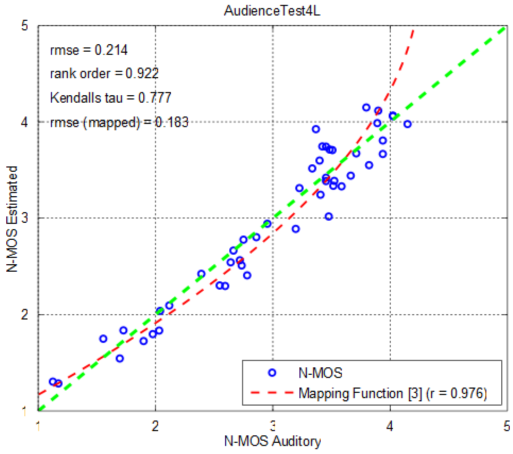
RMSE*:	no Mapping	0,23	0,08	0,11
	1st Ord. Mapping	0,15	0,07	0,11
	3rd Ord. Mapping	**0,11**	**0,08**	**0,11**

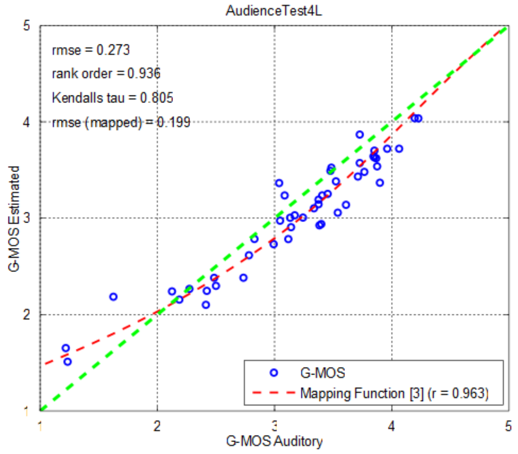
Table 3: Example NB training results: database "Audience – Test 2L"

2.9. Wideband Training

For the retrained wideband mode, a total of seven databases with 387 conditions and 5544 samples were used (some conditions and databases were removed due to inconsistencies). Table 4 exemplarily shows one training database and its prediction performance results.







		S-MOS	N-MOS	G-MOS
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**

Table 4: Example WB training results: database "Audience – Test 4L"

2.10. Validation Process & Prediction Results

The validation of the new retrained model was applied in a similar way as presented in [1], where validation conditions were held back and were not included in the training. These validation conditions are unknown to the model and thus can be regarded as a kind of blind test.

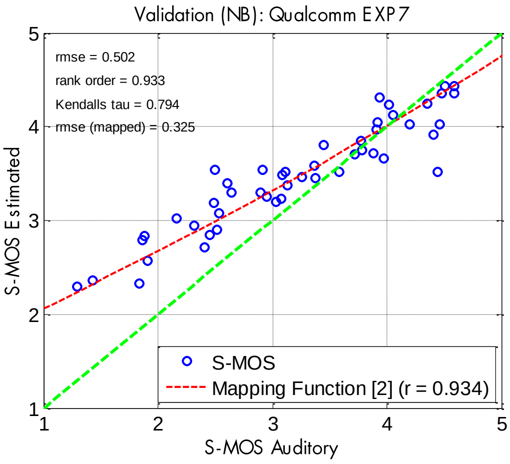
According to table 1, several validation databases were provided without the corresponding subjective results. Only the audio data files were available in order to calculate objective results in the first step. After distributing the objective results, the owners of the databases provided the corresponding subjective results.

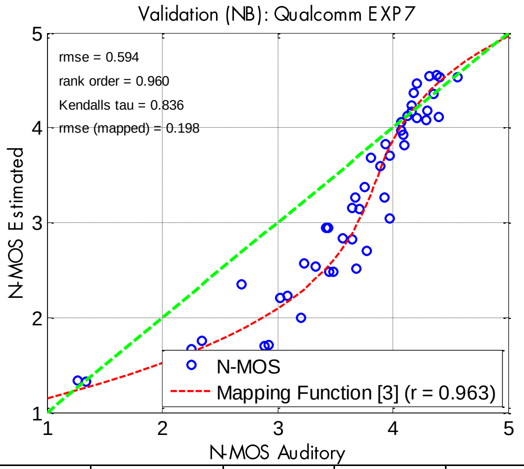
For reasons already described in section 2.7, the output of the new retrained model had to be compared with 1st and 3rd order mapping to compensate database-specific offsets and shifts. In [12], it was agreed on defining only RMSE and RMSE* values (calculated according to [11]) as requirements for the new retrained model; the Pearson correlation coefficient was left out as a performance requirement because it strongly depends on the MOS distribution of a database. The requirements for S-, N- and G-MOS prediction after monotonic 3rd order mapping were defined according to the values shown in table 5.

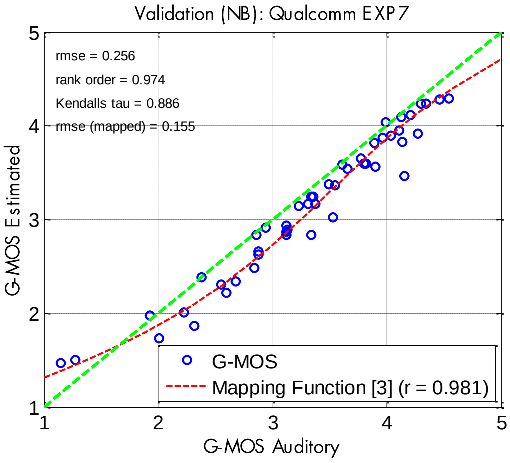
	S-MOS	N-MOS	G-MOS
RMSE	0.40	0.35	0.35
RMSE*	0.35	0.25	0.25

Table 5: Requirements of performance parameters for the retrained predictor

All validation databases (for NB and WB mode) passed these requirements. A detailed list of all evaluation metrics for all validation databases can be found in chapter 8 in [5]. Two validation examples for NB and WB mode are shown in table 6 and table 7.



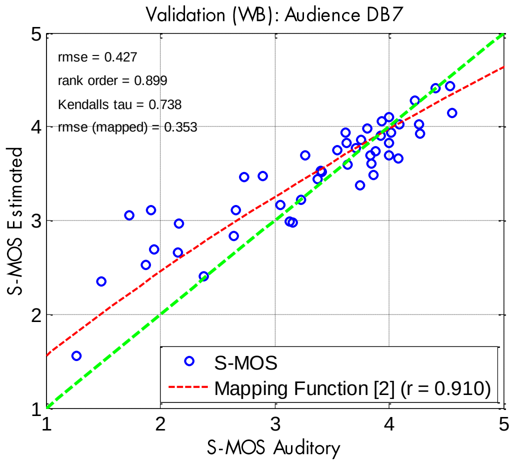


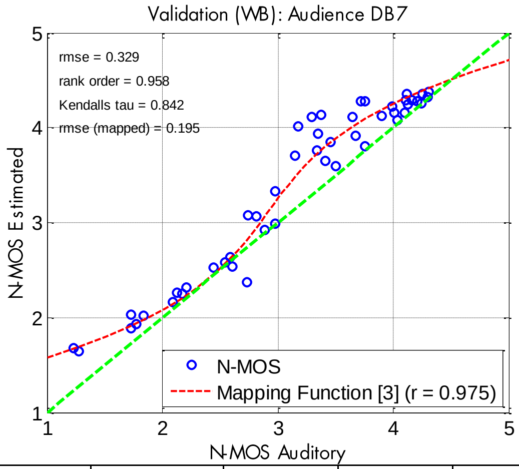


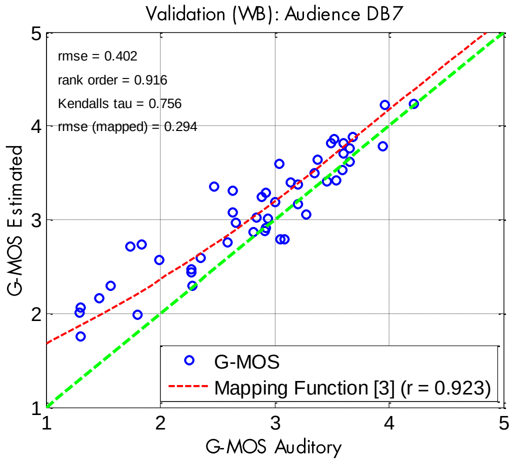
		S-MOS	N-MOS	G-MOS
RMSE:	no Mapping	0,43	0,33	0,40
	1st Ord. Mapping	0,36	0,23	0,30
	3rd Ord. Mapping	**0,34**	**0,19**	**0,29**

RMSE*:	no Mapping	0,33	0,23	0,32
	1st Ord. Mapping	0,25	0,13	0,21
	3rd Ord. Mapping	**0,25**	**0,11**	**0,21**

Table 6: Example of NB validation database







		S-MOS	N-MOS	G-MOS
RMSE:	no Mapping	0,36	0,16	0,26
	1st Ord. Mapping	0,31	0,14	0,18
	3rd Ord. Mapping	**0,23**	**0,13**	**0,15**

RMSE*:	no Mapping	0,25	0,05	0,15
	1st Ord. Mapping	0,22	0,05	0,10
	3rd Ord. Mapping	**0,14**	**0,05**	**0,07**

Table 7: Example of WB validation database

Especially the NB validation in table 6 illustrates the need of 1st and 3rd order mapping functions. Due to the good rank order of the predicted scores, the S-MOS can be compensated with a linear mapping function. The N-MOS prediction shows a more non-linear relation to the subjective data which can be compensated with the 3rd order mapping function.

In contrast, the WB validation in table 7 does not show a strong offset or shift, here only light-weighted monotonic mapping functions are needed to set the objective data into the context of this database.

2.11. **New Standard: TS 103 106**

As a consequence of the work done in 3GPP, the new retrained method was directly transferred into a new ETSI standard: ETSI TS 103 106 [5]. This standard is currently referenced by 3GPP and other standards, e.g. in TS 26.131 and TS 26.132. Within the standard, example audio files and corresponding objective scores can be found as validation data.

3. ACQUA Application

The hardware setup for 3QUEST tests with ACQUA, the Advanced Communication QUality Analysis system of HEAD acoustics, is shown in figure 2. The system represents the standard setup for 3QUEST tests which is identical for EG 202 396-1 and TS 103 106.

Besides the system requirements that are already known from 3QUEST tests according to EG 202 396-1, for testing 3QUEST according to TS 103 106 the following two prerequisites have to be fulfilled:

- The ACQUA Software needs to be version 3.1.200 or higher.
- The ACQUA Single Measurement Descriptors (SMDs) have to be modified accordingly.

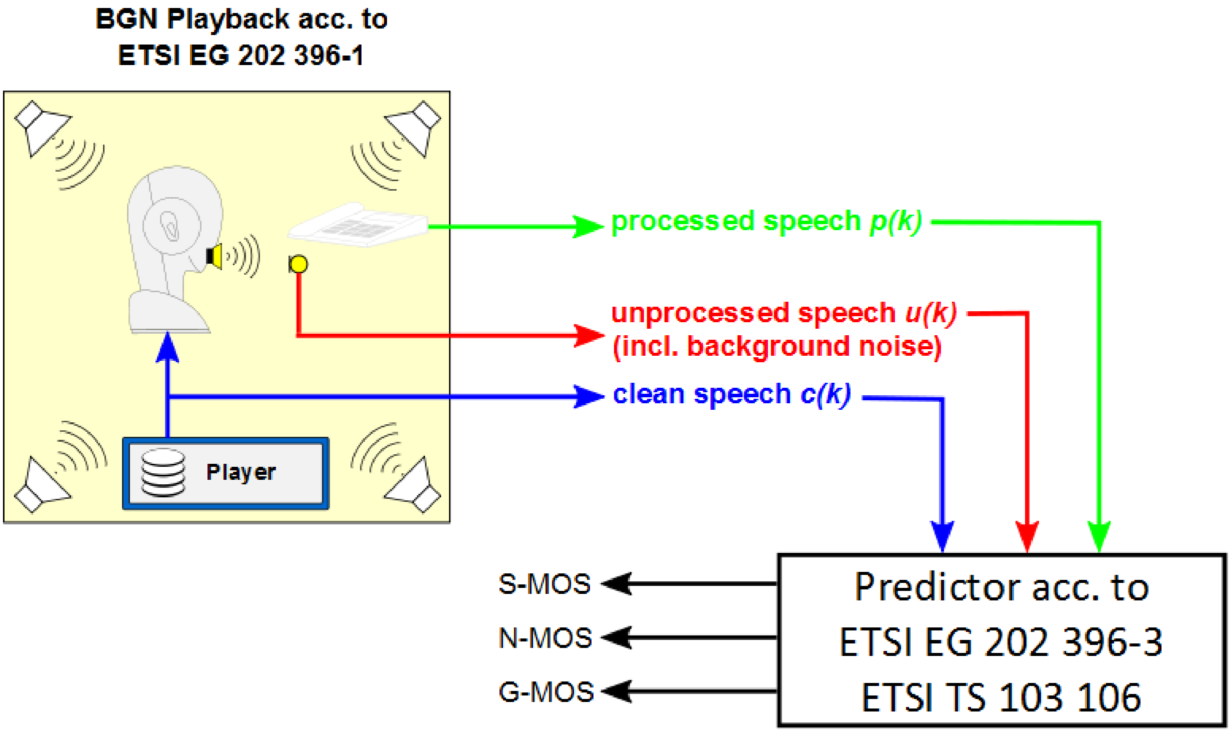
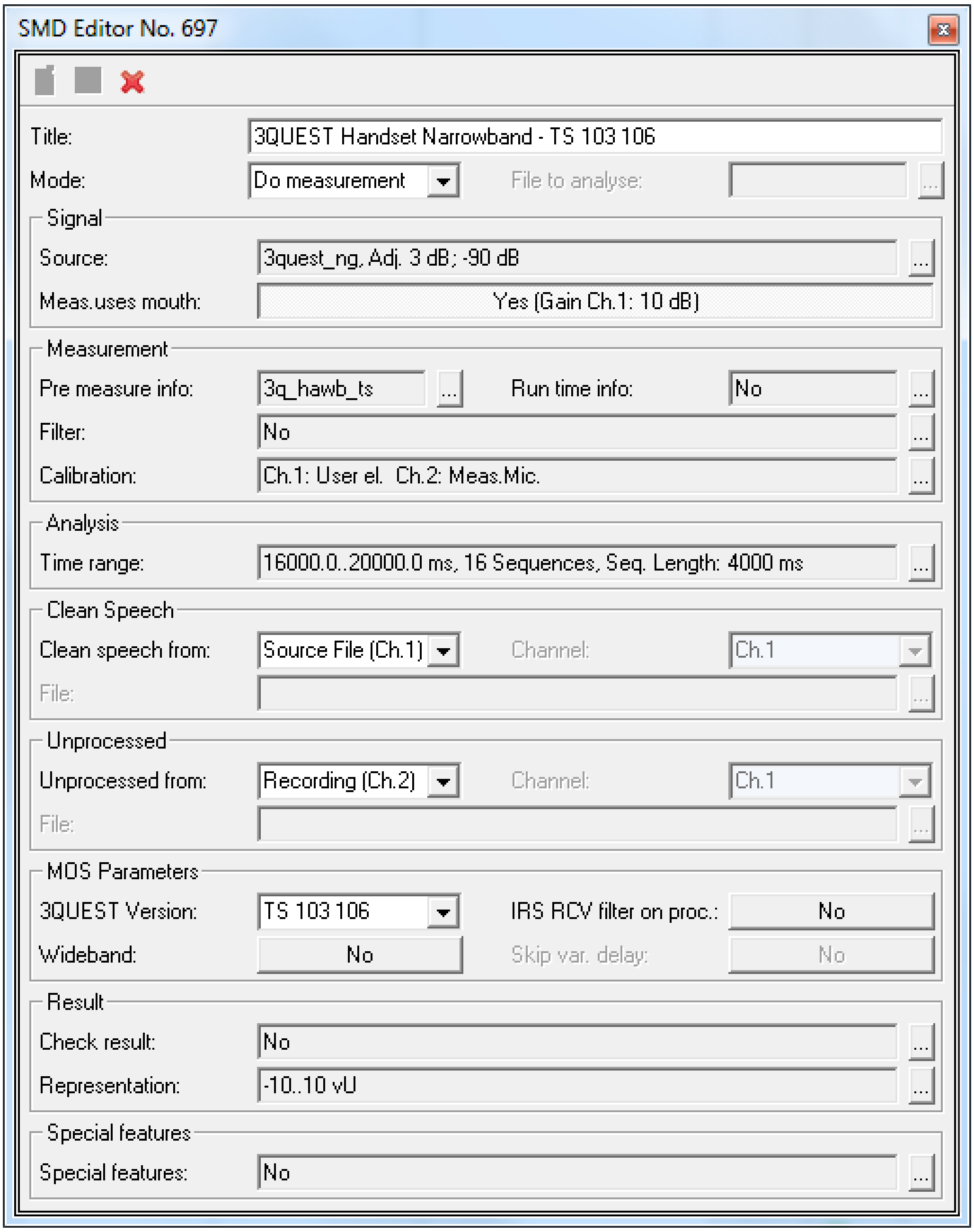


Figure 2: ACQUA setup for 3QUEST tests

3.1. Measurement with ACQUA

Figure 3 demonstrates a sample SMD (Single Measurement Descriptor) of a 3QUEST test for a narrow band application according to TS 103 106. The main content of the SMD entries is explained on the next pages.



SMD Editor No. 697

Title: 3QUEST Handset Narrowband - TS 103 106

Mode: Do measurement — File to analyse:

Signal

Source: 3quest_ng, Adj. 3 dB; -90 dB

Meas.uses mouth: Yes (Gain Ch.1: 10 dB)

Measurement

Pre measure info: 3q_hawb_ts — Run time info: No

Filter: No

Calibration: Ch.1: User el. Ch.2: Meas.Mic.

Analysis

Time range: 16000.0..20000.0 ms, 16 Sequences, Seq. Length: 4000 ms

Clean Speech

Clean speech from: Source File (Ch.1) — Channel: Ch.1

File:

Unprocessed

Unprocessed from: Recording (Ch.2) — Channel: Ch.1

File:

MOS Parameters

3QUEST Version: TS 103 106 — IRS RCV filter on proc.: No

Wideband: No — Skip var. delay: No

Result

Check result: No

Representation: -10..10 vU

Special features

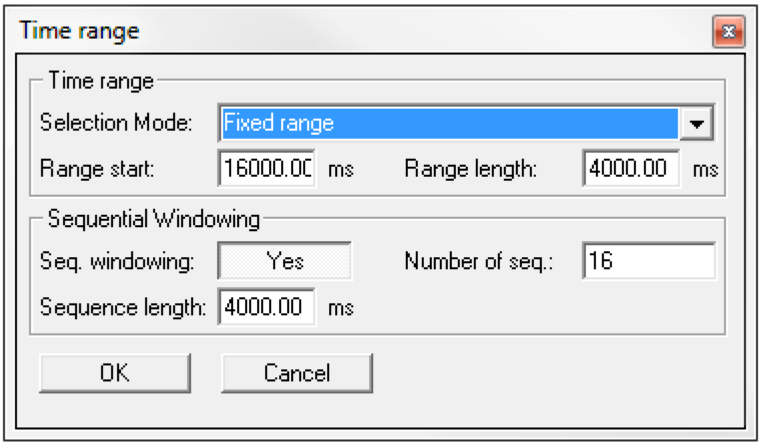
Special features: No

Figure 3: 3QUEST SMD with support of TS 103 106

- **Source**: Although the SMD allows entering all kinds of stimulus, it is strongly recommended to solely operate TS 103 106 tests by using speech according to the so-called *Dynastat speech material* introduced in [5]. Speech samples are made available within the framework of ACOPT 21. As the signal is transmitted via the HATS mouth, it is always a full band signal, thus there is no difference between wide- or narrowband setups with respect to the source signal.

- **Time range** needs to be set according to the source signal (i.e. used adaptation sequences need to be taken into account, see SMD entry *Source*).

Special attention has to be drawn to the *Sequential Windowing* setup of the analysis: The measured time signal can either be analyzed as entire sequence or sentence by sentence of the source file. The final 3QUEST MOS score is then calculated as arithmetic mean of these single speech sequences. The SMD settings are shown by figure 4.



Time range

Time range

Selection Mode: Fixed range

Range start: 16000.0C ms Range length: 4000.00 ms

Sequential Windowing

Seq. windowing: Yes Number of seq.: 16

Sequence length: 4000.00 ms

Figure 4: Submenu *Time range* of 3QUEST SMD

When analyzing via *Sequential Windowing* it needs to be ensured that the *Time range* entries fit to the source signal: Figure 5 shows the used 3QUEST source signal and its *Range length*, *Range start* and *Sequence length*. In this specific case, *Range length* is identical to *Sequence length*, however in general, *Range length* could be shorter than *Sequence length*, depending on the used source signal.

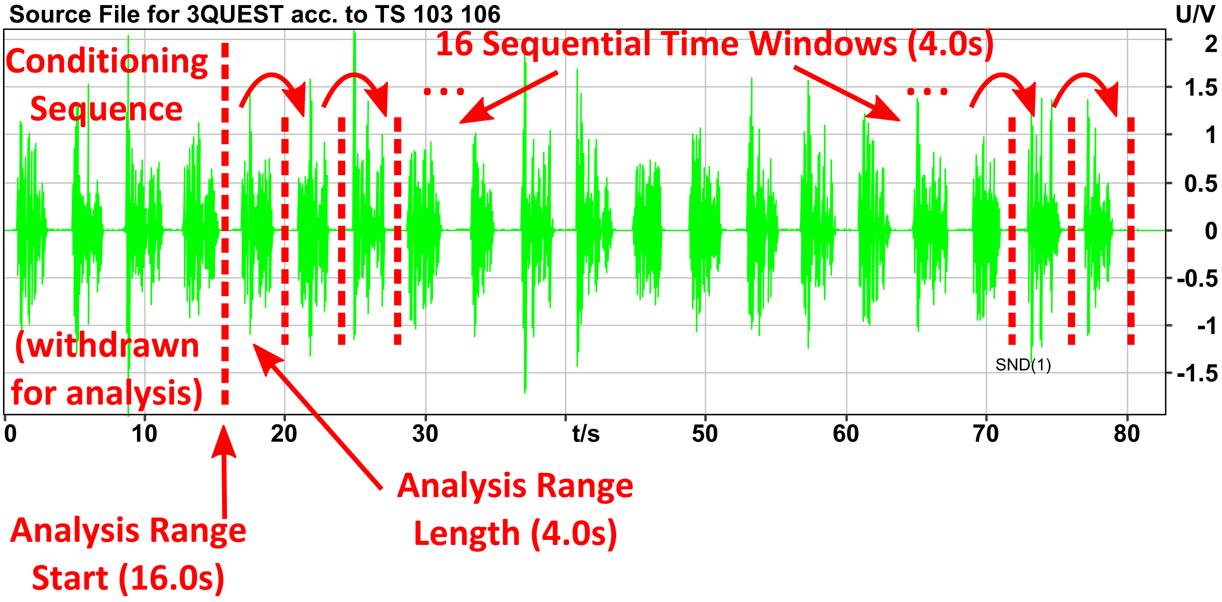
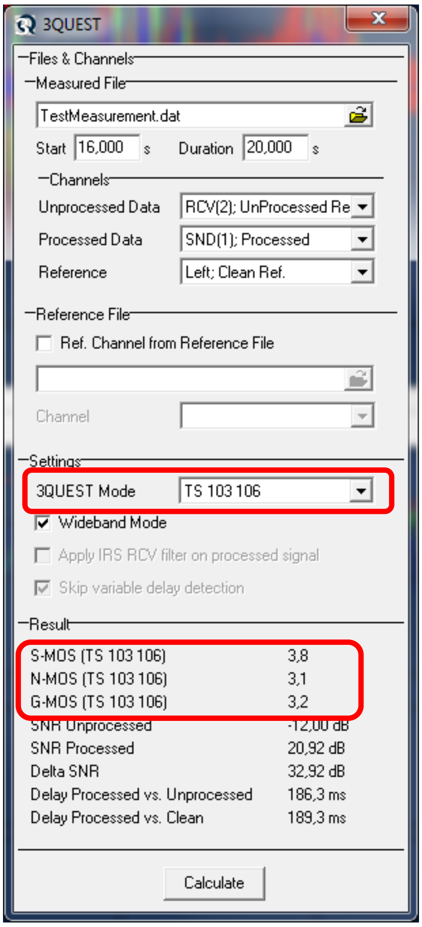


Figure 5: New source file for 3QUEST measurements according to TS 103 106

3.2. Manual Analysis in ACQUAlyzer

Also the manual offline analysis in the ACQUAlyzer (menu *Calculation → 3QUEST*) was modified in version 3.1.200 in order to calculate 3QUEST scores according to EG 202 396-3 and TS 103 106.



Files & Channels

Measured File

TestMeasurement.dat

Start 16,000 s Duration 20,000 s

Channels

Unprocessed Data: RCV(2); UnProcessed Re

Processed Data: SND(1); Processed

Reference: Left; Clean Ref.

Reference File

☐ Ref. Channel from Reference File

Channel

Settings

3QUEST Mode: TS 103 106

☑ Wideband Mode

☐ Apply IRS RCV filter on processed signal

☑ Skip variable delay detection

Result

S-MOS (TS 103 106)	3,8
N-MOS (TS 103 106)	3,1
G-MOS (TS 103 106)	3,2
SNR Unprocessed	-12,00 dB
SNR Processed	20,92 dB
Delta SNR	32,92 dB
Delay Processed vs. Unprocessed	186,3 ms
Delay Processed vs. Clean	189,3 ms

Figure 6: New version of post-analysis

Note that in EG 202 396-3 always the full sequence must be analyzed and yields one set of MOS values. In TS 103 106, the analysis is always performed per sentence and the overall MOS values have to be averaged over all sentences.

Nevertheless, calculating 3QUEST TS over multiple sentences at once also gives results, but these are not determined according to the standards and have not been validated.

4. Comparison Tests

This chapter shortly describes the differences between the classical calculation method according to EG 202 396-3 versus the new one according to TS 103 106. In order to demonstrate typical deviations between the EG 202 396-3 and TS 103 106 method, the following chapter presents results of four common use cases.

4.1. General Remarks

As already described in chapter 3, the new source file which must be used for the new method according to TS 103 106 differs from the classical sequence which was used for EG 202 396-3. Strictly spoken, existing recordings from the EG method cannot be used for the calculation of MOS scores according to the TS method.

However, in several internal studies, it was shown that the speech material for the EG 202 396-3 method yields results which are very close to the ones obtained with the new TS source file. The results are highly correlated when segmenting the old sequence in parts of four seconds so that each sentence is analyzed separately, as it is usually applied in the new TS method.

With this procedure, a comparison between 3QUEST EG and TS can be made by re-using old measurements. A huge amount of state-of-the-art mobile devices were tested in several bandwidth and operational modes.

Note: The following investigations do *not* show the relation between subjective and objective data and is thus not an indicator for the prediction performance of one of the methods!

4.2. Narrowband Handset Tests

For the evaluation of the narrowband mode, 76 existing recordings of 19 mobile devices in handset mode were analyzed with both methods. The four background noises Car, Mensa, Train Station and Road were used for each device.

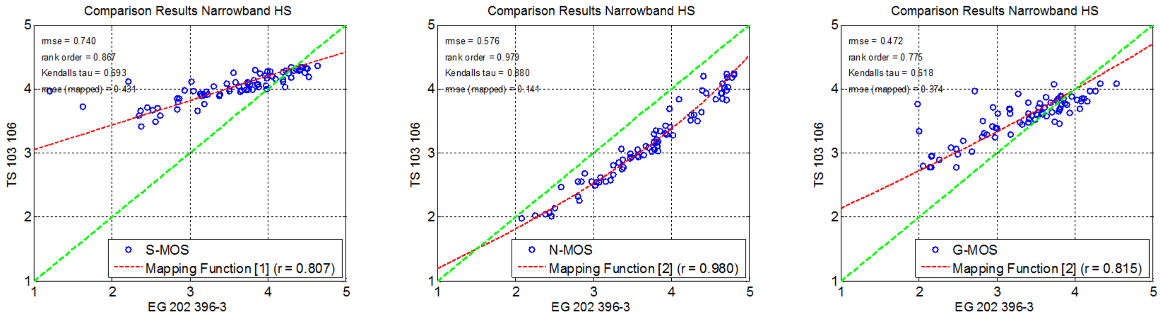


Figure 7: Comparison of S-/N-/G-MOS of EG 202 396-3 vs TS 103 106 (NB HS)

4.3. Narrowband Hands-free Tests

For the evaluation of the narrowband mode with lower SNR conditions, 24 existing recordings of 6 mobile devices in handheld mode were analyzed with both methods. The four background noises Car, Mensa, Train Station and Road were used for each device. Additionally, 16 car hands-free recordings with individual combinations of DUT and car noise were included in the analysis[1].

[1] Note: The training and validation procedure of ETSI TS 103 106 did not include car hands-free applications. The usage of this method for this usage is still under study.

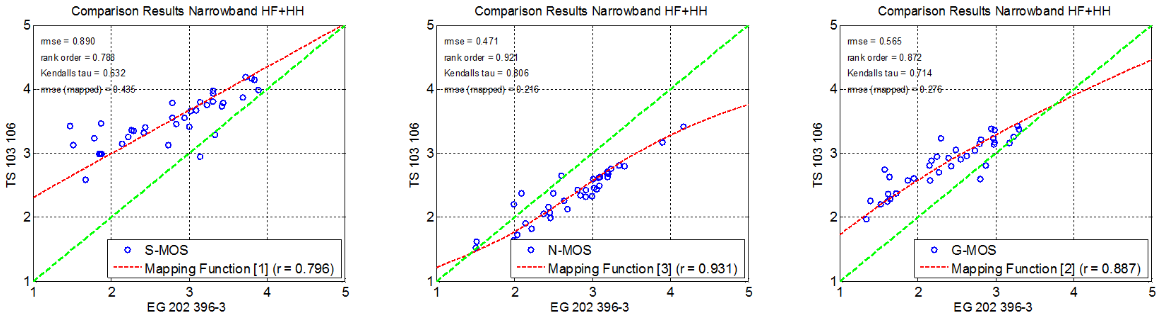


Figure 8: Comparison of S-/N-/G-MOS of EG 202 396-3 vs TS 103 106 (NB HH/HF)

4.4. Wideband Handset Tests

For the evaluation of the wideband mode, 168 existing recordings of 42 mobile devices in handset mode were analyzed with both methods. The four background noises Car, Mensa, Train Station and Road were used for each device.

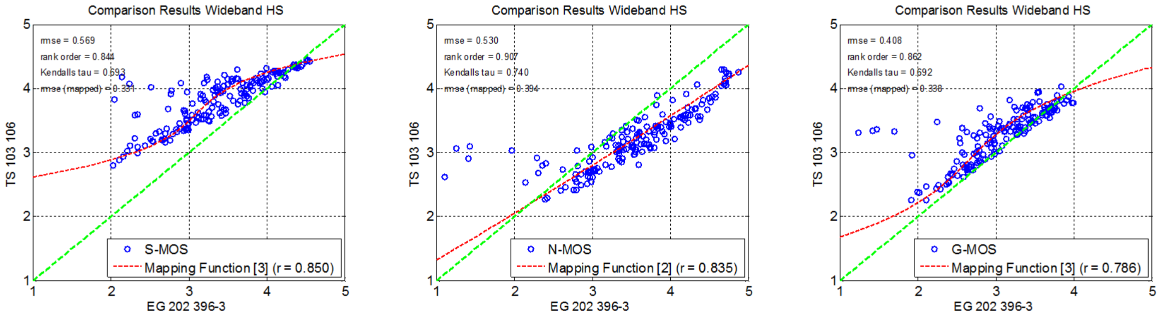


Figure 9: Comparison of S-/N-/G-MOS of EG 202 396-3 vs TS 103 106 (WB HS)

4.5. Wideband Hands-free Tests

For the evaluation of the wideband mode with lower SNR conditions, 72 existing recordings of 18 mobile devices in handheld mode were analyzed with both methods. The four background noises Car, Mensa, Train Station and Road were used for each device.

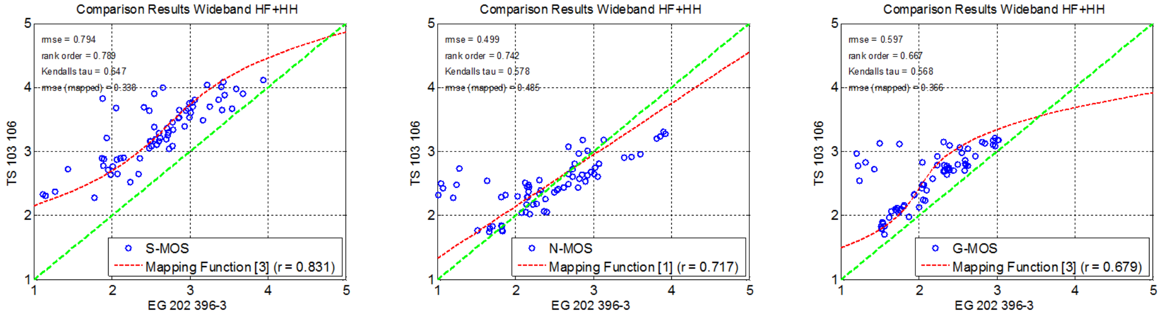


Figure 10: Comparison of S-/N-/G-MOS of EG 202 396-3 vs TS 103 106 (WB HH)

4.6. Statistical Evaluation of Comparison

A typical application of the prediction of S-, N- and G-MOS is the measurement procedure in standards or specifications. Here, certain minimum requirements often have to be fulfilled in order to pass the whole testing. In the past, several specifications and measurement standards included such requirements for the EG method. Either only the G-MOS has to pass a certain value or all three values for speech, noise and global quality have to reach specific thresholds.

Of course, the new TS method can also be used for the specification of minimum requirements. When changing from EG to TS method, the possibly existing requirement values have to be adapted. It is not recommended to just take over these values.

The finding process of requirement values usually aims at passing a certain amount of devices through the test procedure. To find such threshold values, it is a conventional way to calculate a certain percentile of the S-, N- and G-MOS scores of a large test series. With this approach, for example the minimum requirement can be defined as the top 33% percent of a series of modern state-of-the-art devices (66% percentile or 3rd tercile).

In order to provide an overview over the shifts between the EG and TS method, some percentiles p=33% (lower third), p=50% (median) and p=66% (top 33%) of the results presented in sections 4.2 to 4.5 are given in table 8 to table 10. These percentiles should represent rough estimations for three quality classes "good/excellent", "fair" and "poor/bad".

The differences between these metrics are also given; green cells indicate that the TS method yields more optimistic results, red cells indicate that the values for the EG method are higher.

3QUEST EG			
Mode	S-MOS p=33%	S-MOS p=50%	S-MOS p=66%
NB HS	3,22	3,73	3,98
NB HH/HF	2,34	2,89	3,17
WB HS	3,12	3,40	3,71
WB HH/HF	2,48	2,64	2,86
3QUEST TS			
Mode	S-MOS p=33%	S-MOS p=50%	S-MOS p=66%
NB HS	3,88	4,04	4,24
NB HH/HF	3,41	3,59	3,77
WB HS	3,71	3,97	4,12
WB HH/HF	2,97	3,40	3,71
3QUEST Delta TS – EG			
Mode	S-MOS p=33%	S-MOS p=50%	S-MOS p=66%
NB HS	0,65	0,31	0,26
NB HH/HF	1,07	0,70	0,60
WB HS	0,58	0,57	0,41
WB HH/HF	0,49	0,76	0,84

Table 8: Statistical metrics and difference metrics for test series (S-MOS)

Differences for S-MOS for all percentiles in table 8 are significantly larger than zero, which implies that the TS method obtains much more optimistic values. Especially for the hands-free / handheld conditions for NB and WB, the median and the top 33% class increased up to 0.7 MOS. The threshold for the lower 33% class is also increased (up to 1.0 MOS), which indicates that S-MOS is rated constantly higher now with the TS method over all categories of percentiles and operational modes.

3QUEST EG			
Mode	N-MOS p=33%	N-MOS p=50%	N-MOS p=66%
NB HS	3,38	3,78	3,97
NB HH/HF	2,48	2,88	3,06
WB HS	3,33	3,56	3,83
WB HH/HF	2,15	2,35	2,73
3QUEST TS			
Mode	N-MOS p=33%	N-MOS p=50%	N-MOS p=66%
NB HS	2,88	3,17	3,48
NB HH/HF	2,17	2,40	2,63
WB HS	2,97	3,24	3,44
WB HH/HF	2,30	2,47	2,67
3QUEST Delta TS - EG			
Mode	N-MOS p=33%	N-MOS p=50%	N-MOS p=66%
NB HS	-0,50	-0,61	-0,48
NB HH/HF	-0,30	-0,48	-0,44
WB HS	-0,36	-0,32	-0,39
WB HH/HF	0,15	0,12	-0,07

Table 9: Statistical metrics and difference metrics for test series (N-MOS)

In contrast, the S-MOS and the N-MOS values obtained in table 9 with the TS method are significantly lower than for the EG method. Only for the wideband handheld/hands-free class, a slight positive delta of 0.12-0.15 MOS is reached, which is in the range of reproduction precision. The top 33% of N-MOS for the TS method are about 0.5 MOS lower than for the EG score, similar to the median. The differences in the NB mode over the percentile classes are located in a tight range (0.3-0.6 MOS) which indicates that the scores have almost just a constant offset.

3QUEST EG			
Mode	G-MOS p=33%	G-MOS p=50%	G-MOS p=66%
NB HS	3,10	3,50	3,74
NB HH/HF	2,16	2,41	2,80
WB HS	2,81	3,11	3,38
WB HH/HF	1,80	2,09	2,37
3QUEST TS			
Mode	G-MOS p=33%	G-MOS p=50%	G-MOS p=66%
NB HS	3,43	3,62	3,78
NB HH/HF	2,70	2,96	3,13
WB HS	3,18	3,40	3,56
WB HH/HF	2,34	2,64	2,91
3QUEST Delta TS - EG			
Mode	G-MOS p=33%	G-MOS p=50%	G-MOS p=66%
NB HS	0,33	0,12	0,04
NB HH/HF	0,54	0,55	0,33
WB HS	0,37	0,30	0,18
WB HH/HF	0,53	0,55	0,55

Table 10: Statistical metrics and difference metrics for test series (G-MOS)

The large deviations between S- and N-MOS are compensated to a certain extent for the G-MOS prediction. As a result, there still is a positive offset, which means that the TS method always gives more optimistic scores for G-MOS. The values in table 10 show larger differences for hands-free / handheld mode for all classes. To classify the top 33% class for NB and WB handset devices, the differences finally are quite similar (deviation 0.04-0.18 MOS).

4.7. Conclusions

It is difficult to describe the expected behavior of the analyses presented above, because in general two different prediction algorithms are compared against each other. Thus, the plots in figure 7 to figure 10 do not show a good correlation between the EG and TS method.

This is not a surprising result because the TS method was completely retrained to several new databases. In case of a rather strong correlation, there would have been no need for a new standard because also the new results could be reached just by applying a 3rd order mapping function on the EG data.

Nevertheless, the plots show that on the one hand the TS method is a completely new algorithm, which does not necessarily lead to the same prediction results as the EG method. But on the other hand, it shows at least some similarities with the EG method in some quality ranges and test cases.

In consequence, it is even hard to define a rule of thumb which describes the shifts between the TS and EG method: In average, S-MOS is rated significantly higher, N-MOS significantly lower; G-

MOS is also higher than the EG prediction but does not fully compensate its contrary components S- and N-MOS.

In general, the differences between the two methods can be explained by the underlying listening test databases:

- For the EG method, one huge subjective database was used for the training and validation procedure for each NB and WB mode. A subset of conditions (50 in NB mode, 79 in WB mode) was not passed into the training (216 in NB mode, 179 in WB mode) and was used for validation. The prediction of S-, N- and G-MOS thus is within the bounds and quality range of exactly this one database.

- The way for the new TS method was different: According to the test plan described in [6], multiple smaller databases with 60 conditions each were created. The training procedure was applied over several databases as well as the validation. The prediction of S-, N- and G-MOS thus is a kind of average over these databases and results have eventually been mapped (e.g. with a 3rd order mapping function) to fit for several applications.

- Additionally, the background of the EG and TS databases must be considered. The databases for the EG method were created with German (NB) and French (WB) listeners while the TS databases mainly were created in the US with American listeners. Investigations described in [13] revealed that performing an identical listening test once with German and once with American listeners may lead to quite different results. These cultural differences must also be taken into account when comparing EG and TS prediction scores.

5. Summary

This application note gives an overview about the new 3QUEST operational mode according to ETSI TS 103 106 and the differences to the existing method according to ETSI EG 202 396-3. A brief summary of the algorithm modifications is presented and an insight about the development phase and the standardization progress is given in chapter 2.

The updates and new features within the HEAD Analyzer ACQUA program are shown in chapter 3. The measurement of scores according to the new TS method can be conducted with the same set-up and measurement equipment as for the EG method. Only slight modifications in the SMD and the source file have to be applied.

A series of comparison tests between the EG and TS method is presented in chapter 4. It was shown that the methods themselves represent their own listening test databases and are not comparable against each other in general. The choice which method to use for which application also certainly depends on the device to be tested (e.g. 2/3G-, LTE mobile terminals should be tested according to TS 103 106 – VoIP phones should still be tested according to EG 202 396-3). But the choice might also depend on the market (e.g. European vs. American) the device is targeting at.

Another application of EG 202 396-3 still is the testing of car hands-free devices. Up to now, only the EG method is validated for this application, the usability of the new TS 103 106 method in this field is still under study.

6. References

[1] ETSI EG 202 396-3: “Speech Quality performance in the presence of background noise - Part 3: Background noise transmission - Objective test methods”, (02/2011).

[2] ITU-T Recommendation P.835: “Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm”, (11/2003)

[3] Audience Inc., “Correlation of ETSI EG 202 396-3 Objective Speech Quality Measures to P.835 Subjective Test Results”, Tdoc S4-110596, TSG SA WG4 Meeting #65, 15-19 August 2011, Kista, Sweden.

[4] “Adaption of a Prediction Model for Noisy Speech Quality Assessment”, J. Reimes, H.W. Gierlich, G. Mauer, IWAENC 2012, Aachen

[5] ETSI TS 103 106: “Speech quality performance in the presence of background noise: Background noise transmission for mobile terminals-objective test methods”, (02/2013)

[6] “Common subjective testing framework for training of P.835 test predictors”, 3GPP document S4-120542

[7] ITU-T Rec. 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”.

[8] ITU-T Rec. P.830: “Subjective performance assessment of telephone-band and wideband digital codecs”

[9] T. Hastie, R. Tibshirani, J. Friedman: “The Elements of Statistical Learning: Data Mining, Inference, and Prediction”. New York: Springer-Verlag, 2001.

[10] ITU-T Rec. G.191, “Software tools for speech and audio coding standardization“

[11] ITU-T Rec. P.1401 (ex. P.STAT), “Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models”

[12] Orange, Qualcomm Incorporated, "Ext_ATS Permanent document (EATS-6): Requirements for the objective P.835 predictor model performance and P.835 database collection", Tdoc S4-120811, TSG-SA4#69 meeting, 21 – 26 May, 2012, Erlangen, Germany

[13] “Cultural Differences of Speech Quality Perception in the Presence of Background Noise”, H.W. Gierlich, J. Reimes, G. Mauer, DAGA 2012, Darmstadt.