ETSI TS 103 106 V1.5.1 (2018-04)

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-271

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

Important notice

The present document can be downloaded from:
http://www.etsi.org/standards-search

The present document may be made available in electronic versions and/or in print. The content of any electronic and/or print versions of the present document shall not be modified without the prior written authorization of ETSI. In case of any existing or perceived difference in contents between such versions and/or in print, the only prevailing document is the print of the Portable Document Format (PDF) version kept on a specific network drive within ETSI Secretariat.

Users of the present document should be aware that the document may be subject to revision or change of status. Information on the current status of this and other ETSI documents is available at
https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

If you find errors in the present document, please send your comment to one of the following services:
https://portal.etsi.org/People/CommiteeSupportStaff.aspx

Contents

Intellectual Property Rights

Essential patents

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

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

Trademarks

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

Foreword

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

The present document is to be used in conjunction with the ETSI ES 202 396-1 [i.2] and ETSI EG 202 396-3 [i.4]:

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

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 202 396-3 [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] Void.

[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+) (GSM); 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] Void.

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

[i.12] Void.

[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)".

[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:

78KBP	7,8 kHz band-pass according to Recommendation ITU-T G.191
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
EATS	Enhanced Acoustic Test Specification
EXP	Experiment
FB	Fullband
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
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
NTT	Nippon Telegraph and Telephone
OVRL	Overall (speech + noise) Component
PRO	Professional
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
NOTE:	In case of clipping is generated for condition i02 and/or i12, a different signal scaling is recommended (e.g. 73 dB_{SPL} refers to -36 instead of -26 dBov).		

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:

$$\underbrace{\begin{pmatrix} 1 & S_{i01} & N_{i01} & G_{i12} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & S_{i12} & N_{i12} & G_{i12} \end{pmatrix}}_{R_j(Ref.\ set\ j)} \times T_j = \underbrace{\begin{pmatrix} \overline{S_{i01}} & \overline{N_{i01}} & \overline{G_{i12}} \\ \vdots & \vdots & \vdots \\ \overline{S_{i12}} & \overline{N_{i12}} & \overline{G_{i12}} \end{pmatrix}}_{A\ (Avg.\ ref.\ set)} \tag{5.1}$$

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

$$T_j = \left({R_j}^T \times R_j\right)^{-1} \times {R_j}^T \times A \tag{5.2}$$

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:

$$\underbrace{\begin{pmatrix} 1 & S_1 & N_1 & G_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & S_N & N_N & G_N \end{pmatrix}}_{S_j\ (scores\ of\ samples\ of\ database\ j)} \times T_j = \underbrace{\begin{pmatrix} \tilde{S}_1 & \tilde{N}_1 & \tilde{G}_1 \\ \vdots & \vdots & \vdots \\ \tilde{S}_N & \tilde{N}_N & \tilde{G}_N \end{pmatrix}}_{\tilde{S}_j\ (transformed\ scores\ of\ samples\ of\ database\ j)} \tag{5.3}$$

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 Void

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 databases of the wideband mode.

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 analysed sequence should be normalized to 73 dB SPL/-21 dB Pa (for the listening test and the algorithm) as described in ETSI EG 202 396-3 [i.4].

6.4 Modified neural network 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-MOS (and N-MOS as well). The parameters are shown in tables 2 and 3. A detailed description of the calculation for the parameters can be found in ETSI EG 202 396-3 [i.4].

Table 2: Extracted parameters for N-MOS

P_0	$N_{BGN,P}$	P_3	$\sigma(\Delta RA_{BGN,P\text{-}U})$
P_1	$\sigma(RA_{BGN,U})$	P_4	$\mu(RA_{BGN,U})$
P_2	$\sigma(RA_{BGN,P})$	P_5	$\mu(RA_{BGN,P})$

Table 3: Extracted Parameters for S-MOS

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

The calculation of the objective S-MOS in clause 6.5.2 of ETSI EG 202 396-3 [i.4] is performed with the 6 parameters of table 3 in conjunction with a neural network.

Figure 1: Void

Several vectors and matrices are necessary to implement the neural network with regard to the underlying listening test database. For the present document, the training databases differ from the ones described in ETSI EG 202 396-3 [i.4] are provided in the following equations.

For the normalization of the input parameters, different average and standard deviation vectors M_{in} and S_{in} for narrow- and wideband mode are necessary. For wideband, the vectors are provided in equation 6.1, for narrowband in equation 6.2.

$$M_{in,WB} = (0{,}0 \quad 12{,}7309 \quad 4{,}2076 \; -1{,}2456 \quad 0{,}8834 \quad 12{,}25222 \quad 7{,}0541) \qquad (6.1)$$

$$S_{in,WB} = (1{,}0 \quad 11{,}8503 \quad 1{,}2824 \; -0{,}3124 \quad 0{,}2511 \quad 6{,}7091 \quad 5{,}\mathrm{s}) \qquad (6.2)$$

For narrowband, the corresponding input normalization vector are given by equations 6.3 and 6.4.

$$M_{in,NB} = (0{,}0 \quad 13{,}7519 \quad 2{,}0884 \; -0{,}3124 \quad 0{,}2511 \quad 6{,}7091 \quad 5{,}2951) \qquad (6.3)$$

$$S_{in,NB} = (1{,}0 \quad 11{,}4341 \quad 0{,}4047 \quad 0{,}3877 \quad 0{,}3309 \quad 3{,}1189 \quad 2{,}5976) \qquad (6.4)$$

As described in ETSI EG 202 396-3 [i.4], 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. For the present document, the updated matrices are provided in equations 6.5 for wideband and 6.6 for narrowband.

$$\mathbf{H_{WB}} = \begin{pmatrix} -0,4336 & -0,9873 & 0,0091 & -0,0845 & 0,0203 \\ 0,1141 & -0,0004 & -0,7133 & -0,2798 & -1,8189 \\ 1,0265 & 0,5001 & 0,5120 & 0,0537 & 0,1265 \\ -0,8627 & -1,7518 & -0,0374 & -0,2908 & 0,3064 \\ 2,1381 & 0,4190 & 1,0715 & -1,6716 & 0,4973 \\ -1,3933 & 0,5972 & 0,0852 & 0,1977 & 0,2222 \\ -0,3793 & -1,7785 & -0,5306 & -1,7538 & -2,9630 \end{pmatrix} \quad (6.5)$$

$$\mathbf{H_{NB}} = \begin{pmatrix} -0,3608 & -0,3805 & 0,5359 & -1,1131 & -0,1322 \\ 0,7348 & -4,4639 & -1,2552 & 0,3338 & 0,5452 \\ 0,9117 & 2,7177 & 0,8876 & 0,1712 & -2,1279 \\ -0,2383 & 1,7228 & -0,0354 & -1,0284 & 1,0483 \\ 1,4511 & 2,1467 & 1,0010 & 0,7356 & 0,1154 \\ -0,5573 & -0,6137 & -0,2648 & 1,6202 & 0,5966 \\ -3,2194 & -7,9575 & -0,7736 & -0,8676 & 0,1663 \end{pmatrix} \quad (6.6)$$

The five transformed output values of the hidden layer are then passed 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. For the present document, the updated vectors are provided in equations 6.7 for wideband and 6.8 for narrowband.

$$O_{WB} = (0,1777 \ \ -0.2835 \ \ -0,3147 \ \ 0,1837 \ \ -0,3237) \quad (6.7)$$

$$O_{NB} = (0,3832 \ \ -0.5250 \ \ -0,1878 \ \ -0,2674 \ \ -0,1548) \quad (6.8)$$

With these modifications described above, instrumental assessment of S-MOS is completely defined for the context of the present document.

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 6.9:

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

(6.9)

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 6.10:

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

(6.10)

The calculation steps for N-MOS and G-MOS are not modified, only the coefficients for the linear regressions according to equations 6.9 and 6.10 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.