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

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

Reference

RTS/STQ-269

Keywords

noise, quality, speech, testing, transmission

ETSI

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

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

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

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 prevailing version of an ETSI deliverable is the one made publicly available in PDF format at www.etsi.org/deliver.

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:

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

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

Modal verbs terminology

In the present document "**shall**", "**shall not**", "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

1 Scope

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

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

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

The present document is to be used in conjunction with:

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

2 References

2.1 Normative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

Referenced documents which are not found to be publicly available in the expected location might be found at https://docbox.etsi.org/Reference/.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

The following referenced documents are necessary for the application of the present document.

Not applicable.

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

[i.1] ETSI ES 202 396-1: "Speech and multimedia Transmission Quality (STQ); Speech quality performance in the presence of background noise; Part 1: Background noise simulation technique and background noise database".

[i.2] ETSI EG 202 396-3: "Speech and multimedia Transmission Quality (STQ); Speech Quality performance in the presence of background noise Part 3: Background noise transmission - Objective test methods".

[i.3] ETSI TS 103 106: "Speech and multimedia Transmission Quality (STQ); Speech quality performance in the presence of background noise: Background noise transmission for mobile terminals-objective test methods".

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3 Definition of terms, symbols and abbreviations

3.1 Terms

Void.

3.2 Symbols

Void.

3.3 Abbreviations

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

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

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

4 Introduction

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

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

5 Underlying speech databases and preparations

The basis of any perceptually-based measure which models the behaviour of human test persons, are auditory tests. In general, these tests are carried out with naïve test persons, who are asked to rate a certain quality aspect of a presented speech sample. For the evaluation of processed and transmitted noisy speech, the Recommendation ITU-T P.835 [i.5] is a state-of-the-art method for the assessment of speech and noise quality. The listening test procedure described in [i.5] is also the basis for the prediction model.

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

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

6 Model descriptions

6.1 Introduction

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

6.2 Common definitions

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

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

6.3 Model A

6.3.1 Introduction

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

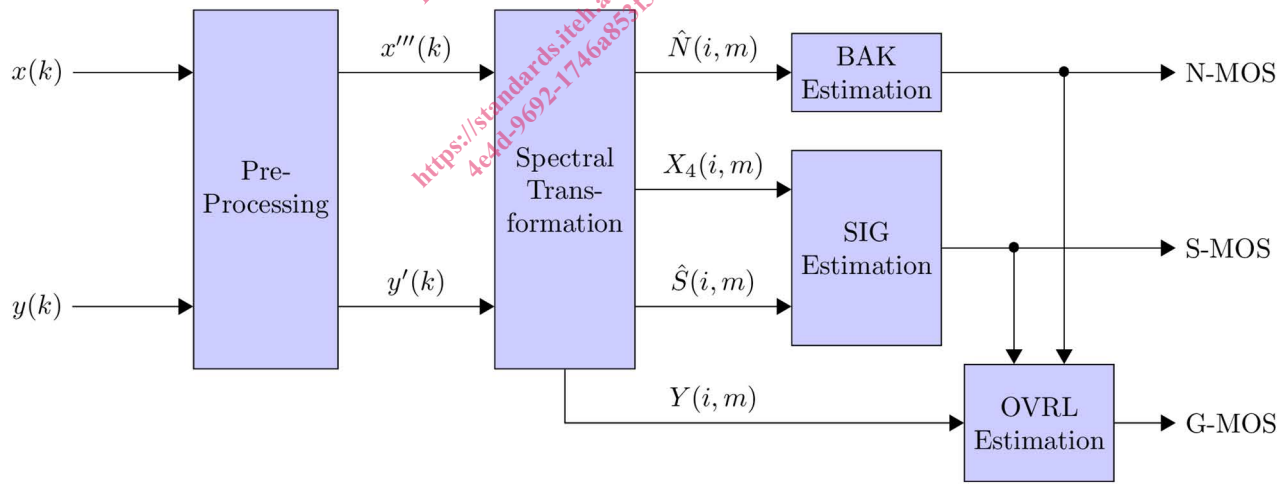

Figure 6.1: Block diagram of model A

6.3.2 Pre-Processing

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

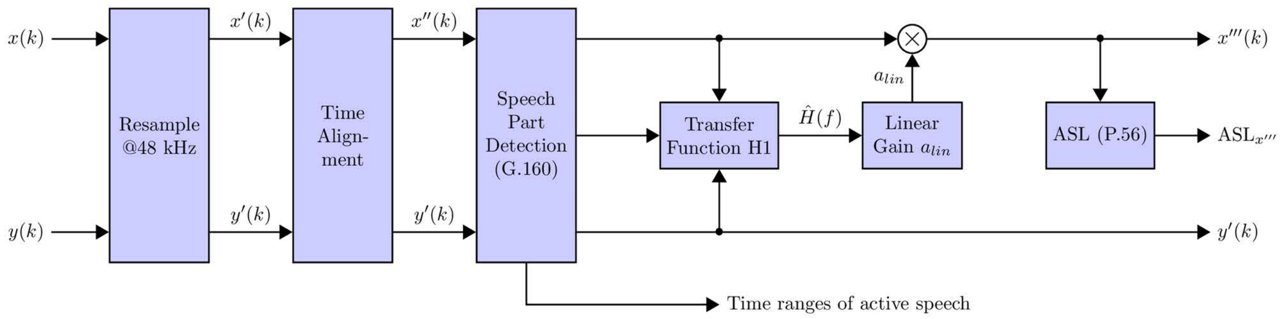

Figure 6.2: Block diagram of model A

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

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

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

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

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

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

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

The next block extracts the active speech parts from the signal x"(k). For this analysis, the first step is to classify energy frames of 10 ms (block-wise, no overlap) according to the method described in [i.9]. The thresholds for the classification are defined relatively to the active speech level [i.7]. As a result, each speech frame is identified either as high (H), medium (M), low (L) or uncertain (U) activity. Frames without activity are either classified as short pauses (P) or silence (S). The speech parts are finally determined as regions excluding frames of type S. The information of the active time ranges is employed in several other blocks which are introduced in the following clauses.

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

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

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