Final draft ETSI EG 202 396-3 V1.2.1 (2008-11)

ETSI Guide

Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality performance in the presence of background noise Part 3: Background noise transmission - Objective test methods

Reference
REG/STQ-00124

Keywords
noise, QoS, quality, speech

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

Individual copies of the present document can be downloaded from:
http://www.etsi.org

The present document may be made available in more than one electronic version or in print. In any case of existing or perceived difference in contents between such versions, the reference version is the Portable Document Format (PDF). In case of dispute, the reference shall be the printing on ETSI printers of the 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
http://portal.etsi.org/tb/status/status.asp

If you find errors in the present document, please send your comment to one of the following services:
http://portal.etsi.org/chaircor/ETSI_support.asp

Contents

Intellectual Property Rights

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

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

Foreword

This ETSI Guide (EG) has been produced by ETSI Technical Committee Speech Processing, Transmission and Quality Aspects (STQ), and is now submitted for the ETSI standards Membership Approval Procedure.

The present document is a deliverable of ETSI Specialized Task Force (STF) 294 entitled: "Improving the quality of eEurope wideband speech applications by developing a performance testing and evaluation methodology for background noise transmission".

The present document is part 3 of a multi-part deliverable covering speech quality performance in the presence of background noise, as identified below:

Part 1: "Background noise simulation technique and background noise database";

Part 2: "Background noise transmission - Network simulation - Subjective test database and results";

Part 3: "Background noise transmission - Objective test methods".

1 Scope

The present document aims to identify and define testing methodologies which can be used to objectively evaluate the performance of narrowband and wideband terminals and systems for speech communication in the presence of background noise.

Background noise is a problem in mostly all situations and conditions and need 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 in the presence of background noise. The present document includes:

- The description of the experts post evaluation process chosen to select the subjective test data being within the scope of the objective methods.
- The results of the performance evaluation of the currently existing methods described in ITU-T Recommendation P.862 [i.16], [i.17] and in TOSQA2001 [i.19] which is chosen for the evaluation of terminals in the framework of ETSI VoIP speech quality test events [i.8], [i.9], [i.10] and [i.11].
- 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 document is to be used in conjunction with:
  - EG 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.
  - EG 202 396-2 [i.2] which describes the simulation of network impairments and how to simulate realistic transmission network scenarios and which contains the methodology and results of the subjective scoring for the data forming the basis of the present document.
  - French speech sentences as defined in ITU-T Recommendation P.501 [i.13] for wideband and English speech sentences as defined in ITU-T Recommendation P.501 [i.13] for narrowband.

2 References

References are either specific (identified by date of publication and/or edition number or version number) or non-specific.

- For a specific reference, subsequent revisions do not apply.
- Non-specific reference may be made only to a complete document or a part thereof and only in the following cases:
  - if it is accepted that it will be possible to use all future changes of the referenced document for the purposes of the referring document;
  - for informative references.

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

For online referenced documents, information sufficient to identify and locate the source shall be provided. Preferably, the primary source of the referenced document should be cited, in order to ensure traceability. Furthermore, the reference should, as far as possible, remain valid for the expected life of the document. The reference shall include the method of access to the referenced document and the full network address, with the same punctuation and use of upper case and lower case letters.

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

2.1 Normative references

The following referenced documents are indispensable for the application of the present document. For dated references, only the edition cited applies. For non-specific references, the latest edition of the referenced document (including any amendments) applies.

Not applicable.

2.2 Informative references

The following referenced documents are not essential to the use of the present document but they assist the user with regard to a particular subject area**.** For non-specific references, the latest version of the referenced document (including any amendments) applies.

[i.1] ETSI EG 202 396-1: "Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality performance in the presence of background noise; Part 1: Background Noise Simulation Technique and Background Noise Database".

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

[i.3] ITU-T Recommendation P.835: "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm".

[i.4] ITU-T Recommendation P.800: "Methods for subjective determination of transmission quality".

[i.5] ITU-T Recommendation P.831: "Subjective performance evaluation of network echo cancellers".

[i.6] Genuit, K.: "Objective Evaluation of Acoustic Quality Based on a Relative Approach", InterNoise '96, Liverpool, UK.

[i.7] ITU-T Recommendation SG 12 Contribution 34: "Evaluation of the quality of background noise transmission using the "Relative Approach"".

[i.8] ETSI 2nd Speech Quality Test Event: "Anonymized Test Report", ETSI Plugtests, HEAD acoustics, T-Systems Nova.

NOTE: Available at: http://www.etsi.org/plugtests/History/2002VOIP2.htm. Also available as ETSI TR 102 648-3.

[i.9] ETSI 3rd Speech Quality Test Event: "Anonymized Test Report "IP Gateways"".

NOTE: Available at: http://www.etsi.org/plugtests/History/2004SQTE.htm.

[i.10] ETSI 3rd Speech Quality Test Event: "Anonymized Test Report "IP Phones"".

[i.11] ETSI 4th Speech Quality Test Event: "Anonymized Test Report "IP Gateways and IP Phones"".

NOTE: Available at: http://www.etsi.org/plugtests/History/2006SQTE.htm.

[i.12] F. Kettler, H.W. Gierlich, F. Rosenberger: "Application of the Relative Approach to Optimize Packet Loss Concealment Implementations", DAGA, March 2003, Aachen, Germany.

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

[i.14] R. Sottek, K. Genuit: "Models of Signal Processing in human hearing", International Journal of Electronics and Communications (AEÜ)" vol. 59, 2005, p. 157-165.

NOTE: Available at: http://www.elsevier.de/aeue.

[i.15] SAE International - Document 2005-01-2513: "Tools and Methods for Product Sound Design of Vehicles" R. Sottek, W. Krebber, G. Stanley.

[i.16] ITU-T Recommendation P.862: "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs".

[i.17] ITU-T Recommendation P.862.1: "Mapping function for transforming P.862 raw result scores to MOS-LQO".

[i.18] ITU-T Recommendation P.862.2: "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs".

[i.19] ITU-T Recommendation SG 12 Contribution 19: "Results of objective speech quality assessment of wideband speech using the Advanced TOSQA2001".

[i.20] ITU-T Recommendation G.722: "7 kHz audio-coding within 64 kbit/s".

[i.21] ITU-T Recommendation G.722.2: "Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)".

[i.22] ITU-T Recommendation P.56: "Objective measurement of active speech level".

[i.23] ITU-T Recommendation P.57: "Artificial ears".

[i.24] M. Spiegel: "Theory and problems of statistics", McGraw Hill, 1998.

[i.25] R.A. Fisher: "Statistical methods and scientific inference", Oliver and Boyd, 1956.

[i.26] M. Kendall: "Rank correlation methods", Charles Griffin & Company Limited, 1948.

[i.27] Sottek, R.: "Modelle zur Signalverarbeitung im menschlichen Gehör, PHD thesis RWTH Aachen, 1993".

[i.28] ITU-T Recommendation P.830: "Subjective performance assessment of telephone-band and wideband digital codecs".

3 Abbreviations

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

ACR	Absolute Comparison Rating
AMR	Adaptive MultiRate
ASL	Active Speech Level

NOTE: According to ITU-T Recommendation P.56 [i.22].

BGN	BackGround Noise
CDF	Cumulative Density Function
CI	Confidence Interval
DB	Data Base
dB SPL	Sound Pressure Level re 20 µPa in dB
G-MOS	Global MOS

NOTE: MOS related to the overall sample.

HP	HighPass
IP	Internet Protocol
IRS	Intermediate Reference System
ITU	International Telecommunication Union
ITU-T	Telecom Standardization Body of ITU
MOS	Mean Opinion Score
MOS-LQSN	Mean Opinion Score - Listening Quality Subjective Noise
MRP	Mouth Reference Point
NI	Network I conditions
NII	Network II conditions
NIII	Network III conditions
NB	NarrowBand
N-MOS	Noise MOS

NOTE: MOS related to the noise transmission only.

NR	Noise Reduction
NR (filter)	Noise Reduction (filter)
PCM	Pulse Code Modulation
PLC	Packet Loss Concealment
RCV	ReCeiVe
RMSE	Random Mean Square Error
S-MOS	Speech MOS

NOTE: MOS related to the speech signal only.

SNR	Signal to Noise Ratio
STF	Specialized Task Force
TOR	Terms Of Reference
VAD	Voice Activity Detection
VoIP	Voice over IP
WB	WideBand

4 Speech signals to be used

As with any objective model, the prediction of speech quality depends on the conditions under which the model was tested and validated (see clauses 6.1 and 8). This dependency also applies to the speech material used in conjunction with the objective model.

The wideband version of the model uses French speech sentences. The near end speech signal (clean speech signal) consists of 8 sentences of speech (2 male and 2 female talkers, 2 sentences each). Appropriate speech samples can be taken from ITU-T Recommendation P.501 [i.13].

The narrowband version of the model uses English speech sentences. The near end speech signal (clean speech signal) consists of 8 sentences of speech (2 male and 2 female talkers, 2 sentences each). Appropriate speech samples can be taken from ITU-T Recommendation P.501 [i.13].

5 Selection of the data within the scope of the wideband objective model: Experts evaluation

5.1 Selection process

The aim of the selection process was to identify those data in the databases described in EG 202 396-2 [i.2] which are consistent with the scope of the objective models to be studied within the present document.

The experts were selected on the based on the definition found in ITU-T Recommendation e.g. P.831 [i.5]: experts are experienced in subjective testing. Experts are able to describe an auditory event in detail and are able to separate different events based on specific impairments. They are able to describe their subjective impressions in detail. They have a background in technical implementations of noise reduction systems and transmission impairments and do have detailed knowledge of the influence of particular implementations on subjective quality.

Their task was to select the relevant conditions within the scope of the model to be developed. Therefore they had to verify the consistency of the data with respect to the following selection criteria:

1) Artefacts others than the ones which should have been produced by the signal processing described in [i.2] e.g. due to the additional amplification required in order to provide a listening level of 79 dB SPL.

2) Inconsistencies within one condition due to the selection of the individual speech samples from the database for subjective evaluation.

3) Inconsistencies within one condition due to statistical variation of the signal processing described in [i.2] leading to non consistent judgements within this condition.

4) Inconsistencies due to ITU-T Recommendation P.56 [i.22] level adjustment process chosen for the complete files including the background noise.

5) Impact of the different listening levels used in the two databases - the French and the Czech database.

As a result of the experts listening test a set of data was selected which is used for the development of the objective model.

In the selection process five expert listeners (not native French/Czech speakers) were involved. Their task was not to produce new judgements, but to check all the samples in the database with respect to the possible artefacts described above.

A playback system with calibrated headphones was used for the test. The headphones used were Sennheiser HD 600 connected to the HEAD acoustics playback system HPS V. The equalization provided by the headphone manufacturer was used since this was the one used in the French and Czech test setup.

All samples could be heard by the experts as often as required in order to get final agreement about the applicability of the data within the terms of reference of the model. There was no limitation in comparing samples to the ones previously heard.

5.2 Results

In general it could be observed that the 4 seconds sample size chosen in the experiment according to ITU-T Recommendation P.835 [i.3] lead to a more difficult task even for expert listeners, especially in the case of non stationary background noises. It is more difficult to identify the nature of the noise itself and then identify in addition possible impairments introduced by the signal processing or by the network impairments. It is very likely that some comparatively high standard deviations seen in the data are caused by these effects.

5.3 French database

In general the French database is in line with the ToR except network condition NII. In network condition NII 1 % packet loss was chosen which is too low for the conditions to be evaluated. Due to the inhomogeneously distributed packet losses there are conditions where no packet loss is audible up to conditions where 5 out of 6 samples show packet loss. Furthermore the packet loss may occur during speech as well as during the noise periods. The impact of the different packet losses is not controlled with respect to their occurrence due to the statistical nature of the packet loss distribution, even within a set of 6 samples used for evaluating one condition. Since packet loss is clearly audible under NIII conditions (3 % packet loss) and much better distributed amongst the different samples the NII conditions are not used within the scope of the objective method. They are either covered by the NI condition (0 % packet loss) or by the NIII conditions. This results in 144 NII conditions which are not retained for the development of the model.

From the 288 NI and NIII conditions 28 conditions are not retained. The main reasons therefore are:

- Not consistent signal levels due to the amplification process.
- Insufficient S/N, speech almost inaudible.

The individual reasons for the samples of these conditions being not retained can be found in table A.1.

In total 260 out of 432 conditions are used as the reference for the objective model. In other words, 60,2 % of the data can be used for the model. The distribution of the ratings is between 1,2 and 4,96 MOS for S-/N-/G-MOS.

5.4 Czech database

For every combination of background noise and speaker gender, a single Czech sentence was used (see table 5.1). The 24 Czech listeners had to rate this single sentence, while the French ratings are a mean value of six different sentences (assessed by 4 listeners each).

Table 5.1: Sentences from the test corpus chosen for the different conditions

Condition	Sentence No.
Lux Car 130kmh Female2	S3
Lux Car 130kmh Male1	S2
Crossroads Female2	S4
Crossroads Male1	S3
Road Noise Female2	S5
Road Noise Male1	S4
Office Noise Female2	S6
Office Noise Male1	S5
Pub Noise Female2	S7
Pub Noise Male1	S6

This leads to a limited representation of the individual background noise conditions especially in the case of time varying background noises. Furthermore the NII conditions were even more critical in judgement compared to the French data since either there was no packet loss at all. Or if there was packet loss all listeners rated this particular packet loss because they all listened to the same sentence for one condition. In the French listening test 6 sentences were listened for one condition which provided a higher variance of the distributed packet loss.

The listening level variation in the Czech database, preserved from previous database processing adds another degree of complexity to the problem. The listening levels are generally lower as within the French database and as compared to the general rules laid down in ITU-Recommendations P.800 [i.4] and P.835 [i.3]. The listening level variation within the Czech database is up to 16 dB. In the experts tests the following conclusions were drawn:

- The conditions AMR NII and G.722 NII (1 % packet loss) were not selected, because in most cases, the sound files had too low packet loss. A distinction between and NI and NII conditions is hardly possible.
- The effect of packet loss in the samples should be audible in AMR NIII and G.722 NIII conditions. Because every single Czech condition consists just of one sentence, the packet loss may not be distributed uniformly in the sample. Therefore, only samples with at least one packet loss in speech *and* background noise (before or after speech) were selected.
- Due to the fact that every Czech sound file has a different level (which depends on codec, noise reduction algorithm, etc.), a minimum level of 69 dB SPL was set (10 dB below the recommended listening level of 79 dB SPL). All conditions below this limit were not retained.
- Analysis of NI conditions:

  a) AMR Codec:
  70 conditions were not retained based on the following selection criteria:

    1) Too low level (54).

    2) Inconsistent BGN level (12).

3) Too low S/N (2).

4) Too low overall level / given listening level not correct (2).

b) G.722 Codec:
19 conditions were not retained based on the following selection criteria:

1) Too low level (15).

2) MOS values irreproducible (4).

c) Selected conditions dependent of BGN: see table 5.2.

Table 5.2: Selected Czech NI conditions

BGN-Condition	Total not retained	Total retained	Selected test samples / MOS available	Selected verification samples / no MOS available
Lux_Car	17	19	10	9
Crossroads	36	0	0	0
Road	17	1	1	0
Office	14	22	16	6
Pub	5	13	10	3

d) Overall NI acceptance: 48 % of NI conditions are useful (22 % AMR, 65 % G.722).

- Analysis of NIII conditions:

a) AMR Codec:
76 conditions were not retained based on the following selection criteria:

1) Too low level (43).

2) Inconsistent packet loss (33).

b) G.722 Codec:
35 conditions were not retained based on the following selection criteria:

1) Too low level (13).

2) Inconsistent packet loss (22).

c) Selected samples dependent of BGN: see table 5.3.

Table 5.3: Selected Czech NIII conditions

BGN-Condition	Total not retained	Total retained	Selected test samples / MOS available	Selected verification samples / no MOS available
Lux_Car	30	6	4	2
Crossroads	30	6	5	1
Road	16	2	2	0
Office	24	12	10	2
Pub	11	7	2	5

d) Overall NIII acceptance: 23 % of NIII conditions are useful (16 % AMR, 35 % G.722).

The list of the selected Czech conditions is found in table A.1.

In total 88 conditions out of 432 (20,4 %) are suited to be used in a further step for checking language dependencies.

5.5 General differences between the databases

The most important differences between the French and the Czech database can be summarized as follows:

- The French and Czech listening samples of one condition do not have the same levels. The French sound files are louder than the Czech ones, in some random tests, the mean of these level differences is given in table A.2, of EG 202 396-2 [i.2]. This may have lead to different ratings for the Czech samples compared to the French samples. This has be regarded especially for further processing of the sound files.

- For every background noise condition, a single Czech sentence was used (see table 5.1). To quantify the last point, the correlation between French and Czech ratings (S-, N- and G-MOS) can be calculated. As shown below, this correlation is very low. It seems that the differences mentioned above are reflected here. Coefficients of correlation (Pearson's equation) are summarized in table 5.4.

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

with:

x MOS Data (Czech)

$\bar{x}$ Mean of MOS Data (Czech)

y MOS Data (French)

$\bar{y}$ Mean of MOS Data (French)

Table 5.4: Comparison of correlation

Over all available ratings (French and Czech, 302 condition each)	**Only selected French MOS Data (NI and NIII conditions, ratings reviewed by experts)** (179 selected French conditions)	**Only Czech and French selected MOS Data (NI and NIII conditions, ratings reviewed by experts)** (59 conditions selected for French *and* Czech)
S-MOS: 0,703 N-MOS: 0,816 G-MOS: 0,668	S-MOS: 0,736 N-MOS: 0,822 G-MOS: 0,776	S-MOS: 0,830 N-MOS: 0,897 G-MOS: 0,871

As shown in the scatter plots below, a slight correlation for the French-optimized data can be noticed, but for a usable correlation, the measurement points are distributed too far away from a (virtual) regression line of best fit (see figures 5.1, 5.3 and 5.5).

If the calculation of the correlation is limited only to the selected data (86 conditions are selected for French *and* Czech speech), the correlation increases for all values, especially for the G-MOS data (see figures 5.2, 5.4 and 5.6).