ETSI TR 103 503 V1.2.1 (2018-10)

Speech and multimedia Transmission Quality (STQ); Procedures for Multimedia Transmission Quality Testing with Parallel Task including Subjective Testing

Reference

RTR/STQ-273

Keywords

assessment, listening quality, parallel task

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 Report (TR) has been produced by ETSI Technical Committee Speech and multimedia Transmission Quality (STQ).

The present document describes auditory test methodologies for the prediction of perceived audio signal quality under parallel task conditions.

Modal verbs terminology

In the present document "**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.

Introduction

Subjective testing of speech quality and intelligibility is standardized at ETSI, ANSI, ITU-T and ITU-R. Tests are performed in defined environments using listening/conversational rigorous procedures (Recommendation ITU-T P.800 [i.16], Recommendation ITU-T P.805 [i.21], Recommendation ITU-T P.835 [i.18], Recommendation ITU-R BS.1534-3 [i.22], Recommendation ITU-R BS.1116 [i.23], etc.), and they require relaxed, fresh, fit and concentrated naive or expert listeners seated comfortably in usually artificially looking listening room/booth.

However, such a test does not correspond to the normal use of the tested technologies. Voice services are often used in sports, driving, work, public transport, or other noisy or less convenient environments. Users are tired, stressed or concentrate on another, often important, task.

In an attempt to bring laboratory tests closer to reality, the so-called dual-task or parallel-task tests are introduced, in these test participants are asked to perform multiple different tasks at the same time.

1 Scope

The present document describes the methods for assessment of subjective audio (including speech) quality and speech intelligibility under parallel task condition. This approach can be used to evaluate the perceived listening quality or speech intelligibility in situations which better mimics real operation of the tested telecommunication equipment or algorithm.

The present document describes possible parallel task generation and scenarios, the test design and reference conditions used to evaluate the quality or intelligibility subjectively.

Several parallel task scenarios are covered:

- Physically oriented.
- Mentally oriented.
- Hybrid.

2 References

2.1 Normative references

Normative references are not applicable in the present document.

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] V. Durin and L. Gros: "Measuring speech quality impact on tasks performance", Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, pp. 2074-2077, 2008.

[i.2] A. Serampalis, S. Kalluri, B. Edwards, and E. Hafter: "Objective measures of listening effort in noise", J. Speech, Lang. Hear. Res., vol. 52, no. October 2009, pp. 1230-1240, 2009.

[i.3] G. P. Sonntag, T. Portele, and F. Haas: "Comparing the comprehensibility of different synthetic voices in a dual task experiment", Proc. Third Work. Speech Synth. Jenolan Caves House, Blue Mt., pp. 5-10, 1998.

[i.4] L. Gros, N. Chateau, and S. Busson: "The impact of real environments on transmitted speech quality judgments", Quality, vol. 0, pp. 45-50, 2003.

[i.5] D. Guse, S. Egger, A. Raake, and S. Moller: "Web-QOE under real-world distractions: Two test cases", 2014 6th Int. Work. Qual. Multimed. Exp. QoMEX 2014, pp. 220-225, 2014.

[i.6] S. L. Beilock, T. H. Carr, C. MacMahon, and J. L. Starkes: "When paying attention becomes counterproductive: Impact of divided versus skill-focused attention on novice and experienced performance of sensorimotor skills", J. Exp. Psychol. Appl., vol. 8, no. 1, pp. 6-16, 2002.

[i.7] J. Holub: "Low Bit-rate Coded Speech Intelligibility - Comparison of Laboratory Test Results and Results of Test with Parallel Task", in Future Forces Forum, 2016.

[i.8] D. L. Strayer and W. A. Johnston: "Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular telephone", Psychol. Sci., vol. 12, no. 6, pp. 462-466, 2001.

[i.9] S. Choi, A. Lotto, D. Lewis, B. Hoover, and P. Stelmachowicz: "Attentional Modulation of Word Recognition by Children in a Dual-Task Paradigm", J. Speech Lang. Hear. Res., vol. 51, no. 4, p. 1042, Aug. 2008.

[i.10] Y.-H. Wu, E. Stangl, X. Zhang, J. Perkins, and E. Eilers: "Psychometric Functions of Dual-Task Paradigms for Measuring Listening Effort", Ear Hear., vol. 37, no. 6, pp. 660-670, 2016.

[i.11] C. Kwak and W. Han: "Comparison of Single-Task versus Dual-Task for Listening Effort", J. Audiol. Otol., Oct. 2017.

[i.12] L. Gros, N. Chateau, and A. Macé: "Assessing speech quality : a new approach Methodology", 2005.

[i.13] K. S. Helfer, J. Chevalier, and R. L. Freyman: "Aging, spatial cues, and single- versus dual-task performance in competing speech perception", J. Acoust. Soc. Am., vol. 128, no. 6, pp. 3625-3633, Dec. 2010.

[i.14] K. Bunton and C. K. Keintz: "The use of a dual-task paradigm for assessing speech intelligibility in clients with Parkinson disease", J. Med. Speech. Lang. Pathol., vol. 16, no. 3, pp. 141-155, Sep. 2008.

[i.15] ITU-T Handbook: "Practical procedures for subjective testing", 2011.

[i.16] Recommendation ITU-T P.800 (08/1996): "Methods for subjective determination of transmission quality".

[i.17] Recommendation ITU-T P.807 (02/2016): "Subjective test methodology for assessing speech intelligibility".

[i.18] Recommendation ITU-T P.835 (11/2003): "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm".

[i.19] Council of Europe (2011): "Common European Framework of Reference for Languages: Learning, Teaching, Assessment" Council of Europe.

[i.20] Recommendation ITU-T P.1400 (03/2013): "Statistical analysis, evaluation and reporting guidelines of quality measurements".

[i.21] Recommendation ITU-T P.805: "Subjective evaluation of conversational quality", Geneva 2007.

[i.22] Recommendation ITU-R BS.1534: "Method for the subjective assessment of intermediate quality levels of coding systems", Geneva 2015.

[i.23] Recommendation ITU-R BS.1116: "Methods for the subjective assessment of small impairments in audio systems", Geneva 2015.

[i.24] Recommendation ITU-T G.711 Amendment 2009: "Pulse Code Modulation (PCM) of Voice Frequencies".

[i.25] STANAG 4591 C3: "The 600 bit/s, 1200 bit/s and 2400bit/s NATO Interoperable Narrow Band Voice Coder", NSA/1025(2008)-C3/4591, NATO Standardization Agency 2008.

[i.26] Recommendation ITU-T G.722.2: "Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)".

[i.27] ETSI TS 126 445: "Universal Mobile Telecommunications System (UMTS); LTE; EVS Codec Detailed Algorithmic Description (3GPP TS 26.445)".

[i.28] 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.29] ITU-T Temporary Document 12rev1: "Statistical evaluation. Procedure for P.OLQA v.1.0.", Berger J, editor, Geneva. 2009.

NOTE: Available at https://www.itu.int/md/T09-SG12-090310-TD-WP2-0012/en.

[i.30] IEEE No. 297™: "IEEE Recommended Practice for Speech Quality Measurements", June 1969.

NOTE: Available at https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7405210.

3 Abbreviations

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

AMR-WB	Adaptive Multirate (coder) - WideBand
ECG	ElectroCardioGraphy
EEG	ElectroEncephaloGraphy
EVS	Enhanced Voice Services (coder)
HMMWV	High Mobility Multipurpose Wheeled Vehicle
MOS	Mean Opinion Score
MRT	Modified Rhyme Test
PC	Personal Computer
PCM	Pulse Code Modulation
QoE	Quality of Experience
SNR	Signal to Noise Ratio
STD	STandard Deviation
VR	Virtual Reality

4 Subjective speech quality assessment, intelligibility and listening effort: existing approaches

4.1 Introduction

Subjective testing of speech quality and intelligibility follows strictly standardized procedures. Tests are performed in defined environments using listening/conversational rigorous procedures (Recommendations ITU-T P.800 [i.16], P.835 [i.18], etc.) and it requires relaxed, fresh, fit and concentrated naive or expert listeners comfortably seated in a listening room/booth with proper acoustic lining to minimize e.g. inherent background noise and room reverberation.

However, such a test does not correspond to normal use of the tested technologies. Voice services are often used during sports, driving, work, etc. Users are tired, stressed or concentrated on another, often important, task.

To bring laboratory tests closer to reality, the so-called dual-task or parallel-task tests are introduced, where test participants are asked to perform multiple different tasks at the same time. The test results obtained during parallel task test differ from regular subjective tests. The differences are sometimes contra-intuitive and cannot be explained e.g. by decreased level of subjects' attention. The parallel task should be designed to distract subjects in a similar way as the activity performed during the real (targeted) situation. Limitations are given by requirements on repeatability, space- and movement- restrictions in the lab, etc.

4.2 Classification of parallel tasks in scientific publications

4.2.1 Current approaches

Parallel tasks found in scientific literature can be divided into three types: Mentally oriented tasks, Physically oriented tasks and Hybrid tasks. Selected available experiments of those three categories are discussed in Table 1.

Table 1: Resource summary

Reference	Test type	Parallel task	Parallel task type	Language
[i.1]	Speech intelligibility	Memorizing digits	Mentally oriented	N/A
[i.2]	Speech intelligibility	Memorizing digits	Mentally oriented	English
[i.3]	Speech intelligibility	Pressing colour buttons	Mentally oriented	German
[i.4]	QoE test	Pressing colour buttons	Mentally oriented	English
[i.5]	QoE test	Traveling in public transport; watching a TV	Mentally oriented; Hybrid	German
[i.6]	Other	Memorizing tones, memorizing words	Mentally oriented	N/A
[i.7]	Speech intelligibility	Laser shooting simulator	Hybrid	English
[i.8]	Other	Telephone call	Hybrid	English
[i.9]	Speech intelligibility	Word repetition, Memorizing digits	Mentally oriented	English
[i.10]	Speech intelligibility	Pressing colour button	Mentally oriented	English
[i.11]	Speech intelligibility	Memorizing sentences, Arithmetic	Mentally oriented	Korean
[i.12]	QoE test	Matching coloured squares	Mentally oriented	N/A
[i.13]	Speech intelligibility	Forward/backward discrimination and speech understanding	Mentally oriented	English
[i.14]	Speech intelligibility	Turning a nut on a bolt	Mentally oriented	English

4.2.2 Mentally oriented tasks

Frequently used mental tasks are memory-related tasks requiring memorization and subsequent repetition of information, most often words or digits. In experiment [i.1], listeners had to identify the letter as prescribed, while remembering the five digits displayed or played before this description. The results of the experiment depend on both the quality of the codec used and the intelligibility of the description, and on the way the numbers are presented and how the conditions are sorted (serial/random). A memory task is also used in other experiments, such as in [i.2], [i.9] and [i.11]. In the first experiment [i.2], the primary test condition consisted in the different levels of noise in the background of test sentences. The listeners had the task of repeating the last word of the sentence heard or trying to guess it if it was not comprehensible. The second task of the listeners was to remember all the last words and repeat them after eight sentences. In the next experiment [i.9], a group of 64 children participated in speech intelligibility test. Half of them were told to pay their primary attention to word repetition and the other half to remember digits. Single-task and dual-task performances were compared. Results showed that significant dual-task decrements were found for digit recall, but no dual-task decrements were found for word recognition. In [i.11] as a parallel-task, subjects were asked to write down the sentence they heard or write down the sum of first and third numbers they heard.

Other types of mental tasks are those that require some computer work. In the second experiment of [i.2], the listeners were asked to repeat the heard sentence or part of it, which they understood (the sentences were played back with different levels of background noise), while watching the computer screen and using the keypad to decide whether the displayed digit is even or odd. Similarly, in experiments [i.3] and [i.12], listeners had to solve simple mathematical examples from the listening input and at the same time press the corresponding key to respond to the different colours displayed on the computer monitor. Experiment [i.3] was primary about comparing different speech synthesis systems. In [i.12] human and synthesized speech with transmission degradation (compression, noise, packet loss) were compared. In both experiments [i.3] and [i.12], the results showed that the worse the quality of speech and thus the clarity of the assignment of the primary task, the longer the reaction times in the secondary task. In experiment [i.12], in the worst-case transmission, some respondents completely omitted the secondary task. In [i.13], authors provided an experiment where younger and older adults were asked to understand a target talker with and without determining how many masking voices were presented in samples time-reversed. In another experiment [i.10], subjects participated in a speech intelligibility test with two similar dual-task paradigms. During the first one, they were asked to press the space bar on the keyboard when they saw any colour on their screens. During the second test, subjects were asked to press a corresponding button for a text colour that appeared on their screens. In experiment [i.5], respondents were asked to search for specific information on a simulated news website (viewing of the site and searched messages were variously delayed), and then evaluate their user experience with a specific setting. In order to bring the experiment closer to reality, respondents also watched TV. The results showed that while watching television, the search took longer time, although the final quality assessment for the condition was the same as in the experiment without a secondary task. Results show that sentence recognitions scores and arithmetic scores decreased as noise increased, while the response time for arithmetic tasks increased as noise increased.

4.2.3 Physically oriented tasks

The physical task usually lies in running, cycling, or other physical or sporting activity. Experiment [i.6] consisted of two parts. In the first part, experienced golfers were asked to put on the training green while listening to a series of tones from the audio player. Their task was to identify and report one particular tone. The results showed that players performed better with an additional listening task than without it. In the second part of the experiment [i.6], the task of the respondents was to lead the soccer ball by slalom from cones while listening to a series of words and identifying and repeating the target word. The group of respondents consisted of experienced footballers and non-players. Experienced players played better in slalom in a parallel task test. The presence of a secondary task and distraction led experienced athletes to better perform automatic and rehearsal moves.

4.2.4 Hybrid tasks

Hybrid tasks require both physical and mental activity. An example may be driving a car or a shooting simulator. In the second part of the experiment [i.5], the respondents also had to search for information on the news site, but this time, the experiment was conducted on public transport. Unlike watching TV, this secondary task did not show up on the experiment's results. In another experiment [i.14], an intelligibility test with a dual-task methodology was performed for subjects with dysarthria related to Parkinson disease. As a parallel task for subjects the turning a nut on a bolt was used. Intelligibility scores for dual-task conditions were lower with significant differences between scores of different tasks. In the experiment [i.8], respondents had to drive the car while handling a telephone call. In contrast to driving without a phone, the driver was significantly more likely to miss the traffic mark. Drivers also had longer reaction times. In the experiment [i.7], the respondents performed the speech intelligibility test in consideration of the codec used and the noise level. The test was first performed under standard laboratory conditions and then again with the addition of a parallel task (shooting simulator). Some tested conditions received higher scores in a parallel test than in a laboratory. It turns out [i.6], [i.7] that some perception and human behaviour mechanisms under load are different from the standard quiescent state.

5 Procedures for subjective testing deploying parallel task

5.1 General considerations

5.1.1 Introduction

The parallel task is a secondary task which test subjects are asked to perform during subjective testing to better mimic real usage situations. The parallel task should be designed to distract subjects in a similar way as an activity performed during the real (targeted) situation. Limitations are given by requirements on repeatability, space- and movement-restrictions in the lab, etc.

5.1.2 Task Class 1 (activity driven)

The selected parallel task should be of one of the types shown in Table 2.