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

TECHNICAL REPORT

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

Copyright Notification

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.

Table 2: Types of the activity driven parallel task

Task Class 1	Descriptions	Examples
Mentally oriented	Subjects perform mental activity which does not significantly influence their physical conditions.	Logical quizzes, math calculations, memory-oriented task, tests in foreign language, VR-based tasks requiring negligible movements.
Physically oriented	Subjects perform physical activity which does not significantly influence their mental conditions. Monitoring of ECG, EEG, blood or saliva tests can be used to objectively measure the amount of physically oriented load.	Exercises: bike riding, running belt, VR based tasks requiring significant movements with only negligible mental load, moving platform, centrifuge, etc.
Hybrid	Subjects perform complex task that requires both physical and mental activity.	Car driving or its simulation, machine operation, aimed shooting or its simulation, complex VR-based tasks, PC gaming, tasting, other psycho-motor tasks: small objects sorting, etc.

5.1.3 Task Class 2 (purpose driven)

Based on potential final usage case the task categories as shown in Table 3 are specified.

Table 3: Types of the purpose driven parallel tasks

Task Class 2	Description	Examples
General	The task is selected to mimic real general usage of the tested technology with no particular use-case expected.	Mobile terminal, general handset, headset testing, general codec testing, general noise suppression algorithm testing.
Purpose oriented	The task mimic certain expected use case.	Public safety, fire brigade or military equipment testing on physically oriented tasks simulating real deployments. Operation centre (airport approach control, military) headset testing using mental task simulating real situations.

5.1.4 Additional comments to Task Class 1 and Task Class 2 classification

For physically and mentally oriented Task Class 1 experiments the risk of unequal subject load effect arises. E.g. in case of physically oriented tasks, subjects with stronger physical constitution are not affected as much as weaker subjects. Therefore, hybrid tasks are preferred for Task Class 2 - General, leaving the applicability area of purely Physical or Mental tasks for Purpose oriented experiments (Task Class 2 - purpose oriented).

Performing testing in other than subjects' native language is considered a case of mentally oriented task according to Task type 1 classification. It is particularly suitable for intelligibility testing. The subjects' language proficiency should be tested prior the subjective testing using language proficiency scale defined in [i.19]: The foreign language levels and their descriptions are shown in Table 4.

Table 4: Foreign language levels and their descriptions

Language level	Level name
A	Basic
B	Independent
C	Proficient user

Unless required by purpose-oriented Task Class 2, subjects of the same language proficiency level (A, B or C) should be used in the subjective test. Language level C is expected not to generate mental load level comparable to subjects classified to Language level A and B.

5.2 Test Environment

5.2.1 Real environment

Only if required by the parallel task nature, real operational environment can be used for subjective testing. In this case, a special attention has to be devoted to acoustic features (headphones, acoustic coupling, environmental reverberation and background noise, etc.) to ensure reliable and repeatable results. All above mentioned parameters have to be reported in the test report.

5.2.2 Lab with simulated parallel task

Using the listening environment defined in Recommendation ITU-T P.800 [i.16] is the preferred way. The parallel task generation is then restricted by space and movement limitations. Caution should be exercised to maintain the required acoustic environment even when during the parallel task generation (e.g. for PC-based simulators, an external or silent PC should be used). The background noise level and reverberation time should be reported if different from the original values of the listening environment.

5.2.3 VR based testing

Virtual reality is a novel mean of parallel task generation. If used, caution should be exercised to maintain the required acoustic features. The background noise level and reverberation time should always be reported together with VR environment parameters (viewing angles, resolution, frame rate).

5.3 Subjective testing procedure

The test procedure should follow established methods stated in Recommendations ITU-T P.800 [i.16], ITU-T P.807 [i.17] and ITU-T P.835 [i.18]. Detailed descriptions and best practices are found also in ITU-T Handbook "Practical procedures for subjective testing" [i.15].

5.4 Result Analysis and Reporting

5.4.1 Introduction

The content and format of results reporting should in general follow chapter B.4.7 of Recommendation ITU-T P.800 [i.16] or chapter 5.4 of Recommendation ITU-T P.835 [i.18] depending on test type. More details can be found in chapter 7 of Recommendation ITU-T P.1400 [i.20]. Those general items (subjective quality per condition, statistical evaluation - STD or confidence interval, etc.) are complemented by the mandatory set of parameters when tested with parallel task as listed in clause 5.4.2.

5.4.2 Special reported items

For subjective tests deploying parallel task, the following parameters are also reported:

- Task Class 1 according to clause 5.1.2 (mentally oriented, physically oriented, hybrid).

- Task Class 2 according to clause 5.1.3 (general, purpose-oriented).

- Statement of mother tongue usage for all subjects or language proficiency level if the tests are performed in a foreign language according to clause 5.1.4 (A, B, C).

- Test environment details as per clause 5.2.

Annex A:
Examples of test scenarios incorporating parallel task

A.1 Example scenario 1 - psychomotor experiment A (hybrid general task)

ENGLISH INSTRUCTIONS [group of 3 subjects]

In this experiment, you will use a professional laser shooting simulator. One of you will play a role of "hunter" while the remaining two will act as "counters". Your roles will be dynamically randomly assigned each 40 s by automated light indicators. The task of the hunter is to aim a handgun at a moving target and shooting at it to achieve as many hits as possible. The task of counters is to count the successful hits of a current shooter. You can make notes about the hits on the paper in front of you. The hits should be counted for each of the shooters separately.

While doing so, you will be listening to sentences pairs via headphones and giving your opinion of the speech quality you hear. For expressing your opinion, a clicker with buttons 1, 2, 3, 4 and 5 will be used. Please, use the scale as follows:

PERCEIVED SPEECH QUALITY

5 - Excellent

4 - Good

3 - Fair

2 - Poor

1 - Bad

[TECHNICAL DETAILS RELATED TO PARTICULAR VOTING PROCEDURE TO BE INSERTED HERE]

You are asked to assess the technical quality of the transmitted speech, means how well is the speech transmitted and reproduced and how much distortion or non-speech signals (e.g. noise) are introduced. Please, do not judge the content of the sentences or speaker voice preferences.

Please, do not discuss your opinions with other test persons before the entire test is over. Each session takes maximally [15] minutes. There will be breaks between the sessions. If you have any question, please, ask the test supervisor immediately. Thank you for keeping your mobile phones switched off (muting the ring of your mobile phone is not enough as it may still interfere with the laboratory equipment). We start with a quick training session containing [five] samples only. Thank you for your help in this experiment!

Do not forget to vote while shooting!

A.2 Example scenario 2 - psychomotor experiment B (hybrid general task)

ENGLISH INSTRUCTIONS [1 subject at a time]

In this experiment, you will use a car driving simulator, following driving scenario given to you by test supervisor [driving from point A to B in a virtual city using the simulated car navigation, following carefully traffic rules]. During your drive you will perform a listening test of the audio quality of sound systems in car. Imagine you are driving a car and listening to music. In the listening test, sound samples with duration of approximately 10 seconds will be played back to you. After listening to a sample, you will be asked to give your judgement of the audio quality of the heard sample on a 9-point scale ranging from 1 (bad) to 9 (excellent).

9 - Excellent

8

7 - Good

6

5 - Fair

4

3 - Poor

2

1 - Bad

The playback of the next sample will start after you say your judgement out loud. Please try to disregard your own personal taste of music while judging and concentrate only on the quality of the perceived sound. We will start with a short training phase, to familiarize you with the procedure of the listening test. The test will take about 45 minutes. Please take a short break after you have listened to half of sound samples.

Following the listening test we would like to ask you for some further information.

Thank you for your participation! Do not forget to vote while driving!

A.3 Example scenario 3 - tasting experiment A (hybrid purpose oriented task)

ENGLISH INSTRUCTIONS

In this experiment, you will taste from the samples (they are a mixture of wheat, sugar and salt in different proportions) and sort them from the saltiest to the sweetest. Put them in order: the sweetest should be in the further left and the saltiest should be in the further right.

While doing so, you will be listening to sentences pairs via headphones and giving your opinion of the speech quality you hear. For expressing your opinion, a clicker with buttons 1, 2, 3, 4 and 5 will be used. Please, use the scale as follows:

PERCEIVED SPEECH QUALITY

5 - Excellent

4 - Good

3 - Fair

2 - Poor

1 - Bad

[TECHNICAL DETAILS FOR THE PARTICULAR VOTING PROCEDURE TO BE INSERTED HERE]

You are asked to assess the technical quality of the transmitted speech, means how well is the speech transmitted and reproduced and how much distortion or non-speech signals (e.g. noise) are introduced. Please, do not judge the content of the sentences or speaker voice preferences.

Please, do not discuss your opinions with other test persons before the entire test is over. Each session takes maximally [15] minutes. There will be breaks between the sessions. If you have any question, please, ask the test supervisor immediately. Thank you for keeping your mobile phones switched off (muting the ring of your mobile phone is not enough as it may still interfere with the laboratory equipment). We start with a quick training session containing [five] samples only. Thank you for your help in this experiment!

Do not forget to vote while tasting!

A.4 Example scenario 4 - tasting experiment B (hybrid purpose oriented task)

ENGLISH INSTRUCTIONS

In this experiment, you have 10 samples to taste, some of them are different - they contain a different proportion of wheat, sugar and salt. Your task will be to find the different samples by tasting them. Once you found them move them to the front of the table.

While doing so, you will be listening to sentences pairs via headphones and giving your opinion of the speech quality you hear. For expressing your opinion, a clicker with buttons 1, 2, 3, 4 and 5 will be used. Please, use the scale as follows:

PERCEIVED SPEECH QUALITY

5 - Excellent

4 - Good

3 - Fair

2 - Poor

1 - Bad

[TECHNICAL DETAILS RELATED TO PARTICULAR VOTING PROCEDURE TO BE INSERTED HERE]

You are asked to assess the technical quality of the transmitted speech, means how well is the speech transmitted and reproduced and how much distortion or non-speech signals (e.g. noise) are introduced. Please, do not judge the content of the sentences or speaker voice preferences.

Please, do not discuss your opinions with other test persons before the entire test is over. Each session takes maximally [15] minutes. There will be breaks between the sessions. If you have any question, please, ask the test supervisor immediately. Thank you for keeping your mobile phones switched off (muting the ring of your mobile phone is not enough as it may still interfere with the laboratory equipment). We start with a quick training session containing [five] samples only. Thank you for your help in this experiment!

Do not forget to vote while tasting!

A.5 Example scenario 5 - stationary bicycle (physically oriented, purpose oriented task)

In this experiment, you will ride a stationary bicycle, maintaining your speed according the device display.

[DETAILS RELATED TO STATIONARY BICYCLE DISPLAY READING TO BE INSERTED HERE].

While doing so, you will be listening to sentences pairs via headphones and giving your opinion of the speech quality you hear. For expressing your opinion, a clicker with buttons 1, 2, 3, 4 and 5 will be used. Please, use the scale as follows:

PERCEIVED SPEECH QUALITY

5 - Excellent

4 - Good

3 - Fair

2 - Poor

1 - Bad

[TECHNICAL DETAILS RELATED TO PARTICULAR VOTING PROCEDURE TO BE INSERTED HERE]

You are asked to assess the technical quality of the transmitted speech, means how well is the speech transmitted and reproduced and how much distortion or non-speech signals (e.g. noise) are introduced. Please, do not judge the content of the sentences or speaker voice preferences.

Please, do not discuss your opinions with other test persons before the entire test is over. Each session takes maximally [15] minutes. There will be breaks between the sessions. If you have any question, please, ask the test supervisor immediately. Thank you for keeping your mobile phones switched off (muting the ring of your mobile phone is not enough as it may still interfere with the laboratory equipment). We start with a quick training session containing [five] samples only. Thank you for your help in this experiment!

Thank you for your participation! Do not forget to vote while riding!

A.6 Example scenario 6 - virtual reality deployment (physically oriented general task)

ENGLISH INSTRUCTIONS

[The text below is displayed prior the test in VR environment on virtual posters placed all around the tester to force body movements].

Welcome!

Today, you will be involved in an experiment designed to evaluate intelligibility of speech.

You will hear 48 samples and with your controllers in your hands pick the word you heard from six options flying all around you. You will need to turn around your head or body to see them all.

Workflow

1. Beep
2. Listen to a sample
3. Look at all choices (all around you)
4. Pick what you heard

Controls - shooting

Move controller with the beam and press highlighted button to shoot.

Annex B: Experiments and studies related to the standard

B.1 Experiment 1

B.1.1 Experiment Description

B.1.1.1 Materials and methods

Two intelligibility tests were performed, using the identical set of distorted speech samples. The first tests followed the standardized methodology and the second one deployed a parallel task. Both tests were performed in an acoustically treated critical listening environment conforming to Recommendation ITU-T P.800 [i.16].

The intelligibility tests were performed following the MRT (Modified Rhyme Test) methodology as described in Recommendation ITU-T P.807 [i.17]. In total, 48 samples were selected from MRT sample list. For the initial preliminary experiment, described in this text, the samples were recorded using voices of two male narrators. Both narrators were native English speakers. The distorted samples were generated then using a network simulator deploying the following coder and background noise options:

- Pulse Code Modulation (PCM) [i.24] at 64 kbit/s (16 samples)
- Low bit rate coder MELPe [i.25] operating at 2,4 kbit/s (16 samples)
- MELPe with the background noise of the interior of a High Mobility Multipurpose Wheeled Vehicle (HMMWV) at SNR = 0 dB (16 samples)

51 subjects (26 female and 25 male) in the age range of 18 to 56 were hired for the tests.

The experiment had the following structure:

- Regular intelligibility tests
- 90 minute break
- Intelligibility test but with a parallel task

The order of regular and parallel task tests was counterbalanced.

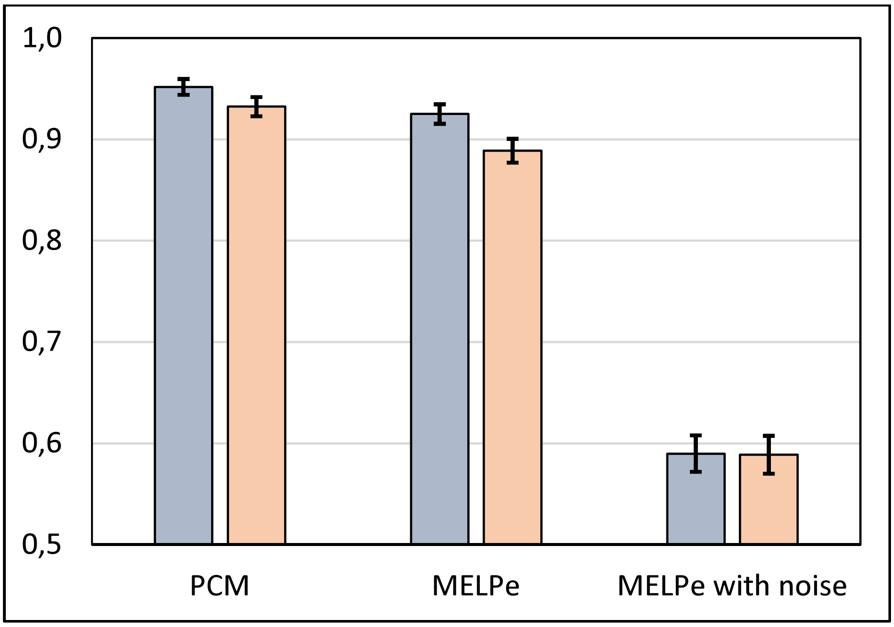
B.1.1.2 Parallel task description and classification

The **Hybrid** (see Table 2) and **General** (Table 3) task was used in this experiment.

The parallel task deployed a professional laser shooting simulator. This part of the test was performed by groups of 3 subjects. One of them always played a role of "shooter" while the remaining ones were "counters". The roles were dynamically randomly assigned each 40 s by automated light indicators, not synchronized with the pace of intelligibility test. All three subjects ("shooter" and both "counters") performed in parallel the intelligibility test. The task (aiming handgun against a moving target and shooting or counting successful hits of another shooter, respectively) generated well defined and highly repeatable psycho-motoric load.

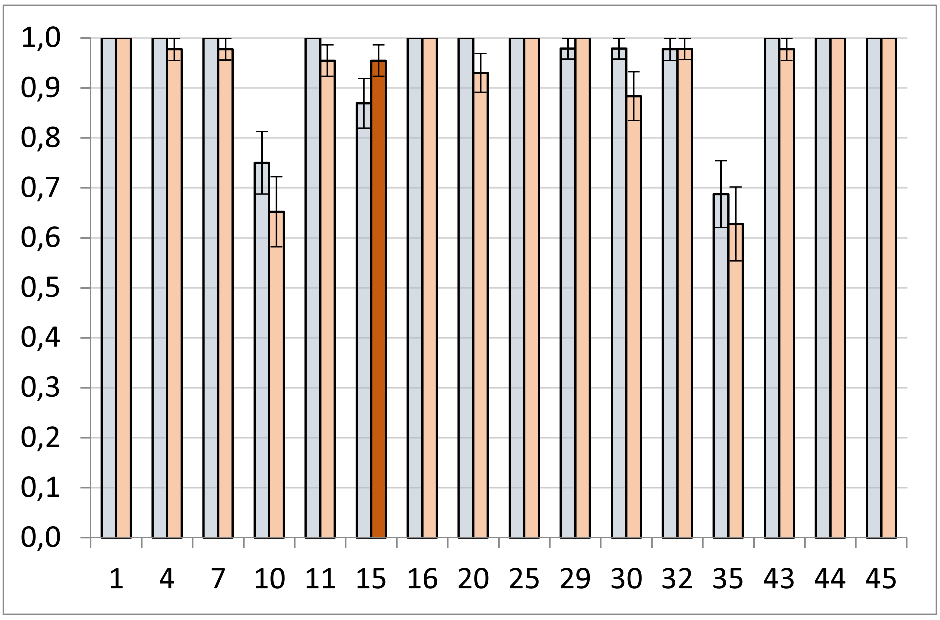
B.1.2 Results

In most cases, intelligibility without parallel task is higher than intelligibility with parallel task However, as can be seen (highlighted samples) in Figure B.2, Figure B.3 and Figure B.4, for 4 of the 48 samples (15, 17, 36 and 42) the results are opposite - meaning their intelligibility increases when the parallel task is introduced.



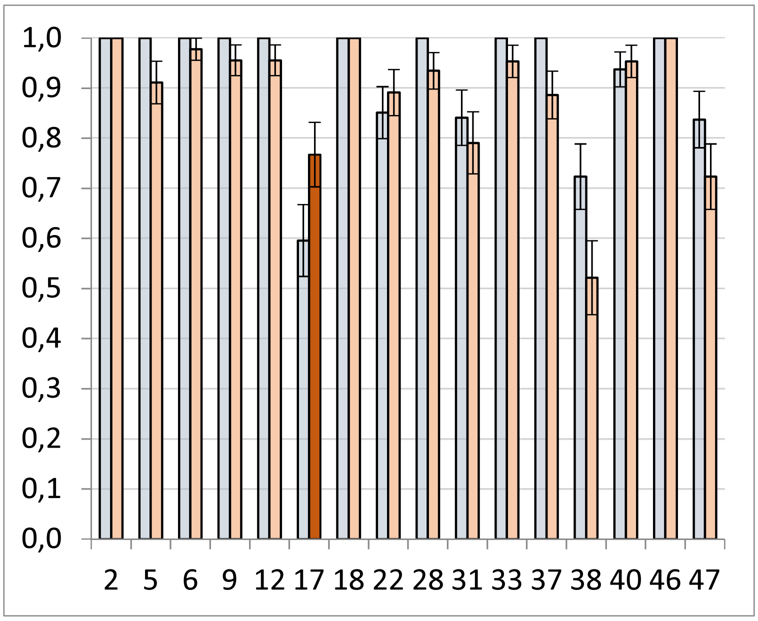
NOTE: Vertical Scale: Intelligibility, uncompensated for random correct votes.
Minimum 750 votes per condition.

Figure B.1: Intelligibility scores of ale coders without (left bar) and with (right bar) a parallel task, missing votes not considered



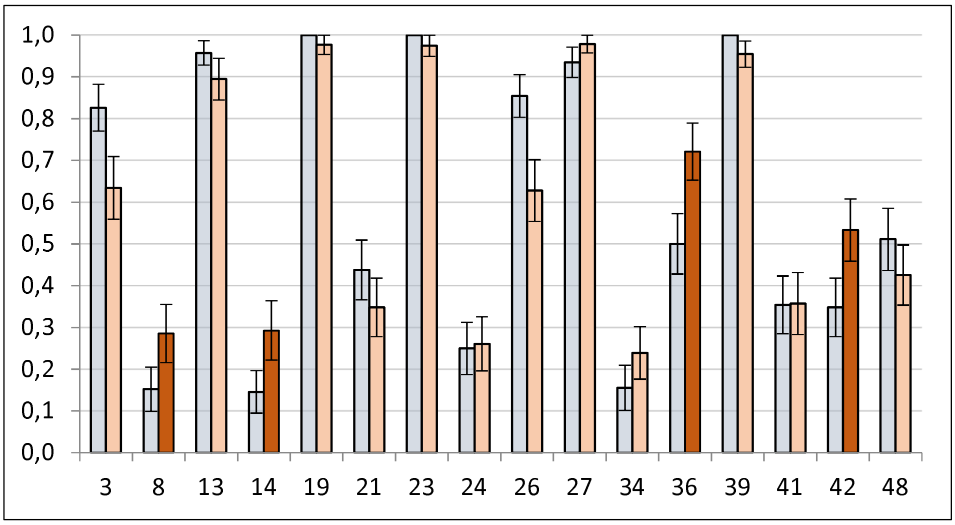
NOTE: Horizontal scale: Sample Number.
Vertical Scale: Intelligibility, uncompensated for random correct votes.
Minimum 43, median 48 votes per sample. Distinctive result highlighted.

Figure B.2: Intelligibility scores of PCM samples without (left bars) and with (right bars) a parallel task, missing votes not considered



NOTE: Horizontal scale: Sample Number.
Vertical Scale: Intelligibility, uncompensated for random correct votes.
Minimum 43, median 48 votes per sample. Distinctive result highlighted.

Figure B.3: Intelligibility scores of MELPe samples without (left bars) and with (right bars) a parallel task, missing votes not considered



NOTE: Horizontal scale: Sample Number.
Vertical Scale: Intelligibility, uncompensated for random correct votes.
Minimum 43, median 48 votes per sample. Distinctive results highlighted.

Figure B.4: Intelligibility scores of MELPe with 0dB HMMWV noise samples without (left bars) and with (right bars) a parallel task, missing votes not considered

B.2 Experiment 2

B.2.1 Experiment description

B.2.1.1 Materials and methods

For data analysis, two subjective tests according to Recommendation ITU-T P.835 [i.18] were held in subjective testing laboratory. They are named as A and B. Test subjects from test A were different from test B. Test A contained 32 subjects and test B included 25 subjects. The gender structure of the listening panels was balanced. The age distribution approximately followed human population age distribution in the range between 18 and 65 years of age (average age: 28,4).

A single English sample set was used in both experiments. The speech sample set was prepared following requirements of Recommendation ITU-T P.800 [i.16] and ITU-T P.835 [i.18]. A selection of Harvard phonetically balanced sentences from the Appendix of IEEE Subcommittee on Subjective Measurements [i.30] was used. The following coders and cases of background noise were used:

Coders:

- AMR WB [i.26]
- EVS [i.27]

Background noises all adopted from [i.28]:

- Cafeteria
- Mensa
- Road
- Pub
- Office
- Car

The background noise was mixed with speech material following Recommendation ITU-T P.835 [i.18], Appendix 1.

The test methodology was based on recommendation Recommendation ITU-T P.835 [i.18]. The concept of this standard is to make subjects listen to the same sample three times: first time for assessing the speech quality, second time - the noise annoyance, and the third time - the overall sample quality.

During test A, a simple P.835 test without any parallel task was performed. During test B, an additional parallel task was included to distract test subjects from fully concentrating on the subjective testing.

B.2.1.2 Parallel task description and classification

The **Hybrid** (see Table 2) and **General** (Table 3) task was used in this experiment.

The parallel task deployed a professional laser shooting simulator. This part of the test was performed by groups of 3 subjects. One of them always played a role of "shooter" while the remaining ones were "counters." The roles were dynamically randomly assigned each 40s by automated light indicators, not synchronized with the pace of intelligibility test. All three subjects ("shooter" and both "counters") performed in parallel the intelligibility test. The task (aiming handgun against a moving target and shooting or counting successful hits of another shooter, respectively) generated well defined and highly repeatable psycho-motoric load.

B.2.2 Results

B.2.2.1 Results overview

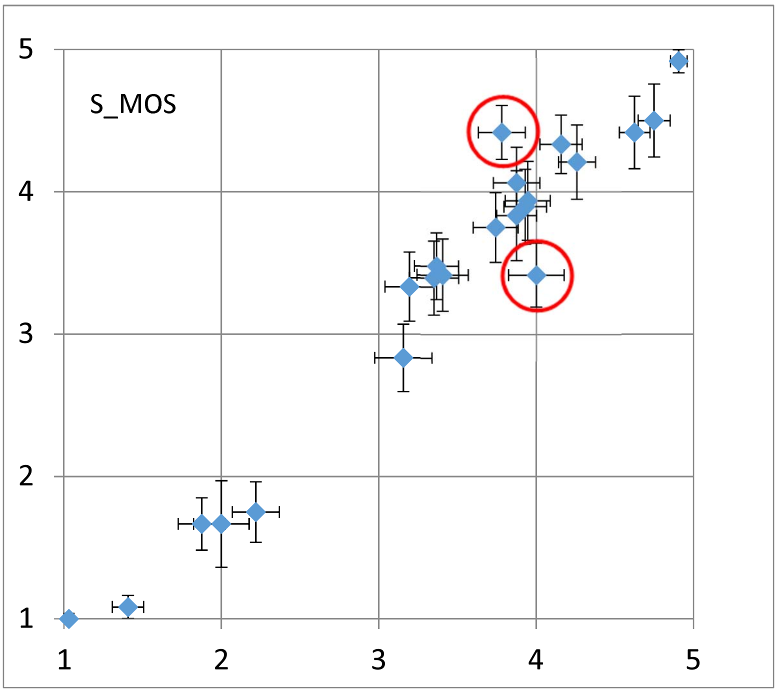
The graphs show that the subjects voted similarly. Correlation values are close to the maximum value of 1. However, as indicated in Figure B.5, certain sample pairs are ranked oppositely with and without a parallel task. For this purpose, pair-wise comparisons [i.29] were performed as described further.

In Figure B.5, there are two interesting points which do not correspond to overall results of the tests. The points are marked with red circles. Both points provide a similar evaluation in the A-tests (3,781 and 4,000) while in the B-tests their rank order is significantly opposite (4,417 and 3,417). By analysis of the sound files for the involved conditions it was concluded that this order swapping is caused by voting mistakes caused by the introduction of the parallel task. The subjects were not able to distinguish properly between speech distortion and strong background noise. This means that some subjects decreased the speech quality score due to background noise even for non-distorted speech and also considered speech distorted by artificial coding artifact as noisy. It indicates that the P.835 methodology is too complex if used with the parallel task of the described type. Not all subjects can correctly assess speech distortion (only) and background noise annoyance (only) in different playouts as required by the P.835, as they are distracted by another task in parallel.

B.2.2.2 Pairwise comparison of each test

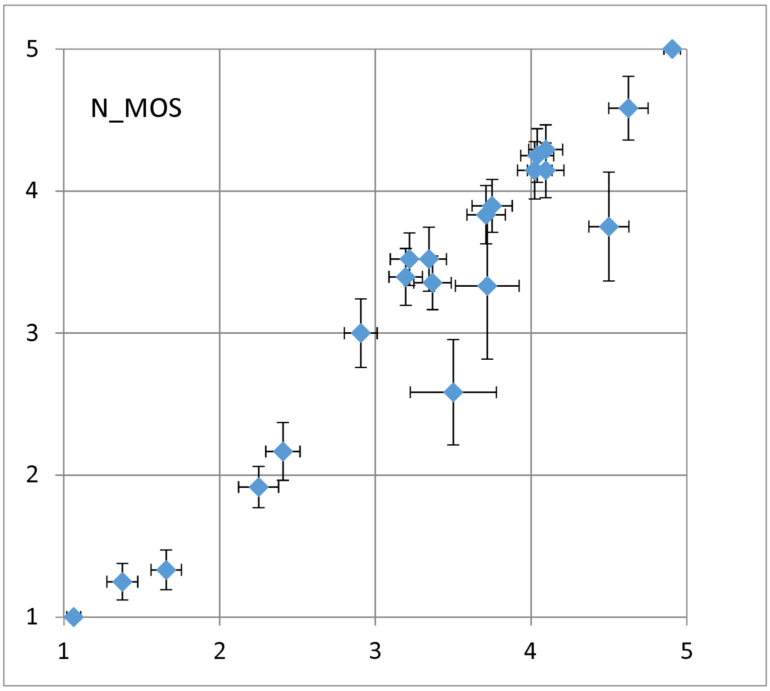
After the data correlations procedure, pairwise comparisons for the tests were evaluated. The comparison was performed in the following way: First, global MOS values of the first test were compared with global MOS values of the second test. Afterward, the absolute difference between each pair of samples was calculated. There were 231 cases (22 datasets).

After the pairwise comparison between Global qualities (G-MOS), ten differences were found which is 4,3 % of all cases. In these cases, users preferred one sample out of the pair without the parallel task but preferred the other one in the pair with the parallel task. Except for one case (the one marked by circles in Figure B.5) statistical analysis has shown those differences are statistically significant only at a confidence level 0,2 (CI80) but statistically insignificant at a confidence level of 0,05 (CI95). More subjects would be needed to obtain statistically more significant data. Although, the single case mentioned above is significant at confidence level 0,05 (CI95).



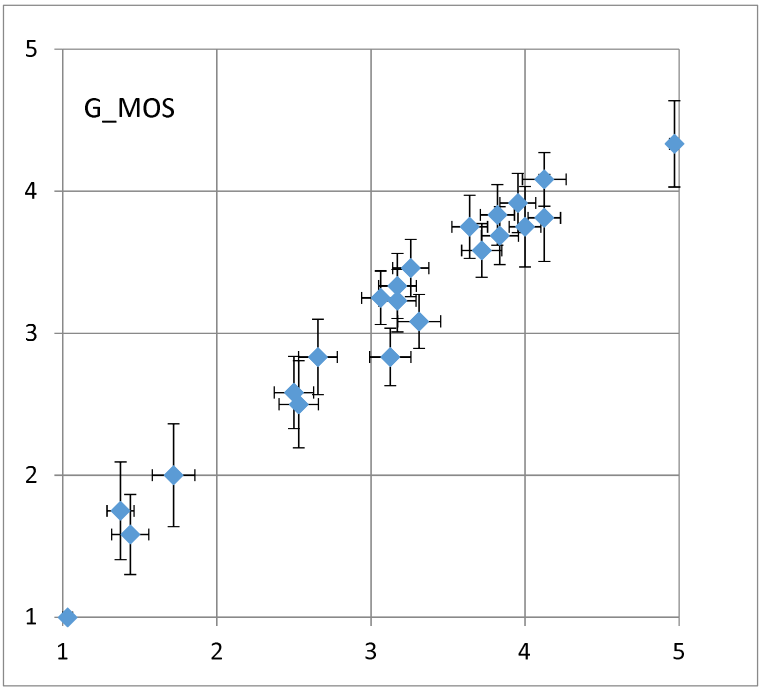
NOTE: Pearson correlation coefficient value is 0,971. During the voting process of speech samples, the subjects voted on speech signal distortion (5 - not distorted to 1 - very distorted).

Figure B.5: Speech MOS (S-MOS) comparison between test A (without the parallel task, horizontal) and test B (with the parallel task, vertical)



NOTE: Pearson correlation coefficient value is 0,982. The subjects voted for background noise annoyance (5 - not noticeable to 1 - very intrusive).

Figure B.6: Background noise (N-MOS) comparison between test A (without the parallel task, horizontal) and test B (with the parallel task, vertical)



NOTE: Pearson correlation coefficient value is 0.989. The subjects were voting for the overall quality of each sample (5 - excellent to 1 - bad).

Figure B.7: Overall quality (G-MOS) comparison between test A (without the parallel task, horizontal) and test B (with the parallel task, vertical)

History

Document history		
V1.1.1	March 2018	Publication
V1.2.1	October 2018	Publication