# A Novel Method of Subjective Sound and Audio Playback User Experience Evaluation in Multitasking Context: Two Case Studies

*Jan Holub[1], Yann Kowalczuk[1],*
*Magnus Schäfer[2], Jan Reimes[2]*

*[1]Czech Technical University in Prague,*
*Faculty of Electrical Engineering, Czech Republic*
*[2]HEAD acoustics GmbH, Herzogenrath, Germany*

## ABSTRACT

Present-day telecommunication devices are rarely utilized by comfortably seated users who do not perform any other parallel task. The typical communication scenario includes walking, driving a car, watching TV, working on a PC, etc., during a conversation. However, transmission quality evaluation has traditionally taken place in ideal laboratory conditions. The authors of this paper have prepared a new standard for subjective transmission quality testing with a parallel task that has been approved as ETSI TR 103 503. The paper summarizes the most widely-used transmission quality testing methods, discusses their disadvantages, and in-troduces a new testing methodology with a parallel task. It also presents two ex-periments performed in

Human Interaction & Emerging Technologies (IHIET-AI 2022)
Artificial Intelligence & Future Applications
https://openaccess.cms-conferences.org/#/publications/book/978-1-7923-8989-4

parallel task scenarios and highlights some differences in human perception in these scenarios.

**Keywords**: Quality of Experience · Parallel task · Psychomotor task · Subjective testing

# INTRODUCTION

Evaluating the quality of speech and multimedia signals is a crucial task in the design of modern telecommunication networks. The importance of this task grows with the increasing complexity and extent of these networks, where the communications chain involves more and more transmission technologies. The quality of multimedia transmission (e.g., voice, music, or video) thus becomes one of a few general measurable parameters for comparing different transmission devices and technologies from aspects that are substantially close to the end user's point of view (Gulliver & Ghinea 2006).

Transmission quality can be evaluated in two primary ways:

- Subjectively, with real users (humans).
- Instrumentally, using a computer algorithm that replicates human perception.

Instrumental tests are simple to perform and are easily repeatable. Most of the widely used objective algorithms, e.g., PESQ (ITU-T P.862 2001), POLQA (ITU-T P.863 2011), or E-model (ITU-T G.107 2015), provide excellent compliance with the results of subjective tests in typical applications. A disadvantage is the compromised reliability and the lower accuracy ratio for atypical applications, or for new methods in coding and compression of the signal on which the algorithm has not already been trained (Dalal 2011).

Subjective tests are further divided, e.g., for speech transmission testing, into listening tests (ITU-T P.800 1996, ITU-T P.835 2003, ITU-R BS.1116 2015, ITU-R BS.1534 2015), and conversational tests ((ITU-T P.800 1996 or ITU-T P.805 2007). A particular type of subjective test is a technique called crowd testing (Chang, Hsu, Hoßfeld & Chen 2018). This method uses crowdfunding practices for QoE testing in multimedia applications.

Since listeners' ratings are used directly for quality assessment, subjective tests provide the most accurate estimate of end-user opinion. They, therefore, are a reference method for other types of tests. The disadvantages of subjective tests are their high cost, and considerable time and organizational difficulties due to the need for many listeners. It is also necessary to provide a precisely defined environment for conducting the tests. The prescribed environment, the technical equipment, and the course of the test are common features of most standardized methods of subjective testing. However, the tested technologies are not always used only in an artificial, comfortable environment with relaxed and focused users. To bring subjective tests closer to the natural use of technologies and based on our experience and previous research (Avetisyan, Holub & Drábek 2018, Schäfer, Holub, Reimes & Drábek 2018), we have drafted a new standard for subjective multimedia transmission quality testing using a parallel task. This new standard has been approved by the European

Telecommunication Standardization Institution (ETSI) as ETSI TR 103 503. It introduces a parallel task to split the listeners' attention.

# STATE OF THE ART

Subjective testing with a parallel task is a current approach that has been used in several experiments and has brought exciting results. A detailed list and additional information can be found in (ETSI TR 103 503 2018). In (Kwak & Han 2017), the listeners had to remember digits displayed on a monitor when performing an intelligibility test. In (Beilock, Carr, MacMahon & Starkes 2002) experienced golfers and footballers showed off their skills while listening to a series of tones and being asked to identify a tone that had been specified in advance. It turned out that the participants achieved better results in their sport when they focused on a parallel task. In (Avetisyan & Holub 2018), the respondents performed a speech intelligibility test under standard laboratory conditions and again with an additional parallel task (a shooting simulator). Some test conditions produced higher scores in a parallel test than in a laboratory. The experiment (Holub, Slavata, Sula & Soares 2018) consisted of two parts. In the first part, the subjects were asked to drive a car simulator, while in the second part, they were asked to sort a variety of samples by taste. In both situations, the subjects performed the parallel task while assessing audio quality. The results obtained in the experiments with a parallel task showed considerable differences from the standard P.800 test. Until recently, the parallel task was an optional technique that was always performed without any requirements or recommendations. (ETSI TR 103 503 2018) drafted by the authors of this article, deals with that gap. It includes the state-of-the-art approaches of various published scientific experiments and necessary recommendations, classifications, types of subjects, and various scenario examples. These can guide researchers when organizing parallel task-based subjective speech quality, speech intelligibility, and listening effort tests. The document describes the methods of subjective audio quality and speech intelligibility assessments under parallel task conditions. It includes various scenarios such as laser shooting simulator, car driving simulator, tasting experiment, stationary bicycle, and virtual reality deployment.

# EXPERIMENTS AND RESULTS

## Experiment I: Comparison of Car Audio Quality with and without a Parallel Task

This section presents the results of subjective quality testing performed on audio data – excerpts of binaural music signals recorded in different car interiors using the car audio system. The quality of the stimuli was subjectively assessed while driving a (simulated) car as the parallel task. The obtained data are compared to an auditory evaluation using identical stimuli, and an identical playback configuration, but

without the parallel task. The results show decreased sensitivity of subjects for samples with significantly compromised quality.

The two tests followed the ITU-T P.800 ACR methodology, using a 9-point Mean Opinion Score (MOS)-like scale. The recorded music signals were sampled at 48-kHz with a 16-bit resolution. The tests were performed in an acoustically-treated critical listening room that conforms to the requirements of P.800 – background noise level of less than 30dB SPL(A) with no significant spectral peaks. A professional digital voting device was used to collect the votes in the experiment without a parallel task. In the experiment with a parallel task, listeners were asked to vote orally. Their selections were recorded and processed offline. The samples were normalized to the loudness of 23 sones (approximately 65 dB(A) for these signals). The .wav files were exported with full-scale corresponding to 100 dB SPL, forming a calibration point for the HW playout loudness calibration. A 9-point scale of 1 (worst)…9 (best) was used. The results have been recalculated to an MOS-like scale with 0.5 resolution (1.0, 1.5, 2.0, 2.5 … 4.5, 5.0). The experiment incorporated 32 listeners (16 males, 16 females) of various nationalities (13xCZ, 8xSK, 3xRU, 2xFR, 2xNZ, 1xUS, 1xGE, 1xTW, 1xDZ). The average age of the listeners was 29.1 years, with a standard deviation of 8.4 years. Several subjects above 50 years of age reported a slight-to-intermediate level of visually induced motion sickness (as known, e.g., from experiments deploying virtual reality). One of the experiments, originally introduced in a different context in (Schäfer, Holub, Reimes & Drábek 2018), indicates that in some cases, a systematic shift was observed in the assessment of mid and lower-quality samples deploying a parallel task (car simulator driving). In contrast, the results for the listening laboratory and the sound car (without parallel task) were very similar.

Another experiment, originally briefly introduced in the context of various parallel task types in (Holub, Slavata, Sula & Soares 2018), shows that subjects were less critical (gave higher scores) of low and medium-quality music recordings. They assessed the recordings considerably closer to 3. A brief overview of the most important observations is given here. Scatter plots of the results per listening condition and per stimulus are given in Fig. 1 and 2, respectively. The per-condition results are listed in Table 1. Comparing the plain and dual-task-based listening tests shows good agreement between the two listening situations for many stimuli. However, one clear trend can be observed: the lower end of the quality scale is not used as frequently when a parallel task is being performed. This has a necessary practical consequence – excellent quality samples (the very upper part of the MOS range) are critically perceived by the user even in the parallel task-based test. This confirms previous findings (Beilock, Carr, MacMahon & Starkes 2002, Kwak & Han 2017), where complex differences between pure laboratory tests and parallel task tests have been identified – differences that cannot be explained (only) by the subjects' loss of attention due to the introduction of a parallel task. A psycho-physiological explanation for this phenomenon lies beyond the scope of this paper.

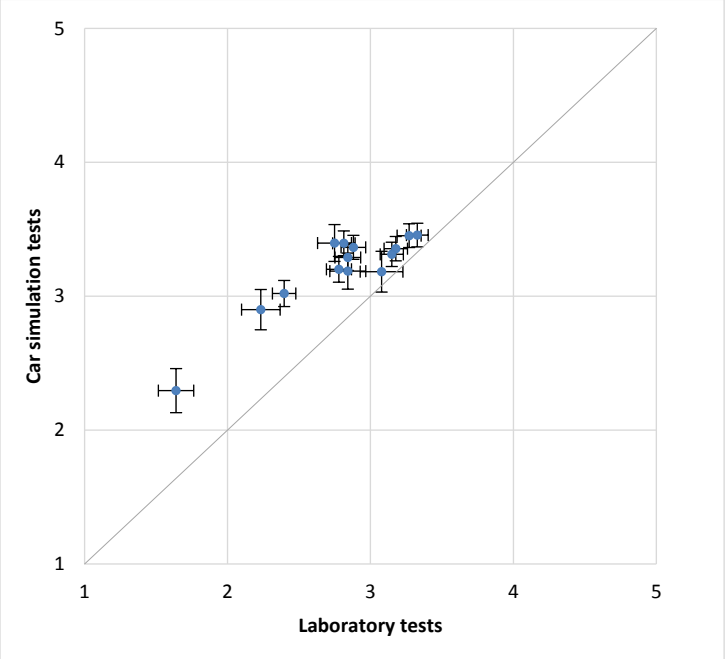


Figure 1 Comparison between the tests in laboratory conditions and simulated car environments. As can be seen from the graph, some MOS values for simulated car environments are significantly higher than MOS values in tests in laboratory conditions.

## Experiment II – ITU-T P.835 deploying a parallel task

For data analysis, two subjective tests were held in a subjective testing laboratory (Avetisyan & Holub 2018). In the following, they are labeled as a Lab test and Parallel test. The subjects of both tests were naïve and were fluent in English (language proficiency B2 and higher according to (Council of Europe 2011). The lab test featured 32 subjects, and the Parallel test 25 subjects.

A single set of English speech samples was used in both experiments. The speech sample set was prepared following all relevant requirements of the ITU-T P.800 and P.835 standards and contained 22 conditions. Contemporary coders (various bit-rates of EVS and AMR-WB) and selected cases of background noise (cafeteria, road, etc.) were used to create a balanced set of realistic speech samples with reasonably uniform coverage of the quality range.

The test methodology was based on ITU-T P.835. The principle of this standard is to make subjects listen to the same sample 3 times: first to assess the speech quality, then to assess the noise annoyance, and finally the overall sample quality. The lab test followed the P.835 procedure without any parallel task. During the Parallel test, an additional parallel task was included to distract the test subjects from full

concentration on subjective testing. A professional laser shooting simulator (Simway) was used as a simple parallel task (aiming and shooting towards a randomly moving target). This was performed in the following way: a panel of 3-4 subjects evaluated the samples. However, at any given time, one of them was a "shooter," and the other two or three were "counters." The "shooter's" task was to shoot as many in-game ducks as he/she could, and the task of the "counters" was to count each shot duck. The roles were assigned randomly, with a light-bulb identifying the shooter. The samples were played out in random order, using different randomization for each listening panel.

The tests were conducted in low-reverberation listening rooms, fully conforming to the requirements of ITU-T P.800 (reverberation time below 500ms, background noise below 30dB SPL (A) without significant spectral peaks). By further data processing, the corresponding MOS were obtained separately for Speech quality (SIG), Noise annoyance (BAK), and Overall sample quality (OVRL). The subjects had to vote for speech signal distortion (1 – very distorted to 5 – not distorted). Then, the subjects voted for background noise annoyance (1 – very intrusive to 5 – not noticeable). Finally, during the third part, the subjects voted for the overall quality of each sample (1 – bad to 5 – excellent). Figures 3 – 5 present the correlations between the SIG, BAK, and OVRL values. The values are highly correlated.

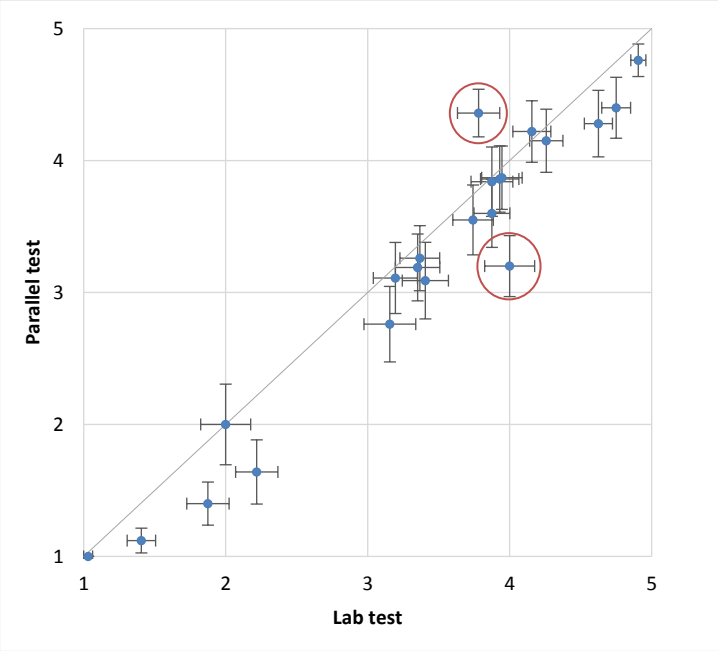


Figure 2 Speech MOS (SIG) of Lab test and Parallel test. X and Y axes show the Mean Opinion score values.

The SIG (S-MOS) comparison between the Lab and Parallel tests is shown in Fig. 3. Its correlation value is 0.971. In Fig. 3, there are two points (conditions 14 and 16) where the results of the two tests differ significantly (marked as red circles). These two samples provide a similar auditory result in the Lab test (approx. 3.8 and 4.0), while in the Parallel test, there is a clear difference in the results, and their rank order is opposite (approx. 4.4 vs. 3.2). This means that if such a test is used to select better technology from these two, the ranking order will differ completely with and without a parallel task. Assuming that the results with a parallel task were closer to the real-life experience, a selection based on a test without a parallel task would lead to inferior performance. Similar to the results for Experiment I (car driving), our results - MOS (SIG) in particular – confirm the previous findings that subjective results differ after the introduction of a parallel task.

# CONCLUSIONS

A new TR of the European Telecommunication Standardization Institute ETSI TR 103 503 was approved in March 2018. It is a novel allowing subjective multimedia testing procedures (e.g., speech quality testing, speech intelligibility testing, audio quality testing, video streaming visual quality testing, etc.) to be run under parallel task conditions, specifying the types of parallel tasks and classifying these into three basic categories (mental, physical, hybrid). These tests complement traditional standardized laboratory procedures, performed in defined environments using rigorous listening/conversational procedures and requiring relaxed, fresh, fit, and focused naive or expert listeners, comfortably seated in a listening chamber to minimize background noise and room reverberation.

The parallel task-based test procedure better mimics the daily use of the tested technologies, e.g., voice services are sometimes used while driving or working, etc., and their users are stressed, tired, or concentrating on another, usually important, task. The parallel task is designed to place an additional load on subjects, comparable to the activity performed during the real targeted situation, without losing the test repeatability that can be achieved in a laboratory environment. The only limitations are due to requirements for laboratory equipment, load task repeatability, space and movement restrictions, or safety.

Two different experiments, performed both in the traditional way and with a parallel task, have been presented, and the results have been discussed. Identified differences between our results and regular laboratory tests demonstrate the importance of parallel task-based standardized procedures for tests on future and emerging technology.

# REFERENCES

Avetisyan H & Holub J 2018, ‘Subjective speech quality measurement with and without parallel task: Laboratory test results comparison’, *PLoS One*, vol. 13, no. 7, pp. 1–8.

Avetisyan H, Holub J & Drábek T 2018, 'Low Bit-rate Coded Speech Intelligibility Tested with Parallel Task', *ACTA ACUSTICA UNITED WITH ACUSTICA*, 104(4), pp. 678-684.

Beilock SL, Carr TH, MacMahon C & Starkes JL 2002, '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.

Council of Europe 2011,'Common European Framework of Reference for Languages: *Learning, Teaching, Assessment (CEFR)'*.

Dalal, AC 2011, 'User-perceived quality assessment of streaming media using reduced feature sets," *ACM Trans. Internet Technol.*, vol. 11, no. 2, pp. 1–32.

ETSI TR 103 503 2018, 'Speech and multimedia Transmission Quality (STQ); Procedures for Multimedia Transmission Quality Testing with Parallel Task including Subjective Testing', *European Telecommunication Standardization Institution*, Sophia Antipolis, pp. 1–17.

Gulliver, SR & Ghinea, G 2006, 'Defining user perception of distributed multimedia quality', *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 2, no. 4, pp. 241–257, 2006.

Holub J, Slavata O, Sula J & Soares L 2018, 'Subjective audio quality testing, with tasting and car driving as parallel tasks', *IEEE Access*, vol. 6, pp. 1–1.

Chang HS, Hsu CF, Hoßfeld T & Chen KT 2018, 'Active Learning for Crowdsourced QoE Modeling', *IEEE Trans. Multimed.*, vol. 20, no. 12, pp. 3337–3352.

ITU-T P.862 2001, 'Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs' *Int. Telecommun. Union*, Geneva, vol. 862, p. 862.

ITU-T P.863 2011, 'Perceptual Objective Listening Quality Assessment' *Int. Telecommun. Union*, Geneva, vol. 863.

ITU-T G.107 2015, 'The E-model: a computational model for use in transmission planning', *Int. Telecommun. Union*, Geneva

ITU-T P.800 1996, 'Methods for subjective determination of transmission quality' *Int. Telecommun. Union*, Geneva

ITU-T P.835 2003, 'Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm', *Int. Telecommun. Union*, Geneva.

ITU-R BS.1116 2015, 'Methods for the subjective assessment of small impairments in audio systems," *Int. Telecommun. Union*, Geneva.

ITU-R BS.1534 2015, 'Method for the subjective assessment of intermediate quality level of audio systems', *Int. Telecommun. Union*, Geneva.

ITU-T P.805 2007, 'Subjective evaluation of conversational quality', *Int. Telecommun. Union*, Geneva.

Kwak C, Han W 2017, 'Comparison of Single-Task versus Dual-Task for Listening Effort', *J. Audiol. Otology,* 22(2), pp.69-74.

Schäfer M, Holub J, Reimes J & Drábek T 2018 'Subjective Testing of Car Audio Systems With and Without Parallel Task', *Proceedings of DAGA 2018: 44. Deutsche Jahrestagung für Akustik*, Munich, Germany, pp. 326-327.