Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Measurement

Parallel Task in Subjective Audio Quality and Speech Intelligibility Assessments

Doctoral Thesis

Mgr. Hakob Avetisyan

Ph.D. programme: P2612 - Electrical Engineering and Information Technology
Branch of study: 2601V006 - Measurement and Instrumentation
Supervisor: Prof. Ing. Jan Holub, Ph.D.

Prague, January 2021

Thesis Supervisor:
Prof. Ing. Jan Holub, Ph.D.
Department of Measurement
Faculty of Electrical Engineering
Czech Technical University in Prague
Technická 2
160 00 Prague 6
Czech Republic

Declaration

I hereby declare I have written this doctoral thesis independently and quoted all the sources of information used in accordance with methodological instructions on ethical principles for writing an academic thesis. Moreover, I state that this thesis has neither been submitted nor accepted for any other degree.

In Prague, January 2021

.....
Mgr. Hakob Avetisyan

Abstract

This thesis deals with the subjective testing of both speech quality and speech intelligibility, investigates the existing methods, record their main features, as well as advantages and disadvantages. The work also compares different tests in terms of various parameters and provides a modern solution for existing subjective testing methods.

The first part of the research deals with the repeatability of subjective speech quality tests provided in perfect laboratory conditions. Such repeatability tasks are performed using Pearson correlations, pairwise comparison, and other mathematical analyses, and are meant to prove the correctness of procedures of provided subjective tests. For that reason, four subjective speech quality tests were provided in three different laboratories. The obtained results confirmed that the provided tests were highly repeatable, and the test requirements were strictly followed.

Another research was done to verify the significance of speech quality and speech intelligibility tests in communication systems. To this end, more than 16 million live call records over VoIP telecommunications networks were analyzed. The results confirmed the primary assumption that better user experience brings longer call durations. However, alongside the main results, other valuable conclusions were made.

The next step of the thesis was to investigate the parallel task technique, existing approaches, their advantages, and disadvantages. It turned out that the majority of parallel tasks used in tests were either physically or mentally oriented. As the subjects in most cases are not equally trained or intelligent, their performances during the tasks are not equal either, so the results could not be compared correctly.

In this thesis, a novel approach is proposed where the conditions for all subjects are equal. The approach presents a variety of tasks, which include a mix of mental and physical tasks (laser-shooting simulator, car driving simulator, objects sorting, and others.). Afterward, the methods were used in several subjective speech quality and speech intelligibility tests. The results indicate that the tests with parallel tasks have more realistic values than the ones provided in laboratory conditions.

Based on the research, experience, and achieved results, a new standard was submitted to the European Telecommunications Standards Institute with an overview, examples, and recommendations for providing subjective speech quality and speech intelligibility tests. The standard was accepted and published under the number ETSI TR 103 503.

Keywords: Subjective testing, parallel task, psychomotor task, speech quality, speech intelligibility

Abstrakt

Tato disertační práce se zabývá subjektivním testováním jak kvality řeči, tak i srozumitelnosti řeči, prozkoumává existující metody, určuje jejich základní principy a podstaty a porovnává jejich výhody a nevýhody. Práce také porovnává testy z hlediska různých parametrů a poskytuje moderní řešení pro již existující metody testování.

První část práce se zabývá opakovatelností subjektivních testování provedených v ideálních laboratorních podmínkách. Takové úlohy opakovatelnosti se provádí použitím Pearsonové korelace, porovnání po párech a jinými matematickými analýzami. Tyto úlohy dokazují správnost postupů provedených subjektivních testů. Z tohoto důvodu byly provedeny čtyři subjektivní testy kvality řeči ve třech různých laboratořích. Získané výsledky potvrzují, že provedené testy byly vysoce opakovatelné a testovací požadavky byly striktně dodrženy.

Dále byl proveden výzkum pro ověření významnosti subjektivních testování kvality řeči a srozumitelnosti řeči v komunikačních systémech. Za tímto účelem bylo analyzováno více než 16 miliónů záznamů živých hovorů přes VoIP telekomunikační sítě. Výsledky potvrdily základní předpoklad, že lepší uživatelská zkušenost působí delší trvání hovorů. Kromě dosažených hlavních výsledků však byly učiněny další důležité závěry.

Dalším krokem disertační práce bylo prozkoumat techniku paralelních zátěží, existující přístupy a jejich výhody a nevýhody. Ukázalo se, že většina paralelních zátěží používaných v testech byla buď fyzicky, nebo mentálně orientovaná. Jelikož subjekty ve většině případů nejsou stejně fyzicky nebo mentálně zdatní, jejich výkony během úkolů nejsou stejné, takže výsledky nelze správně porovnat.

V této disertační práci je navržen nový přístup, kdy jsou podmínky pro všechny subjekty stejné. Tento přístup představuje celou řadu úkolů, které zahrnují kombinaci mentálních a fyzických zátěží (simulátor laserové střelby, simulátor řízení auta, třídění předmětů apod.). Tyto metody byly použity v několika subjektivních testech kvality řeči a srozumitelnosti řeči. Závěry naznačují, že testy s paralelními zátěží mají realističtější výsledky než ty, které jsou prováděny v laboratorních podmínkách.

Na základě výzkumu, zkušeností a dosažených výsledků byl Evropskému institutu pro normalizaci v telekomunikacích předložen nový standard s přehledem, příklady a doporučeními pro zajištění subjektivních testování kvality řeči a srozumitelnosti řeči. Standard byl přijat a publikován pod číslem ETSI TR 103 503.

Klíčová slova: Subjektivní testování, paralelní zátěž, psychomotorická úloha, kvalita řeči, srozumitelnost řeči

Acknowledgements

First of all, I would like to sincerely thank and express my appreciation to my supervisor Prof. Jan Holub for his unconditional support and guidance with my thesis and studies.
Next, my studies and research would be much harder without the help and assistance of my colleagues and our academic personnel. Special thanks to Prof. Haasz, Prof. Ripka, Prof. Šmíd, doc. Roztočil, Dr. Svatoš, Dr. Slavata, Dr. Sobotka, Ing. Drábek, Dr. Bruna, Ing. Hanuš, Ing. Pospíšil and Ing. Kubeš. Big thanks to Dean's office headed by Prof. Páta, and faculty administrative staff, especially Mrs. Kroutilíková, Mrs. Florianová, Mgr. Sankotová and Mrs. Kočová.
I want to express my deepest gratitude to the Czech Government with the Ministry of Education, Youth and Sports, The Centre for International Cooperation in Education (DZS), Czech Embassy in Armenia with the former Ambassador Petr Mikyska for this great opportunity to live and study in the wonderful country of the Czech Republic. My appreciation to the Institute for Language and Preparatory Studies of Charles University in Poděbrady for their contribution in my knowledge of Czech language, which made my integration in the Czech society much smoother.
This research was possible thanks to the Grant Agency of the Czech Technical University in Prague, grant No. SGS17/191/OHK3/3T/13

List of Tables

All the tables included in individual articles are listed directly in the articles.

List of Figures

All the figures included in individual articles are listed directly in the articles.

List of Acronyms

ACD	Average Call Duration
ACR	Absolute Category Rating
AMR	Adaptive Multi-Rate audio codec
CCR	Comparison Category Rating
CDR	Call Detail Records
CI95	The 95% Confidence Interval
CMR	Critical Minute Ratio
CS-ACELP	Conjugate Structure Algebraic-Code-Excited Linear Prediction
DALT	Diagnostic Alliteration Test
DCR	Degradation Category Rating
DMCT	Diagnostic Medial Consonant Test
DRT	Diagnostic Rhyme Test
ETSI	European Telecommunications Standards Institute
EVS	Enhanced Voice Services Codec
G-MOS/OVRL	Overall sample quality MOS
HMMWV	High Mobility Multipurpose Wheeled Vehicle
ITU	International Telecommunication Union
MELPe	Mixed-Excitation with Linear Predictive enhanced
MOS	Mean Opinion Score
MRT	Modified Rhyme Test
N-MOS/BAK	Noise annoyance MOS
PCM	Pulse Code Modulation
PESQ	Perceptual Evaluation of Speech Quality
POLQA	Perceptual Objective Listening Quality Analysis
QoE	Quality of Experience
RMS	Root Mean Square
RMSE	Root Mean Squared Error
RTP	Real-time Transport Protocol
SIP	Session Initiation Protocol
S-MOS/SIG	Speech quality MOS
SNR	Signal-to-Noise Ratio
STD	Standard Deviation of Arithmetical Mean
VoIP	Voice over Internet Protocol

Contents

Chapter 1

Introduction

1.1 Motivation

In the age of telecommunication and automation, technologies progress dramatically and with incredibly fast speeds. Since it is strongly believed that better call quality brings to longer call durations [1], each generation of devices has various advanced parameters and functions designed to have a higher quality audio signal processing and noise suppression. To achieve these goals, various objective and subjective tests are provided to investigate, compare and improve the audio quality and speech intelligibility of emerging mobile technologies according to ITU-T recommendations P.800 [2], P.835 [3], P.807 [4], and others.

1.2 Objective Tests

Objective methods are aimed to replace test subjects using applicable psycho-acoustic modeling and to compare clean and distorted speech samples algorithmically. This assessment is based on the physical parameters of the transmission channel. Most of the commonly used objective algorithms, e.g., PESQ [5], POLQA [6], or E-model [7], provide excellent compliance with the results of subjective tests in typical applications. The disadvantages of this method are the worse 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 [8]. Unlike objective tests, subjective tests are believed to provide more accurate results, but they are also more demanding regarding time, equipment, effort, and price.

1.3 Subjective Tests

1.3.1 Subjective speech quality testing

Subjective speech quality testing or Quality of Experience (QoE) testing is designed to collect subjective opinions (votes) from human test subjects following standardized procedures specified, e.g., in [2].

ITU-T Recommendation P.835 [3] depicts methods for measuring speech quality for noisy (and partially de-noised) speech. Its main application is a comparison of different noise suppression algorithms. The advantage of the P.835 is that it makes it possible to evaluate the speech quality and noise levels separately. Parameters of test environments are adopted from ITU-T Recommendation P.800. Subjects evaluate tested samples on a five-point scale separately for speech quality, noise annoyance, and overall quality of samples.

Both objective and subjective speech quality tests outputs are often mapped to the subjective quality Mean Opinion Score (MOS). Terms S-MOS, N-MOS, and G-MOS are adopted from ETSI TS 103 106 [9] and ETSI EG 202 396-3 [10]. These terms replace in the further text the original SIG, BAK, and OVRL ratings used in [3].

1.3.2 Subjective speech intelligibility testing

In audio communication, intelligibility is being used to measure the clarity of speech in various conditions. It has a direct impact on the amount of information transferred by the communication act among the communicating parties. Intelligibility tests [4], [11], [12] are designed to evaluate the human (or automated listener) ability to understand the meaning of spoken words. In most cases, the output of objective and subjective speech intelligibility tests is represented by a percentage of total correct votes by the subjects for each sample.

Multiple methods and their modifications were developed [13] and are discussed next:

DRT - Diagnostic Rhyme Test

The DRT measures the intelligibility of speech over communication systems. The test materials contain 96 rhyming monosyllable word pairs (e.g., veal-feel) that were selected to differ in the initial consonant. During the test, the listener is asked which of the two rhyming words presented was spoken.

DMCT - Diagnostic Medial Consonant Test

The DMCT is a variation of-the DRT with test materials consisting of 96 bi-syllable word pairs (e.g., stopper-stocker) selected to differ in only their intervocalic consonant.

DALT - Diagnostic Alliteration Test

The DALT is another variation of the DRT. Ninety-six monosyllable word pairs (e.g., pack-pat) selected to differ in their final consonant only.

MRT - Modified Rhyme Test

The MRT [14] was standardized by ANSI to measure the intelligibility of communication systems. Its test material consists of 50 rhyming monosyllable word sets of 6 words (e.g., pin, sin, tin, fin, din, win) selected with half to differ in the initial consonant and the other half in the final consonant. As in the DRT case, the listener is asked which of the six presented rhyming words was spoken.

1.4 Scope and Objective

This thesis is concentrated on subjective speech quality and subjective speech intelligibility test techniques and related parallel tasks, which are meant to simulate real-life conditions for the proposed tests.

To answer the most important questions of this thesis, the following hypotheses were formulated:

1. Subjective tests provided in different laboratories are repeatable if test requirements are strictly followed.
2. In VoIP communications, higher call qualities lead to longer call durations in average.
3. Subjective speech intelligibility values are lower when subjects perform additional (parallel) tasks instead of fully concentrating on the conversation.
4. Subjective speech quality test results performed with parallel task provide higher vote dispersion (and thus RMS) while keeping the MOS values (arithmetical mean) unchanged.

1.5 Structure of the Thesis

This doctoral thesis is written in the format of a thesis by publication approved by the Dean of Faculty of Electrical Engineering and by the Directive for dissertation theses defense, Article 1.

The thesis presents publications relevant to the topic of the thesis as individual chapters. Each chapter begins with a summary section, where the main topic, conclusion, and contribution of the research work is explained.

The whole thesis is divided into nine parts with corresponding chapters.

Chapter 2 analyzes already existing approaches, State of the Art mechanisms, their advantages, and disadvantages.

Chapter 3 deals with a repeatability verification of inter-lab and intra-lab subjective speech quality assessment results.

The aim of Chapter 4 was to confirm call duration dependency on the quality of VoIP calls.

Chapter 5 brings the idea of the parallel task in subjective speech intelligibility testing.

In Chapter 6, the parallel task is used in subjective speech quality measurement.

Chapter 7 presents the implementation of the parallel task in the Czech language.

The conclusion and summary, as well as hypotheses verification and future work, are provided in Chapter 8

Chapter 2

State of the Art

2.1 Introduction to Recommendations

Various standardized methods for performing subjective tests were designed for listening tests and for conversation. They are also different in terms of listeners selection or rating scales. The most commonly used standards for subjective tests are:

2.1.1 ITU-T Recommendation P.800

ITU-T P.800 [2] includes a set of methods for the quality of speech transmission evaluation. For covered methods, the standard defines test room parameters (dimensions, reverberation, noise levels, etc.). The listeners are allowed to participate the tests once a specified period of time. Some regularly used methods are:

- Absolute Category Rating (ACR) is a listening test where the subjects evaluate the sample on a scale from 1 (bad) to 5 (excellent) without knowledge of the reference signal.
- The Conversation-opinion test is a conversational test, where two subjects evaluate the quality of the conversation on the same scale as in the ACR.
- Comparison Category Rating (CCR), Degradation Category Rating (DCR), and Quantal-Response Detectability Tests are methods that use various rating scales and various ways of comparing a degraded signal with a reference.

2.1.2 ITU-T Recommendation P.835

ITU-T P.835 [3] describes methods for evaluating speech quality in the event of noisy (and partially de-noised) speech. It is used, e.g., to compare various noise suppression algorithms. The methodology makes it possible to evaluate speech quality and noise levels separately.

The test environment parameters are adopted from ITU-T P.800. Listeners evaluate tested samples on a five-point scale.

2.1.3 ITU-R Recommendation BS.1116

Recommendation ITU-R BS.1116-3 [15] serves to assess minor faults in audio systems and for comparing technologies with high transmission quality. The "double-blind triple-stimulus with hidden reference" method uses a comparison of a reference sample with two tested samples, where one of the tested samples is a hidden reference. Unlike the P.800 series, this standard suggests using of expert listeners for more reliable fault detection. The listener evaluates interference in the tested sample on a scale from 1 (very annoying) to 5 (imperceptible).

The standard prescribes the characteristics of the audio equipment that is used, and also the dimensions and parameters of the test room.

2.1.4 ITU-R Recommendation BS.1534 MUSHRA

Recommendation ITU-R BS.1534 [16] "Multiple Stimuli with Hidden Reference and Anchor" is similar to BS.1116, but designed to test systems with lower transmission quality and greater disturbances. The use of expert listeners makes it possible to achieve more accurate results with fewer evaluators [17]. Researchers in [18] found no difference in the ratings of naive and expert listeners when there were only timbre artifacts in the tested signal. In contrast, [19] shows that spatial artifacts are more rigorously evaluated by expert listeners. Listeners have access to all the tested samples at the same time, and samples can be freely compared with the reference. In the test set, there is also one hidden reference, and there are one or two anchors (intentionally degraded/filtered references). Listeners evaluate samples on a scale of 0-100 points, often using sliders in a computer program.

2.1.5 Crowdtesting

Crowdtesting is a method that uses some crowdfunding practices for QoE testing in multimedia applications. Each tester/user evaluates the quality of the transmission himself in his own environment. The advantage is a large and diverse panel of internationally geographically distributed users in realistic user settings, and a reduction in cost and in organizational demands. The problem is to ensure that the evaluation is reliable. Advanced statistical methods are required to identify reliable user ratings and to ensure high data quality. The method is described in more detail in [20], [21], and [22].

2.1.6 Comparison of Recommendations

Table 2.1 describes commonly used recommendations, their advantages and disadvantages.

Table 2.1: Comparison of various recommendations, their advantages and disadvantages

Recommendation	Description	Advantages	Disadvantages
ITU-T P.800	Listening and conversational tests	Traditional and proven method	Less sensitive to small impairments
ITU-T P.805	Conversational tests	Allows for delay and echo evaluation	Complex setup, real-time network simulator required
ITU-T P.835	Listening tests with separate speech and noise quality judgement	Results show smaller variance than with P.800 for noisy and/or denoised samples	Time consuming (each sample played three times)
ITU-R BS.1116	Listening tests for small impairments judgement	Sensitive to small impairments	Expert listeners required
ITU-R BS.1534 MUSHRA	Listening tests with hidden reference and anchor	Sensitive to intermediate level of impairments	Time consuming (triple stimulus listening)
Crowdtesting	Deploying larger amount of (remote) listeners	Cost and time efficient	Complex measures needed to exclude unreliable responses

2.2 Problem Definition

The main issue of subjective tests is the fundamental philosophy of currently used testing methods. They suggest that the test subjects are comfortably seated in a test room (usually anechoic or semi-anechoic) and are entirely focused on listening to the tested material. However, in real life, the users usually perform multiple tasks at once (such as talking on the phone and working on a PC, walking or driving a vehicle, or monitoring a screen with airplane location and approach situation while communicating with the airplane pilot on the radio-link).

Most of the existing approaches and related standardized recommendations assume that the method of laboratory testing with subjects fully concentrated on subjective tests provides the most robust results, which is not always true.

To bring the above stated subjective tests results closer to the real-life conditions, additional psychomotor (or parallel) tasks are implemented into those tests. This technique is aimed to distract the users' attention from the main subjective testing procedure.

2.3 Defined Types of Parallel Tasks

In scientific literature, parallel-task techniques can be divided into three types: Mentally oriented tasks, Physically oriented tasks and Hybrid tasks. Some of available experiments are discussed in Table 2.2.

Table 2.2: Resource Summary

Reference	Test type	Parallel task	Parallel task type	Language
[23]	Speech intelligibility	Memorizing digits	Mentally oriented	N/A
[24]	Speech intelligibility	Memorizing digits	Mentally oriented	English
[25]	Speech intelligibility	Word repetition; Memorizing digits	Mentally oriented	English
[26]	Speech intelligibility	Memorizing sentences, Arithmetic	Mentally oriented	Korean
[27]	Speech intelligibility	Pressing color buttons	Mentally oriented	German
[28]	QoE test	Matching colored squares	Mentally oriented	N/A
[29]	Speech intelligibility	Forward / backward discrimination and speech understanding	Mentally oriented	English
[30]	Other	Memorizing tones; memorizing words	Physically oriented	N/A
[31]	Speech intelligibility	Turning a nut on a bolt	Mentally oriented	English
[32]	Other	Telephone call	Hybrid	English
[33]	QoE test	Tasting; Car driving	Hybrid	English

2.3.1 Mentally Oriented Tasks

Among observed experiments with mentally oriented parallel-tasks, memory-related tasks requiring subsequent repetition and memorization of information are used the most frequently. For instance, in the experiment [23], the test subjects had to identify prescribed letters while remembering the five digits displayed or played before the description. The results of this experiment depend on both the quality of the used codec and the intelligibility of the description and on the way the numbers are presented and how the conditions are sorted (serial/random). Memorization tasks are also used in [24], [28], and [29]. In the experiment [24], the primary test condition consisted of the different levels of noise in the test sentence background. The listeners' task was to repeat the last word of the sentence or try to guess it if it was unintelligible. The second task for the listeners was to repeat the last words after eight sentences. The next experiment [28] included a group of 64 children for a

speech intelligibility test. A task for half of them was to remember digits, and for the other half – to pay their primary attention to word repetition. Single-task and dual-task performances were compared. Results show that significant dual-task decrements were found for digit recall, but no dual-task decrements were found for word recognition. In [29], subjects, as a parallel task, had to write down the sentence they heard or write down the sum of first and third numbers they heard.

In experiments [25] and [30], listeners were asked to solve simple mathematical examples while pressing the corresponding keys to respond to different colors displayed on a screen. [25] was primarily about comparing different speech synthesis systems. In [30], human and synthesized speech with transmission degradation (compression, noise, packet loss) were compared. In both experiments [25] and [30], the results showed that the worse the quality of speech (and thus the clarity of the assignment of the primary task) is, the longer the reaction times in the secondary task. In [30], under the worst-case transmission, some subjects completely omitted the secondary task. In the next experiment [31], younger and older adults had to understand the target talker with and without determining how many masking voices were presented in samples time reserved.

2.3.2 Physically Oriented Tasks

Physically oriented tasks usually include activities like running, cycling, or other physical or sporting exercises. For example, the experiment [26] consisted of two parts. The first part included professional golfers as subjects that performed the golf-putting task on a carpeted indoor putting green. At the same time, they had to listen to a series of tones from the audio player and to identify and report a particular tone. The results showed that the subjects performed better with an additional listening task than without it. In the second part, the respondents' task was to lead the soccer ball by slalom from cones while listening to a series of words, meanwhile identifying and repeating the target word. The group of respondents consisted of experienced football players and non-players. Experienced players played better in the slalom in a parallel-task test. The presence of a secondary task and distraction led experienced athletes to better perform automatic and rehearsal moves.

2.3.3 Hybrid Tasks

Hybrid tasks include both physical and mental activities. In the experiment [27], subjects had to drive a car in a simulated driving environment while handling a telephone call.

In contrast to driving without a phone, the driver was significantly more likely to miss traffic marks while telephoning. They also had longer reaction times. The next experiment [33] deals with two different hybrid parallel tasks. In the first part, the subjects' task was to

drive a car on a PC-based car-driving simulator while assessing a music loudness of various car sound systems. The next part of the experiment was focused on a sense of taste of subjects. Several substances of different tastes, namely salt, sugar, wheat, and sweeteners, were chosen as the main ingredients. The task was to differentiate among samples with different proportions of these ingredients. For the first part of the experiment, results show that the subjects were less critical of low-quality and medium-quality music recordings. However, high-quality recordings were identified with no sensitivity degradation. For the next part, similarly to the driving task, the quality in medium-quality audios is not sensitively perceived when the subject is under a load.

Chapter 3

Comparison of Different Laboratory Test Results of Subjective Speech Quality Measurement

3.1 Summary

At the very beginning of the research, it was necessary to understand the basics of subjective speech quality testing, as well to confirm the repeatability of the results among the test provided by different laboratories. For this purpose, four different tests running the same set of speech samples were provided in 3 laboratories deploying ITU-T P.835 methodology. Two of the laboratories were located in the USA (Mountain View, CA and Boulder, CO, USA) and one in Europe (Prague, Czech Republic). After that, the test results have been compared in terms of Pearson correlation, RMSE, RMSE*, and numbers of pairwise comparisons. The results of the tests show the level of inter-lab and intra-lab repeatability with the deployment of identical test speech samples, and in strictly followed rules and requirements of ITU-T P.800 and ITU-T P.835. This confirms that the performed tests were highly repeatable.

3.2 Publication

The detailed information about the research and test results are presented below in the article [34].

Int J Speech Technol (2017) 20:69–74
DOI 10.1007/s10772-016-9389-6

Subjective speech quality measurement repeatability: comparison of laboratory test results

Jan Holub[1] · **Hakob Avetisyan**[1] · **Scott Isabelle**[2]

Received: 9 September 2016 / Accepted: 21 October 2016 / Published online: 3 November 2016

Abstract This article reports on a multi-lab subjective listening experiment aiming at inter-lab and intra-lab test results repeatability verification. An identical set of speech samples corresponding to contemporary networks has been tested by three independent labs deploying ITU-T P.835 methodology. The tests results have been compared regarding Pearson correlation, RMSE, RMSE* and numbers of opposite pair-wise comparisons. The results show the level of inter-lab and intra-lab repeatability in the case of identical test speech samples utilization and thus confirm the subjective tests are highly repeatable in case they follow recommendation requirements strictly. The tests also show differences in results in case subject expectations are set differently using a wider set of test speech samples (as presented in one of the labs).

Keywords Mean opinion score · Speech quality · Subjective testing · Test repeatability · Test reproducibility

1 Introduction

Nowadays technology is being developed with incredibly high speed. New gadgets are appearing every month with different characteristics and functions. A substantial part of them process voice data and each of them is required to provide as high quality as possible. For this purpose, audiovisual subjective and objective tests are being performed to analyze and improve the audiovisual quality of future products. Subjective speech quality testing is based on collecting subjective opinions (votes) from human test subjects following standardized procedures as specified e.g. in ITU-T P.800. (1996). Objective methods attempt to replace human test subjects with relevant signal processing procedures, comparing clean and distorted speech samples algorithmically. Their final output parameter is often mapped into subjective quality scale (MOS) (ITU-T P.863 2014) The difference between subjective and objective speech quality assessment is that subjective tests are more reliable but compared to objective tests also more demanding in terms of time, effort and price. To be sure that test results are relevant they are often held in different laboratories in parallel and afterwards the results agreement is checked. However, in such repeated experiments, there is usually one or more parameters that differ lab-to-lab, typically language used and deployed test subject nationality are varied (Goodman and Nash 1982). The purpose of our testing was to understand the level of inter-lab and intra-lab test results repeatability while keeping all parameters as identical as possible, including the language and nationality aspects.

In Chapter II, description of the experiment with basic information about test locations and dates, considered standards, number conditions of tested subjects, used equipment, etc. is introduced.

Chapter III shows data analysis on obtained results of mean opinion scores (MOS) of speech quality, noise annoyance, and overall quality (as per ITU-T P.835 2003) and correlations between each couple of tests were calculated. Afterwards, pairwise comparisons between each couple of tests were performed without and with consideration of CI 95 coefficient interval.

✉ Hakob Avetisyan
avetihak@fel.cvut.cz

[1] Department of Measurement, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic

[2] Boston University, Boston, USA

Chapter IV contains the conclusion and final remarks about tests.

2 Experiment description

For data analysis four subjective tests were held in three various laboratories. They are named as A, B and C. Two of them were located in the US (Mountain View, CA and Boulder, CO, USA) and one in Europe (Prague, Czech Republic). Test A was performed in December 2014. B1 and B2 tests were held in April and July, 2015 respectively, and C tests were held in May, 2015. B1 and B2 tests were performed in the same laboratory however, subjects for B2 test were different from B1 test.

To be able to verify the inter-lab and intra-lab test results repeatability, the same sample set had to be used in all experiments. The speech sample set was prepared following all relevant requirements of ITU-T P.800 (1996) and P.835 standards, i.e. two male and two female talkers, studio quality recording apparatus with high SNR, anechoic recording environment, test sentences have been selected from Harvard set of phonetically balanced sentences (Appendix of IEEE Subcommittee on Subjective Measurements 1969). The final selection contained 22 conditions, each condition being represented by at least 16 sentences (4 different sentences spoken by four different speakers). Contemporary coders (wideband versions of Adaptive Multi-Rate audio codec (AMR) (3GPP TS 26 071) and Codec for Enhanced Voice Services (EVS) (2015) and selected cases of background noise (cafeteria, road, etc.) have been used to create a balanced set of realistic speech samples with reasonably uniform coverage of quality range (see Figs. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12). It shall be noted that data set A contained a larger set of conditions, for a total of 60, of which 22 were tested in tests B and C. Only those 22 common conditions are compared in this article, however, the fact the A tests were originally performed using wider set of conditions can create bias between A and others tests.

The test methodology was based on ITU-T P.835 standard. This procedure is particularly suitable for samples processed by noise cancelling algorithms or coders where a certain part of background noise is removed but the speech itself is partially corrupted, too. The basic idea of the P.835 is to repeat the assessment of each speech sample three times, asking the subjects to focus on different aspect of the sample quality during each playout: For the first half of samples, the subjects are asked to focus on speech quality only during the first playout, to noise annoyance during the second playout and to overall sample quality during the last third playout. For the second half of samples, the first playout is to judge noise annoyance, the speech quality is

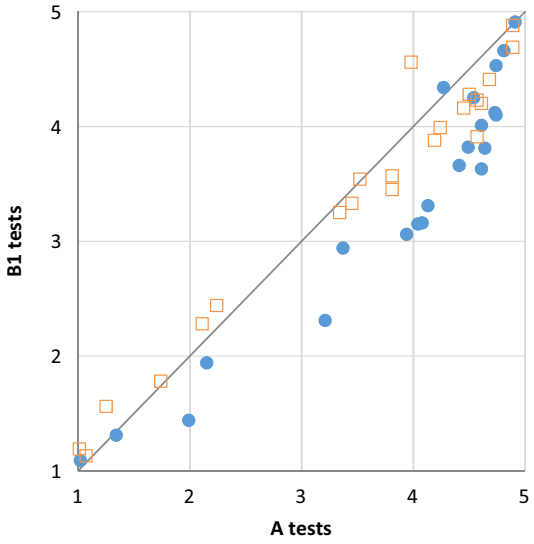


Fig. 1 Speech MOS (S-MOS) and noise annoyance MOS (N-MOS) of A and B1 tests. *Blue circles* show dependencies of S-MOS values. *Orange squares* show dependencies of N-MOS values. (Color figure online)

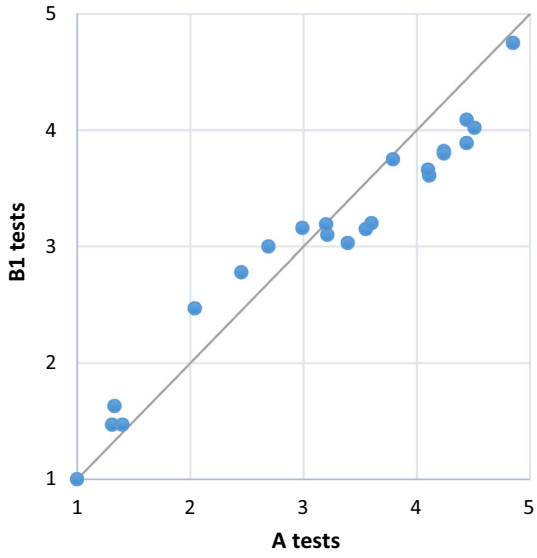


Fig. 2 Overall quality MOS (G-MOS) of A and B1 tests

judged during the second playout and, similarly to the first half of samples, the third playout is for judging the overall sample quality. The samples are played out in random order using different randomization for each listening panel.

The study (Pinson et al. 2012) shows that 24 is the minimal number of subjects when tests are performed in a controlled environment and 35 in the case of the public environment or narrow range of audiovisual quality. In our case (controlled environment), each condition has been presented to 32 listeners (4 panels with eight listeners per panel). All of them were naive native US English speakers. Also, the tests performed in Prague laboratory used native

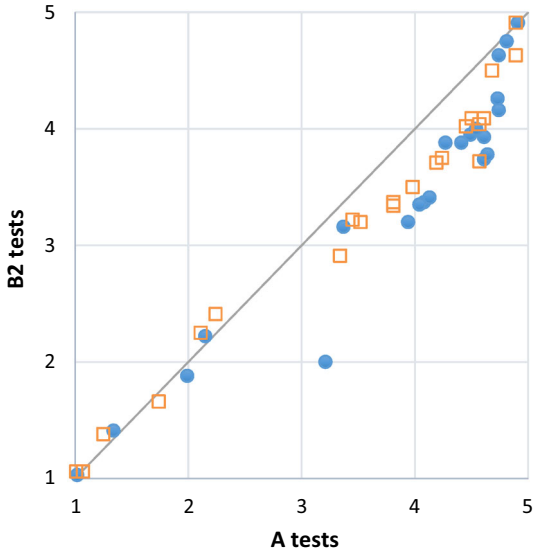


Fig. 3 Speech MOS and noise annoyance MOS of A and B2 tests. *Blue circles* show dependencies of S-MOS values. *Orange squares* show dependencies of N-MOS values. (Color figure online)

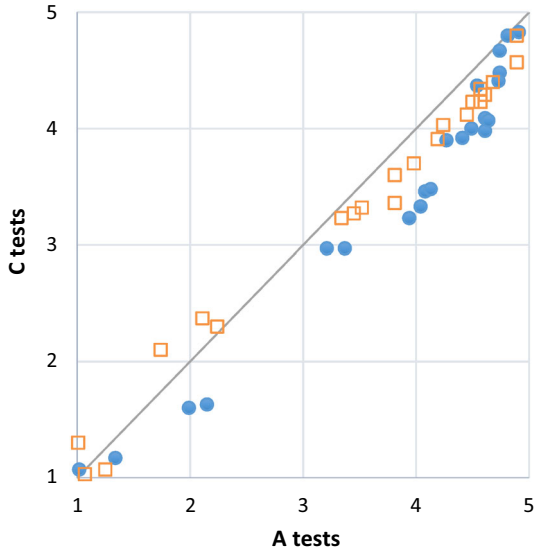


Fig. 5 Speech MOS and noise annoyance MOS of A and C tests. *Blue circles* show dependencies of S-MOS values. *Orange squares* show dependencies of N-MOS values. (Color figure online)

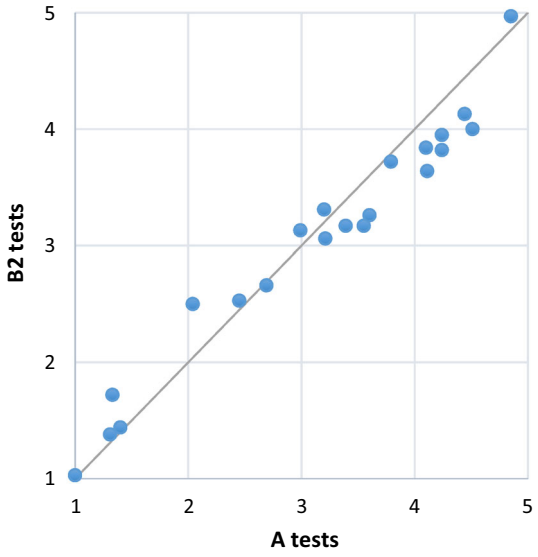


Fig. 4 Overall quality MOS (G-MOS) of A and B2 tests

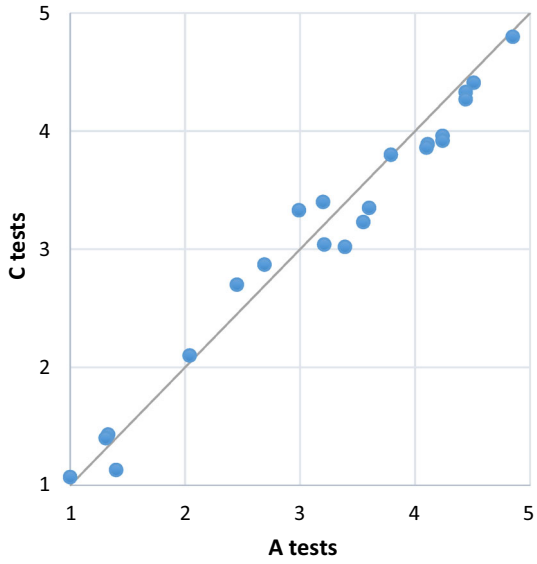


Fig. 6 Overall quality MOS (G-MOS) of A and C tests

American listeners recruited from expats and other US citizens living in Prague or temporarily visiting the city. Thorough recruiting procedure assures no subject lived outside the U.S.A. for more than 6 months prior the test.

The test participants were recruited by the listening labs using their common acquisition procedures, with the ratio between male and female listeners between 40:60% and 60:40%. The age distribution approximately followed human population age distribution in the range between 18 and 65 years of age.

For the sound reproduction, Sennheiser HD 280 PRO professional headphones have been used.

A professional voting device has been used to collect the votes. The tests were conducted in low-reverberation listening rooms conforming to requirements of P.800 in full (reverberation time below 500 ms, background noise below 30 dB SPL (A) without significant peaks in spectra).

3 Data analysis and results

By further data processing, the corresponding MOS are obtained separately for speech quality (S-MOS), noise annoyance (N-MOS) and overall sample quality (G-MOS).

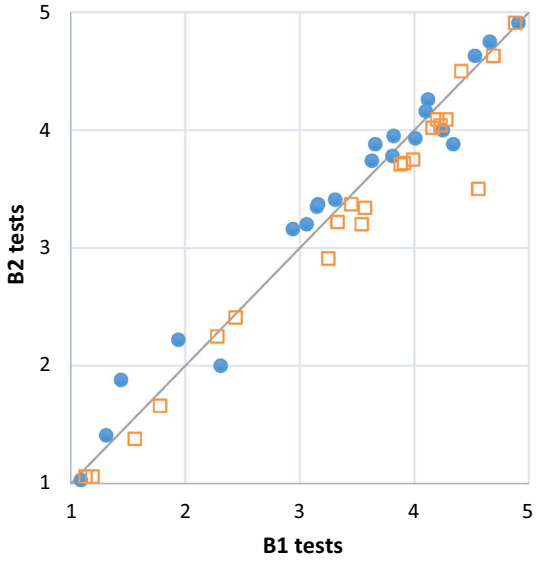


Fig. 7 Speech MOS and noise annoyance MOS of B1 and B2 tests. *Blue circles* show dependencies of S-MOS values. *Orange squares* show dependencies of N-MOS values. (Color figure online)

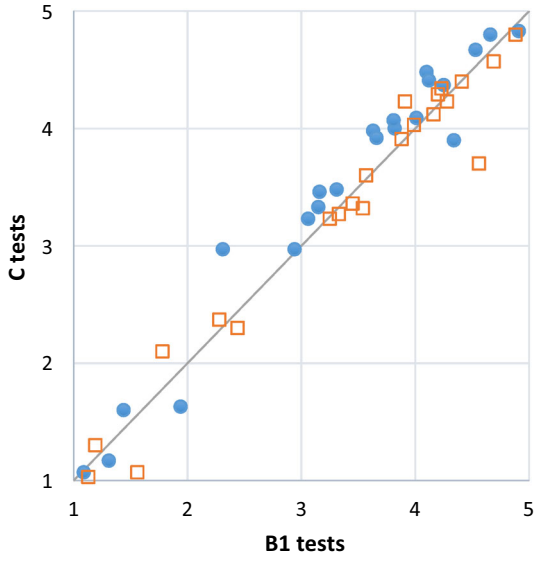


Fig. 9 Speech MOS and noise annoyance MOS of B1 and C tests. *Blue circles* show dependencies of S-MOS values. *Orange squares* show dependencies of N-MOS values. (Color figure online)

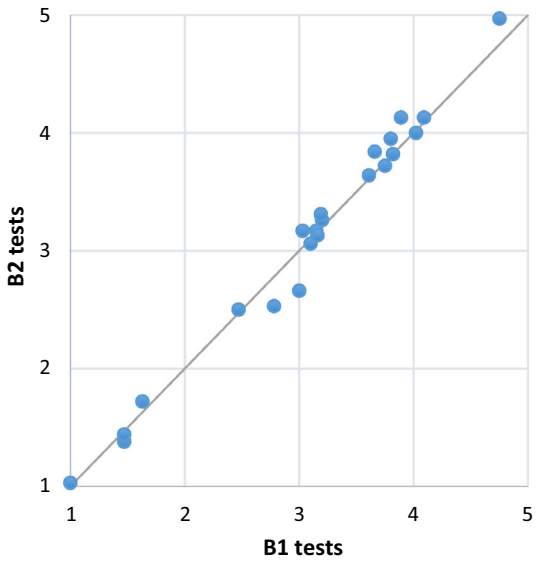


Fig. 8 Overall quality MOS (G-MOS) of B1 and B2 tests

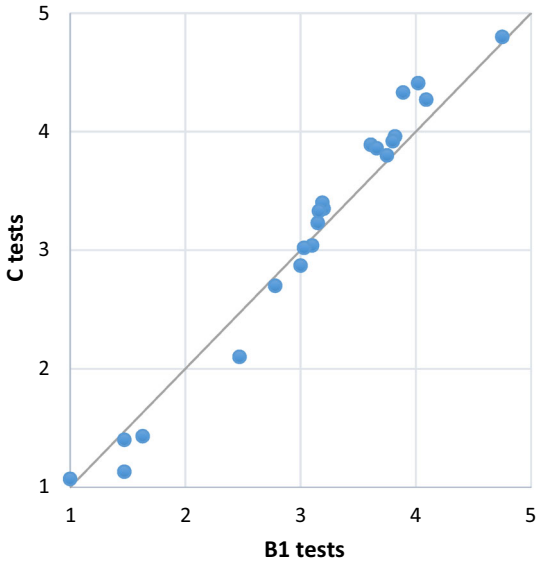


Fig. 10 Overall quality MOS (G-MOS) of B1 and C tests

The terms S-MOS, N-MOS, and G-MOS are adopted from ETSI TS 103 106 (2014) and ETSI EG 202 396-3 (2008). These terms replace in the further text the original SIG, BAK, and OVRL ratings used in P.835. Correlations between test MOS are provided in Tables 1, 2, and 3.

In all tests involving A laboratory, the A results in MOS categories above 3 seem to be systematically higher in S-, N- and G-MOS. It was already mentioned that data set A contained a larger set of conditions, for a total of 60, of which 22 were tested also in tests B and C. We assume that the fact the A tests were originally performed using a wider set of conditions creates the above-mentioned bias between A and others tests, setting the user expectations lower than in the other experiments.

Speech and noise MOS correlations between A and B1 tests are shown in Fig. 1. Their values are 0.953 and 0.982 respectively.

Figure 2 shows global MOS correlations between A and B1. Its value is 0.976.

In Fig. 3, speech and noise MOS correlations between A and B2 tests are shown. Their correlation values are 0.954 and 0.985 respectively.

Figure 4 shows global MOS correlations between A and B2. Its value is 0.9812.

Speech and noise MOS correlations between A and C tests are shown in Fig. 5. Their values are 0.98 and 0.992 respectively.

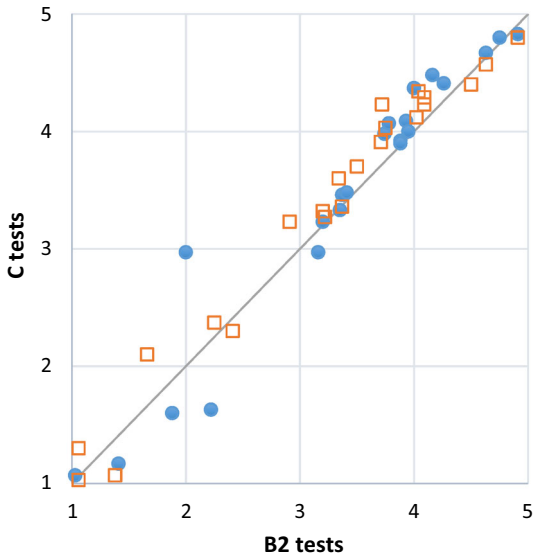


Fig. 11 Speech MOS and noise annoyance MOS of B2 and C tests. *Blue circles* show dependencies of S-MOS values. *Orange squares* show dependencies of N-MOS values. (Color figure online)

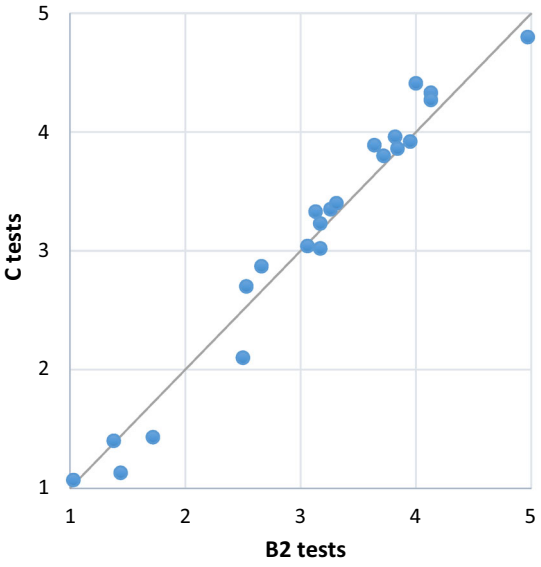


Fig. 12 Overall quality MOS (G-MOS) of B2 and C tests

Table 1 Correlations between S-MOS of tests

	A	B1	B2	C
A	1	0.9529	0.9535	0.9800
B1	0.9529	1	0.9833	0.9794
B2	0.9535	0.9833	1	0.9684
C	0.9800	0.9794	0.9684	1

Figure 6 shows global MOS correlations between A and C. Its value is 0.986.

Speech and noise MOS correlations between B1 and B2 tests are shown in Fig. 7. Their values are 0.983 and 0.981 respectively.

Figure 8 shows global MOS correlations between B1 and B2. Its value is 0.992.

Speech and noise MOS correlations between B1 and C tests are shown in Fig. 9. Their values are 0.979 and 0.977 respectively.

Figure 10 shows global MOS correlations between B1 and C. Its value is 0.989.

Speech and noise MOS correlations between B2 and C tests are shown in Fig. 11. Their values are 0.968 and 0.985 respectively.

Figure 12 shows global MOS correlations between B2 and C. Its value is 0.986.

As can be seen in graphs, subjects' votes were highly similar, and correlation values were extremely close to its maximum value.

All the correlations of speech qualities, noise annoyances, and overall qualities are introduced in tables below.

Table 1 represents correlations between speech qualities (S-MOS) of tests.

Table 2 Correlations between N-MOS of tests

	A	B1	B2	C
A	1	0.9816	0.9852	0.9922
B1	0.9816	1	0.9810	0.9768
B2	0.9852	0.9810	1	0.9853
C	0.9922	0.9768	0.9853	1

Table 3 Correlations between G-MOS of tests

	A	B1	B2	C
A	1	0.9762	0.9812	0.9857
B1	0.9762	1	0.9918	0.9893
B2	0.9812	0.9918	1	0.9861
C	0.9857	0.9893	0.9861	1

Table 4 Pairwise comparison between G-MOS for each pair of tests

	A	B1	B2	C
A	0	14 (6.1%)	19 (8.2%)	11 (4.8%)
B1	14 (6.1%)	0	9 (4.8%)	6 (2.6%)
B2	19 (8.2%)	9 (4.8%)	0	12 (5.2%)
C	11 (4.8%)	6 (2.6%)	12 (5.2%)	0

Table 5 Average CI 95 of the tests

	Average CI 95: S-MOS	Average CI 95: N-MOS	Average CI 95: G-MOS
A	0.13	0.13	0.14
B	0.13	0.11	0.11
C	0.16	0.14	0.13

Table 6 Pairwise comparison between G-MOS for each pair of tests with consideration of CI 95

	A	B	C
A	0	0 (0%)	0 (0%)
B	0 (0%)	0	0 (0%)
C	0 (0%)	0 (0%)	0

Table 2 represents correlations between noise annoyance scores (N-MOS) of tests.

Table 3 represents correlations between global qualities (G-MOS) of tests.

3.1 Pairwise comparisons of each test

After data correlations, pairwise comparisons for each couple of tests were evaluated. The principle of comparison was the following: First, each global MOS value of each test was compared with all remaining global MOS values of the same test. Afterwards, absolute differences (of every couple of tests) were calculated. Totally there were 231 cases (22 datasets).

Table 4 shows results of pairwise comparison between global qualities (G-MOS) for each pair of tests. First numbers show the quantity of identified differences. Numbers in brackets show the difference percentage.

3.2 Pairwise comparisons with consideration of subjective test confidence intervals

Due to the differences in pairwise comparisons between the tests, pairwise comparisons with consideration of confidence intervals (CI 95) have been done. In Table 5 average confidence intervals of the tests are presented. Laboratory B is represented only by its first test run (previously presented as B1).

During those comparisons, absolute differences of G-MOS of each pair of tests have been calculated.

RMSE* (ITU-T TD12rev1 2009) (root mean squared error with suppressed influence of subjective testing uncertainty) analysis shows the differences between the CI 95% interval borders of the pair (zero if the CI 95% intervals overlap each other).

The CI 95 tests show that all the existing differences in all pairwise comparisons have disappeared. The results are shown in Table 6.

4 Conclusion

The results of data analysis of four subjective tests made in three different laboratories show the level of inter-lab and intra-lab repeatability in case of identical test speech samples are used and confirm the subjective tests are highly repeatable in case P.800 and P.835 rules and requirements are strictly followed. The tests also show minor differences in results in case subject expectations are set differently using a wider set of test speech samples presented in one of the labs.

Acknowledgements Authors thank Andrew Catellier and Stephen Voran at the United States Department of Commerce's Institute for Telecommunication Sciences in Boulder Colorado for providing the test premises and test subjects and also for valuable discussions related to this project.

References

Appendix of IEEE Subcommittee on Subjective Measurements IEEE Recommended Practices for Speech Quality Measurements. (1969). IEEE Transactions on Audio and Electroacoustics. Vol 17, pp. 227–246.

European Telecommunications Standards Institute. (2008). Speech processing, transmission and quality aspects (STQ); speech quality performance in the presence of background noise part 3: Background noise transmission—Objective test methods. European Telecommunications Standards Institute, ETSI EG 202 396-3.

European Telecommunications Standards Institute. (2014). Speech and multimedia transmission quality (STQ); speech quality performance in the presence of background noise: Background noise transmission for mobile terminals—Objective test methods. European Telecommunications Standards Institute, ETSI TS 103 106.

Goodman, D. J., & Nash, R. D. (1982). Subjective quality of the same speech transmission conditions in seven different countries. *IEEE Transactions on Communications, 30*(4), 642–654.

3GPP TR 26.952. (2015). Codec for Enhanced Voice Services (EVS); Performance Characterization.

3GPP TS 26 071. Mandatory speech CODEC speech processing functions; AMR speech Codec; General description

ITU-T Rec. P.800. (1996). Methods for subjective determination of transmission quality, Series P: Telephone transmission quality, ITU, Geneva, am. 1998.

ITU-T Rec. P.835. (2003). Methods for objective and subjective assessment of quality, Series P: Telephone transmission quality, Telephone Installations, Local Line Networks, ITU, Geneva.

ITU-T Rec. P.863. (2014). Methods for objective and subjective assessment of speech quality, Series P: Terminals and subjective and objective assessment methods, ITU, Geneva.

ITU-T TD12rev1. (2009). Statistical evaluation. Procedure for P.OLQA v.1.0, SwissQual AG (Author: Jens Berger), ITUTSG12 Meeting, Geneva, Switzerland, March 10–19, 2009.

Pinson, M. H., Janowski, L., Pepion, R., Huynh-Thu, Q., Schmidmer, C., Corriveau, P., et al. (2012). The influence of subjects and environment on audiovisual subjective tests: An international study. *IEEE Journal of Selected Topics in Signal Processing, 6*(6), 640–651.

Chapter 4

Call Duration Dependency on the Quality of VoIP Calls Analysis

4.1 Summary

To verify the importance of providing subjective speech quality and speech intelligibility tests in communication systems, it was necessary to present that users tend to talk longer by their mobile devices if their call quality is higher. Call duration serves as an input parameter in many network- and service-models [35]. It is affected by various factors [36], particularly by the amount of information to be exchanged, social circumstances [37], gender of the call parties [38] or their nationalities [39] and is generally of great interest when examining large amounts of network data [40], [41]. For this purpose, over 16 million live call records over VoIP telecommunications networks deploying ITU-T G.711 [42], ITU-T G.729 [43], and AMR-NB 12.2k [44] codecs have been analyzed with an average call duration of 220s. Distributions of call duration dependency on Critical Minute Ratio and call duration dependency on the average MOS were considered. Further analysis showed that the results were not as expected, and the average call duration appeared a non-monotonic function of call quality. It turns out that alongside the assumption that better user experience brings longer call durations is valid only for moderate to high quality, and lower quality also yields longer call durations, which contradicts common expectations.

4.2 Publication

Since the results were surprising and meaningful to the scientist community, they were published in the article [45].

Analysis of the Dependency of Call Duration on the Quality of VoIP Calls

Jan Holub, Michael Wallbaum, Noah Smith, and Hakob Avetisyan

***Abstract*—This letter analyses call detail records of 16 million live calls over Internet-protocol-based telecommunications networks. The objective is to examine the dependency between average call duration and call quality as perceived by the user. Surprisingly, the analysis suggests that the connection between quality and duration is non-monotonic. This contradicts the common assumption, that higher call quality leads to longer calls. In light of this new finding, the use of average call duration as an indicator for (aggregated) user experience must be reconsidered. The results also impact modeling of user behavior. Based on the finding, such models must account for quality since user behavior is not fully inherent, but also depends on external factors like codec choice and network performance.**

***Index Terms*—ACD, call detail record, call duration, Internet protocol, VoIP, voice quality, speech codecs, telephony.**

I. Introduction

IT IS widely assumed that longer call durations indicate better call quality. Indeed, this dependency between call quality and average call duration (ACD) was reported for a mobile network in 2004 [1]. However, the study was conducted in times when most of the mobile calls were charged based on their duration so that users were motivated to keep calls as short as possible. Other conditions, such as the transition to IP-based packet switching and the introduction of new codecs, have also changed, which motivates a second look at the relation between call quality and duration.

Telephone calls carried over IP networks are affected by technical impairments, influencing the users' subjective perception of the call. Common technical impairments include coding distortion, packet loss, packet delay and its variations (jitter). The relation between the amount of each impairment and the final quality as perceived by a service user is not simple, as impairments can mask each other; or two impairments, each unnoticeable by itself, can multiply their effect and become subjectively annoying.

Monitoring systems analyzing live calls in telecommunications networks apply algorithmic models that attempt to estimate the subjective quality based on objective measurements of selected technical impairments. Commercial monitoring products for voice over IP (VoIP) services often use derivatives of the E-model defined in ITU-T G.107 [2] to estimate call quality. In some parts of the industry, e.g., in international wholesale business, the ACD is used as a cost-effective indicator of subjective call quality. The underlying assumption is that higher call duration means better user experience.

Call duration, meaning the time difference between call establishment and call termination, serves as input parameter in many network- and service-models [3]. It is influenced by a number of factors [4], particularly by the calling and called party situation, amount of information to be exchanged, social circumstances [5], gender of the call parties [6] or their nationalities [7] and is generally of great interest when examining large amounts of network data [8], [9].

II. Background

This letter is based on call detail records (CDR) produced by a commercial non-intrusive VoIP monitoring system measuring the quality of real calls in the network of a communication service provider. The system method analyses the Session Initiation Protocol (SIP) signaling messages as well as the flow of Real-time Transport Protocol (RTP) packets, their interarrival times and the information contained in the protocol headers. For every five-second segment of each RTP flow, the system generates a quality summary with several hundred metrics. Each summary contains basic information, such as the source/destination IP addresses, the used codec, as well as details about packet losses, interarrival times and the estimated quality. Estimates for the subjective quality are calculated using the E-Model with information about the packet loss, jitter, and the used codec as input. The E-Model yields an R-factor value for every time slice, which is mapped to an estimated MOS (Mean Opinion Score) value. MOS is the commonly used metric for subjective call quality. Finally, the system marks 'critical' five-second segments which suffer from burst loss or excessive jitter. Specifically, a five-second segment is marked as 'critical' if more than three packets are lost in sequence or if the packet interarrival time exceeds the packet rate by 40 ms or more.

The monitoring system's CDRs describe the characteristics of each call from the signaling and media quality perspective. They summarize the five-second data, e.g., by storing the minimum, average and maximum R-factor and MOS for each call direction. Other quality metrics provided by the CDRs are described in the next section.

III. Data Set Characteristics

The data was provided as a database of CDRs, with each database record corresponding to one call in the network. The following parameters were used for the analysis: call duration, used audio codec, critical minute ratio (CMR) per media direction and average R-factor/MOS per media direction. CMR is

Manuscript received January 3, 2018; revised February 6, 2018; accepted February 9, 2018. Date of publication February 15, 2018; date of current version August 21, 2018. The associate editor coordinating the review of this paper and approving it for publication was S. Zhou. *(Corresponding author: Jan Holub.)*

J. Holub and H. Avetisyan are with the Department of Measurement, FEE, Czech Technical University, CZ-166 27 Prague, Czech Republic (e-mail: holubjan@fel.cvut.cz; avetihak@fel.cvut.cz).

M. Wallbaum and N. Smith are with Voipfuture GmbH, 20097 Hamburg, Germany (e-mail: mwallbaum@voipfuture.com; nsmith@voipfuture.com).

Digital Object Identifier 10.1109/LWC.2018.2806442

TABLE I
DISTRIBUTION OF CODECS IN THE ANALYZED DATA SET

Codec	Number of CDRs	Percentage
G.711	10,998,417	68.1%
G.729	4,108,152	25.4%
AMR-NB 12.2k	1,044,612	6.5%

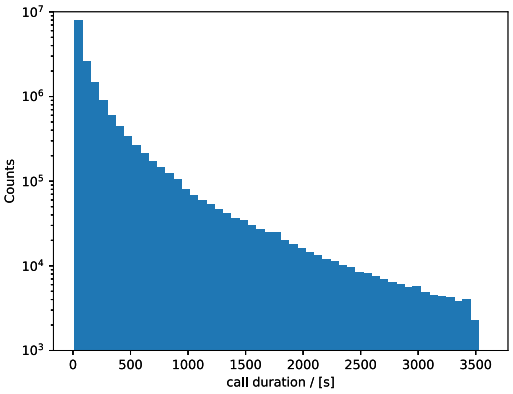


Fig. 1. Distribution of the call duration in the analyzed data set.

TABLE II
BEST QUALITY ACHIEVED AND ACD PER CODEC

Codec	Max. R-Factor	Max. MOS	ACD/[s]
G.711	93	4.41	262
AMR-NB 12.2k	86	4.23	210
G.729	82	4.10	181

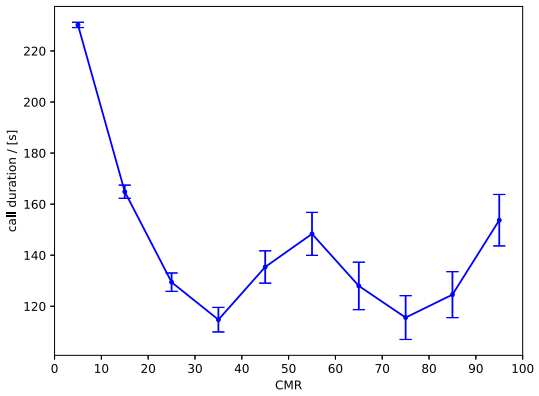


Fig. 2. Distribution of call durations depending on Critical Minute Ratio.

a proprietary metric provided by the monitoring system. It is calculated as the ratio of 'critical' five-second segments over all segments. For example, a CMR of 10% states that one out of ten time slices were affected by critical loss or jitter.

All quality data is available separately for media streams sent by the calling party (A-party) and the called party (B-party). However, the analysis only considers the worst direction per call, i.e., the direction with smallest R-factor and greatest CMR.

The raw data set contains nearly 30 million mobile, international and domestic calls. The calls used 22 different types of codecs including G.711, G.729, G.722, G.723.1 and various modes of AMR-NB and AMR-WB. The majority of these codecs were however used so rarely that the respective CDRs were excluded from further analysis. The following CDRs were not considered for the analysis:

- calls which did not use the top three codecs G.711, G.729 or AMR-NB 12.2k,
- calls which lasted less than 10 s or more than one hour,
- calls where at least one direction was impacted by duplicate packets.

At the end of filtering process, the data contained more than 16 million CDRs. The codec distribution of the data set is shown in Table I.

Figure 1 shows the distribution of call durations. It closely matches a log-normal distribution, that is often used to model call duration distributions. The ACD over the entire (filtered) data set is 220 s.

IV. DATA ANALYSIS

A. Dependency on Codec Quality

The three audio codecs in the analyzed data set are all narrowband codecs. G.711 [10] is a widely used codec based on pulse code modulation. The audio signal is sampled at 8 kHz using 8 bits per sample, which leads to a net bitrate of 64 kbit/s. The samples are companded by one of two logarithmic functions, called A-law and μ-law. G.711 μ-law is mainly used in North America and Japan; A-law is preferred in the rest of the world. No distinction is made in the following, as all characteristics relevant for this letter are identical.

The second codec G.729 [11] uses conjugate structure algebraic-code-excited linear prediction (CS-ACELP) to compress speech frames of 10 ms. Annexes to the basic G.729 define several variants and extensions, e.g., for discontinued transmission or alternative bit rates. The analyzed data set contains only the variant compliant to G.729 Annex A and B – other variants are currently not used by commercial VoIP services. The sampling frequency is 8 kHz with 16 bits per sample. An encoded frame consumes 10 bytes yielding a net bitrate of 8 kbit/s.

The AMR codec described in [12] is a narrowband audio compression scheme using an ACELP coding scheme. It offers multiple bit rates ranging from 4.75 to 12.2 kbit/s and is widely used in GSM and UMTS. The only rate that is included in the data set is AMR-NB 12.2k.

Table II shows how the ACD depends on the calls' main codec. To mask out the impact of network conditions on the ACD, the table only considers calls with CMR=0%, i.e., without any critical packet loss or jitter.

The data shows that ACD correlates with the best possible user experience that can be achieved by a codec. For example, calls using G.711 are more than 40% longer on average than calls using G.729. The higher a codec's maximum R-factor/MOS, the higher the ACD.

B. Dependency on Transport Quality

The codec employed by a call is one technical aspect that can be controlled by a communication service provider. The other technical parameter is the amount of packet loss and

TABLE III
DEFINITION OF CATEGORIES OF SPEECH TRANSMISSION QUALITY

R-factor	MOS	Speech quality	User satisfaction
≥ 90	≥ 4.34	Best	Very satisfied
≥ 80	≥ 4.03	High	Satisfied
≥ 70	≥ 3.60	Medium	Some users dissatisfied
≥ 60	≥ 3.10	Low	Many users dissatisfied
≥ 50	≥ 2.58	Poor	Nearly all users dissatisfied

jitter, i.e., the network's transport performance. Figure 2 shows the impact of the CMR on the ACD. The data points represent the window centers and the window radius is five; error bars correspond to one standard deviation.

The ACD drops sharply from 230 s to 114 s for CMR=35%. This drop could be expected as more time slices are impacted by packet loss and jitter, which has a negative impact on the user experience. The severity of the drop is however surprising.

It should be noted that the CMR measures the distribution of severe impairments over the duration of a call. This is not equivalent to measuring the overall amount or intensity of packet loss and jitter. For example, RTP streams which lose three packets in sequence every five seconds, have a CMR of 100%. At a packet rate of 20 ms this corresponds to a packet loss ratio of about 1%, which is not much by conventional wisdom. In contrast, a call where half of one stream's packets are lost in sequence would yield a CMR of only 50%.

The sharp drop implies that even low impairment levels have a significant impact on the ACD. Beyond CMR=35% the ACD shows unexpected behavior as it rises again to 148 s for CMR=55%, drops down to 115 s for CMR=75% and finally rises to 153 s. Since voice quality depends on the level of packet loss and jitter, this behavior apparently contradicts the findings in [1] and common industry assumptions, namely that call duration is a monotonic function of the user experience. Yet, the CMR is a metric describing the technical quality of RTP streams, not the actual user experience. The next section looks into the connection between user experience and ACD.

C. Dependency on Estimated User Experience

As discussed before, the user experience is estimated using the E-model [2], which yields an R-factor value. This value is in the range of 0 to 100, where 0 represents extremely bad quality and 100 very high quality. Table III, based on G.107 [2], relates the E-model ratings R and their corresponding MOS to categories of speech quality and user satisfaction. Note that this mapping is only valid on the narrowband scale, i.e., when only narrowband codecs such as G.711, G.729 and AMR-NB are considered.

Figure 3 shows the ACD as a function of the average MOS of a call's worst stream. The average is calculated from the individual MOS values of the worst stream's five-second time slices. The data points represent the window centers with a window radius of 0.5.

The ACD drops from 222 s for MOS=4.25 to an absolute minimum of 133 s only to rise again to 190 s for MOS=1.25. It must be underlined that the absolute low of the ACD at MOS=2.25 is located just below the area of the MOS scale

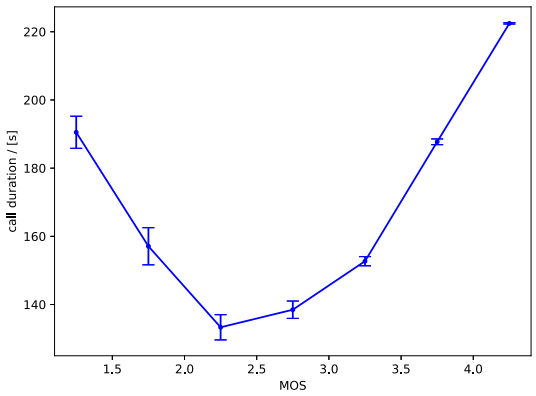


Fig. 3. Distribution of call durations depending on the average MOS.

where - according to Table III - 'nearly all users [are] dissatisfied'. For MOS < 2.58 one can assume that practically all call parties are dissatisfied. A rising ACD despite increasingly dissatisfied users is an unexpected result.

V. DISCUSSION

The previous section showed that the average duration of millions of VoIP calls generally depends on call quality. This is true for different definitions of quality, i.e., when defining quality via the main codec, the networks' transport performance (CMR) and the user experience (MOS).

The common assumption, substantiated in [1], is that average call duration is a monotonic function of quality, i.e., better user experience leads to longer durations. For moderate to high call quality this assumption is confirmed by the underlying data. Yet, the results for low call quality are surprising, since call durations also increase as quality gets worse. This suggests that the simplistic presumption about the link between quality and ACD may no longer be true. At least for contemporary VoIP-based services, the ACD appears to be a non-monotonic function of call quality. For example, given Figure 3 one cannot decide whether an ACD of 190 s indicates horrible or good user experience.

This unexpected finding has practical impact on the telecommunications industry. Specifically, mobile and international wholesale service providers may currently be working under wrong assumptions. Both types of service providers frequently have to deal with low-quality calls, e.g., because of poor air interface conditions or because of problematic routes to developing countries. The non-monotonic dependency of call duration and quality renders the ACD useless as a measure of user experience; service operators that solely rely on ACD as quality metric are likely to make wrong decisions. For example, actual mobile network issues may go undetected, when the ACD is relatively high or – even worse – international traffic is switched to a route with seemingly better quality. Interestingly, in wholesale business knowledge about the non-monotonic behavior could even be exploited for financial benefit. As the ACD is considered the main quality metric for routes (with impact on price), it may pay off to deliberately degrade moderate-quality routes to increase the ACD.

Consequently, other metrics, such as CMR or the average MOS, are needed to complement ACD.

Another area that is impacted by the findings of this letter is modeling of user behavior in terms of the call duration distribution, e.g., as described in [13]. Obviously, such models must also account for quality since the behavior of a user or user group is not entirely inherent to the user (group), but also depends on external factors like codec choice and network transport quality. New user models must consider the unexpected behavior of the ACD.

There are multiple potential reasons for the observed ACD increase for heavily compromised quality. One of them is the obvious need of word and sentence repetition when communicating over bad channels. Another one might be the Lombard effect [14] – an involuntary tendency to speak louder and slower under uncomfortable listening situations. Also, high packet loss and jitter may lead to stronger channel coding schemes and deeper de-jitter buffer adoption which adds delay to the communication path. If the call parties avoid double-talk situations, such delay is directly added to the overall call duration, multiplied by twice the number of role swaps (talker/listener) during the call [15]. Conversations in English language exhibit 103 swaps on average during a three minute call [16]. Assuming the additional swap delay increases by 60 ms, then a call, which would last three minutes under perfect conditions, is prolonged by 12.4 s through this effect only. In practice, the need for repetitions under adverse conditions will increase the number of swaps, which further adds to the call duration.

VI. Conclusion

This letter explored the dependency between the average duration and quality of VoIP calls. More than 16 million live calls were analyzed by a commercial non-intrusive monitoring system. The main contribution of this letter is that it reveals a surprising non-monotonic dependency between call quality and average call duration. The common assumption, that better user experience leads to increasing call duration, is only valid for moderate to high call quality. Under this condition the findings of an earlier study [1] could be confirmed, although charging models, traffic types, quality measurement methods and geographical regions differ. However, under conditions of low quality the dependency changes, i.e., lower quality yields longer call durations, which contradicts common expectations. The inflection point roughly corresponds to quality that dissatisfies virtually all users.

The following conclusions can be drawn:

- The ACD is not an indicator for user experience.
- Codec choice and network transport performance influence the call duration.

Further studies with traffic using a wider variety of codecs need to be performed, to confirm that the codec quality influences the average call duration even for unconventional codecs. Specifically, traffic using wideband codecs needs to be examined, to determine if there is a quality saturation beyond which call duration is not impacted. Furthermore, the dominating factors of IP transport quality on call duration need to be analyzed, i.e., which impairment patterns lead to changes in the ACD. Finally, empirical studies need to investigate the reasons for increasing call duration under adverse conditions.

Acknowledgment

The authors would like to thank the operator for providing data for this project. The authors also would like to thank their colleagues Jan Bastian, Fabio Isabettini and Lucas Coutinho for feedback and support.

References

[1] J. Holub, J. Beerends, and R. Smid, "A dependence between average call duration and voice transmission quality: Measurement and applications," in *Proc. Wireless Telecommun. Symp.*, 2004, pp. 75–81.

[2] *The E-Model: A Computational Model for Use in Transmission Planning*, Int. Telecommun. Union, Geneva, Switzerland, ITU Recommendation G.107 (06/15), Jun. 2015.

[3] Z. Yang and Z. Niu, "Load balancing by dynamic base station relay station associations in cellular networks," *IEEE Wireless Commun. Lett.*, vol. 2, no. 2, pp. 155–158, Apr. 2013.

[4] V. D. Blondel *et al.*, "A survey of results on mobile phone datasets analysis," *EPJ Data Sci.*, vol. 4, no. 1, p. 10, Dec. 2015.

[5] Y. Dong, J. Tang, T. Lou, B. Wu, and N. V. Chawla, *How Long Will She Call Me? Distribution, Social Theory and Duration Prediction*. Berlin, Germany: Springer-Verlag, 2013, pp. 16–31.

[6] G. Friebel and P. Seabright, "Do women have longer conversations? Telephone evidence of gendered communication strategies," *J. Econ. Psychol.*, vol. 32, no. 3, pp. 348–356, Jun. 2011.

[7] D. Goodman and R. Nash, "Subjective quality of the same speech transmission conditions in seven different countries," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 7. Paris, France, 1982, pp. 984–987.

[8] J. Kim *et al.*, "Modeling cellular network traffic with mobile call graph constraints," in *Proc. IEEE Win. Simulat. Conf. (WSC)*, Phoenix, AZ, USA, Dec. 2011, pp. 3165–3177.

[9] M. Seshadri *et al.*, "Mobile call graphs: Beyond power-law and lognormal distributions," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min. (KDD)*, Las Vegas, NV, USA, 2008, pp. 596–604.

[10] *Pulse Code Modulation (PCM) of Voice Frequencies*, Int. Telecommun. Union, Geneva, Switzerland, ITU Recommendation G.711 (11/88), Nov. 1988.

[11] *Coding of Speech at 8 kbit/s Using Conjugate Structure Algebraic-Code-Excited Linear Prediction (CS-ACELP)*, Int. Telecommun. Union, Geneva, Switzerland, ITU Recommendation G.729 (06/12), Jun. 2012.

[12] "Mandatory speech CODEC speech processing functions; AMR speech codec; general description," 3GPP, Sophia Antipolis, France, Rep. 26.071, 1999.

[13] P. O. S. Vaz de Melo, L. Akoglu, C. Faloutsos, and A. A. F. Loureiro, *Surprising Patterns for the Call Duration Distribution of Mobile Phone Users*. Berlin, Germany: Springer-Verlag, 2010, pp. 354–369.

[14] S. A. Zollinger and H. Brumm, "The Lombard effect," *Current Biol.*, vol. 21, no. 16, pp. R614–R615, 2011.

[15] J. Holub and O. Tomiska, "Delay effect on conversational quality in telecommunication networks: Do we mind?" in *Wireless Technology: Applications, Management, and Security*, S. Powell and J. P. Shim, Eds. Boston, MA, USA: Springer, 2009, pp. 91–98. [Online]. Available: https://doi.org/10.1007/978-0-387-71787-6_6, doi: 10.1007/978-0-387-71787-6_6.

[16] "Speech and multimedia transmission quality (STQ); adaptation of the ETSI QoS model to better consider results from field testing LQO and delay," ETSI, Sophia Antipolis, France, Rep. ETSI TR 103 121, Rev. 1.1.1, Mar. 2013.

Chapter 5

Parallel Task in Subjective Speech Intelligibility Testing

5.1 Summary

In this step, the primary attention was paid to subjective speech intelligibility tests. These tests are used to evaluate the intelligibility of recorded speech samples, usually transmitted over communication technologies to be assessed or compared. This article demonstrates that the assumption claiming that the speech intelligibility in laboratory testing conditions is higher than in parallel task conditions is not always true. Achieving in some aspect better results using a parallel task has been reported in [46],[47]. Since subjective tests provided in laboratory conditions don't reflect the real-life environmental conditions, a newly proposed hybrid parallel-task methodology was Implemented in the tests deploying ITU-T P.807 for comparison. The proposed parallel-task deployed a laser-shooting simulator and was consisted of two types of subjects: "shooter" and "counters." The roles were dynamically randomly assigned every 40s by automated light bulbs. Totally, 51 subjects participated in the tests assessing 48 samples modified with various background noises and coders. There were certain samples where intelligibility values were counterintuitive, which proves that tests provided in laboratory conditions cannot be considered as an etalon of speech intelligibility testing, and parallel-task techniques are highly recommended for subjective speech intelligibility evaluation. The results should be considered as a subject for further study in terms of counterintuitive results.

5.2 Publication

The full study with all details and conditions are mentioned in the article [48].

ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 104 (2018) 678 – 684
DOI 10.3813/AAA.919207

Low Bit-rate Coded Speech Intelligibility Tested with Parallel Task

Hakob Avetisyan, Tomáš Drábek, Jan Holub
Department of Measurement 13138, Faculty of Electrical Engineering, Czech Technical University in Prague, Technická 2, CZ 16627 Prague 6, Czech Republic. holubjan@fel.cvut.cz

Summary
A speech intelligibility test in realistic environments has been designed and performed deploying 51 subjects. A single set of speech samples distorted by various background noises and low bit-rate coding techniques has been used - without and with additional (parallel) psychomotor tasks. The addition of psychomotor tasks to the test simulates realistic environments. The differences in intelligibility test results between standard laboratory and parallel task test methodology were identified. Overall, the intelligibility of regular tests was mainly higher than for tests with a parallel task. Intelligibility results of certain samples have been found counterintuitive, which is possibly explained by a different mode of thinking under parallel task conditions.

PACS no. 43.71.Gv

1. Introduction

In speech communication, intelligibility is being used to measure the clearness of speech in different conditions. Intelligibility directly influences the amount of information transferred by the communication act among the communicating parties, having either human or machine at either side of the communication chain. Methods of subjective and objective intelligibility assessment are of great interest in the speech expert community [1, 2]. This article deals with subjective assessment methods used to evaluate the intelligibility of recorded speech samples, usually transmitted over communication technologies to be assessed or compared. As in any subjective tests, an exact reproduction of the test results is not guaranteed in general, and inter-test comparisons require the inclusion of several reference conditions at least. The results depend on the individual subject responses. Application of appropriate statistical methods [3] allows then for statistically significant assessment of, e.g., compared technologies or reliability statement when a minimum intelligibility threshold is defined for the tests. The main question, however, remains in existing fundamental philosophy of all currently used test methods: Test subjects are comfortably seated in a (usually anechoic or semi- anechoic) test room and are fully focused on listening to the tested samples. In real communication scenarios, the users are usually performing multiple tasks in parallel (driving the car while talking by phone, observing a monitor where airplane location and approach situation is displayed while actively using radio-communication link with the airplane pilot, etc.).

Most of the existing test methods and related standardized recommendations are based on the simple assumption that the case of laboratory testing with subjects fully concentrated to the intelligibility test provides the most sensitive results; meaning that the test samples will achieve the highest intelligibility scores compared to any real scenario when the users are distracted by performing other tasks in parallel. This article demonstrates this simple assumption is not always true, formulates explanatory hypothesis and sketches possible utilization of this fact. Multiple experiments including parallel task (sometimes referred as "dual task") have been performed on impaired subjects [4] or children [5] which do not aim to represent the regular working population. The regular population is usually tested for a relationship between listening effort and parallel task introduction [6, 7, 8, 9].

Achieving in some aspect better results under the parallel task condition has been reported in some other domains in several psychological studies [10, 11, 12]. However, the effect has not yet been demonstrated for speech intelligibility testing. The remain of this article is structured as follows: Section 2 includes main information about intelligibility testing and various types of tests.

Section 3 deals with the problem definition. Section 4 contains all the necessary information about experimental design, parameters and the standards used. In Section 5, data analysis and full obtained results are presented. Also, the hypothesis possibly explaining the results is formulated there.

In Section 6, the conclusion of the article and information about future work is presented.

Received 10 November 2017,
accepted 4 June 2018.

2. Intelligibility testing

Intelligibility is affected by the level and quality of speech signal, the type, and level of background noise, etc. Intelligibility tests [13, 14] have been designed to evaluate the ability of a human or automated listener to understand a meaning of spoken words. [14] defines speech intelligibility as: "a measure of the effectiveness of understanding speech." Two principally different assessment methods may be used: Subjective assessment, based on the use of (human or artificial) speakers and (human) listeners, or Objective assessment based on physical parameters of the transmission channel. Multiple methods and their modifications were developed [15] and are discussed next.

2.1. DRT – Diagnostic Rhyme Test

DRT measures the intelligibility of speech over communication systems [13]. The test materials contain 96 rhyming monosyllable word pairs (e.g., veal-feel) that were selected to differ in the initial consonant. During the test, the listener is asked which of the two rhyming words presented was spoken.

2.2. DMCT – Diagnostic Medial

Consonant Test

The DMCT is a variation of the DRT with test materials consisting of 96 bi-syllable word pairs (e.g., stopper-stocker) selected to differ in only their intervocalic consonant.

2.3. DALT – Diagnostic Alliteration

Test

The DALT is another variation of the DRT. 96 monosyllable word pairs (e.g., pack-pat) selected to differ in their final consonant only.

2.4. MRT – Modified Rhyme Test

The MRT was standardized by ANSI to measure the intelligibility of communication systems. Its test material consists of 50 rhyming monosyllable word sets of 6 words (e.g., pin, sin, tin, fin, din, win) selected with half to differ in the initial consonant and the other half in the final consonant. As in DRT case, the listener is asked which of the six presented rhyming words was spoken.

3. Problem Definition

The common feature of the methods mentioned above is the fact the tests are performed in precisely defined and carefully maintained laboratory environments where the acoustic background, room reverberation, playout equipment and headphones, listener age and gender structure and similar parameters are usually strictly defined or monitored and reported. The test listeners are usually seated comfortably and are able to fully attend to the test procedure and listening material.

As such a rigid laboratory environment is far from realistic situations where, telecommunication equipment is usually used, the question of the practical value of the regular intelligibility tests as described in Section 2 arises. To bring the test procedure (and results) closer to the real usage situation, a parallel task was introduced to distract test subjects from full concentration on the intelligibility testing itself.

4. Experiment Description

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, designed to occupy both mental and physical resources of the test subjects. Both tests were performed in an acoustically treated critical listening environment conforming to ITU-T P.800 [16]. Its reverberation time is 185 ms, and background noise is less than 30 dB SPL(A) without peaks in frequency spectra. Closed circumaural headphones (Sennheiser HD280Pro) with no additional frequency compensation and a professional distribution amplifier were used for sample playout. For collecting the subjects' votes, a professional voting device was used.

The intelligibility tests were performed following the MRT (Modified Rhyme Test) methodology as described in ITU-T P.807 [13]. In total, 48 samples were selected from MRT sample list, and for practical reasons (voting device limitation), only five answer options were used instead of original six. For the initial preliminary experiment, described in this text, the samples were recorded using voices of two male narrators: Narrator 1 (24 samples) and Narrator 2 (24 samples). 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) [17] at 64kbit/s (16 samples) and a low bit-rate coder MELPe [18] (Mixed-Excitation with Linear Predictive enhanced) operating at 2.4kbit/s (32 samples). 16 MELPe samples (out of those 32) were previously mixed with the background noise of the interior of a High Mobility Multipurpose Wheeled Vehicle (HMMWV) at Signal-to-Noise ratio SNR = 0 dB. All coded samples were originally recorded by both narrators. Then, the randomization of the sample playout order has been performed and samples were picked up in a way to avoid two consequent samples being spoken by the same narrator (N1-N2-N1-N2-...). Thus, the numbers of samples in each codec category were not balanced between both narrators. Diotic presentation level of 73 dB SPL (A) corresponded to −26 dBoV in the digital recording.

51 subjects (26 female and 25 male) in the age range of 18–56 were hired for the tests. Most of them (68%) were native English speakers; the others were selected from fluent and highly proficient non-native English speakers. This combination of native and non-native but highly proficient subjects corresponds to the estimated language structure

of U.S. professional organization chosen as an example [19]. The experiment consisted of two parts. First, regular intelligibility tests were performed. One session lasted approximately 12 minutes. This test was followed by the 90-minute break when a different subjective visual low-complexity task was assigned to the subjects – watching a set of short movies. This part was used to make subjects relaxed and give full focus to a different task.

Then the second part - repeating the intelligibility test but with a parallel task this time - was performed. 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, as opposed to purely physical (e.g., stationary bicycle) or purely mental (e.g., mathematical quizzes) tasks described in [20]. This kind of task was chosen to minimize sensitivity to inequality of physical or mental skills of subjects. During the voting process, the subjects were required to vote within a five second "voting break" following each sample playout. Before each voting break, a short beep (440 Hz, 60 dB SPL(A), 300 ms) sounded to indicate the start of the voting break. Any answer provided outside of the voting break was not counted.

5. Data analysis and results

In this section, results of intelligibility tests both for the laboratory and the parallel task are presented. As the test arrangement allowed for missing votes (no option selected for the given word playout), missing votes are not considered at all so total number of votes is not constant (up to 5% of votes for the laboratory experiment and up to 9% of votes for the parallel task experiment were not acquired).

In Figure 1, intelligibility without and with parallel task scores per sample, without consideration of missing votes is shown for samples coded by PCM coder. As expected, in most cases, intelligibility without parallel task is higher than intelligibility with parallel task, likely because subjects are able to focus completely on the intelligibly task without the distraction of a second task .

Figure 2 depicts intelligibility of all PCM condition (across all 16 samples shown in Fig.1). In both figures, the x-axis represents each test sample, and the y-axis represents the intelligibility.

Similarly, Figure 3 and 4 report MELPe (without any background noise) intelligibility results (Figure 3 shows results per sample while Figure 4 shows overall intelligibility of noise-free MELPe condition). Finally, Figure 5 reports per-sample results for MELPe samples affected by

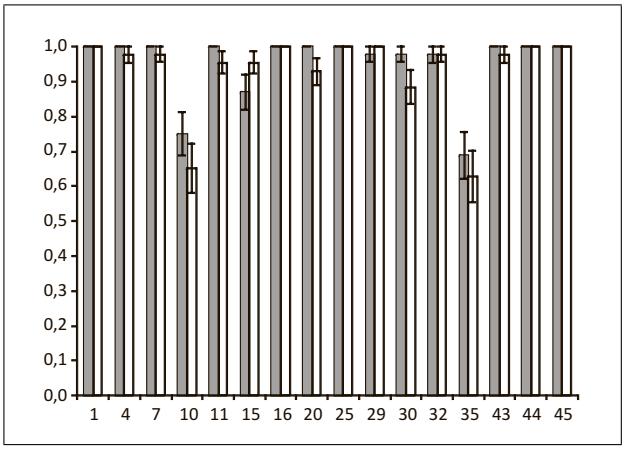


Figure 1. Intelligibility scores of PCM samples without (left bars) and with (right bars) a parallel task, missing votes not considered. Horizontal scale: Sample Number. Vertical Scale: Intelligibility, uncompensated for random correct votes. Minimum 43, median 48 votes per sample.

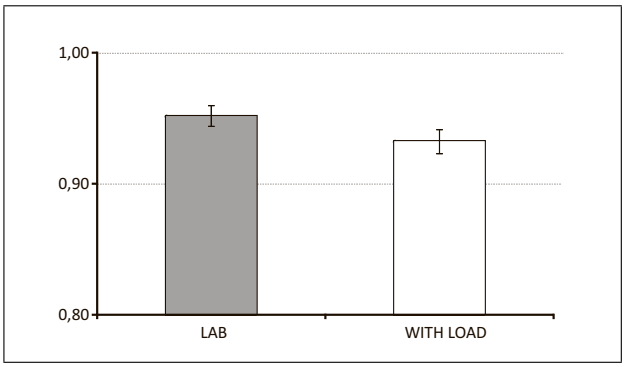


Figure 2. Intelligibility scores of PCM condition without (left bar) and with (right bar) a parallel task, missing votes not considered. Vertical Scale: Intelligibility, uncompensated for random correct votes. Minimum 750 votes per condition.

Table I. Statistical results of intelligibility using various combinations codecs and noise.

	PCM	MELPe	MELPe/HMMWV	Overall
LAB INTEL	95,2%	92,5%	59,0%	82,2%
PAR.T. INTEL	93,2%	88.9%	58,9%	80,5%

background noise at 0 dB SNR and Figure 6 depicts overall intelligibility of all those noisy MELPe samples together.

Overall results are mostly as expected: with the parallel task, the intelligibility decreases with parallel task introduction.

Table I shows the statistical results of intelligibility for PCM, MELPe codec, MELPe codec with HMMWV noise and the overall intelligibility both for the laboratory and the parallel task environment. However, as can be seen in Figure 1, Figure 3 and Figure 5, for 9 of the 48 samples (8, 14, 15, 17, 22, 27, 34, 36, and 42) the results are counter-intuitive - meaning their intelligibility increases when the

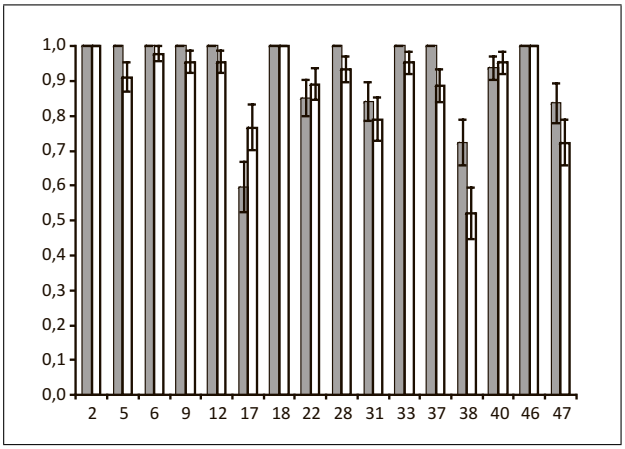


Figure 3. Intelligibility scores of MELPe samples without (left bars) and with (right bars) a parallel task, missing votes not considered. Horizontal scale: Sample Number. Vertical Scale: Intelligibility, uncompensated for random correct votes. Minimum 43, median 48 votes per sample.

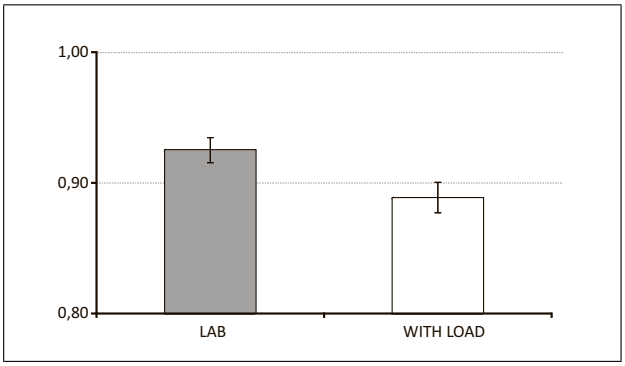


Figure 4. Intelligibility scores of MELPe condition without (left bar) and with (right bar) a parallel task, missing votes not considered. Vertical Scale: Intelligibility, uncompensated for random correct votes. Minimum 750 votes per condition.

Table II. Samples with counterintuitive results and their parameters.

Sample	SNR (dB)	Coder	Narrator
8	0	MELPe	2
14	0	MELPe	2
15	>50	PCM	1
17	>50	MELPe	1
22	>50	MELPe	2
27	0	MELPe	1
34	0	MELPe	2
36	0	MELPe	2
42	0	MELPe	2

parallel task is introduced. The parameters of these samples are presented in Table II.

The percentage of correct answers without and with consideration of missing votes both for intelligibility without and with the parallel task are presented in Table III.

As two male narrators were used to record the clean (studio-quality) speech samples, per narrator analysis was

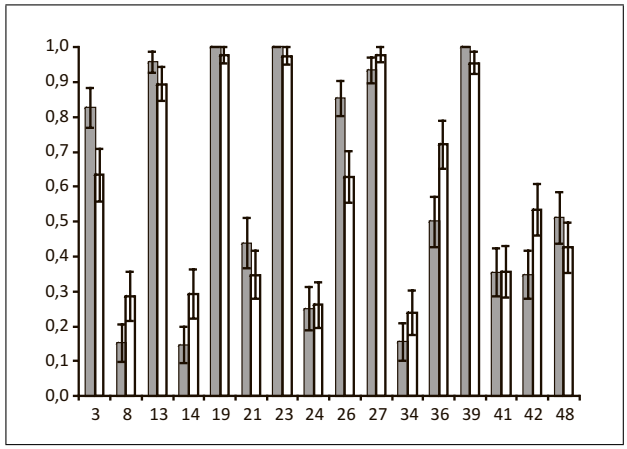


Figure 5. Intelligibility scores of MELPe with 0dB HMMWV noise samples without (left bars) and with (right bars) a parallel task, missing votes not considered. Horizontal scale: Sample Number. Vertical Scale: Intelligibility, uncompensated for random correct votes. Minimum 43, median 48 votes per sample.

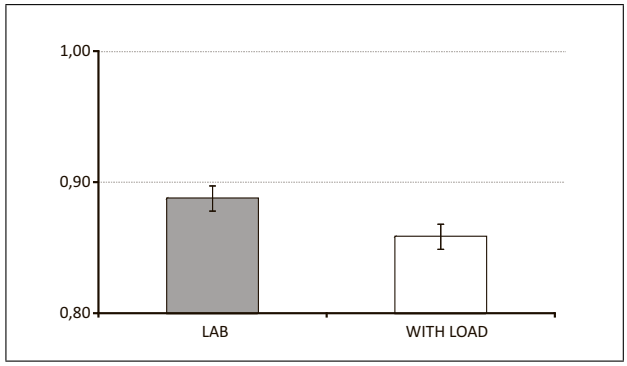


Figure 6. Intelligibility scores of MELPe with 0dB HMMWV noise without (left bar) and with (right bar) a parallel task, missing votes not considered. Vertical Scale: Intelligibility, uncompensated for random correct votes. Minimum 750 votes per condition.

Table III. The percentage of correct answers of counterintuitive samples without consideration of missing votes both for intelligibility without and with parallel task.

Sample	Lab w/out missing votes (%)	Par.task w/out missing votes (%)
8	14	27
14	16	28
15	88	96
17	60	78
22	84	90
27	94	98
34	15	22
36	51	70
42	37	56

also performed. Figure 7 shows intelligibility results of samples recorded by Narrator 1 and Figure 8 indicates the overall intelligibility of all samples recorded by Narrator 1. The same for Narrator 2 is shown in Figures 9 and 10. It

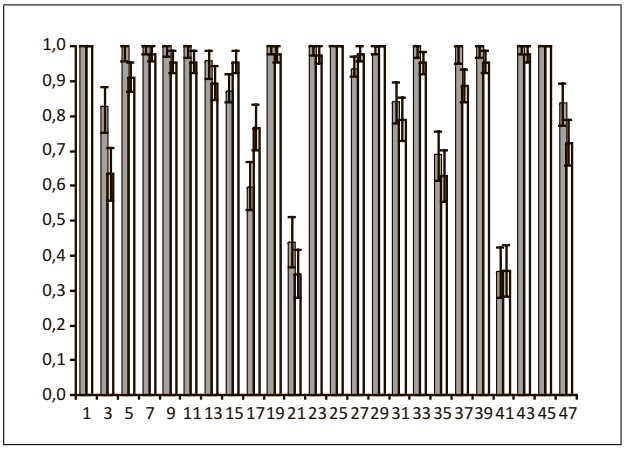


Figure 7. Intelligibility scores of samples recorded by Narrator 1 and processed with codecs and noise without (left bars) and with (right bars) a parallel task, missing votes not considered. Horizontal scale: Sample Number. Vertical Scale: Intelligibility, uncompensated for random correct votes. Minimum 43, median 48 votes per sample.

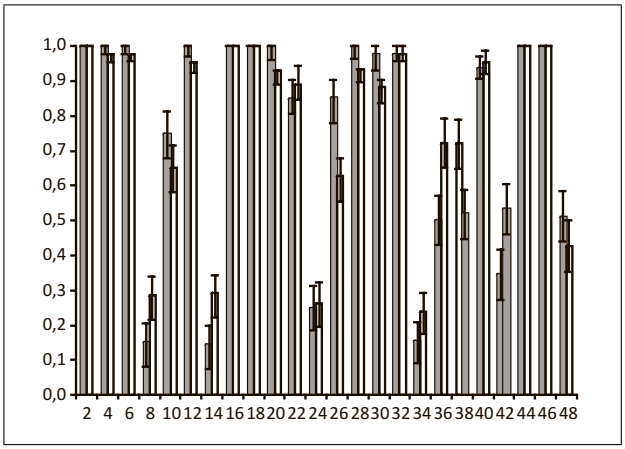


Figure 9. Intelligibility scores of samples recorded by Narrator 2 and processed with codecs and noise without (left bars) and with (right bars) a parallel task, missing votes not considered. Horizontal scale: Sample Number. Vertical Scale: Intelligibility, uncompensated for random correct votes. Minimum 43, median 48 votes per sample.

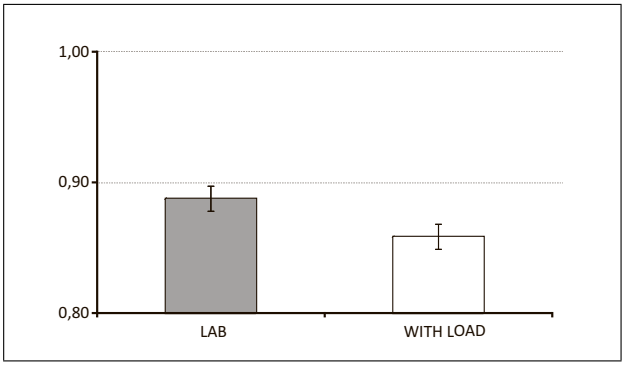


Figure 8. Intelligibility scores of samples recorded by Narrator 1 and processed with codecs and noise without (left bar) and with (right bar) a parallel task, missing votes not considered. Vertical Scale: Intelligibility, uncompensated for random correct votes. Minimum 1100 votes per condition.

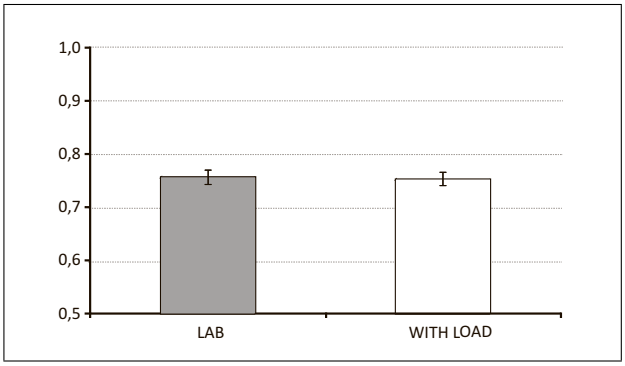


Figure 10. Intelligibility scores of samples recorded by Narrator 2 and processed with codecs and noise without (left bar) and with (right bar) a parallel task, missing votes not considered. Vertical Scale: Intelligibility, uncompensated for random correct votes. Minimum 1100 votes per condition.

is interesting to see that while the intelligibility of Narrator 1 samples drops with the introduction of the parallel task, samples recorded by Narrator 2 provide quite similar intelligibility levels both for laboratory and parallel task test versions. In all Figures 1–10, error bars show Standard Deviation of Arithmetical Mean (=Intelligibility STD), corresponding approximately to one half of 95% Confidence Interval. In Figure 1,3,5,7 and 9, minimum 43 and maximum 51 votes (median 48) are used to calculate one intelligibility score. In Figure 2,4,6 at least 750 votes are used to calculate one intelligibility score. In Figure 8 and 10, at least 1100 votes are used to calculate one intelligibility score.

It should be noted that 67% of the counterintuitive samples are affected by heavy background noise which brings their intelligibility close to random threshold (20% in this case of five possible answer options). Except for one sample (sample No.15), all counterintuitive samples are coded by low bit-rate MELPe coder.

As already mentioned, other experiments in different domains describe achieving better results with a parallel task [10, 11, 12], [21]. While [10, 11, 12] report experimental results without attempting to explain them, [21] suggests an interesting concept, giving a possible explanation for counterintuitive results shown above.

Kahnemann in [21] introduces a concept of two characters in human brains, so-called System 1 and System 2. System 1 includes automatic, involuntary and effortless automatic mental operations (ex.: driving, simple counts like 2x2, etc.) while System 2 slows down for reasoning, computing, and logical thinking, not jumping to quick, intuitive solutions. Sometimes, those two systems provide different results, conflicting with each other.

The laboratory testing can be considered as a typical situation evaluated by the System 2 character – subjects are entirely concentrated on the intelligibility task, analyzing the perceived stimuli by their full and careful attention. On the contrary, the parallel task arrangement may force

the tester to use System 1 for intelligibility tests while being fully engaged in the parallel task activity that occupies their System 2 capacity. This would potentially explain the better intelligibility of certain test words for certain subjects under the parallel task condition – their intuition (System 1) prevails over their analytical and critical thinking (System 2). This effect perhaps arises for some listeners only, depending on their past experience which cannot be consciously (System 2) recalled and utilized.

6. Conclusions and future work

A new methodology for speech intelligibility subjective testing was proposed and demonstrated. The addition of parallel psychomotor tasks brings the test conditions closer to a real environment, thus making the test results more realistic. 51 subjects participated in intelligibility tests in conditions of the laboratory environment and parallel task. Various background noises and coders were used during the tests. Demonstration experiment results are presented.

The differences in intelligibility test results between standard laboratory and parallel task test methodology were identified. Overall, the intelligibility of regular tests was mainly higher than for tests with a parallel task. However, intelligibility results of certain samples have been found counterintuitive, which is possibly explained by a different mode of thinking under parallel task conditions and should be considered as a subject for further study. The demonstrated results justify the need for further experiments deploying parallel task for speech intelligibility and in general audio quality subjective tests. Better understanding of which stimuli and which parallel task types (if any) may lead to higher intelligibility can improve communication, e.g., between military personnel or air-approach control dispatchers, including critical and emergency situations.

In the future, it is planned to continue the study with different stimuli and parallel task types to mimic other real scenarios of communication equipment usage.

Methods

Our experiment involved human participants and had been approved by Head of Department Advisory Committee at the Faculty of Electrical Engineering, Czech Technical University in Prague under the No. 82/2017. All experiments were performed in accordance with ethical principles of Declaration of Helsinki. All involved subjects provided their written informed consent prior the experiment. There are no subject identifying details (HIPAA) in our contribution.

Acknowledgement

This study was supported by Internal Grant of Czech Technical University "Comparison of speech intelligibility between native and non-native English speakers in laboratory and simulated battlefield environments" under the Number SGS17/191/OHK3/3T/13. Authors would like to thank Dr. Stephen Voran from NTIA ITS for his valuable hints and final text revision and mesaqin.com sro (Ltd.) for providing test equipment, test premises and test subjects for this study.

References

[1] P. Zhu, F. Mo, J. Kang: Relationship between chinese speech intelligibility and speech transmission index under reproduced general room conditions. Acta Acustica united with Acustica **100** (2014) 880–887.

[2] S. Jørgensen, J. Cubick, T. Dau: Speech intelligibility evaluation for mobile phones. Acta Acustica United With Acustica **101** (2015) 1016–1025.

[3] Ergonomics – Assessment of speech communication. ISO Standard 9921:2003.

[4] K. Bunton, C. K. Keintz: The use of a dual-task paradigm for assessing speech intelligibility in clients with Parkinson disease. J Med Speech Lang Pathol **16** (2008) 141–155.

[5] S. Choi, A. Lotto, D. Lewis, B. Hoover, P. Stelmachowicz: Attentional modulation of word recognition by children in a dual-task paradigm. J Speech Lang Hear Res. **51** (2008) 1042–54.

[6] K. S. Helfer, J. Chevalier, R. L. Freyman: Aging, spatial cues, and single- versus dual-task performance in competing speech perception. J Acoust Soc Am. **2010** (128) 3625–3633.

[7] C. Kwak, W. Han: Comparison of single-task versus dual-task for listening effort. J Audiol Otol. (2017).

[8] A. Sarampalis, S. Kalluri, B. Edwards, E. Hafter: Objective measures of listening effort: effects of background noise and noise reduction. J Speech Lang Hear Res. **52** (2009) 1230–40.

[9] Y. H. Wu, E. Stangl, X. Zhang, J. Perkins, E. Eilers: Psychometric functions of dual-task paradigms for measuring listening effort. Ear Hear. **37** (2016) 660–670.

[10] D. Navon, D. Gopher: On the economy of the human-processing system. Psychol. Rev. **86** (1979) 214–255.

[11] C. D. Wickens: Processing resources and attention. – In: Multiple Task Performance. D. L. Damos (ed.). Taler & Francis, Ltd., Bristol, 1991, 3–34.

[12] S. L. Beilock, T. H. Carr, C. MacMahon, J. L. Starkes: When paying attention becomes counterproductive: Impact of divided versus skill-focused attention on novice and experienced performance of sensorimotor skills. Journal of Experimental Psychology: Applied **8** (2002) 6–16.

[13] ITU-T Rec. P.807: Subjective test methodology for assessing speech intelligibility. Series P: Terminals and subjective and objective assessment methods. Geneva, 2016.

[14] Method for measuring the intelligibility of speech over communication systems. ANSI/ASA Standard S3.2-2009.

[15] Dynastat Inc.: Summary of speech intelligibility testing methods. http://www.dynastat.com/Speech%20Intelligibility.htm.

[16] ITU-T Rec. P.800: Methods for subjective determination of transmission quality. Series P: Telephone transmission quality. ITU, Geneva, 1996, am. 1998.

[17] ITU-T Rec. G.711: Pulse code modulation (PCM) of voice frequencies. Series G: Transmission systems and media, digital systems and networks. ITU, Geneva, 1988, am. 2009.

[18] 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.

[19] M. McIntosh, S. S.: Non-citizens in the enlisted U.S. military. CRM D0025768.A2/Final, CAN Analysis, November 2011.

[20] A. E. Hill, B. J. Davidson, D. G. Theodoros: The performance of standardized patients in portraying clinical scenarios in speech-language therapy. Int. J. Lang. Commun. Disord. **48** 613–624.

[21] D. Kahneman: Thinking, fast and slow. Farrar, Straus and Giroux, New York, 2011.

Chapter 6

Parallel Task in Subjective Speech Quality Measurement

6.1 Summary

After the successful implementation of the parallel-task technique in subjective speech intelligibility tests, the next step was to implement it in subjective speech quality assessment tests. Two sets of subjective tests deploying ITU-T P.835 were provided to this end. Both tests were provided in Prague, the Czech Republic in July of 2015 and in January of 2017 with 32 and 25 subjects respectively. Contemporary coders AMR WB [49] and EVS [50] and selected cases of background noises (adopted from [51]) were used to create a balanced set of realistic speech samples. The scenario of the parallel-task was the same as in the previous test: a laser-shooting simulator with ducks as targets. The "shooters" and "counters" were distributed with the same logic as in the previous test. However, this time, the subjects' task was to assess three values: Speech quality, noise annoyance, and overall quality of samples. Afterward, the results were compared using Pearson correlations, and pairwise comparisons were performed. The resulting analysis indicated voting mistakes because of loss of subjects' concentration due to parallel task introduction.

6.2 Publication

All the information about the tests and results analysis are described in the article [52].

PLOS ONE

RESEARCH ARTICLE

Subjective speech quality measurement with and without parallel task: Laboratory test results comparison

Hakob Avetisyan, Jan Holub*

Department of Measurement, Faculty of Electrical Engineering, Czech Technical University, Prague, Czech Republic

* holubjan@fel.cvut.cz

Citation: Avetisyan H, Holub J (2018) Subjective speech quality measurement with and without parallel task: Laboratory test results comparison. PLoS ONE 13(7): e0199787. https://doi.org/10.1371/journal.pone.0199787

Editor: Gavin Kearney, University of York, UNITED KINGDOM

Received: December 13, 2017

Accepted: June 13, 2018

Published: July 2, 2018

Data Availability Statement: All relevant data underlying the findings of this study (speech samples, raw votes, results) are available as Supporting Information files and also from protocols.io under the following DOI: dx.doi.org/10.17504/protocols.io.nwwdffe.

Funding: This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS17/191/OHK3/3T/13. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

This paper focuses on a novel methodology of subjective speech quality measurement and repeatability of its results between laboratory conditions and simulated environmental conditions. A single set of speech samples was distorted by various background noises and low bit-rate coding techniques. This study aimed to compare results of subjective speech quality tests with and without a parallel task deploying the ITU-T P.835 methodology. Afterward, tests results performed with and without a parallel task were compared using Pearson correlation, CI95, and numbers of opposite pair-wise comparisons. The tests show differences in results in the case of a parallel task.

Introduction

Each generation of mobile phones has different advanced features and characteristics designed to have a better quality of voice processing and noise suppression. For this purpose, various subjective and objective tests are performed to analyze, compare and improve the audio quality emerging mobile technologies. Subjective speech quality testing is designed for collecting subjective opinions from human test subjects deploying standardized procedures as specified, e.g., in [1]. Objective methods [2] are used to replace test subjects using psycho-acoustic modeling, comparing clean and distorted speech samples algorithmically. Outputs from these two method groups are often mapped to the subjective quality scale Mean Opinion Score (MOS) [2]. Comparing subjective and objective quality tests, subjective tests are believed to provide more accurate results but are also more demanding regarding time, equipment, effort and price. The main point, however, is the fundamental philosophy of currently used test methods: Test subjects are seated in anechoic or semi-anechoic test room and are fully focused on listening to the tested material. In real life, the users are usually performing multiple tasks at once (such as talking on the phone while working on PC, walking or even driving a car, or visually monitoring a screen where airplane location and approach situation is displayed while communicating on radio-link with the airplane pilot).

This paper deals with a novel technique of subjective testing with an implementation of a parallel task which simulates a real environmental situation. The reported experiment aims to

Competing interests: Mesaqin.com Ltd. provided the test equipment, test premises and test subjects for this study. This service has been performed voluntarily, at no costs and no obligations and does not alter our adherence to PLOS ONE policies on sharing data and materials. There are no restrictions on sharing of data and materials.

verify if the ITU-T P.835 [3] methodology is suitable for parallel task incorporation, to identify potential differences in human perception under a parallel task situation and to demonstrate their impact to speech quality perception. Previously, comparisons between tests in laboratory conditions (without a parallel task) were performed, which didn't show any crucial differences between tests performed in different laboratories [4–6].

The paper is structured as follows. After the Background section, the experiment description is given, providing information about methods, tested samples and equipment used. Next, we provide data analysis of measured speech quality, noise annoyance, and overall quality (as per ITU-T P.835 [3]) and compare results with and without a parallel task. Alongside Pearson correlation coefficient and CI95 uncertainty intervals, pairwise comparisons between each couple of tests are also provided. The final section contains conclusions and motivations for future work.

Background

ITU-T Recommendation P.835 [3] describes methods for evaluating speech quality in noisy (and partially de-noised) speech. A typical example of its application is a comparison of different noise suppression algorithms. The P.835 methodology makes it possible to evaluate speech quality and noise levels separately. Test environment parameters are adopted from ITU-T P.800 [1]. Listeners evaluate tested samples on a five-point scale. This procedure is particularly suitable for samples processed by noise canceling algorithms that remove certain part of background noise but also corrupt the speech itself. Therefore, the principle of P.835 is to repeat the assessment of each speech sample three times, requiring the subjects to focus on a different aspect of the sample quality during each assessment. For the first half of samples, the subjects are asked to focus on speech quality only during the first playback, noise annoyance during the second playback and overall sample quality during the third (last) playback. For the second half of samples, subjects are asked to judge noise annoyance during the first playback, speech quality during the second playback and, identical to the first half of samples, overall sample quality for the third playback. The order of sample presentation is randomized.

The existing test methods and recommendations are based on the intuitive assumption [7] that the case of laboratory testing where the subjects are fully concentrated to the test procedure provides the most sensitive results compared to any real scenario when the users are distracted by performing other tasks. Multiple experiments that including a dual task have been performed on impaired [8] or child [9] subjects and do not focus on the average adult population. Subjects acquired from the common population are usually tested for a relationship between listening effort and dual task introduction [10–15].

Experiment description

For data analysis, two subjective tests were held in subjective testing laboratory based in Prague, Czech Republic. They are named as A and B. Test A was performed in July 2015 and test B in January 2017. Test subjects from test A were different from test B. Test A contained 32 subjects and test B included 25 subjects. The test subjects were hired by professional listening lab service using social media advertisements. A mixture of subjects' nationalities has been used (American, British, German, French, Czech, and Slovak). The exact nationality distribution is shown in Table 1. The English language proficiency of non-English participants was higher than average as verified by a short written English quiz, preceding the subjective testing. The written quiz was selected due to its short duration; despite the fact it is not an optimal means of assessing the ability to understand the spoken language. However, language understanding is not a necessary condition for speech quality assessment as demonstrated in [16].

Table 1. Nationality distribution in tests A and B.

	U.S.	British	German	French	Czech	Slovak	TOTAL
Test A	2	2	3	4	15	6	**32**
Test B	1	2	2	3	12	5	**25**

https://doi.org/10.1371/journal.pone.0199787.t001

The gender structure of the listening panels was balanced–test A included 16 male and 16 female test subjects while test B included 13 male and 12 female subjects. 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 [1] and [3]. Original studio recordings were spoken by native professional English speakers (two male, two female voices). A selection of Harvard phonetically balanced sentences from the Appendix of IEEE Subcommittee on Subjective Measurements was used. Contemporary coders AMR WB [17] and EVS [18] and selected cases of background noise (Cafeteria, Mensa, Road, Pub, Office, Car, all adopted from [19]) were used to create a balanced set of realistic speech samples that covered a full coverage of quality. The background noise was mixed with speech material following ITU-T P.835 [3] Appendix 1. The final sample selection contained 22 conditions. Table 2 details the samples used.

The test methodology was based on recommendation ITU-T P.835. As already discussed in the Background section, 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. As required by P.835, half of the test was performed in speech-noise-overall and the other half noise-speech-overall orders. MOS scores were obtained separately for Speech quality (S-MOS), Noise annoyance (N-MOS) and Overall sample quality (G-MOS). The terms S-MOS, N-MOS, and G-MOS, are adopted from ETSI TS 103 106 [20] and ETSI EG 202 396–3 [21]. These terms replace in the further text the original SIG, BAK, and OVRL ratings used in [3].

Table 2. Test sample conditions.

Sample type (coder, bit rate)	Noise type acc. to [19]	SNR (dB) acc. to [3]	Number of samples
AMR WB 12,65k	Cafeteria	14,8	8
AMR WB 12,65k	Mensa	19,5	8
AMR WB 12,65k	Road	7,9	8
AMR WB 12,65k	Pub	8,4	8
AMR WB 12,65k	Office	25,3	8
EVS WB 13,2k	Cafeteria	14,8	8
EVS WB 13,2k	Mensa	19,5	8
EVS WB 13,2k	Road	7,9	8
EVS WB 13,2k	Pub	8,4	8
EVS WB 13,2k	Office	25,3	8
Reference 1–5	n/a	n/a	10
Reference 6,10	Car	0	4
Reference 7,11	Car	12	4
Reference 8,12	Car	24	4
Reference 9	Car	36	2

https://doi.org/10.1371/journal.pone.0199787.t002

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.

Both mental and physical parallel tasks are used in existing experiments [8–15]. To avoid the problem of generated load inequity for differently physically or mentally developed subjects, we designed a combined parallel task, incorporating both physical and mental efforts: A simple game deploying a professional laser shooting simulator (Simway) was used. Always a group of three subjects was evaluating the samples; however, at any given time one of them was a "shooter," and other two were "counters." The "shooter's" task was to shoot as many in-game ducks as they could, and the "counters'" task was to count every single shot duck. The turn of the shooter was changed randomly using a light-bulb indicating who the current shooter was. The three bulbs (one in front of each subject) were operated by a random number generator always ensuring only one lamp was on, and every 40 seconds another lamp activated. The reason for swapping the roles was the shooting simulator limitation–only one single shooter is allowed at a time. Running the test separately for each subject, with each subject only as a shooter, would be extremely time-consuming. The compromising solution was to assign the "shooter" role randomly among three subjects, all of them assessing the speech samples in parallel. The samples were played out in random order using a different randomization for each listening panel.

Materials and methods

Our experiment involved human participants and has been approved by Advisory Committee of the Dean of Faculty of Electrical Engineering, Czech Technical University in Prague, decision letter dated April 17th, 2015. All experiments were performed in accordance with the Declaration of Helsinki and relevant local guidelines and regulations. All involved subjects provided their written informed consent prior the experiment. There are no subject identifying details (HIPAA) in our contribution.

For the sound reproduction, Sennheiser HD 600 professional headphones were used. Votes were collected using a professional voting device. The used low-reverberation listening rooms conformed to requirements of [1]. Its reverberation time was 185ms and background noise level below 30dB SPL (A) without significant peaks in spectra.

All test results and their evaluation are available as supporting information files and also at protocols.io under dx.doi.org/10.17504/protocols.io.nwwdffe

Results and data analysis

In S1, S2 and S3 Figs, the correlations between S-MOS, N-MOS, and G-MOS values are presented. The values are highly correlated. Nevertheless, there are interesting values worth mentioning.

Speech MOS (S-MOS) comparison between A and B tests are shown in S1 Fig. Its 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), as shown in Table 3.

In the second part, the subjects were voting for background noise annoyance (5 –not noticeable to 1 –very intrusive). S2 Fig shows noise annoyance MOS correlations between A and B. Its Pearson correlation coefficient value is 0,982.

Finally, during the third part, the subjects were voting for the overall quality of each sample (5 –excellent to 1 –bad). For the second half of each experiment, the order of second and third voting was swapped as required by P.835. In S3 Fig, overall quality MOS correlations between A and B tests are shown. The Pearson correlation coefficient is 0.989.

Table 3. Test questions as per ITU-T P.835.

Opinion Score	Speech signal rating scale	Background noise rating scale	The overall quality rating scale
5	Not distorted	Not noticeable	Excellent
4	Slightly distorted	Slightly noticeable	Good
3	Somewhat distorted	Noticeable but not intrusive	Fair
2	Fairly distorted	Somewhat intrusive	Poor
1	Very distorted	Very intrusive	Bad

https://doi.org/10.1371/journal.pone.0199787.t003

In S1 Fig, 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 we conclude 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.

The graphs show that the subjects voted similarly. Correlation values are close to the maximum value of 1. However, as indicated in S1 Fig, certain sample pairs are ranked oppositely with and without a parallel task. For this purpose, pair-wise comparisons [22] were performed as described further.

Pairwise comparison of each test

After the data correlations procedure, pairwise comparisons for the tests were evaluated. The comparison was performed in 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 S1 Fig and described in the section Results and Data analysis) 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).

Table 4 includes information about the average Confidence Intervals of each type of MOS for both tests. CI95 increases with parallel task introduction.

Table 4. Average CI95 of each test.

	Average CI95: S-MOS	Average CI95: N-MOS	Average CI95: G-MOS
A	0,133	0,117	0,113
B	0,155	0,137	0,148

https://doi.org/10.1371/journal.pone.0199787.t004

Conclusion and motivation for future work

A novel subjective testing methodology has been designed and demonstrated. The purpose of the parallel task during subjective testing was to bring the test results closer to realistic conditions. In total, 57 subjects participated in 2 different tests with and without implementation of the parallel task.

Pearson correlations between tests were calculated, and positions of values of subjects' votes were plotted in graphs. Due to non-consistent values, pair-wise comparisons were performed, and ten differences were found.

Although the test results were highly correlated, certain conditions indicate different pair rankings after the parallel task is introduced. The resulting analysis indicated voting mistakes because of loss of subjects' concentration due to parallel task introduction. Therefore, we conclude that ITU-T P.835 methodology is too complicated to be combined successfully with a complex parallel task as described here.

In the future, it is planned to continue the investigation, experimenting with less complex parallel task within P.835 context or using different methodology (e.g., ITU-T P.800) for the existing parallel task. Also, standardization effort will be initiated to define parallel task subjective testing as a logical counterpart to traditional laboratory subjective speech quality tests.

Supporting information

S1 Fig. Speech MOS (S-MOS) of A and B tests. Both axes have the values of MOS (1–5).
(TIF)

S2 Fig. Noise annoyance MOS (N-MOS) of A and B tests. Both axes have the values of MOS (1–5).
(TIF)

S3 Fig. Overall quality MOS (G-MOS) of A and B test. Both axes have the values of MOS (1–5).
(TIF)

S1 File. Subjective data.xlsx. Detailed results of A and B experiments.
(XLSX)

S2 File. Samples. Speech samples used for both the A and B experiments.
(ZIP)

Acknowledgments

This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS17/191/OHK3/3T/13.

Authors would like to thank Mesaqin.com Ltd. for providing the test equipment, test premises and test subjects for this study. This service has been performed voluntarily, at no costs and no obligations and does not alter our adherence to PLOS ONE policies on sharing data and materials.

Author Contributions

Conceptualization: Jan Holub.

Formal analysis: Hakob Avetisyan.

Investigation: Hakob Avetisyan.

Methodology: Jan Holub.

Project administration: Jan Holub.

Supervision: Jan Holub.

Validation: Hakob Avetisyan.

Writing – original draft: Hakob Avetisyan.

Writing – review & editing: Jan Holub.

References

1. ITU-T Rec. P.800. Telephony transmission quality, Methods for subjective determination of transmission quality. International Telecommunication Union, Geneva. 1996
2. ITU-T Rec. P.863. Perceptual Objective Listening Quality Assessment. International Telecommunication Union, Geneva. 2011
3. ITU-T Rec. P.835. The subjective test methodology for evaluating speech communication systems that include noise suppression algorithm. International Telecommunication Union, Geneva, 2003
4. Holub J, Avetisyan H, Isabelle S. Subjective speech quality measurement repeatability: comparison of laboratory test results. Int J Speech Technol [Internet]. 2017; 20(1):69–74. Available from: http://link.springer.com/10.1007/s10772-016-9389-6
5. Goodman D, Nash R. Subjective quality of the same speech transmission conditions in seven different countries. In: ICASSP '82 IEEE International Conference on Acoustics, Speech, and Signal Processing [Internet]. Institute of Electrical and Electronics Engineers; p. 984–7. Available from: http://ieeexplore.ieee.org/document/1171565/
6. Arifianto D, Sulistomo TR. Subjective evaluation of voice quality over GSM network for quality of experience (QoE) measurement. In: 2015 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS) [Internet]. IEEE; 2015;p. 148–52. Available from: http://ieeexplore.ieee.org/document/7432755/
7. Cote N. Integral and Diagnostic Intrusive Prediction of Speech Quality. Springer 2011, ISBN 978-3-642-18462-8
8. Bunton K, Keintz CK. The Use of a Dual-Task Paradigm for Assessing Speech Intelligibility in Clients with Parkinson Disease. J Med Speech Lang Pathol. 2008 Sep 1; 16(3):141–155 PMID: 21637738
9. Choi S, Lotto A, Lewis D, Hoover B, Stelmachowicz P. Attentional modulation of word recognition by children in a dual-task paradigm. J Speech Lang Hear Res. 2008 Aug; 51(4):1042–54. https://doi.org/10.1044/1092-4388(2008/076) PMID: 18658070
10. Helfer KS, Chevalier J, Freyman RL. Aging, spatial cues, and single- versus dual-task performance in competing speech perception. J Acoust Soc Am. 2010 Dec; 128(6):3625–33. https://doi.org/10.1121/1.3502462 PMID: 21218894
11. Kwak C, Han W. Comparison of single-task versus dual-task for listening effort. J Audiol Otol. 2017 Oct 17. https://doi.org/10.7874/jao.2017.00136 PMID: 29036758
12. Wu YH, Stangl E, Zhang X, Perkins J, Eilers E. Psychometric functions of dual-task paradigms for measuring listening effort. Ear Hear. 2016 Nov/Dec; 37(6):660–670 https://doi.org/10.1097/AUD.0000000000000335 PMID: 27438866
13. Navon D, and Gopher D. On the economy of the human-processing system. Psychol. Rev. 86; 1979; 214–255.
14. Wickens CD 1991. Processing resources and attention. In Multiple Task Performance (ed. Damos D. L.), pp. 3–34. Taler & Francis, Ltd., Bristol.
15. Beilock S Carr L, MacMahon T H, &Starkes, J L C. When paying attention becomes counterproductive: Impact of divided versus skill-focused attention on novice and experienced performance of sensorimotor skills. Journal of Experimental Psychology: Applied, 2002; 8, 6–16. PMID: 12009178
16. Schinkel-Bielefeld N, Zhang J; Qin Y; Leschanowsky A K; Fu S. Perception of Coding Artifacts by Non-native Speakers—A Study with Mandarin Chinese and German Speaking Listeners, February 2018; JAES Volume 66 Issue 1/2 pp. 60–70; January 2018, https://doi.org/10.17743/jaes.2017.0042
17. ITU-T G.722.2, Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB), International Telecommunication Union, Geneva 2013

18. ETSI TS 126 445, Universal Mobile Telecommunications System (UMTS); LTE; EVS Codec Detailed Algorithmic Description, European Telecommunication Standardization Institution, Sophia-Antipolis, 2014
19. 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, European Telecommunication Standardization Institution, Sophia-Antipolis, 2008
20. ETSI TS 103–106. European Telecommunications Standards Institute. Speech and multimedia Transmission Quality (STQ); Speech quality performance in the presence of background noise: Background noise transmission for mobile terminals-objective test methods. European Telecommunication Standardization Institution, Sophia-Antipolis, 2014
21. ETSI EG 202-396-3, Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality performance in the presence of background noise Part 3: Background noise transmission—Objective test methods. European Telecommunication Standardization Institution, Sophia-Antipolis, 2008
22. ITU-T TD12rev1. Statistical evaluation. Procedure for P.OLQA v.1.0. Berger J, editor. International Telecommunication Union, Geneva. 2009.

6.3 Supplementary materials

Since the corresponding journal includes supporting materials in a complementary file, they are added in this subsection.

Mentioned **S1 File.Subjective data.xlsx** and **S2 File.Samples** were not included in this thesis because of their big sizes and can be found here:
`https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0199787`

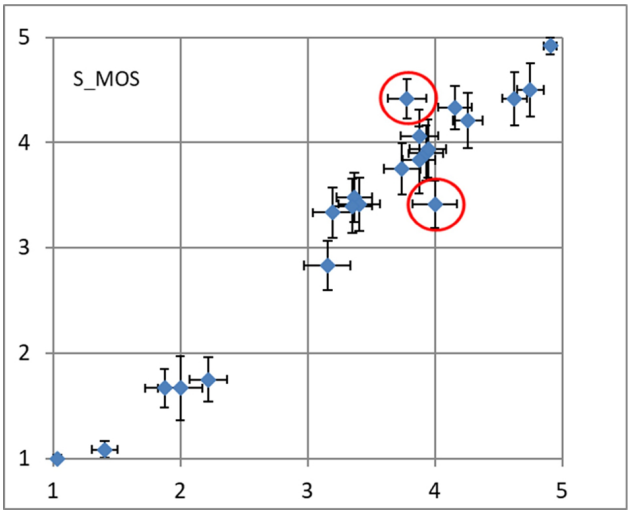


Figure 6.1: (Corresponding to S1 Fig) Speech MOS(S-MOS) of A and B tests. Both axes have the values of MOS(1–5).

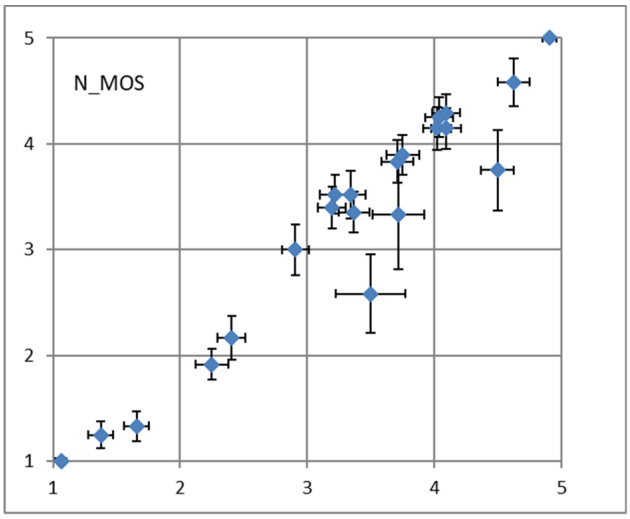


Figure 6.2: (Corresponding to S2 Fig) Noise annoyance MOS (N-MOS) of A and B tests. Both axes have the values of MOS (1–5).

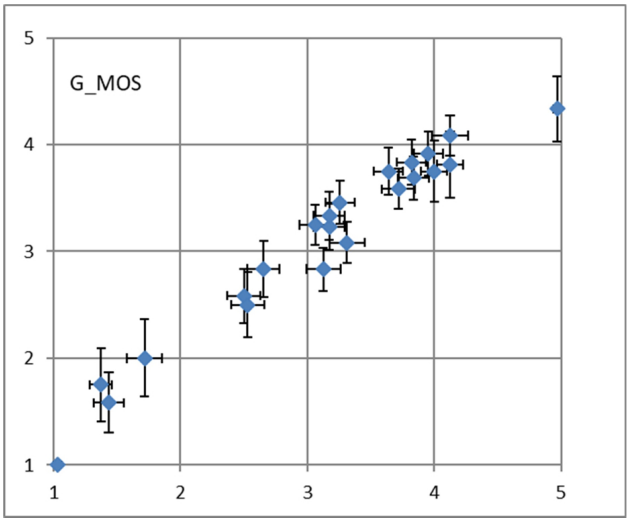


Figure 6.3: (Corresponding to S3 Fig) Overall quality MOS (G-MOS) of A and B test. Both axes have the values of MOS (1–5).

Chapter 7

Testing the Speech Intelligibility for Czech Language

7.1 Summary

After a deep investigation of subjective speech quality testing and speech intelligibility testing with parallel task in the English language, we decided to go further and provide intelligibility tests in the Czech language. For this purpose, 40 new samples in the Czech language were recorded using two male and two female narrators. Speech samples have been recorded in conformance to P.800 using two male and two female talkers with different pitch positions in frequency spectra. Single words, as proposed by [53], were recorded. A new software running on LabVIEW programming environment has been created according to ITU-T P.807 requirements. A new laser shooting simulator was used for the parallel task with a simple game of shooting range. In total, 45 (without parallel-task test) and 70 (in the parallel-task test) Czech nationals participated in the tests in the age range of 18 – 56. As before, these results also had counterintuitive values. However, all of them were statistically insignificant since they were in ranges of the Standard Deviation of Arithmetical Mean. Nevertheless, this was another step for improving voice transmission quality and noise suppression algorithms.

7.2 Publication

The related publication is available in the Journal of Audio Engineering Society [54].

H. Avetisyan, J. Holub and O. Slavata, "Low Bit-Rate Coded Speech Intelligibility Testing in Czech Language Using Parallel Task"
J. Audio Eng. Soc., vol. 68, no. 5, pp. 1–8, (2020 May.).
DOI: https://doi.org/10.17743/jaes.2020.0008

Low Bit-Rate Coded Speech Intelligibility Testing in Czech Language Using Parallel Task

HAKOB AVETISYAN, JAN HOLUB, AND OLDŘICH SLAVATA
(avetihak@fel.cvut.cz) (holubjan@fel.cvut.cz) (slavao1@fel.cvut.cz)

Czech Technical University in Prague, Prague, Czech Republic

This article deals with subjective tests of speech intelligibility. A set of samples in the Czech language, recorded by four different narrators, was distorted with different noise levels and encoded by a low bit-rate encoder. The subjective test consisted of two parts. The first part (45 participants) proceeded according to the ITU-T Recommendation P.807 - Subjective test methodology for assessing speech intelligibility. The second part (70 participants) included an additional (parallel) psychomotor task deploying a laser-shooting simulator, in which subjects had the roles of shooters and counters. The purpose of the parallel task is to bring the testing closer to the real use of technology. Significant differences have been found in the results of the intelligibility of samples from different speakers. There were also differences in evaluation with and without a parallel task. Samples from male narrators have a significantly higher intelligibility score in the standard laboratory test but also show a greater decrease in intelligibility after engaging a parallel task.

0 INTRODUCTION

Speech intelligibility [1] is one of the fundamental parameters measured on newly designed communication equipment, speech coding algorithms and circuits, and overall communication channels. The speech intelligibility is usually of interest for speech channels with compromised speech quality (for example, when low-bit rate coders or multiple speech coding [transcoding] are deployed). This parameter and its optimization also play a crucial role in special telecommunications and radiocommunications used by military or public safety services.

Speech intelligibility can be measured either by subjective testing (auditory measurement) or by objective algorithmic procedures. This article focuses on the first option: a subjective test using a panel of human listeners. It is structured as follows:

Sec. 1 includes primary information about intelligibility testing and various types of tests.

Sec. 2 deals with the problem definition.

Sec. 3 contains all the necessary information about experimental design, parameters, and used standards.

In Sec. 4, information about data analysis and full obtained results is presented.

Sec. 5 includes a discussion about the results.

In Sec. 6, the conclusion of the article and information about future work is presented.

1 STATE OF THE ART

Different methods with various modifications for intelligibility testing have been developed, including Diagnostic Rhyme Test (DRT), Diagnostic Medical Consonant Test (DMCT), Diagnostic Alliteration Test (DALT), Modified Rhyme Test (MRT) [2], [3], or so-called matrix sentence test [4].

Intelligibility tests are determined by the tested type of speech: sentences, meaningful words, nonsense words, rhyme-words, or a limited vocabulary (spell-alphabet, digits).

The intelligibility measurement is a useful diagnostic tool for communication systems optimization and designing and training of the predictive (algorithmic) intelligibility measurements [5], [6].

The common denominator of all standardized intelligibility tests is the required environment of silent and anechoic or semianechoic listening chamber or listening booth with defined acoustic properties. Test subjects (listeners) are seated in this test environment, and their only task is listening to each sample and assessing its meaning by marking (on a paper form or more often by means of personal computer or tablet) the answer considered as correct [6].

Such test a environment contributes to testing sensitivity and repeatability but differs significantly from consequent regular use of the tested communication equipment.

Users of this equipment drive cars, operate other devices, watch TV, or perform other parallel activity during a conversation on the phone. It was shown that such multitask scenario changes human perception in a hardly predictable way and should be considered during the laboratory testing phase by parallel task introduction [6]. According to the European Telecommunications Standards Institute (ETSI) recommendation ETSI TR 103 503 [7], parallel tasks are divided into three types: mentally oriented tasks, physically oriented tasks, and hybrid tasks. Typical examples of mentally oriented parallel tasks are mentioned, for instance, in [8] and [9], in which listeners had to memorize digits simultaneously performing an intelligibility test. Physically oriented parallel tasks include various physical exercises that can be performed in the laboratory, e.g., riding the stationary bike [7]. Hybrid tasks include both mental and physical tasks. Hybrid tasks used in speech intelligibility have been used in [2] and [10] and in subjective speech-quality assessment (has been studied in the parallel task context) used in [11]

2 PROBLEM DEFINITION

Speech intelligibility testing with a deployment of a parallel task is used for various languages, usually for Germanic and Romanic language groups (English, German, French, Italian). However, no materials have been found about tests in the Slavic language group (Czech, Polish, Russian, Serbian, etc.). The communication within many organizations in these countries (including civil aviation, police, and army) is provided in their native languages. Thus, it was decided to apply this method also to this language group. The Czech language was chosen because of its availability for research purposes, including the number of subjects, source materials, and the location of the research providers.

This research aimed to answer the following questions raised in the context of speech intelligibility testing with the parallel task:

- Do the changes in measured intelligibility identified after parallel task introduction differ for different tested conditions (different coders, signal-to-noise ratios, etc.)?
- Is there any talker or talker gender dependency of the above for low bit-rate coders?
- Are there any cases of counterintuitive test results (the intelligibility evaluated under parallel task condition is better than when measured by regular intelligibility test), as was observed in [2]?

3 EXPERIMENT DESCRIPTION

The tests have been performed in an environment with acoustical conditions strictly compatible with ITU-T P.800 [12]. Its reverberation time is below 200 ms, and background noise less than 30 dB SPL(A) without peaks in frequency spectra. Speech samples have been recorded in conformance to P.800 using two male and two female talkers with different pitch positions in frequency spectra. Single words, as proposed by [13], were recorded. For sample playout, a high-quality amplifier and a professional near-field audio/system with compensated and calibrated frequency response has been used. Subjects' votes have been collected by numerical keyboards using a proprietary voting system programmed for the required purposes.

The intelligibility tests were performed following Modified Rhyme Test (MRT) methodology as described, e.g., in ITU-T P.807 [6] (Subjective test methodology for assessing speech intelligibility, Geneva, February 2016). In total, 40 new samples in the Czech language have been recorded and implemented into the MRT sample list, and for practical reasons, only five answer options have been used instead of the original six. For the experiment, described in this text, the samples have been recorded using voices of two male and two female narrators, conditionally named as M1, F1, M2, and F2. The samples have been distorted using a network simulator deploying the following coder and background noise options: PCM [14] at 8-bit 16 kSa/s (16 samples) and MELPe [15] at 2.4 kSa/s (32 samples). Sixteen MELPe samples have been previously mixed with the background noise of interior High Mobility Multipurpose Wheeled Vehicle (HMMWV) noise at a signal-to-noise ratio (SNR) of 0dB [16].

3.1 Apparatus

An external audio sound card (Fireface UCX by RME) and wide-band digitally compensated near-field monitor loudspeaker Dynaudio BM5 have been used. The reverberation time of the test room with a special acoustic lining was less than 500 ms and background noise below 20 dB SPL(A) with no peaks in its frequency spectra.

As a parallel task, a simple game has been used deploying professional laser shooting simulator (VRHunter).

A new software running on LabVIEW programming environment has been created according to ITU-T P.807 requirements.

The whole software part was running on a Windows 10 operating system.

3.2 Stimuli

The list of 40 used Czech words and their rhyme options has been designed following principles given in [6]. The samples have been derived from high-quality studio recordings of professional native Czech speakers. PCM [14] and MELPe [15] coders have been selected as typical examples of regular narrow-band and modern narrow-band coders. The only background noise used was HMMWV noise at 0 dB SNR. Its frequency spectrum is shown in Fig. 1. This noise type and level represents a typical military communication scenario with stationary background noise.

Detailed information about every sample is shown in Table 1.

3.3 Subjects

Forty-five subjects for tests in a laboratory environment and 70 subjects for tests with the deployment of a parallel

Table 1. SAMPLE DESCRIPTION

Sample	Narrator	Codec/noise	Sample	Narrator	Codec/noise	Sample	Narrator	Codec/noise
1	M1	PCM	**15**	M2	MELPe/HMMWV	**29**	M1	MELPe
2	F1	MELPe	**16**	F2	PCM	**30**	F1	MELPe/HMMWV
3	M2	MELPe/HMMWV	**17**	M1	MELPe	**31**	M2	PCM
4	F2	PCM	**18**	F1	MELPe/HMMWV	**32**	F2	MELPe
5	M1	MELPe	**19**	M2	PCM	**33**	M1	MELPe/HMMWV
6	F1	MELPe/HMMWV	**20**	F2	MELPe	**34**	F1	PCM
7	M2	PCM	**21**	M1	MELPe/HMMWV	**35**	M2	MELPe
8	F2	MELPe	**22**	F1	PCM	**36**	F2	MELPe/HMMWV
9	M1	MELPe/HMMWV	**23**	M2	MELPe	**37**	M1	PCM
10	F1	PCM	**24**	F2	MELPe/HMMWV	**38**	F1	MELPe
11	M2	MELPe	**25**	M1	PCM	**39**	M2	MELPe/HMMWV
12	F2	MELPe/HMMWV	**26**	F1	MELPe	**40**	F2	PCM
13	M1	PCM	**27**	M2	MELPe/HMMWV			
14	F1	MELPe	**28**	F2	PCM			

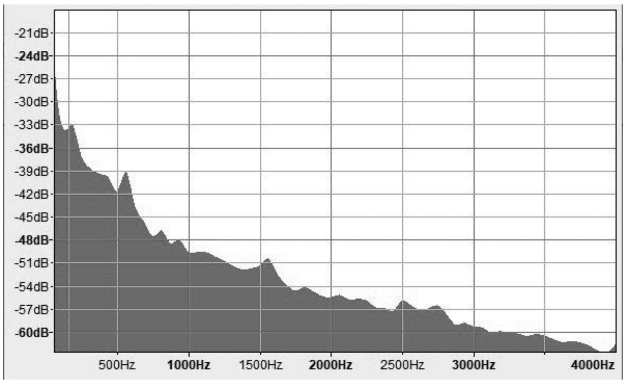


Fig. 1. Frequency spectrum of the High Mobility Multipurpose Wheeled Vehicle (HMMWV) noise.

task in the age range of 18–56 participated in the tests. All subjects were native Czech speakers. Since it was expected that the results of parallel task-driven tests could have a larger dispersion, it was decided to increase the number of subjects for the parallel task deployed test so that the standard deviation of the arithmetical mean of both tests could be comparable. This does not affect the overall results of the tests.

This experiment involved human participants and had been approved by Head of Department Advisory Committee at the Faculty of Electrical Engineering, Czech Technical University in Prague under the No. 82/2017. All experiments were performed in accordance with the ethical principles of the Declaration of Helsinki. All involved subjects provided their written informed consent prior to the experiment. There are no subject identifying details (HIPAA) in our contribution.

3.4 Procedure

The testing process has been divided into two sessions. In the first session, regular intelligibility tests have been performed. This was followed by a 60-minute break when a different subjective (visual) low-complexity task had been assigned to the subjects. It was used to make the subjects feel relaxed and focused on a different task.

In the next session, the same intelligibility test has been performed but with a parallel task implementation. This part of the test was performed by groups of three to four subjects. One of them always played the role of ”shooter” while the remaining ones were ”counters.” Every 30 seconds, the roles were dynamically randomly assigned by a random number generator, which was allocated near the screen with voting options. The task (aiming handgun against a moving target and shooting or counting successful hits of another shooter, respectively) generated well-defined and highly repeatable psychomotoric load, contrary to too physically (e.g., stationary bicycle) or too mentally (e.g., mathematical quizzes) oriented tasks as described in [17]. This kind of task has been chosen to avoid inequality of physical or mental skills of subjects.

To exclude possibilities of biasing the results by the subjects' memories, the order of the two tests has been swapped for half of the subjects. Also, during the second session, all samples and their options have been randomized, minimizing the possibility to memorize the samples. Also, the correct answers have never been disclosed to the subjects.

Each sample output value can be classified as a success (a subject voted for a correct option) or a failure (a subject voted for an incorrect option), which means that there is a possibility for subjects to vote randomly for given options and there always will be a 20% chance of guessing the correct answer [18]. Models based on Bernoulli trials and the underlying binomial distribution are usually used. The probability of success is specified by the parameter R. The intelligibility R is defined then as a maximum likelihood estimate (E) of R for any group of N trials resulting in S successes, which aligns well with intuition [19]:

$$R = E = \frac{S}{N} \tag{1}$$

Sometimes, to compensate for random results as described above, the approach described in [20] is used. Then, for five voting options case, the compensated intelligibility value is calculated:

$$R_{comp} = \frac{5}{4}(E - \frac{1}{5}) \tag{2}$$

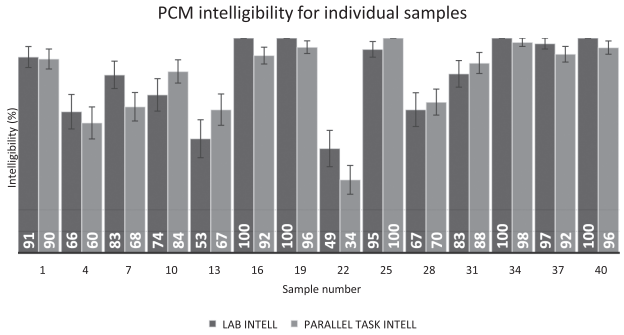


Fig. 2. Intelligibility scores of PCM samples without (left bars) and with (right bars) a parallel task. Error bars show the standard deviations of the arithmetical mean.

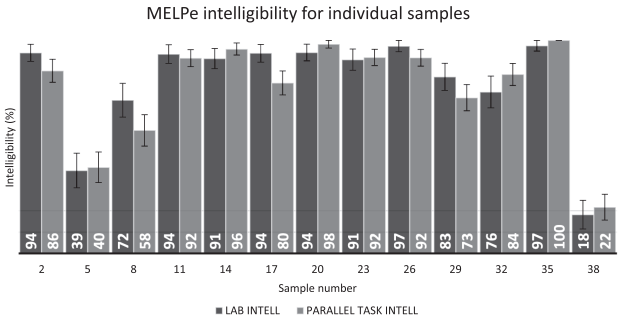


Fig. 3. Intelligibility scores of MELPe samples without (left bars) and with (right bars) a parallel task. Error bars show the standard deviations of the arithmetical mean.

In this case, the first option (1) has been used to express the measured results, and the compensation has not been performed.

The voting process has been realized in the following way. First, the sample has been played, which has been followed by a short beep (440 Hz, 20 dB SPL(A), 300 ms) indicating the start of the voting period. During this period, which lasts 5 seconds, the subjects had to choose their preferred answer. Then the next sample has been played, and so on.

4 DATA ANALYSIS AND RESULTS

In this section, the results of the intelligibility tests both for laboratory and parallel task conditions are presented. As the test arrangement allowed for missing votes (no option selected for the given sample), missing votes haven't been been considered, so the total number of votes is not constant.

In the following graphs, intelligibility scores for individual samples and average values for all coders and narrators are presented. Left columns indicate speech intelligibility without and right columns with parallel task implementation.

In Fig. 2, intelligibility scores in percentages without and with parallel task scores for samples coded by PCM coder are shown. In most of the cases, the intelligibility without parallel task is higher than the intelligibility with parallel task. This drop is not statistically significant on commonly chosen confidentiality levels ($\alpha = 0.05$).

Similarly, Fig. 3 depicts intelligibility results for the MELPe codec (with no background noise) for individual samples. Seven samples (out of all 13) indicate, similarly to the English language experiment [2], an increase of intelligibility measured during the parallel task execution. However, none of them is statistically significant for the given number of subjects and chosen confidence level ($\alpha = 0.05$). The most interesting sample (No. 32) shows an intelligibility increase from initially 76% (without parallel task) to 84% (with the parallel task).

With the same principle, MELPe/HMMWV codec intelligibility influenced by background noise at 0dB SNR is shown in Fig. 4. Two samples (No. 21 and 27) show statistically significant ($\alpha = 0.05$) intelligibility decrease caused

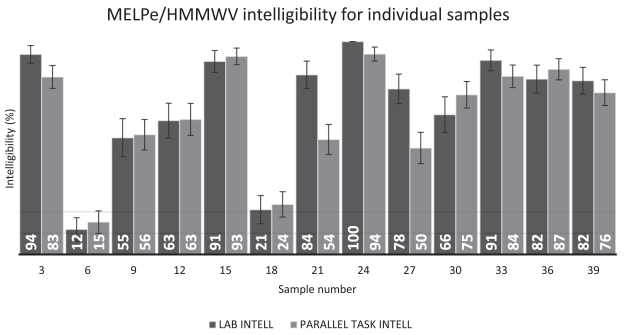


Fig. 4. Intelligibility scores of MELPe/High Mobility Multipurpose Wheeled Vehicle (HMMWV) samples without (left bars) and with (right bars) a parallel task. Error bars show the standard deviations of the arithmetical mean.

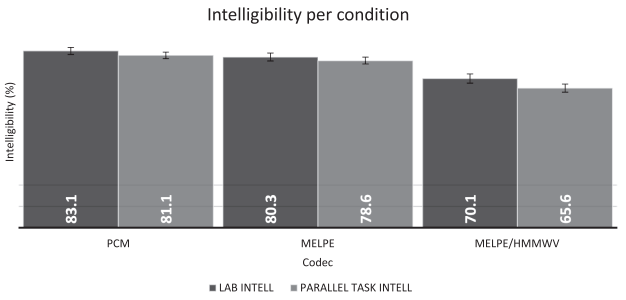


Fig. 5. Average intelligibility scores per condition without (left bars) and with (right bars) a parallel task. Error bars show the standard deviations of the arithmetical mean.

by parallel task introduction, the most influenced sample is No. 21 (from 84% to 54%–).

Fig. 5 includes average intelligibility values for all used codecs. The first two bars show the average intelligibility of all samples with PCM condition (over all 14 samples shown in Fig. 2). The drop of approximate 2% caused by parallel task introduction is not statistically significant ($\alpha = 0.05$). The second two bars show the average MELPe condition intelligibility, which indicates a statistically insignificant drop of 1.7% caused by the parallel task introduction. Finally, the overall MELPe/HMMWV condition intelligibility (last two bards) is reduced by 4.5%, which is, however, not statistically significant on the chosen confidence level ($\alpha = 0.05$).

Table 2. Statistical results of intelligibility using various combinations of codecs and noise (%).

	PCM	MELPe	MELPe/HMMWV
LAB INTELL ± STD (%)	83.1 ± 1.68	80.3 ± 1.88	70.1 ± 2.18
PAR.T. INTELL ± STD (%)	81.1 ± 1.52	78.6 ± 1.63	65.6 ± 1.91

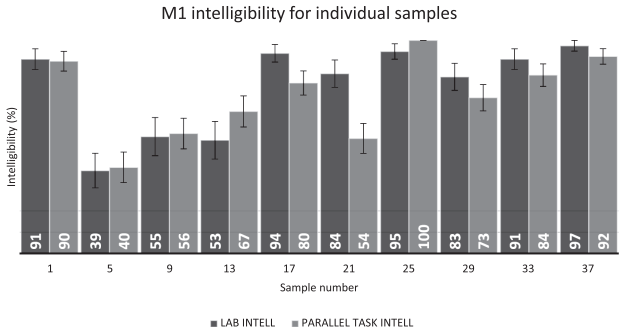


Fig. 6. Intelligibility scores of samples by the narrator M1 without (left bars) and with (right bars) a parallel task. Error bars show the standard deviations of the arithmetical mean.

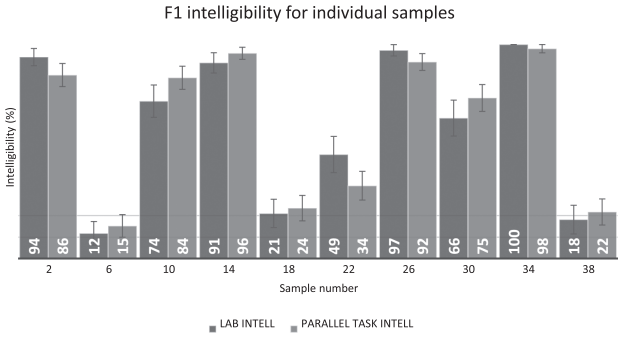


Fig. 7. Intelligibility scores of samples by the narrator F1 without (left bars) and with (right bars) a parallel task. Error bars show the standard deviations of the arithmetical mean.

In all cases, the average intelligibility without a parallel task is higher than with parallel task. The highest intelligibility values are obtained using PCM codec and the lowest values are obtained using the MELPe codec with HMMWV noise. The biggest intelligibility drop has been noticed in samples coded by MELPe with HMMWV background noise (about 4.5%), however, it is still not statistically significant. Statistical intelligibility results using various combinations of codecs and noise with their average standard deviations are presented in Table 2.

As four different narrators (two male and two female) were used to record the clean (studio quality) speech samples, per narrator analyses have also been performed. Next, individual and average values of intelligibility for different narrators are shown.

Fig. 6 shows intelligibility scores per sample for the narrator M1.

Results for the narrator F1 are shown in Fig. 7.

Next, results for the narrator M2 are presented in Fig. 8.

Finally, the results for the narrators F2 are presented. They are shown in Fig. 9.

Finally, average intelligibility values for all narrators are presented in Fig. 10.

After the analysis of the narrator-dependent results, the following conclusions can be made:

- Among all four narrators, the narrator M2 has the most significant drop between laboratory conditions and with parallel task deployment (5.2%).

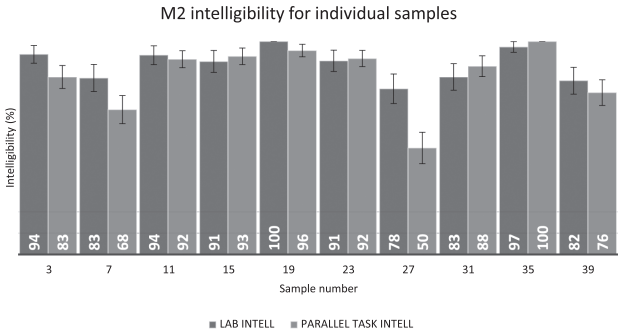


Fig. 8. Intelligibility scores of samples by the narrator M2 without (left bars) and with (right bars) a parallel task. Error bars show the standard deviations of the arithmetical mean.

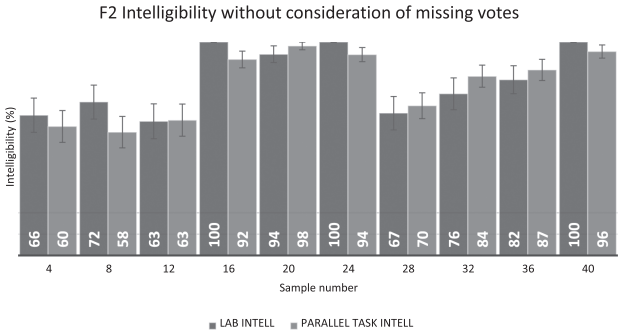


Fig. 9. Intelligibility scores of samples by the narrator F2 without (left bars) and with (right bars) a parallel task. Error bars show the standard deviations of the arithmetical mean.

Table 3. Statistical results of intelligibility for various narrators (%).

	M1	F1	M2	F2
LAB INTELL ± STD (%)	78.8 ± 2.20	62.1 ± 2.61	89.1 ± 1.66	82.1 ± 2.06
PAR.T. INTELL ± STD (%)	73.8 ± 2.00	62.5 ± 2.22	83.9 ± 1.68	80.9 ± 1.79

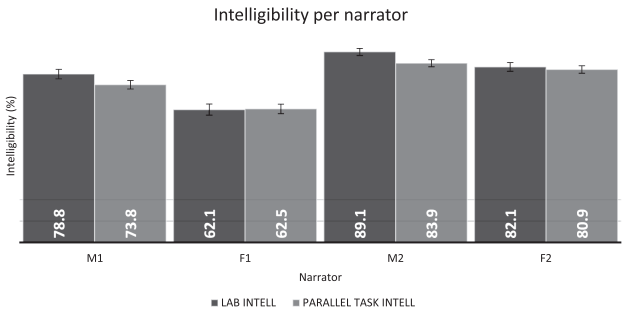


Fig. 10. Average intelligibility scores per narrator without (left bars) and with (right bars) a parallel task. Error bars show the standard deviations of the arithmetical mean.

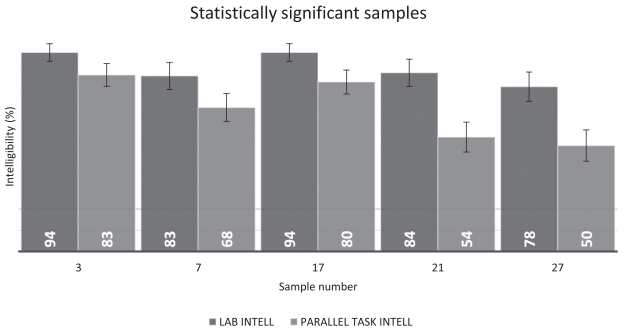


Fig. 11. Statistically significant samples for the tests without (left bars) and with (right bars) a parallel task. Error bars show the standard deviations of the arithmetical mean.

Table 4. The percentage of correct answers of counterintuitive samples both for intelligibility without and with parallel task.

SAMPLE	LAB INTELL (%)	PAR.T. INTELL (%)
6	11.8	15.2
9	54.8	56.3
13	53.1	66.7
20	94.3	98.2
28	66.7	70.2
30	65.6	75.0
32	75.8	84.0
35	97.4	100.0
36	82.4	87.0
38	18.2	21.7

- Intelligibility values for the narrator F1 were almost unchanged with an insignificant lead of intelligibility with a parallel task deployment.
- Samples recorded by male narrators have a slightly higher decrease of intelligibility after parallel task introduction, while intelligibility changes in samples recorded by female narrators are almost unnoticeable.
- The average intelligibility value of male narrators is higher than the average female intelligibility value.

Table 3 includes statistical intelligibility results with their deviations for every narrator.

In general, the expectations about results have been justified: with the parallel task, intelligibility decreases with increased load. In some cases (e.g., samples 6, 10, 13, etc.), the intelligibility with the parallel task was higher than without it. However, those counter-intuitive differences are not statistically significant on the 0.05-significance level—see the displayed standard deviation of the arithmetical mean (Intelligibility STD).

In Figs. 2-10, error bars show the standard deviations of the arithmetical mean (= Intelligibility STD), corresponding to approximately one-half of the 95% confidence interval. As it was mentioned above, in some cases, intelligibility was counterintuitive. Table 4 includes intelligibility results of counterintuitive samples.

4.1 Statistically Significant Values

Fig. 11 shows the statistically significant samples. This values represent the cases where the intelligibility without parallel task is significantly higher than the intelligibility with a parallel task (including the standard deviation of the arithmetical mean).

Table 5. χ^2 values per sample.

Sample number	3	7	17	21	27
χ^2	4.74	4.94	6.47	11.57	7.75

Table 6. χ^2 values per condition.

Condition	PCM	MELPe	MELPe/HMMWV
χ^2	24.8	33.48	31.23

To eliminate the type 1 error in the results, statistical tests were performed to reject the null hypothesis. For this purpose, the χ^2 test was applied as it was described in [18].

Tables 5 and 6 show the χ^2 values per sample for statistically significant results and per condition.

The function of the cumulative distribution is presented in [21], [22], and [23], and since all χ^2 values were higher than recommended value of 3.841, the null-hypothesis can be rejected excluding the type 1 error probability.

These results once more prove the fact that the intelligibility testing with an implementation of a parallel task has more realistic results than the testing in the laboratory conditions.

5 DISCUSSION

The results of our experiment indicate that the speech intelligibility decreases in the case of psychomotor parallel task testing for the given coders (PCM, MELPe) and noise conditions. The experiment also shows MELPe coder performance for communication in the Czech language (no Slavic language has been tested during MELPe development phase, as stated in [15]). Even though more test subjects would be needed to provide statistically significant differences on the 95% (two sigma) confidence level, the following observations can be made already on the achieved 67% (one sigma) confidence level:

- The amount of intelligibility drop with the parallel task is higher for samples with background noise than for noiseless samples. Therefore, we suggest that speech samples with background noise should be tested with the parallel task to bring the results closer to real communication scenarios as the laboratory test results not deploying a parallel task may lead to unnaturally high (optimistic) intelligibility results.
- The amount of intelligibility drop with the parallel task is different for male and female voices. Female voices' intelligibility drop caused by parallel task introduction in our case is much less than for male voices. A more distant position of the female voice in frequency spectra from that of tested noise may explain this observation but given the amount of data available it remains a hypothesis only.

6 CONCLUSION AND FUTURE WORK

This research justifies the necessity of implementation of a parallel task for performing subjective speech intelligibility tests. The two tests which were performed according to ETSI recommendation showed that speech intelligibility drops with the addition of psychomotor tasks, which brings test conditions closer to a real environment, thus making results more realistic.

For the testing purposes, new Czech samples were recorded using two male and two female narrators. Forty-five subjects (in the first part) and 70 subjects (in the second part) of Czech nationality participated in intelligibility tests in conditions of the laboratory environment and with the deployment of the parallel task. Different background noises and coders were used in the tests.

The results show the following:

- The most stable tested condition was MELPe with overall intelligibility values of 80.3% and 78.6% for laboratory conditions and with parallel task, respectively. On the contrary, the biggest decrease of intelligibility was noticed in MELPe/HMMWV condition (70.1% and 65.6%).
- Even though the overall intelligibility of samples recorded by male narrators was higher than those from female narrators, they also had a bigger intelligibility drop between two tests. However, this results cannot be applied to other studies as well. The reason is that the number of the narrators (two males and two females) is too small to consider the results as significant.
- In some cases, the intelligibility value in the test with the parallel task was higher than in the test with laboratory conditions, similarly to results reported in the English language. However, these differences were not statistically significant for the given number of test subjects.

In the future, it is planned to continue the study with different stimuli and parallel task types to mimic other real scenarios of communication equipment usage.

7 ACKNOWLEDGMENT

This study was supported by Internal Grant of Czech Technical University under the Number SGS17/191/OHK3/3T/13. The authors would like to thank mesaqin.com s.r.o. (Ltd.) for providing the samples, test premises, and test subjects for this study.

8 REFERENCES

[1] H. J. M. Steeneken, "The Measurement of Speech" (2002).

[2] H. Avetisyan, T. Drábek, and J. Holub, "Low Bit-Rate Coded Speech Intelligibility Tested With Parallel Task," *Acta Acustica united with Acustica*, vol. 104, no. 4, pp. 678–684 (2018 Jul.), doi: 10.3813/AAA.919207. https://doi.org/10.3813/AAA.919207.

[3] D. A. Hicks, T. Letowski, and M. D. Rao, "A Comparison of Speech Intelligibility Results Between the Call-sign Acquisition Test and the Modified Rhyme Test," presented at the *117th Convention of the Audio Engineering Society* (2004 Oct.), convention paper 6285.

[4] B. Kollmeier, A. Warzybok, S. Hochmuth, M. A. Zokoll, V. Uslar, T. Brand, and K. C. Wagener, "The Multilingual Matrix Test: Principles, Applications, and Comparison Across Languages: A Review," *International Journal of Audiology*, vol. 54, no. sup2, pp. 3–16 (2015 May), doi: 10.3109/14992027.2015.1020971. https://doi.org/10.3109/14992027.2015.1020971.

[5] J. A. Beracoechea, S. Torres-Guijarro, L. García, F. J. Casajús-Quirós, and L. Ortiz, "Subjective Intelligibility Evaluation in Multiple-Talker Situation for Virtual Acoustic Opening-Based Audio Environments," *J. Audio Eng. Soc.*, vol. 56, no. 5, pp. 339–356 (2008 May).

[6] ITU-T Rec. P.807, "Subjective Test Methodology for Assessing Speech Intelligibility," *ITU-T Recommendation* (2016).

[7] ETSI TR 103 503, "Speech and Multimedia Transmission Quality (STQ); Procedures for Multimedia Transmission Quality Testing With Parallel Task Including Subjective Testing," *ETSI*, pp. 1–17 (2018).

[8] V. Durin and L. Gros, "Measuring Speech Quality Impact on Tasks Performance," *Proc. Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2074–2077 (2008).

[9] A. Sarampalis, S. Kalluri, B. Edwards, and E. Hafter, "Objective Measures of Listening Effort: Effects of Background Noise and Noise Reduction," *Journal of Speech, Language, and Hearing Research*, vol. 52, no. 5, pp. 1230–1240 (2009 Oct.), doi: 10.1044/1092-4388(2009/08-0111.

[10] J. Holub, J. Sula, L. Soares, and O. Slavata, "Subjective Audio Quality Testing, With Tasting and Car Driving as Parallel Tasks," *IEEE Access*, vol. 6, pp. 60769–60775 (2018), doi: 10.1109/ACCESS.2018.2873568.

[11] H. Avetisyan and J. Holub, "Subjective Speech Quality Measurement With and Without Parallel Task: Laboratory Test Results Comparison," *PLOS ONE*, vol. 13, no. 7, p. e0199787 (2018 Jul.), doi: 10.1371/journal.pone.0199787.

[12] ITU-T Rec. P.800, "Methods for Subjective Determination of Transmission Quality," *ITU-T Recommendation* (1996).

[13] D. Tihelka and J. Matoušek, "The Design of Czech Language Formal Listening Tests for the Evaluation of TTS Dystems" (2004).

[14] ITU-T Rec. G.711, "Pulse Code Modulation (PCM) of Voice Frequencies," *International Telecommunication Union, Recommendation G.711*, vol. Series G (1993).

[15] M. Street and J. S Collura, "Test and Selection of the Future NATO Narrow Band Voice Coder" (2001).

[16] A. Al-Noori and P. Duncan, "Robust Speaker Recognition in Noisy Conditions by Means of Online Training with Noise Profiles," *J. Audio Eng. Soc.*, vol. 67, no. 4, pp. 174–189 (2019 Apr.), doi: 10.17743/jaes.2019.0004.

[17] A. E. Hill, B. J. Davidson, and D. G. Theodoros, "The Performance of Standardized Patients in Portraying Clinical Scenarios in Speech-Language Therapy," *International Journal of Language & Communication Disorders*, vol. 48, no. 6, pp. 613–624 (2013 Nov.), doi: 10.1111/1460-6984.12034.

[18] S. D. Voran, "Speech Codec Intelligibility Testing in Support of Mission-Critical Voice Applications for LTE" (2015 Sep.).

[19] N. L. Johnson, A. W. Kemp, and S. Kotz, *Univariate Discrete Distributions,* 3rd ed. (Wiley-Interscience, 2005).

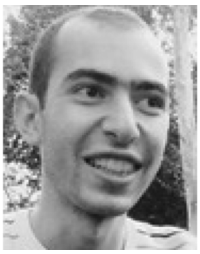
[20] ANSI/ASA, *ANSI/ASA S3.2-2009 Method for Measuring the Intelligibility of Speech over Communication Systems* (Standards Secretariat c/o Acoustical Society of America, New York, New York, 2009).

[21] E. L. Crow, F. Davis, and M. Maxfield, *Statistics Manual* (Dover, New York, New York, 1960).

[22] A. Mood, F. Graybill, and D. Boes, *Introduction to the Theory of Statistics* (McGraw-Hill, New York, New York, 1974).

[23] D. Borowiak, R. V. Hogg, and E. A. Tanis, "Probability and Statistical Inference," *Technometrics*, vol. 31, no. 3, p. 391 (1989 Aug.), doi: 10.2307/3556162.

THE AUTHORS



Hakob Avetisyan

Jan Holub

Oldřich Slavata

Hakob Avetisyan received his MSc degree in Communication Systems from the State Engineering University of Armenia in 2014. He is currently working toward the PhD degree with the Department of Measurement, Faculty of Electrical Engineering, Czech Technical University in Prague.

•

Jan Holub received his PhD degree from the Czech Technical University in Prague in 1999 and became a professor in 2016. He is currently the head of the Department of Measurement, Faculty of Electrical Engineering, Czech Technical University in Prague.

•

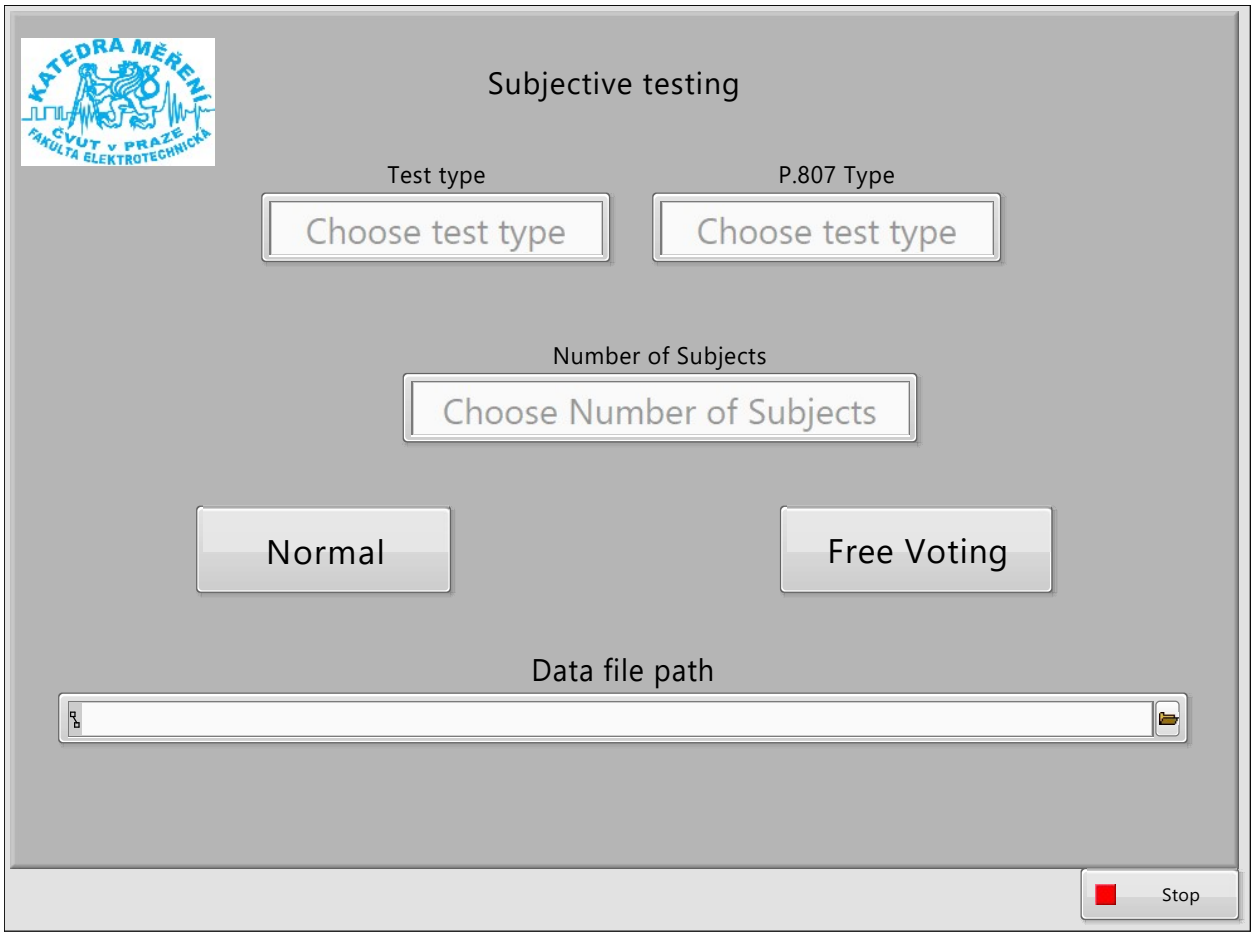
Oldřich Slavata received his PhD degree from the Czech Technical University in Prague in 2017. He is currently working as an assistant professor in the Department of Measurement, Faculty of Electrical Engineering, Czech Technical University in Prague.

7.2.1 Supplementary material

The following materials demonstrate the front panel and block-diagram of newly designed software in LabVIEW environment. For the optimal demonstration in the thesis work, only the main file screenshots are provided. The original program files will be added separately, in an archive.

Main.vi
G:\My Drive\LabView Data\Subjective testing\Main.vi
Last modified on 15-Jan-21 at 12:38
Printed on 19-Jan-21 at 15:07

Main.vi

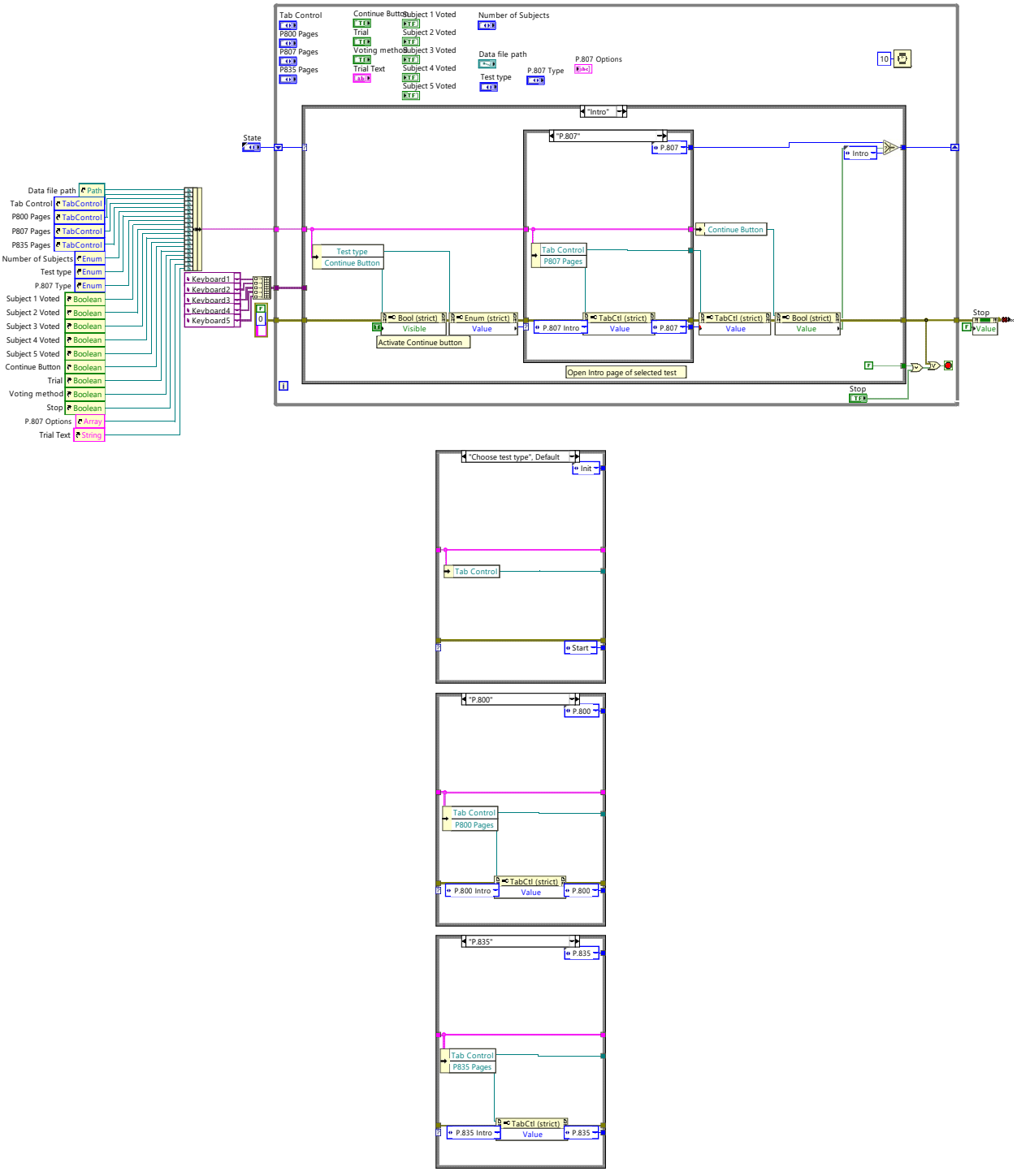


Main.vi

G:\My Drive\LabView Data\Subjective testing\Main.vi

Last modified on 15-Jan-21 at 12:38

Printed on 19-Jan-21 at 15:07

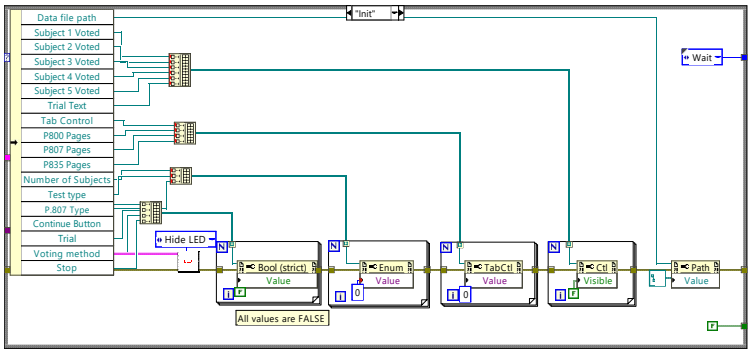


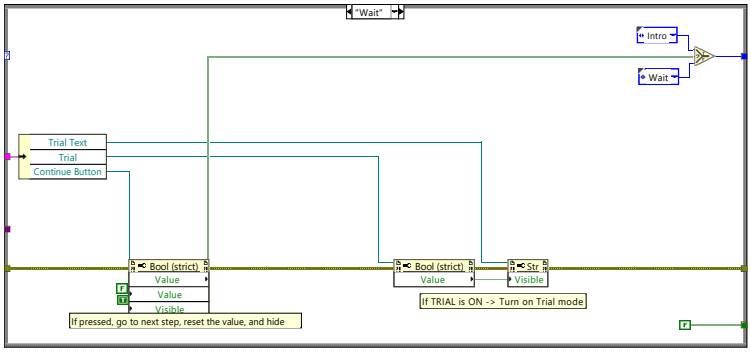
Main.vi

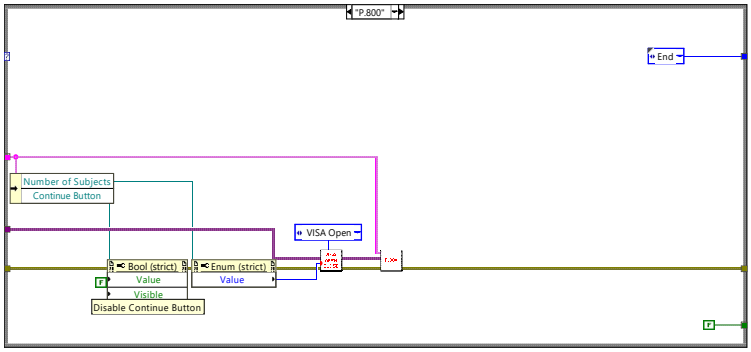
G:\My Drive\LabView Data\Subjective testing\Main.vi

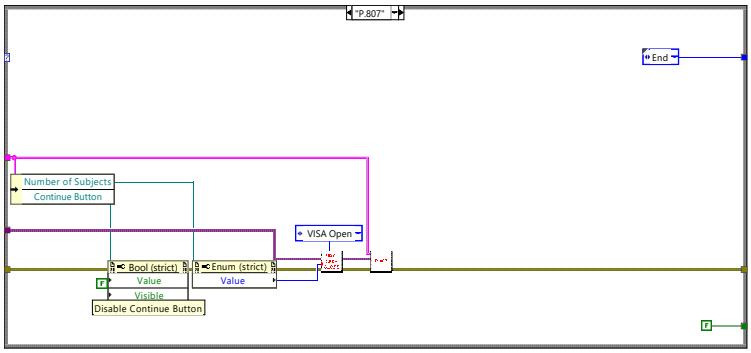
Last modified on 15-Jan-21 at 12:38

Printed on 19-Jan-21 at 15:07

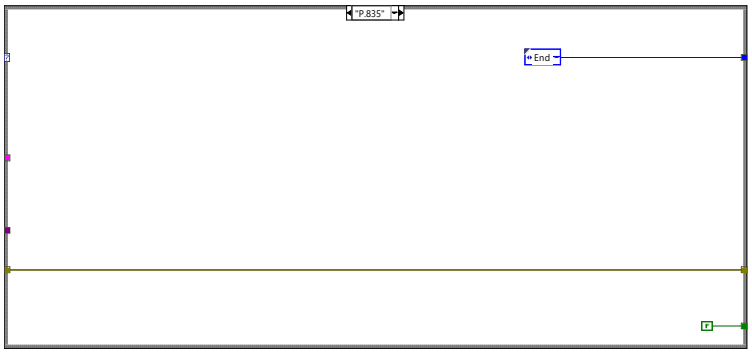


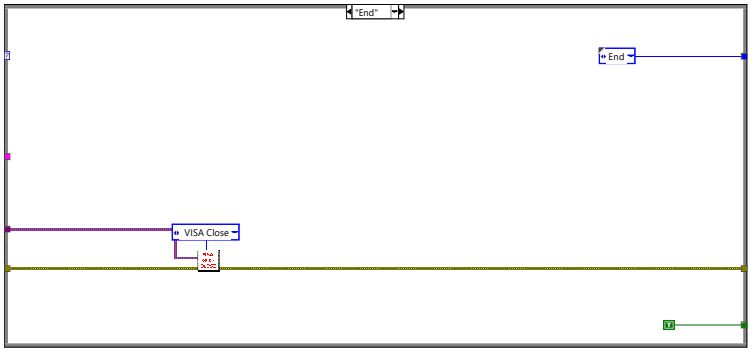






Main.vi
G:\My Drive\LabView Data\Subjective testing\Main.vi
Last modified on 15-Jan-21 at 12:38
Printed on 19-Jan-21 at 15:07





Chapter 8

Conclusion

8.1 Summary of the thesis and hypotheses verification

This thesis was proposed to deal with the subjective speech quality and speech intelligibility tests, investigate existing testing methods, record their main features, as well as advantages and disadvantages. The work also compared different tests in terms of various parameters and provided a modern solution for existing subjective testing methods.

In the beginning, four different tests were provided based on ITU-T P.835 recommendation. This was made to confirm the repeatability of the results among the tests provided in different laboratories. The results showed that the tests were highly correlated, which confirms that the test rules and requirements were strictly followed. Thus, the **hypothesis 1**, stating that "subjective tests provided in different laboratories are repeatable if test requirements are strictly followed" **was confirmed**.

Next, an analysis of over 16 million live call records over VoIP telecommunications networks was performed to verify the importance of providing subjective speech quality and speech intelligibility tests in communication systems. It turns out that for lower qualities cause longer call durations. This may contradict common expectations; however, this is a result of the subjects repeating themselves during the calls since the lower intelligibility of their words. The **hypothesis 2 was confirmed** with the addition that besides higher qualities, lower qualities also cause longer call durations.

In the next part, new speech intelligibility tests were proposed deploying a newly introduced hybrid parallel task. It deployed a laser-shooting simulator consisting of the roles of a "shooter" and "counters", which were dynamically randomly assigned every 40 seconds. In certain samples, the intelligibility values were counterintuitive, which proved that tests provided in laboratory conditions cannot be considered as absolutely correct. This way, the formulated **hypothesis (3) was rejected** for specific samples, where the intelligibility values for tests with parallel task were higher than in the tests without parallel task.

Using the same parallel task methodology, new sets of tests were provided to verify its role in the subjective speech quality measurement. The tests were performed according to ITU-T P.835 methodology. The subjects' task was to assess three values: speech quality, noise annoyance, and overall quality of the samples. The results indicated subjects' lack of concentration due to the parallel task activity, since voting mistakes were done during the tests. The **hypothesis 4 was rejected**, since the voting result differences between laboratory and parallel task conditions were not distributed evenly.

The next important move was to implement the parallel task methodology in the speech intelligibility tests in the Czech language. For this purpose, 40 new Czech samples were recorded using two male and two female narrators in conformance to ITU-T P.800. A new software running on LabVIEW programming environment was designed according to ITU-T P.807 recommendation. As before, this tests also had counterintuitive results. However, those results were statistically insignificant since they were in ranges of standard deviations of arithmetical mean. The only significant results were the ones where the intelligibility in the laboratory conditions was higher than during the parallel task.

8.1.1 Main Contributions

The main contributions in the field mentioned in this thesis are as follows:

- This thesis analyzes the existing subjective speech quality and speech intelligibility testing standards, requirements, and procedures.
- Various tests provided in laboratory conditions have been compared in terms of repeatability, using correlations, pairwise analysis, and RMSE* (Root Mean Squared Error with the suppressed influence of subjective testing uncertainty) analysis.
- The significance of the provided tests was evaluated, proving that the call duration is dependent on the quality of calls – the higher is the quality, the long calls last.
- A novel method for providing subjective speech quality and speech intelligibility tests was proposed, which is more suitable for such tests in terms of simulation of a real-life environment. A new hybrid parallel-task technique was implemented in the current subjective test standards. This type of task is determined to occupy subjects' minds during the testing process avoiding unequal conditions among the subjects (mentally and physically). It turned out that some test samples had counterintuitive values, which is proof that the parallel task better simulates real-life situations.
- A new recommendation was submitted to the European Telecommunications Standards Institute (ETSI) and approved by the number ETSI TR 103 503 [55].

- New software in the LabVIEW programming environment was developed to make the testing process more comfortable and more precise. The software corresponds to existing subjective speech quality and speech intelligibility testing recommendations.
- The proposed method and the developed software were used in subjective speech intelligibility tests in the Czech language.

8.1.2 Author Publications List

All the achieved results were presented in different conferences, as well as published in impacted journals, which are indexed in Scopus and/or Web of Science lists.

Journal Publications Presented in the Thesis

- Holub, J.; Avetisyan, H.; Isabelle, S. "Subjective speech quality measurement repeatability: comparison of laboratory test results." in International Journal of Speech Technology. 2017, 20(1), 69-74. ISSN 1381-2416. https://doi.org/10.1007/s10772-016-9389-6

  Author share: Holub, J. - 33.5%; Avetisyan, H. - 33.5%, Isabelle, S. - 33%

- Holub, J.; Wallbaum, M.; Smith, N.; Avetisyan, H. "Analysis of the Dependency of Call Duration on the Quality of VoIP Calls" in IEEE Wireless Communications Letters. 2018, 7(4), 638-641. ISSN 2162-2337. https://doi.org/10.1109/LWC.2018.2806442

  Author share: Holub, J. - 25%; Wallbaum, M. - 25%; Smith, N. - 25%; Avetisyan, H. - 25%,

- Avetisyan, H.; Drábek, T.; Holub, J. "Low Bit-rate Coded Speech Intelligibility Tested with Parallel Task." in ACTA ACUSTICA UNITED WITH ACUSTICA. 2018, 104(4), 678-684. ISSN 1610-1928. https://doi.org/10.3813/AAA.919207

  Author share: Avetisyan, H. - 35%; Drábek, T. - 20%; Holub, J. - 45%

- Avetisyan, H.; Holub, J. "Subjective Speech Quality Measurement with and without Parallel Task: Laboratory Test Results Comparison." in PLoS ONE. 2018, 13(7) ISSN 1932-6203. https://doi.org/10.1371/journal.pone.0199787

  Author share: Avetisyan, H. - 60%; Holub, J. - 40%

- Avetisyan, H.; Holub, J.; Slavata, O. "Low Bit-rate Coded Speech Intelligibility Testing in Czech Language using Parallel Task." in Journal of Audio Engineering Society. 2020, ISSN 1549-4950. https://doi.org/10.17743/jaes.2020.0008

  Author share: Avetisyan, H. - 34%; Holub, J. - 33%; Slavata, O. - 33%

Conference Publications

- Drábek, T.; Avetisyan, H. "Prototype of Automated Device for Measurement of a Light Vector." at Proceedings of the 20th International Scientific Student Conference POSTER 2016, Prague, Czech Republic

  Author share: Drábek, T. - 50%; Avetisyan, H. - 50%

- Avetisyan, H.; Bruna, O.; Holub, J. "Overview of existing algorithms for emotion classification. Uncertainties in evaluations of accuracies." at 2016 Joint IMEKO TC1-TC7-TC13 Symposium, Berkeley, USA

  Author share: Avetisyan, H. - 33.3%; Bruna, O. - 33.3%; Holub, J. - 33.3%

- Bruna, O.; Avetisyan, H.; Holub, J. "Emotion models for textual emotion classification." at 2016 Joint IMEKO TC1-TC7-TC13 Symposium, Berkeley, USA

  Author share: Bruna, O. - 33.3%; Avetisyan, H. - 33.3%; Holub, J. - 33.3%

- Avetisyan, H. "Subjective Speech Quality Measurement: Comparison of Laboratory Test Results and Results of Test with Parallel Task" at ETSI Workshop on Multimedia Quality in Virtual, Augmented or other Realities. European Telecommunications Standards Institute, 2017. ETSI STQ Workshops. Sophia Antipolis, France

  Author share: Avetisyan, H. - 100%

- Mayilyan, H.; Poghosyan, S.; Avetisyan, H. "Educational augmented reality systems: Benefits of implementation and government support." at 4th International Conference of the Virtual and Augmented Reality in Education, Budapest, Hungary

  Author share: Mayilyan, H. - 34%; Poghosyan, S. 33%; Avetisyan, H. - 33%

Publications Not Presented in the Thesis

- Holub, J.; Slavata, O.; Avetisyan, H. "ETSI TR 103 503: Speech and Multimedia Transmission Quality (STQ); Procedures for Multimedia Transmission Quality Testing."

8.2 Future work

In the future, it is planned to continue the study with different stimuli and parallel task types to mimic other real-life scenarios of communication equipment usage.

Other interesting testing perspectives are:

- Testing in various languages (German, Chinese, Russian, etc.)

- Testing using various parallel tasks scenarios (e.g. virtual reality tasks, car driving simulations, plane flying simulations, etc.)
- Testing video QoE (based on ITU-T P.910) and listening effort (ITU-T P.85)
- Improving the existing testing software

Bibliography

[1] J. Holub, J. Beerends, and R. Smid, "A dependence between average call duration and voice transmission quality: measurement and applications", *Wireless Telecommunications Symposium*, pp. 75–81, 2004. DOI: 10.1109/WTS.2004.1319562. [Online]. Available: http://ieeexplore.ieee.org/document/1319562.

[2] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality", *ITU-T Recommendation*, 1996. [Online]. Available: https://www.itu.int/rec/T-REC-P.800-199608-I.

[3] ITU-T Rec. P.835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm", *ITU-T Recommendation*, 2003. [Online]. Available: https://www.itu.int/rec/R-REC-P.835/en.

[4] ITU-T Rec. P.807, "Subjective test methodology for assessing speech intelligibility", *ITU-T Recommendation*, 2016. [Online]. Available: https://www.itu.int/rec/T-REC-P.807/en.

[5] ITU-T Rec. P.862, *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*. [Online]. Available: https://www.itu.int/rec/T-REC-P.862 (visited on 09/11/2020).

[6] ITU-T Rec. P.863, "Perceptual objective listening quality assessment", *ITU-T Recommendation*, pp. 1–82, 2014. [Online]. Available: https://www.itu.int/rec/T-REC-P.863.

[7] ITU-T Rec. G.107, "The E-model: a computational model for use in transmission planning", *International Telecommunication Union, Recommendation G.107*, vol. Series G, 2015. [Online]. Available: http://handle.itu.int/11.1002/1000/12505.

[8] A. C. Dalal, "User-Perceived quality assessment of streaming media using reduced feature sets", *ACM Transactions on Internet Technology*, vol. 11, no. 2, pp. 1–32, 2011, ISSN: 15335399. DOI: 10.1145/2049656.2049660. [Online]. Available: https://dl.acm.org/doi/10.1145/2049656.2049660.

[9] ETSI TS 103 106, "Speech and multimedia Transmission Quality (STQ); Speech quality performance in the presence of background noise: Background noise transmission for mobile terminals-objective test methods", Tech. Rep., 2018. [Online]. Available: https://www.etsi.org/deliver/etsi_ts/103100_103199/103106/01.05.01_60/ts_103106v010501p.pdf.

[10] ETSI EG 202 396-3, "Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality performance in the presence of background noise Part 3: Background noise transmission - Objective test methods", *ETSI*, 2011. [Online]. Available: https://www.etsi.org/deliver/etsi_eg/202300_202399/20239603/01.02.01_50/eg_20239603v010201m.pdf.

[11] ANSI/ASA S3.2-2009, "Method for Measuring the Intelligibility of Speech over Communication Systems", New York N.Y., 2009, [Online]. Available: https://www.worldcat.org/title/american-national-standard-method-for-measuring-the-intelligibility-of-speech-over-communication-systems/oclc/122313526.

[12] J. A. Beracoechea, S. Torres-Guijarro, L García, F. J. Casajús-Quirós, and L Ortiz, "Subjective Intelligibility Evaluation in Multiple-Talker Situation for Virtual Acoustic Opening-Based Audio Environments", *J. Audio Eng. Soc*, vol. 56, no. 5, pp. 339–356, 2008. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=14388.

[13] Dynastat Inc., *Summary of Speech Intelligibility Testing Methods*. [Online]. Available: https://www.dynastat.com/SpeechIntelligibility.htm.

[14] D. A. Hicks, T. Letowski, and M. D. Rao, "A Comparison of Speech Intelligibility Results Between the Callsign Acquisition Test and the Modified Rhyme Test", in *Audio Engineering Society Convention 117*, 2004. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=12942.

[15] ITU-R Rec. BS.1116, "Methods for the subjective assessment of small impairments in audio systems", *ITU-R Recommendation*, 2015. [Online]. Available: https://www.itu.int/rec/R-REC-BS.1116/en.

[16] ITU-R Rec. BS.1534, "Method for the subjective assessment of intermediate quality level of audio systems", *ITU-R Recommendation*, 2015. [Online]. Available: https://www.itu.int/rec/R-REC-BS.1534/en.

[17] J. S. Lee, "On designing paired comparison experiments for subjective multimedia quality assessment", *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 564–571, 2014, ISSN: 15209210. DOI: 10.1109/TMM.2013.2292590. [Online]. Available: https://ieeexplore.ieee.org/document/6675876.

[18] N. Schinkel-Bielefeld, N. Lotze, and F. Nagel, "Audio quality evaluation by experienced and inexperienced listeners", *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3246–3246, 2013, ISSN: 0001-4966. DOI: 10.1121/1.4805210. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.4805210.

[19] F. Rumsey, S. Zielinski, R. Kassier, and S. Bech, "Relationships between experienced listener ratings of multichannel audio quality and naïve listener preferences", *The Journal of the Acoustical Society of America*, vol. 117, no. 6, pp. 3832–3840, 2005, ISSN: 0001-4966. DOI: 10.1121/1.1904305. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.1904305.

[20] B. Gardlo, S. Egger, M. Seufert, and R. Schatz, "Crowdsourcing 2.0: Enhancing execution speed and reliability of web-based QoE testing", in *2014 IEEE International Conference on Communications, ICC 2014*, IEEE Computer Society, 2014, pp. 1070–1075, ISBN: 9781479920037. DOI: 10.1109/ICC.2014.6883463. [Online]. Available: https://ieeexplore.ieee.org/document/6883463.

[21] S. Egger-Lampl, J. Redi, T. Hoßfeld, M. Hirth, S. Möller, B. Naderi, C. Keimel, and D. Saupe, "Crowdsourcing quality of experience experiments", in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10264 LNCS, Springer Verlag, 2017, pp. 154–190. DOI: 10.1007/978-3-319-66435-4_7. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-66435-4_7.

[22] F. Ribeiro, D. Florêncio, C. Zhang, and M. Seltzer, "CROWDMOS: An approach for crowdsourcing mean opinion score studies", in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2011, pp. 2416–2419, ISBN: 9781457705397. DOI: 10.1109/ICASSP.2011.5946971. [Online]. Available: https://ieeexplore.ieee.org/document/5946971.

[23] V. Durin and L. Gros, "Measuring speech quality impact on tasks performance", *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2074–2077, 2008, ISSN: 19909772. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2008/i08_2074.html.

[24] A. Serampalis, S. Kalluri, B. Edwards, and E. Hafter, "Objective measures of listening effort in noise", *Journal of Speech, Language, and Hearing Research*, vol. 52, no. October 2009, pp. 1230–1240, 2009, ISSN: 1092-4388. DOI: 10.1044/1092-4388(2009/08-0111). [Online]. Available: https://pubs.asha.org/doi/10.1044/1092-4388%282009/08-0111%29.

[25] G. P. Sonntag, T Portele, and F Haas, "Comparing the comprehensibility of different synthetic voices in a dual task experiment", *Proceedings of the Third Workshop on Speech Synthesis, Jenolan Caves House, Blue Mountains*, pp. 5–10, 1998. [Online]. Available: https://www.isca-speech.org/archive_open/ssw3/ssw3_005.html.

[26] 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.", *Journal of Experimental Psychology: Applied*, vol. 8, no. 1, pp. 6–16, 2002, ISSN: 1076-898X. DOI: 10.1037//1076-898X.8.1.6. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/1076-898X.8.1.6.

[27] D. L. Strayer and W. A. Johnston, "Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular telephone.", *Psychological Science*, vol. 12, no. 6, pp. 462–466, 2001, ISSN: 0956-7976. DOI: 10.1111/1467-9280.00386. [Online]. Available: https://journals.sagepub.com/doi/10.1111/1467-9280.00386.

[28] S. Choi, A. Lotto, D. Lewis, B. Hoover, and P. Stelmachowicz, "Attentional Modulation of Word Recognition by Children in a Dual-Task Paradigm", *Journal of Speech Language and Hearing Research*, vol. 51, no. 4, p. 1042, 2008, ISSN: 1092-4388. DOI: 10.1044/1092-4388(2008/076). arXiv: NIHMS150003. [Online]. Available: http://jslhr.pubs.asha.org/article.aspx?doi=10.1044/1092-4388(2008/076).

[29] C. Kwak and W. Han, "Comparison of Single-Task versus Dual-Task for Listening Effort", *Journal of Audiology and Otology*, 2017, ISSN: 2384-1621. DOI: 10.7874/jao.2017.00136. [Online]. Available: http://ejao.org/journal/view.php?doi=10.7874/jao.2017.00136.

[30] L. Gros, N. Chateau, and A. Macé, "Assessing speech quality : a new approach Methodology", 2005. [Online]. Available: https://www.researchgate.net/publication/228421237_Assessing_speech_quality_a_new_approach.

[31] K. S. Helfer, J. Chevalier, and R. L. Freyman, "Aging, spatial cues, and single- versus dual-task performance in competing speech perception.", *The Journal of the Acoustical Society of America*, vol. 128, no. 6, pp. 3625–33, 2010, ISSN: 1520-8524. DOI: 10.1121/1.3502462. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.3502462.

[32] K. Bunton and C. K. Keintz, "The Use of a Dual-Task Paradigm for Assessing Speech Intelligibility in Clients with Parkinson Disease", *Journal of medical speech-language pathology*, vol. 16, no. 3, pp. 141–155, 2008, ISSN: 1065-1438. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3104935.

[33] J. Holub, J. Sula, L. Soares, and O. Slavata, "Subjective audio quality testing, with tasting and car driving as parallel tasks", *IEEE Access*, vol. 6, pp. 1–1, 2018, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2018.2873568. [Online]. Available: https://ieeexplore.ieee.org/document/8481357/.

[34] J. Holub, H. Avetisyan, and S. Isabelle, "Subjective speech quality measurement repeatability: comparison of laboratory test results", *International Journal of Speech Technology*, vol. 20, no. 1, pp. 69–74, 2017, ISSN: 1381-2416. DOI: 10.1007/s10772-016-9389-6. [Online]. Available: http://link.springer.com/10.1007/s10772-016-9389-6.

[35] Z. Yang and Z. Niu, "Load Balancing by Dynamic Base Station Relay Station Associations in Cellular Networks", *IEEE Wireless Communications Letters*, vol. 2, no. 2, pp. 155–158, 2013, ISSN: 2162-2337. DOI: 10.1109/WCL.2012.121812.120797. [Online]. Available: http://ieeexplore.ieee.org/document/6399495/.

[36] V. D. Blondel, A. Decuyper, G. Krings, D Chakraborty, S Mukherjea, A. Nanavati, A Joshi, M. González, V Colizza, S Moritz, S Zhao, Y Zheng, M Macy, D Roy, and M Alstyne, "A survey of results on mobile phone datasets analysis", *EPJ Data Science*, vol. 4, no. 1, p. 10, 2015, ISSN: 2193-1127. DOI: 10.1140/epjds/s13688-015-0046-0. [Online]. Available: http://www.epjdatascience.com/content/4/1/10.

[37] Y. Dong, J. Tang, T. Lou, B. Wu, and N. V. Chawla, "How Long Will She Call Me? Distribution, Social Theory and Duration Prediction", in, Springer, Berlin, Heidelberg, 2013, pp. 16–31. DOI: 10.1007/978-3-642-40991-2_2. [Online]. Available: http://link.springer.com/10.1007/978-3-642-40991-2_2.

[38] G. Friebel and P. Seabright, "Do women have longer conversations? Telephone evidence of gendered communication strategies", *Journal of Economic Psychology*, vol. 32, no. 3, pp. 348–356, 2011, ISSN: 01674870. DOI: 10.1016/j.joep.2010.12.008. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S016748701000142X.

[39] D. Goodman and R. Nash, "Subjective quality of the same speech transmission conditions in seven different countries", in *ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 7, Institute of Electrical and Electronics Engineers, 1982, pp. 984–987. DOI: 10.1109/ICASSP.1982.1171565. [Online]. Available: http://ieeexplore.ieee.org/document/1171565/.

[40] J. Kim, V. S. A. Kumar, A. Marathe, G. Pei, S. Saha, and B. S. Subbiah, "Modeling cellular network traffic with mobile call graph constraints", in *Proceedings of the 2011 Winter Simulation Conference (WSC)*, IEEE, 2011, pp. 3165–3177, ISBN: 978-1-4577-2109-0. DOI: 10.1109/WSC.2011.6148015. [Online]. Available: http://ieeexplore.ieee.org/document/6148015/.

[41] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskove, "Mobile Call Graphs: Beyond Power-law and Lognormal Distributions", in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '08, New York, NY, USA: ACM, 2008, pp. 596–604, ISBN: 978-1-60558-193-4. DOI: 10.1145/1401890.1401963. [Online]. Available: http://doi.acm.org/10.1145/1401890.1401963.

[42] ITU-T Rec. G.711, "Pulse code modulation of (PCM) of Voice frequencies", *International Telecommunication Union, Recommendation G.711*, vol. Series G: 1993. [Online]. Available: https://www.itu.int/rec/T-REC-G.711.

[43] ITU-T Rec. G.729, "Coding of speech at 8 kbit/s using conjugate structure algebraic-code-excited linear-prediction (CS-ACELP)", *International Telecommunication Union, Recommendation G.729*, vol. Series G, 2012. [Online]. Available: https://www.itu.int/rec/T-REC-G.729.

[44] ETSI TS 126 071, "Universal Mobile Telecommunications System (UMTS); Mandatory Speech Codec speech processing functions AMR Speech Codec; General Description", *ETSI*, pp. 1–14, 1999. [Online]. Available: https://www.etsi.org/deliver/etsi_ts/126000_126099/126071/13.00.00_60/ts_126071v130000p.pdf.

[45] J. Holub, M. Wallbaum, N. Smith, and H. Avetisyan, "Analysis of the Dependency of Call Duration on the Quality of VoIP Calls", *IEEE Wireless Communications Letters*, vol. 7, no. 4, pp. 638–641, 2018, ISSN: 2162-2337. DOI: 10.1109/LWC.2018.2806442. [Online]. Available: https://ieeexplore.ieee.org/document/8292832/.

[46] D. Navon and D. Gopher, "On the economy of the human-processing system", *Psychological Review*, vol. 86, no. 3, pp. 214–255, 1979, ISSN: 0033295X. DOI: 10.1037/0033-295X.86.3.214. [Online]. Available: https://content.apa.org/record/1979-27744-001.

[47] C. D. Wickens and I. U. A. T. U. E.-P. R. LAB., *Processing Resources in Attention, Dual Task Performance, and Workload Assessment*. Defense Technical Information Center, 1981. [Online]. Available: https://books.google.cz/books?id=yQ2WNwAACAAJ.

[48] H. Avetisyan, T. Drábek, and J. Holub, "Low Bit-rate Coded Speech Intelligibility Tested with Parallel Task", *Acta Acustica united with Acustica*, vol. 104, no. 4, 2018. [Online]. Available: https://www.ingentaconnect.com/content/dav/aaua/2018/00000104/00000004/art00013.

[49] ITU-T Rec. G.722.2, "Wideband coding of speech at around 16 kbit/s using Adaptive Multi-rate Wideband (AMR-WB)", *International Telecommunication Union*, 2003. [Online]. Available: https://www.itu.int/rec/T-REC-G.722.2/en.

[50] ETSI TS 126 445, "Universal Mobile Telecommunications System (UMTS); Codec for Enhanced Voice Services (EVS); Detailed algorithmic description", pp. 0–92, 2016. [Online]. Available: https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx.

[51] 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", *ETSI*, 2008. [Online]. Available: https://www.etsi.org/deliver/etsi_eg/202300_202399/20239601/01.02.02_60/eg_20239601v010202p.pdf.

[52] H. Avetisyan and J. Holub, "Subjective speech quality measurement with and without parallel task: Laboratory test results comparison", *PLOS ONE*, vol. 13, no. 7, G. Kearney, Ed., e0199787, 2018, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0199787. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0199787.

[53] D. Tihelka and J. Matoušek, "The design of Czech language formal listening tests for the evaluation of TTS systems", 2004. [Online]. Available: https://dspace5.zcu.cz/handle/11025/16978.

[54] H. Avetisyan, J. Holub, and O. Slavata, "Low Bit-Rate Coded Speech Intelligibility Testing in Czech Language Using Parallel Task", *Journal of the Audio Engineering Society*, vol. 68, no. 4, pp. 284–291, 2020, ISSN: 15494950. DOI: 10.17743/jaes.2020.0008. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=20734.

[55] ETSI TR 103 503, "Speech and multimedia Transmission Quality (STQ); Procedures for Multimedia Transmission Quality Testing with Parallel Task including Subjective Testing", *ETSI*, pp. 1–17, 2018. [Online]. Available: https://www.etsi.org/deliver/etsi_tr/103500_103599/103503/01.01.01_60/tr_103503v010101p.pdf.