

COVER PAGE

Document downloaded by @DAEL

Sun Nov 29 01:40:40 2026
For personal use

When automatic English translation is provided, only the original document is authentic.
The EAA cannot be held responsible of any translation error

Bibliographical reference

Low Bit-rate Coded Speech Intelligibility Tested with Parallel Task, Hakob Avetisyan, Tomáš Drábek and Jan Holub, Acta Acustica **vol. 104** (Number 4), 2018, pp. 678-684

DOI
https://doi.org/10.3813/AAA.919207

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 radiocommunication 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.

© S. Hirzel Verlag · EAA

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

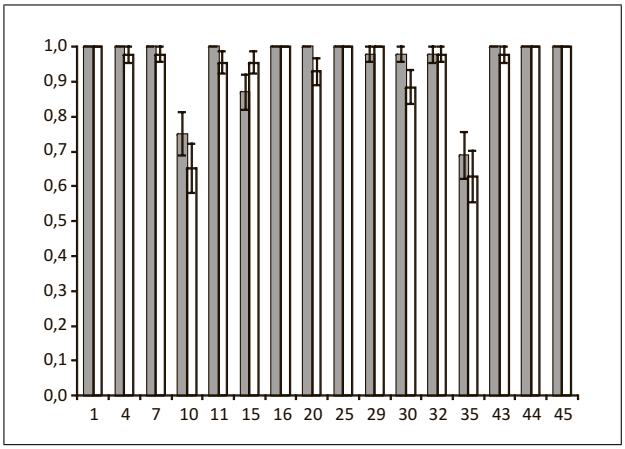


Figure 1. Intelligibility scores of PCM samples without (left bars) and with (right bars) a parallel task, missing votes not considered. Horizontal scale: Sample Number. Vertical Scale: Intelligibility, uncompensated for random correct votes. Minimum 43, median 48 votes per sample.

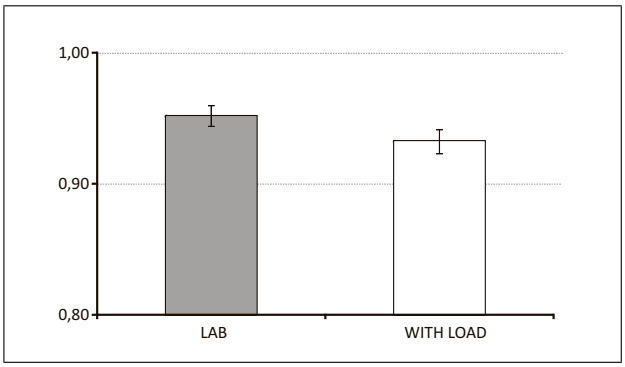


Figure 2. Intelligibility scores of PCM condition without (left bar) and with (right bar) a parallel task, missing votes not considered. Vertical Scale: Intelligibility, uncompensated for random correct votes. Minimum 750 votes per condition.

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

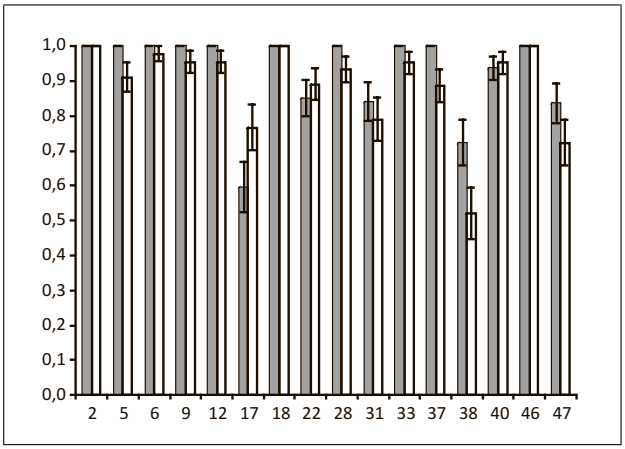


Figure 3. Intelligibility scores of MELPe samples without (left bars) and with (right bars) a parallel task, missing votes not considered. Horizontal scale: Sample Number. Vertical Scale: Intelligibility, uncompensated for random correct votes. Minimum 43, median 48 votes per sample.

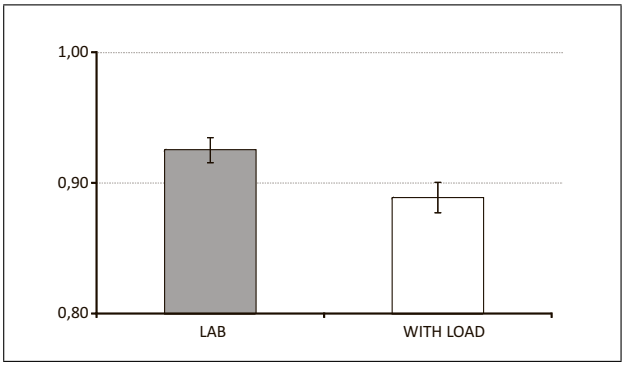


Figure 4. Intelligibility scores of MELPe condition without (left bar) and with (right bar) a parallel task, missing votes not considered. Vertical Scale: Intelligibility, uncompensated for random correct votes. Minimum 750 votes per condition.

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

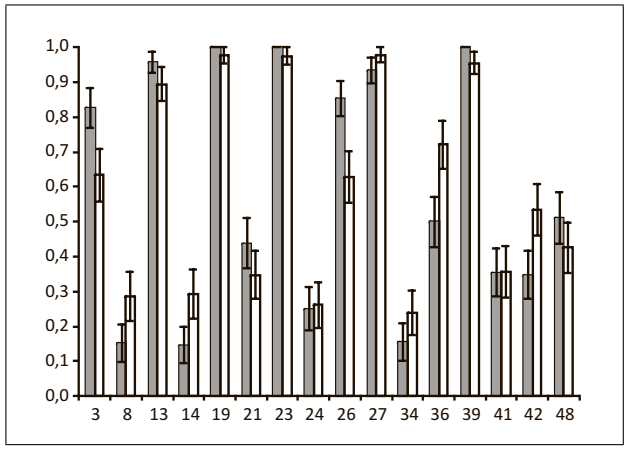


Figure 5. Intelligibility scores of MELPe with 0dB HMMWV noise samples without (left bars) and with (right bars) a parallel task, missing votes not considered. Horizontal scale: Sample Number. Vertical Scale: Intelligibility, uncompensated for random correct votes. Minimum 43, median 48 votes per sample.

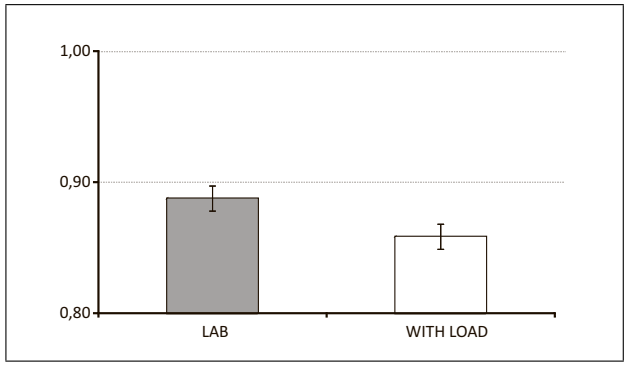


Figure 6. Intelligibility scores of MELPe with 0dB HMMWV noise without (left bar) and with (right bar) a parallel task, missing votes not considered. Vertical Scale: Intelligibility, uncompensated for random correct votes. Minimum 750 votes per condition.

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

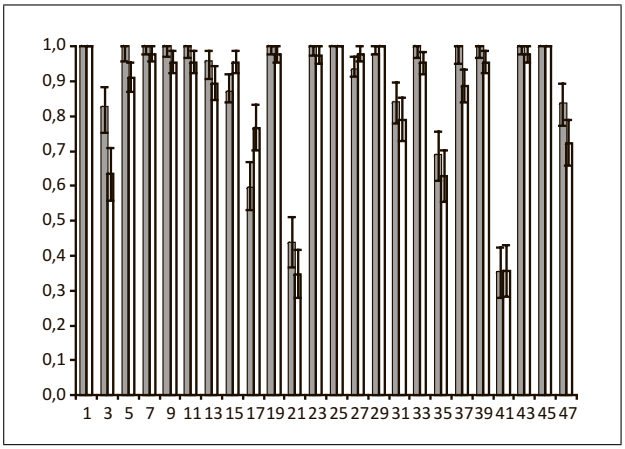


Figure 7. Intelligibility scores of samples recorded by Narrator 1 and processed with codecs and noise without (left bars) and with (right bars) a parallel task, missing votes not considered. Horizontal scale: Sample Number. Vertical Scale: Intelligibility, uncompensated for random correct votes. Minimum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

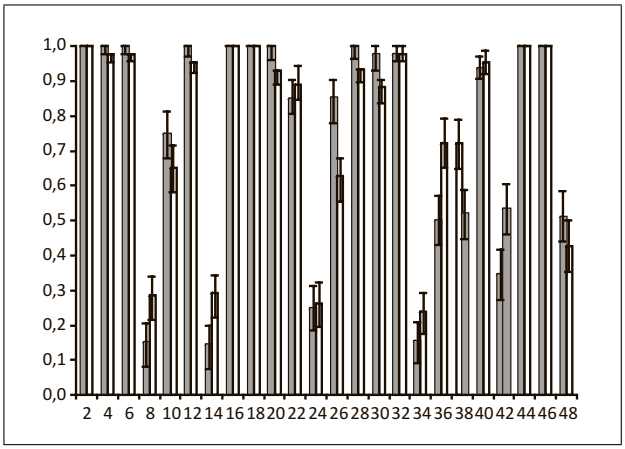


Figure 9. Intelligibility scores of samples recorded by Narrator 2 and processed with codecs and noise without (left bars) and with (right bars) a parallel task, missing votes not considered. Horizontal scale: Sample Number. Vertical Scale: Intelligibility, uncompensated for random correct votes. Minimum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

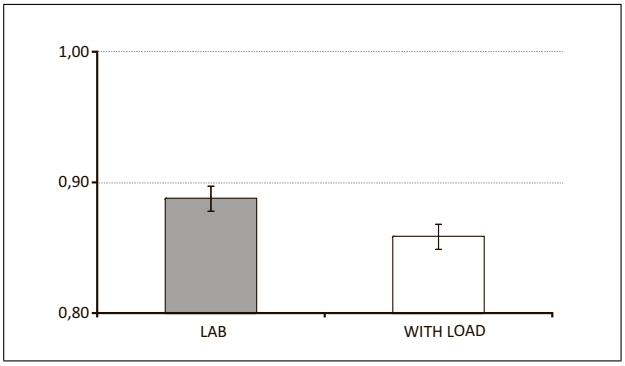


Figure 8. Intelligibility scores of samples recorded by Narrator 1 and processed with codecs and noise without (left bar) and with (right bar) a parallel task, missing votes not considered. Vertical Scale: Intelligibility, uncompensated for random correct votes. Minimum 1100 votes per condition.

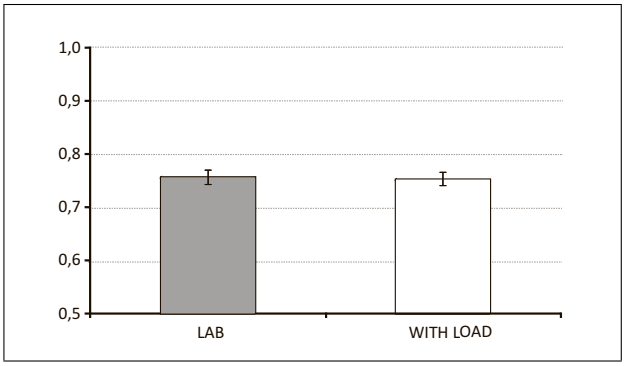


Figure 10. Intelligibility scores of samples recorded by Narrator 2 and processed with codecs and noise without (left bar) and with (right bar) a parallel task, missing votes not considered. Vertical Scale: Intelligibility, uncompensated for random correct votes. Minimum 1100 votes per condition.

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.