PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

A comparison of cognitive performance and listening effort test procedures

Alexander Dickschen[1]; Andreas Liebl[2]; Stefan Bleeck[3]

[1] Fraunhofer-Institute for Building Physics, Stuttgart, Germany

[2] Department of Psychology, HSD Hochschule Döpfer GmbH, Cologne, Germany

[3] Institute of Sound and Vibration Research, University of Southampton, United Kingdom

ABSTRACT

Workers in open-plan offices are often objected to irrelevant speech, which leads to decreased cognitive performance. With acoustical treatments, such as sound masking, the cognitive impact can be reduced, but there is no objective procedure to assess this improvement in real office environments. For in-situ evaluation, a robust test procedure is required.

The serial recall test used in laboratories requires masking levels to be higher than the speech concealed, which seldom is the case in real offices. Therefore, a dual-task paradigm was investigated, which measures latencies for auditory stimuli presented in noise and silence. The method, originating from audiology for listening effort measurement, showed significant differences even for little masked, clearly intelligible speech. The underlying cognitive processes leading to those differences in response times need to be understood.

Experiments with 21 normal hearing participants were conducted to investigate the memory component and the auditory identification process in the listening effort experiment. A comparison to the serial recall test was carried out. The findings suggest an arousal effect at work, with faster replies for higher masking in serial recall and listening effort test. The two tests seem to analyse different cognitive aspects.

Keywords: serial recall, reaction time, listening effort

1. INTRODUCTION

Open-plan offices are the favoured design concept for many offices nowadays due to high spatial efficiency, flexibility and claimed increase exchange of information. The major downside of this concept is the poor acoustical performance, leading to distractions, difficulties to concentrate and fatigue. Overheard conversations of colleagues are the main source of distraction.

In psychology, the effects of fluctuating sounds such as speech on cognitive performance are well researched and termed *irrelevant sound effect* (ISE). They affect especially the short-term memory. The related test procedure is a serial recall of digits (SR), which is a widespread laboratory experiment. The more predominant the fluctuation, the higher the error rate in SR. A model linking decreased performance in serial recall tests to the speech transmission index (STI) can be found in [1]. STI is a technical parameter of speech intelligibility reaching from 0 for no speech intelligibility at all to 1 for perfect intelligibility. The error rate in SR remains high, as long as STI is high (> 0.5). The major transition from full to no impact of irrelevant sounds happens roughly between STI=0.4 and STI=0.3. So, if the transition is of interest, only a narrow band of ΔSTI = 0.1 is relevant. For higher STI, SR shows a ceiling effect. To understand the transition band and impact of highly intelligible sounds, other approaches would be helpful. Short term memory performance is only one of the fundamental cognitive functions currently serving as proxy for overall cognitive performance in the office context. However, there are more domains affected by irrelevant sound, which are harder to analyse in tests, i.e. privacy. Still, the field would benefit from a broader variety of test procedures available.

In audiology, many users of hearing aids complain of fatigue, as the cognitive effort to identify

[1] alexander.dickschen@ibp.fraunhofer.de, https://orcid.org/0000-0002-7001-9642

signals remain high despite amplification of the signal by the hearing aid. For this reason, test procedures to measure *listening effort* are developed. In [2] a procedure is presented to quantify the effort by response time (RT) measurements and significant effects for highly intelligible sounds are found. The procedure is based on digits of a telephone hearing test combined with a simple arithmetic task. This procedure is further investigated in the experiments presented.

It is unclear, which underlying cognitive process in the experiment presented in [2] causes the response times to vary between conditions. In order to gain better understanding, firstly, a replicate of the original experiment was set up. Secondly, two variations were created to thirdly and finally compare the findings to SR results of the same subjects.

2. TEST PROCEDURES AND HYPOTHESIS

A scheme of the original listening effort experiment can be found in Figure 1. The dual task paradigm is based on triplets of digits from a telephone hearing test [3]. To create different listening conditions, the digits are masked by noise in some of the conditions. The difference in *response time* (ΔRT) between Condition A without noise added and Condition B with masking noise serves as indicator, how much the listening effort increases. Four different signal to noise ratios (SNR) were chosen, in line with the original experiment: SNR1=-6 dB, SNR2=-1 dB, SNR3=+4 dB (Condition B in Figure 1) and SNR4=+∞ (that is digits in silence, Condition A in Figure 1). The lower the SNR, the lower the intelligibility of the digits sounded. The spectrum of the masking noise resembles the spectrum of the digits used. The digits were played back with 60 dB(A) over all conditions in all variations and the masking sound adjusted according to the sound condition. Overall, four different tests were used for the investigation:

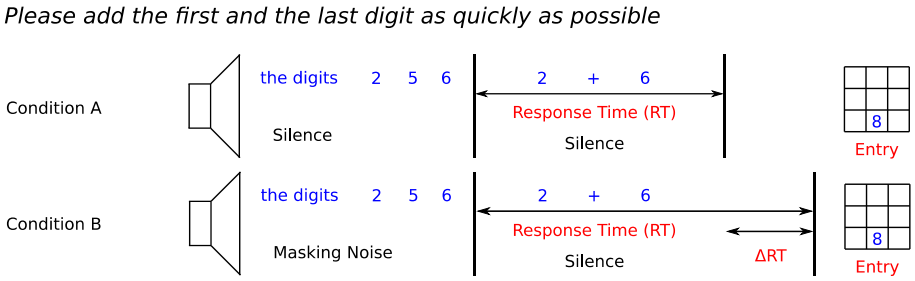

Figure 1 – Scheme of replicated listening effort test by [2]

2.1 Original arithmetic listening effort task (HA)

The original listening effort task (termed Houben Arithmetic (HA) in the following) was to sum the first and the last of the three digits, entering the result as quickly as possible, ignoring the second digit sounded. From the end of the third digit sounded to the entry, the response time was measured. The result of this easy arithmetic task can be checked. This helps to confirm that the task was carried out correctly and to reject wrong entries.

HA contains both, an identification component of the digit, which can be prolonged by lowering the SNR and a memory component, remembering the first digit, whilst ignoring the second digit. It remains unclear, to which extend the memory component contributes to prolonged RT in masked conditions. In the experiment, the task was replicated to compare it with the variations created.

2.2 Variation on the memory component (HM)

To find out, if the memory component has a significant effect on reaction times in HA, the task was changed. Always the same digit (one) was added to the last of the three digits sounded, a process that resembles counting. What remains is the identification of the last digit, as counting in a range to 9 is practiced from early age and should create very little cognitive effort. Thus, it is hypothesized, that RT will be shorter than in HA and show the same trend as HA for different SNRs.

2.3 Variation on the identification component (HV)

To investigate the impact of the identification component in HA, the triplet of digits is presented

visually. The masking noise is presented aurally, despite it having no impact on the identification of visually presented digits. From this variation, the core RT of the arithmetic task should remain. RT should be faster than in HA over all four conditions and show the same trend as HA.

2.4 Serial recall task (SR)

The serial recall task was presented aurally under the same masking conditions as the listening effort experiment. Nine digits were sounded and recalled by the test subjects. There was no retention phase after the last digit. RT for each digit and the error rate were recorded. The masking sound did not fluctuate and thus no irrelevant sound effect (ISE) was expected. Without ISE, the error rate is the same in all four SNR-conditions, unless prolonged identification processes lead to higher error rates. Further, the response times from the last digit sounded to the entry of each digit are recorded. The goal is the comparison of RT between SR and listening effort tasks.

Table 1 shows the number of repeats and specifies the conditions performed by each of the subjects. It also shows the levels of the digits and masking sounds used plus the resulting SNR.

Table 1 – Experimental items performed by each subject

Item	Test	Condition	Repeats	L_{Signal} dB(A)	L_{Masker} dB(A)	SNR dB
1	HA	Practise	10	60		$+\infty$
2	HA	SNR1	40	60	66	-6
3	HA	SNR2	40	60	61	-1
4	HA	SNR3	40	60	56	$+4$
5	HA	Silence	40	60		$+\infty$
6	HM	Practise	10	60		$+\infty$
7	HM	SNR1	40	60	66	-6
8	HM	SNR2	40	60	61	-1
9	HM	SNR3	40	60	56	$+4$
10	HM	Silence	40	60		$+\infty$
11	HV	Practise	10	–		$+\infty$
12	HV	SNR1	40	–	66	-6
13	HV	SNR2	40	–	61	-1
14	HV	SNR3	40	–	56	$+4$
15	HV	Silence	40	–		$+\infty$
16	SR	Practise	5	60		$+\infty$
17	SR	SNR1	20	60	66	-6
18	SR	SNR2	20	60	61	-1
19	SR	SNR3	20	60	56	$+4$
20	SR	Silence	20	60		$+\infty$

2.5 Participants, equipment and methods

The experiments were conducted in the laboratory at Fraunhofer-IBP in Stuttgart, Germany, with German as experimental language. The 21 subjects (15 female, 6 male, 21-29 years, average 24 years) were students from Hohenheim University, collecting experimental hours as part of their program. The number of subjects was oriented on group sizes to find the ISE in laboratory experiments and is higher than the 12 subjects in [2]. The experiment took 90 minutes without break. The sound signals were presented on headphones (Sennheiser HD280 pro), calibrated with the noise signal used in the experiments to 60 dB(A). The instructions and stimuli were presented by PsyScope (B57), running on Macintosh computers. The order of the tests and conditions were randomized, each beginning with a training sequence, followed by 40 repeats (SR: 20) in each condition, see Table 1. The entries were made by mouse clicks on screen.

2.6 Statistical analysis

Response time data is not normally distributed. For this reason, the data was transformed with the natural logarithm and normality was checked with the Shapiro-Wilk test. The analysis of the data sets was carried out in SPSS with one- and two-way repeated-measures ANOVA, followed by Bonferroni

and post-hoc tests.

3. RESULTS

3.1 Listening effort tests

The raw data was viewed and outliers removed (RT < 100 ms and RT > mean + 2 standard deviations) based on [4]. One data set was removed as there were too many wrong answers. Raw and corrected data is shown in Figure 1, with the test conditions on the X-axis and response time RT on the Y-axis. From this overview, it can be observed, that on average, the original listening effort task HA took the participants longer to perform than HM and HV. One would expect, that lower SNR (i.e. SNR = -6 dB) leads to longer RT compared to digits played in silence (indicated by S in Figure 1), as perceived listening effort is higher at low SNR. However, the trend expected cannot be observed.

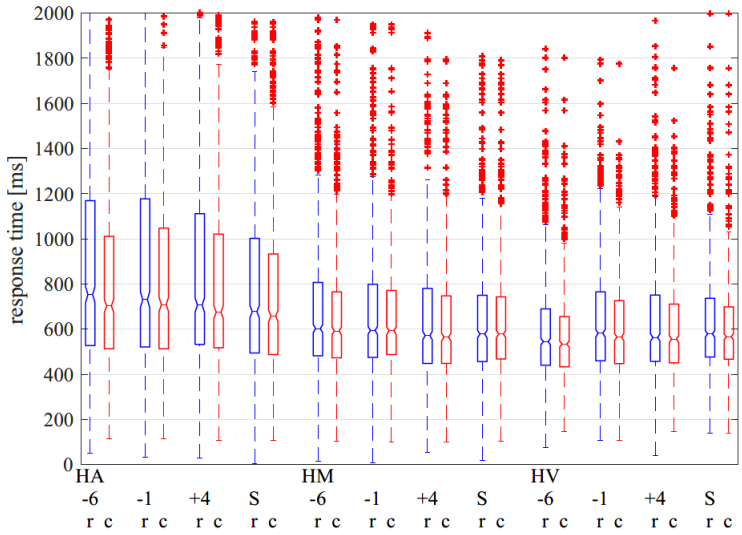

Figure 1 – Data raw and corrected from listening effort tests HA, HM and HV for all SNR conditions

The HA results are compared with findings in [2], see Figure 2 (left). Both trend and spread are different from published data in [2]. Whilst in the original publication (magenta line) all four conditions lead to statically significant differences, none of the findings of the replicate (blue line) is statistically different, $F(3,57) = 1.69$, $p > .05$, Mauchley's test confirms sphericity for all data sets. Further, the trend reported in [2] could not be replicated.

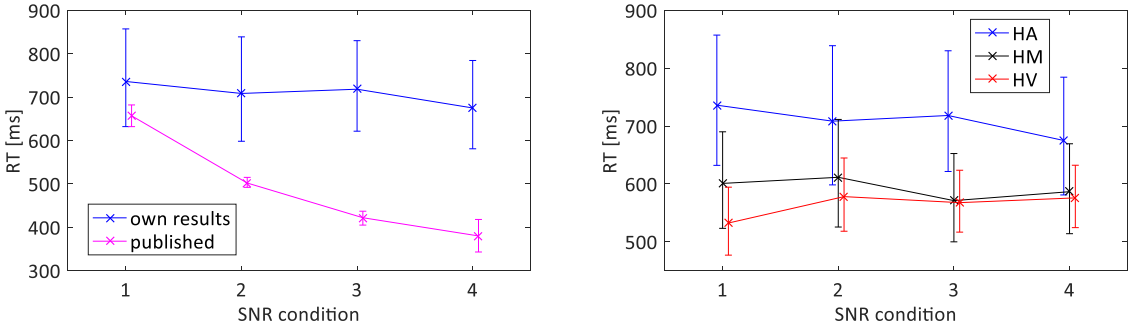

Figure 2 – RT of HA compared to published results (left) and for test variations (right) with 95% CI,

SNR1=-6 dB, SNR2=-1 dB, SNR3=+4 dB, SNR4=+∞

When the task is reduced to identify and add 1 to the last digit sounded in HM, RT decreases to around 600 ms over all four masking conditions. The statistical test reveals that none of the SNR-levels is significantly different, $F(3,57) = 1.13$, $p > .05$. The results can be found in black in Figure 3 (right).

The visual presentation of the digits HV lead to the red curve in Figure 3. Here, the ANOVA results

F(3,57) = 3.56, p = .018 are significant. The Bonferroni post-hoc test reveal that SNR1 = -6 dB is significantly different to SNR3 = +4 dB, p = 0.037. This finding contradicts the hypothesis that all four SNR levels should show the same RT, as the listening effort is the same in each case.

3.2 Serial recall

First, the correct answers in SR were counted over 20 repeats performed. The averaged results are shown in Figure 5 (left). The results show, that for digits 5 – 8 only 30-50% are correct answers, whereas the first and last digit is correct in roughly 90% of answers. Those are typical results in SR tests. SNR as independent variable is spherical, as Mauchley's test confirms. A repeated-measure one-way ANOVA shows that there are no significant differences in correct answers for the different levels of SNR, F(3,60) = 1.45, p = .237. This result was expected based on the non-fluctuating character of the masking sound. Apparently, the identification process of digits does not have any effect on the level of right answers.

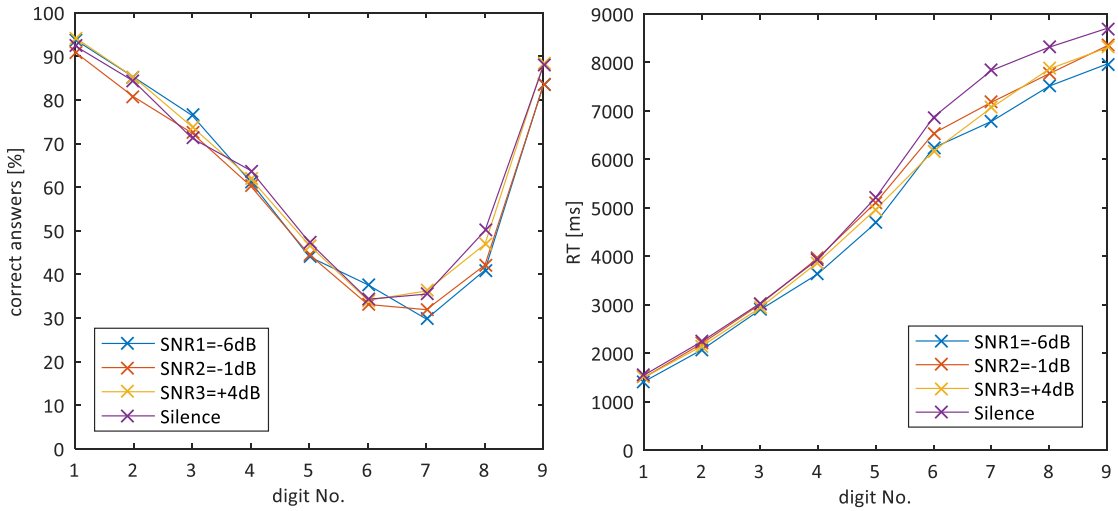

Figure 5 – Serial Recall test, Left: correct answers, Right: response time of correct items

Second, the RT in SR appears to be much longer than in listening effort tests. For the first entry, almost all participants took longer than 1000 ms, whereas in listening effort tests, the entry took between 500 ms and 800 ms. There are several ideas, why this could be: The subjects may lose track of how many digits are sounded and take time to realize, the series has ended. Further, it could be that memorizing the digits still continues after the last digit has been sounded. Whatever the reasons for the delay, there is a difference to the listening effort test variations.

Third, the response time is analysed. As two subjects showed much longer RT, they were excluded from the data set under investigation. The remaining 19 data sets with 20 repeats for each SNR-condition were tested for normal distribution for each digit (Shapiro-Wilk test), which mostly was the case, p > .05. The averages were plotted in Figure 5 (right). A one-way ANOVA was utilized to find out, if there are significant differences between the four SNR-levels for each of the digits. Table 2 shows the results both for Mauchley's test and a ANOVA for every digit. Significant levels were reached for digits 6-9. The Bonferroni post-hoc test could identify only one significant difference for digit 9, where SNR1=-6 dB is significantly different from the unmasked condition, p= .035. There are five interactions, which are close to significant levels, p < .10.

Fourth, it is worth noting, that the subjects performed slowest for digits played with no masking sound and fastest in lowest intelligibility of the digits. Like in the HV experiment above (see Figure 3, right, red line), this contradicts the idea that masked sounds necessarily increase latencies.

Table 2 – Test results for serial recall

Digit	Mauchley's test of sphericity				Analysis of Variance				
	χ^2	df	Sig.	Grh.-G ϵ	df SNR	df Error	F	Sig.	η_p^2
1	7.906	5	.162		3	54	1.289	.287	.067
2	7.220	5	.205		3	54	1.550	.212	.079
3	4.312	5	.506		3	54	0.674	.572	.036
4	4.577	5	.470		3	54	0.754	.525	.040
5	13.007	5	.024	.748	2.245	40.414	1.053	.365	.055
6	7.779	5	.169		3	54	2.913	**.043**	.139
7	10.384	5	.065		3	54	4.325	**.008**	.194
8	5.660	5	.341		3	54	3.529	**.021**	.164
9	12.306	5	.031	.753	2.258	40.635	3.529	**.034**	.164

4. DISCUSSION

The findings of [2] could not be replicated. There are no significant differences for SNR-conditions in HA, which were expected to be found. This is surprising and a close comparison of the two setups was performed to identify potential reasons. The overall level of presentation in [2] was kept constant at 70 dB(A) over all conditions in [2], whereas in the experiments presented, the digits were presented at 60 dB(A) over all conditions and the masking sounds had different levels. We picked this way of presentation to opt out differences in hearing threshold affecting the intelligibility between SNR conditions, as we did not perform a separate hearing test. With fixed level of the digits and a variation of SNR needed, the presentation levels were between 60 dB(A) and 67 dB(A). The other difference concerns the statistical analysis, which has no impact on trends found in descriptive statistics.

In HM, response times decrease as hypothesized (compare HM to HA in Figure 3) due to reduced task complexity. No significant difference between SNR conditions can be found. The trend is flat in HM as in HA. As no change over SNR can be observed in both experiments, no conclusions can be drawn, which role the memory component plays in this variation.

Further, it was hypothesized that if auditory identification was removed, there would be no difference between the SNR-conditions in HV. However, it was found that SNR1 = -6 dB differs significantly form the condition SNR3 = +4 dB in the HV experiment. This can only be explained by the different level of the masking sound presented as the only difference between the conditions.

The logic applied in the listening effort experiment that masking increases the response latencies due to higher cognitive effort is contradicted by the results shown above. The participants took longer to perform the serial recall test in silence than they did in partly masked conditions. Further, the first entry was made later in SR, despite the masking sound being identical and the recall did not involve further cognitive tasks, such as summing digits.

An alternative interpretation could be adopted, that several mechanisms are at work: While identification of the digits is harder for low SNR, the time lost is balanced out by higher overall arousal level of the nervous system of the subjects [5]. If this applies, it would mean that RT is very sensitive to perceived loudness level changes. To confirm this idea, the test needs to be re-run with equal sound level or even better, equal loudness, over all SNR conditions. However, for practical applications in which the level cannot be set, the test procedure would not be suitable.

5. CONCLUSION

The research idea was to find a test procedure to objectively assess impact of noise on the cognition of listeners. The dual task paradigm investigated was reported to find significant differences in response time even for very little masked signals, which is an important aspect in open-plan offices. These findings of the initial experiment could not be replicated. From the set of experiments it appears that there is a level-based effect at work, as faster responses appear in a visual representation of the listening effort experiment and in serial recall for the signals with lowest intelligibility and the highest levels in the set. The set of experiments could be repeated in future with equal sound level over all conditions to rule out different arousal levels possibly leading to faster response times. For serial recall in aural presentation, the error rate did not change over the SNR-levels. The two test procedures appear to investigate different cognitive aspects, as response times are very different. This does not rule out that test procedures based on reaction times might be helpful to investigate the impact of irrelevant sound on office staff. Further research to understand the cognitive mechanisms at work in the listening

effort tasks presented would benefit both, office staff and the hearing impaired.

ACKNOWLEDGEMENTS

Many thanks to Hörtech gGmbH, who supplied us with the German recordings of the digits required for the experiments complimentary. The material presented is part of [6]. The project was carried out in cooperation between Fraunhofer IBP and ISVR Southampton.

REFERENCES

1. Hongisto V. A model predicting the effect of speech of varying intelligibility on work performance. Indoor Air. 2005 Dec;15(6):458–68.
2. Houben R, van Doorn-Bierman M, Dreschler WA. Using response time to speech as a measure for listening effort. Int J Audiol 2013;52(11):753–761.
3. Smits C, Kapteyn TS, Houtgast T. Development and validation of an automatic speech-in-noise screening test by telephone. Int J Audiol 2004 Jan;43(1):15–28.
4. Whelan R. Effective analysis of reaction time data. Psychol Rec. 2008;58(3):475-482.
5. Van Gemmert AW, Van Galen GP. Stress, neuromotor noise, and human performance: A theoretical perspective. J Exp Psychol Hum Percept Perform. 1997;23(5):1299.
6. Dickschen AM. Development of a combined test procedure for listening effort and cognitive performance, Master Thesis. University of Southampton; 2016.