Validation of a simple response-time measure of listening effort

Carina Pals[a)]
Research School of Behavioral and Cognitive Neurosciences, Graduate School of Medical Sciences, University of Groningen, Groningen, The Netherlands
c.pals@alumnus.rug.nl

Anastasios Sarampalis and Hedderik van Rijn
Department of Psychology, University of Groningen, Groningen, The Netherlands
a.sarampalis@rug.nl, hedderik@van-rijn.org

Deniz Başkent
Department of Otorhinolaryngology/Head and Neck Surgery, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
d.baskent@umcg.nl

Abstract: This study compares two response-time measures of listening effort that can be combined with a clinical speech test for a more comprehensive evaluation of total listening experience; verbal response times to auditory stimuli (RT_{aud}) and response times to a visual task (RTs_{vis}) in a dual-task paradigm. The listening task was presented in five masker conditions; no noise, and two types of noise at two fixed intelligibility levels. Both the RTs_{aud} and RTs_{vis} showed effects of noise. However, only RTs_{aud} showed an effect of intelligibility. Because of its simplicity in implementation, RTs_{aud} may be a useful effort measure for clinical applications.

[DOS]
Date Received: February 15, 2015 **Date Accepted:** August 12, 2015

1. Introduction

Speech understanding heavily depends on the cognitive processing required to interpret the (degraded) speech signal in everyday listening environments,[1] perhaps even more so for hearing-impaired individuals. Measures of listening effort (LE) can therefore complement traditional speech intelligibility measures by providing additional information about the listening experience.[2] Different methods have been suggested for quantifying LE, ranging from subjective self-report,[3] to behavioral measures, such as memory tasks,[4] speech response-times (RTs)[5–7] or dual-task paradigms,[8,9] and physiological measures, such as pupillometry.[10] An easy-to-administer method for measuring LE could be a valuable tool in research and clinical settings.

The current study compares two behavioral measures of LE that can be combined with the traditional clinical speech intelligibility test; the dual-task paradigm and verbal RTs to a speech task. Dual-task paradigms are an established method for quantifying LE[8,9] and are based on the assumption that cognitive resources are limited and shared across tasks.[11,12] The resources needed for the primary task reduce the resources available for the secondary task.[13] Therefore, when the primary task is given precedence, secondary task performance is assumed to indirectly reflect the processing demands of the primary task. The verbal response times to auditory stimuli (RTs_{aud}), proposed as early as in the 1960s as a tool for discriminating between seemingly comparable speech communication systems,[5] and later used to quantify hearing device benefit,[6,7] reflect cognitive processing time and index the cognitive effort required to interpret and respond to an incoming auditory signal.[6,7]

In this study, a speech intelligibility task similar to clinical tests used in the Netherlands was performed either by itself to provide the RT_{aud}, or simultaneously with a secondary visual rhyme-judgment task[8] to provide visual response-times (RT_{vis}). To manipulate listening effort and intelligibility separately, and based on previous observations that LE can vary depending on the noise type,[10] participants listened to sentences in quiet, and in two different types of noise, each at two different intelligibility levels.

[a)]Author to whom correspondence should be addressed.

2. Methods

2.1 Participants

Nineteen native Dutch speakers (age = 18 to 25 years; mean = 19 years), all students of University of Groningen, participated in exchange for partial course credit. Exclusion criteria were self-reported dyslexia or other language or learning disabilities, and pure tone thresholds above 20 dB hearing level at any of the audiometric frequencies (250 Hz to 6 kHz). The study was approved by the local ethical committee.

2.2 Stimuli

The speech stimuli used for the listening task were taken from the female speaker set of the Vrije Universiteit (VU) corpus.[14] The corpus consists of 39 balanced lists of 13 conversational Dutch sentences, each 8 to 9 syllables long. A random subset of 24 lists was used per participant, two lists for each experiment or training block. A steady-state, speech-shaped noise (SSN; provided with the VU corpus) and an eight-talker babble in English were used as background noises. The sentences were presented in both noise types, each at two signal-to-noise ratios (SNRs), resulting in two levels of intelligibility; approximately 79% or near ceiling (NC).

Individual SNRs to achieve 79% intelligibility were determined for each participant at the start of the experiment using sentences from the same corpus that were not included in the main experiment. This was done separately for SSN and babble following a three-down-one-up adaptive procedure,[15] which typically results in 79% accuracy. Each sentence-in-noise was presented at an overall level of 70 dB A. The first sentence was played repeatedly until the sentence was correctly understood, starting at −8 dB SNR and increasing the SNR in steps of 4 dB. After this, the adaptive procedure ran for eight reversals at a step size of 2 dB. The resulting mean SNRs from last eight reversals that were used in the experiment were as follows: SNR = −1.20 dB (SD = 1.00) for SSN and SNR = 2.30 dB (SD = 1.10) for babble. A pilot experiment showed that increasing the 79% SNR by 5 dB resulted in NC speech understanding, and this was therefore used as the SNR for the NC intelligibility conditions.

For the secondary, visual rhyme-judgment task, pairs of Dutch monosyllabic words[8] were displayed in large, black capital letters on a white background, one above another, horizontally centered on a computer monitor placed ~60 cm from the participant. Each letter was approximately 7 mm wide and 9 mm high, with 12 mm vertical whitespace between the words.

2.3 Experimental procedure

Before the start of the main experiment two cognitive tests were administered: the symbol search test from the Wechsler Adult Intelligence Scale (WAIS),[16] to measure cognitive processing speed, and the standard computerized version of the reading span test (RST),[17] to measure working memory capacity.

The experimental procedure consisted of 2 training blocks and 11 experimental blocks. Training consisted of one single-task rhyme-judgment task and one dual-task combining the listening task and the rhyme-judgment task. The experimental blocks consisted of six single-task blocks; five times a listening task, and one visual rhyme-judgment task; and five dual-task blocks combining the listening task and the rhyme-judgment task. The listening tasks, in both single and dual task, were presented in five listening conditions: in no noise and in two noise types (babble and SSN) both at two intelligibility levels (79%, NC). Presentation order of the experimental blocks was counterbalanced using a Latin-square design.

In the listening task, participants listened to sentences and repeated them out loud. The sentence recordings were on average 1.8 s in duration and were presented 8 s apart, giving the participants 6.2 s between sentences to respond. The responses were recorded for later scoring of RTs_{aud} and accuracy. The RTs_{aud} were calculated from the offset of the stimulus, as logged by the experimental program, to the onset of the verbal response, as marked by a native Dutch speaker upon visual inspection of the recorded waveform in Audacity. A second native Dutch speaker re-scored a random sample of the recordings to test for inter-rater reliability (Pearson's $r > 0.99$).

In the secondary, visual, rhyme-judgment task, participants pressed one of two buttons as fast as possible to indicate whether two words rhymed or not. Chance of a rhyming pair was 50%. The words were presented on a monitor for a maximum of 2.7 s, or until the participant responded. In case no key was pressed, a "miss" was logged. A fixation cross appeared for a randomly varied interval between 0.5 and 2.0 s between stimuli.

For the dual task, the listening task and the visual rhyme-judgment task were presented simultaneously, but with independent timing to prevent expectation-driven preparation.[8] Note that this meant that the secondary-task stimuli could be presented during or between auditory stimuli.

3. Results

The left panel of Fig. 1 shows the speech intelligibility results in percentage of sentences correctly repeated, and confirms that the desired intelligibility levels were achieved.

The middle panel of Fig. 1 shows the dual-task RTs_{vis} per condition, with average single-task RT_{vis} included as a baseline. Data from incorrect secondary-task trials were excluded from the analysis. Because of the nature of the rhyme-judgment task, with the number of trials depending both on response speed and response accuracy, the number of secondary task trials varied per participant per condition. As ANOVAs are less suitable for analyses based on different number of trials per cell, linear mixed-effects (LME) models were used (lme4-package version 1.1–7; lmerTest-package version 2.0-11) to analyze the RT_{vis} data. As the RT_{vis} were not normally distributed, we log-transformed the response times and excluded reaction times below 0.35 and over 2 s (1.80% of all trials), yielding a reasonably normal $lnRTs_{vis}$ distribution (assessed using QQNorm).

The model of the dual-task $lnRT_{vis}$ results took into account all experimental manipulations; the overall effect of the presence or absence of noise, and for speech in noise, the effects of intelligibility and of noise type. Furthermore, visual stimulus timing (either during or in between the auditory presentation of sentences) and participants' WAIS and RST scores were included as factors. Random intercepts and slopes were included for all within-subject factors, and for stimulus timing.[18] A random intercept for sentence ID was not included, as no sentence can be assigned to RTs_{vis} responses recorded in-between auditory stimuli. Two different contrast-coding strategies were used to reflect the experiment design. The difference between noise and quiet was treatment coded, setting quiet to zero and noise to one. The contrasts between SSN and babble and between 79% and NC intelligibility were effect coded, setting one of the 2 to -0.5 and the other to 0.5. The p-values reported are obtained using the Satterthwaite approximation as reported by the lmerTest package.

The model of the $lnRT_{vis}$ is summarized in the top half of Table 1. The intercept corresponds to the average $lnRT_{vis}$ for speech in quiet, and is estimated at 0.323, although, due to large variance it was not significant ($\beta = 0.323$, SE $= 0.221$, $t = 1.465$, $p = 0.162$). The model shows an effect of Noise, estimated at $\exp(0.323 + 0.041) - \exp(0.323) = 0.7$ s ($\beta = 0.041$, SE $= 0.013$, $t = 3.174$, $p = 0.005$) when compared to the intercept. For speech in noise, the effects of noise type and intelligibility were not significant, nor was the interaction between noise type and intelligibility. RT_{vis} were significantly longer for secondary task trails presented simultaneously with an auditory stimulus than for trials in-between auditory stimuli, the effect in $lnRT_{vis}$ was estimated at 0.055 ($\beta = 0.055$, SE $= 0.009$, $t = 6.161$, $p < 0.001$). From the two cognitive measures collected before the experiment, only the WAIS score showed significant predictive value; the effect of WAIS score on $lnRT_{vis}$ is estimated at -0.007 ($\beta = -0.007$,

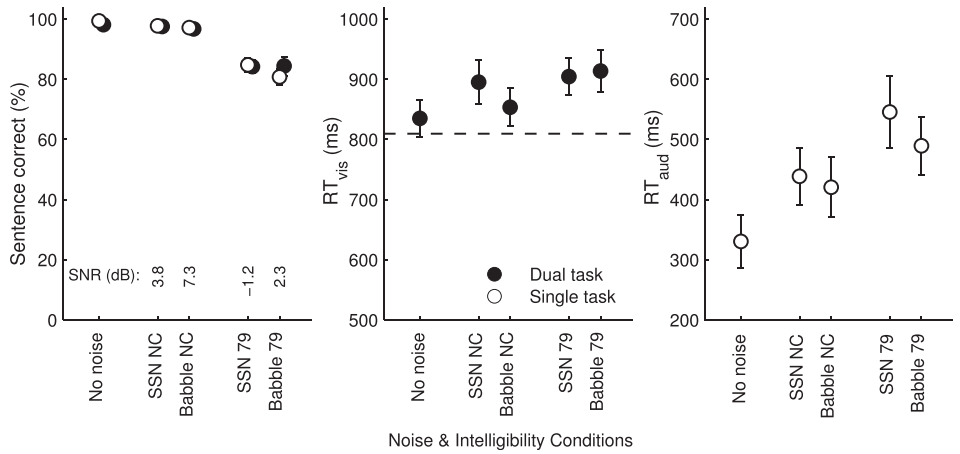

Fig. 1. Left panel: Mean intelligibility in % sentences correctly repeated on the listening task in dual task (closed circles) and single task (open circles). Middle panel: Mean dual-task RTs_{vis} in ms, with single-task RT_{vis} performance indicated by the dashed reference line. Right panel: Mean single-task RTs_{aud} in ms. In all panels, the error bars show ±1 standard error.

Table 1. Summary of the LME model for Dual-task RTs_{vis} (top half) and the Single-task RTs_{aud} (bottom half). The intercept estimates the RT_{vis} for no noise. *Noise* lists the average effect for speech in noise compared to no noise. Effects of Intelligibility, NoiseType, and their interaction are only present in Noise and estimated relative to Noise (signified by "N:"). Asterisks denote significance level: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

	Estimate (ms)	Standard Error	df	*t* value	Pr(>|t|)
Dual-task $lnRT_{vis}$ model					
(Intercept)	323.82	221.09	16.24	1.465	0.162
Noise	41.97	13.23	18.16	3.174	0.005 **
N:Intelligibility	25.61	16.93	17.96	1.513	0.148
N:NoiseType	−1.18	14.54	17.86	−0.081	0.936
N:Intel:NoiseType	16.76	20.19	19.19	0.830	0.417
Timing	55.64	9.03	26.00	6.161	< 0.001 * **
WAIS	−7.67	3.59	16.14	−2.138	0.048 *
RST	−3.61	2.17	16.06	−1.667	0.115
Single-task RT_{aud} model					
(Intercept)	556.82	191.12	16.27	2.913	0.010 *
Noise	131.09	20.85	17.89	6.284	< 0.001 ***
N:Intelligibility	72.21	17.46	17.77	4.137	< 0.001 ***
N:NoiseType	−24.72	12.66	17.77	−1.952	0.067
N:Intel:NoiseType	−6.34	22.73	17.78	−0.279	0.783
WAIS	−4.57	3.09	15.87	−1.480	0.158
RST	−0.48	1.87	15.90	−0.259	0.799

SE = 0.004, $t = -2.138$, $p = 0.048$), suggesting on average lower RT_{vis} for participants with a higher score on the WAIS symbol search.

The right panel of Fig. 1 shows the average RT_{aud} per listening condition. Only RTs_{aud} for sentences that were repeated correctly were included in the analysis, therefore, similar to the dual-task RT_{vis} data, the RT_{aud} data contained unequal numbers of trials per cell depending on speech recognition accuracy. RTs_{aud} were analyzed using the same methodology as the dual-task RTs_{vis}. The RT_{aud} were approximately normally distributed for durations up to 1 s duration, with a skewed tail above 1 s. RTs_{aud} of over 1 s were therefore excluded from the analysis (1.85% of all trials). All factors relevant to the RT_{aud} were included as fixed effects, and a maximal random effects structure was used, accounting for individual intercepts and slopes for all within subject factors, as well as random intercepts for sentence ID.

The results of the model are summarized in the bottom half of Table 1. In quiet listening conditions, the verbal response was estimated to start 557 ms after stimulus offset ($\beta = 556.82$, SE = 191.12, $t = 2.913$, $p = 0.010$). In noise, averaged across the noise conditions, RT_{aud} were significantly longer by 131 ms ($\beta = 131.09$, SE = 20.85, $t = 6.284$, $p < 0.001$) implying an average RT_{aud} in noise of 688 ms. The average RTs_{aud} for speech in noise at 79% intelligibility was 72 ms longer than at NC intelligibility ($\beta = 72.21$, SE = 17.47, $t = 4.137$, $p < 0.001$) suggesting that the average RT_{aud} in noise at NC intelligibility was 652 ms, and the average RT_{aud} in noise at 79% intelligibility was 724 ms. The effect of noise type was not significant, suggesting that RTs_{aud} averaged over both intelligibility levels was no different for speech in SSN compared to babble. Finally, the interaction between noise type and intelligibility was not significant either. The cognitive measures taken before the experiment, the WAIS and the RST, were both included in the model as factors, however neither showed a significant effect.

4. Discussion

The goal of this study was to compare RT_{aud} and RT_{vis} for suitability as measures of LE, especially as a complementary test next to a speech intelligibility test. Speech intelligibility, RTs_{aud} (for a simple speech intelligibility task), and RTs_{vis} (on a secondary rhyme-judgment task in a dual-task paradigm) were measured in five listening conditions: in no noise, and in SSN and babble, each at 79% and NC sentence intelligibility. Both RTs_{vis} and RTs_{aud} showed a clear effect of the presence of noise, similar to what literature suggests. However, RTs_{aud} showed a significant effect of intelligibility, while the RTs_{vis} did not.

The dual-task is a powerful tool for understanding the challenges listeners face in every day settings when combining speech communication with other tasks, or for

showing the consequences of increased LE on cognition.[8,9] Hockey[19] proposed that individual differences in coping strategies in demanding situations result in differences in the total amount of resources allocated to the tasks at hand. Dual-task measures have been suggested to reflect the proportion of the allocated resources needed for the primary task, while physiological measures, such as pupillometry, can reflect the magnitude of resource allocation.[20] It could well be that an increase in dual-task demands results in allocation of more resources to the combination of tasks, therefore not showing a difference in the proportional use of the allocated resources. However, if the goal is to find a measure suitable for clinical purposes, physiological measures would present drawbacks as they require expensive equipment and the procedures can be cumbersome.

The single-task RTs_{aud} showed a significant difference between the two intelligibility levels while the dual-task RTs_{vis} did not. On top of this, the RTs_{aud}, as measured in this experiment, have several advantages over the dual task for potential use in clinical settings and with a wide range of patients, for example, children and elderly. The RT_{aud} can be collected from recordings made during a simple speech-understanding test, already used in clinics, without the need for additional tests or expensive equipment. While the patient listens to sentences and repeats them out loud, the RT_{aud} can be collected by recording the responses for offline analysis, using software for automated speech onset detection,[21] or online using a simple, inexpensive voice-activated trigger. With its ease of implementation, RT_{aud} seems to be a good candidate for a measure of LE, complementing speech tests, in research and clinical settings.

Acknowledgments

The authors gratefully thank Marica Baldessarini for her help with the execution of these experiments, Filiep van Poucke for commenting on an earlier version of this manuscript, and Esmée van der Veen, Floor Burgerhof, and Maraike Coenen for their assistance. This research was supported by Cochlear Ltd, Dorhout Mees, Stichting Steun Gehoorgestoorde Kind, the Heinsius Houbolt Foundation, a Rosalind Franklin Fellowship from the University of Groningen, the Netherlands Organization for Scientific Research (NWO, VIDI Grant 016.096.397), and is part of the research program of the University Medical Center Groningen: Healthy Aging and Communication.

References and links

[1]S. Stenfelt and J. Rönnberg, "The Signal-Cognition interface: Interactions between degraded auditory signals and cognitive processes," Scand. J. Psychol. **50**(5), 385–393 (2009).

[2]R. McGarrigle, K. J. Munro, P. Dawes, A. J. Stewart, D. R. Moore, J. G. Barry, and S. Amitay, "Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper,' " Int. J. Audiol. **53**, 433–440 (2014).

[3]M. Rudner, T. Lunner, T. Behrens, E. S. Thorén, and J. Rönnberg, "Working memory capacity may influence perceived effort during aided speech recognition in noise," J. Am. Acad. Audiol. **23**(8), 577–589 (2012).

[4]P. M. Rabbitt, "Recognition: Memory for words correctly heard in noise," Psychonomic Sci. **6**(8), 383–384 (1966).

[5]M. H. Hecker, K. N. Stevens, and C. E. Williams, "Measurements of reaction time in intelligibility tests," J. Acoust. Soc. Am. **39**(6), 1188–1189 (1966).

[6]T. Baer, B. C. J. Moore, and S. Gatehouse,"Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: Effects on intelligibility, quality, and response times," J. Rehabil. Res. Dev. **30**(1), 49–72 (1993).

[7]S. Gatehouse and J. Gordon, "Response times to speech stimuli as measures of benefit from amplification," Brit. J. Audiol. **24**(1), 63–68 (1990).

[8]C. Pals, A. Sarampalis, and D. Başkent, "Listening effort with cochlear implant simulations," J. Speech Lang. Hear. Res. **56**, 1075–1084 (2013).

[9]A. Sarampalis, S. Kalluri, B. Edwards, and E. Hafter, "Objective measures of listening effort: Effects of background noise and noise reduction," J. Speech Lang. Hear. Res. **52**(5), 1230–1240 (2009).

[10]T. Koelewijn, A. A. Zekveld, J. M. Festen, and S. E. Kramer, "Pupil dilation uncovers extra listening effort in the presence of a single-talker masker," Ear Hear. **33**(2), 291–300 (2012).

[11]A. D. Baddeley and G. Hitch, "Working memory," Psychol. Learn. Motiv. **8**, 47–89 (1974).

[12]D. Kahneman, "Attention and effort," in *Measurement* (Prentice-Hall, Englewood Cliffs, NJ, 1973).

[13]M. Nijboer, N. A. Taatgen, A. Brands, J. P. Borst, and H. van Rijn, "Decision making in concurrent multitasking: Do people adapt to task interference?," PloS One **8**(11), e79583 (2013).

[14]N. J. Versfeld, L. Daalder, J. M. Festen, and T. Houtgast, "Method for the selection of sentence materials for efficient measurement of the speech reception threshold," J. Acoust. Soc. Am. **107**(3), 1671–1684 (2000).

[15]H. Levitt, "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**(2), 467–477 (1971).

[16]D. Wechsler, *Wechsler Adult Intelligence Scale* [Dutch version], 4th ed. (WAIS-IV-NL, 2012), (Pearson, Amsterdam, the Netherlands).
[17]M. Van den Noort, P. Bosch, M. Haverkort, and K. Hugdahl, "A Standard Computerized Version of the Reading Span Test in Different Languages," Eur. J. Psychol. Assess. **24**(1), 35–42 (2008).
[18]D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," J. Mem. Lang. **68**(3), 255–278 (2013).
[19]G. R. Hockey, "Compensatory control in the regulation of human performance under stress and high workload; a cognitive-energetical framework," Biol. Psychol. **45**(1–3), 73–93 (1997).
[20]C. Karatekin, J. W. Couperus, and D. J. Marcus, "Attention allocation in the dual-task paradigm as measured through behavioral and psychophysiological responses," Psychophysiology **41**(2), 175–185 (2004).
[21]P. A. Jansen and S. Watter, "SayWhen: An automated method for high-accuracy speech onset detection," Behav. Res. Methods **40**(3), 744–751 (2008).