Citation/Reference	Annelies Devesse, Astrid van Wieringen, Jan Wouters AVATAR Assesses Speech Understanding and Multitask Costs in Ecologically Relevant Listening Situations Ear Hear 2019; XX; 00-00 (online July 25, 2019)
Archived version	Postprint version: the content is identical to the content of the published paper, but without the final typesetting by the publisher
Published version	Devesse et al 2019 Ear Hear
DOI	10.1097/AUD.0000000000000778
Journal homepage	Ear Hear
Author contact	annelies.devesse@kuleuven.be

AVATAR Assesses Speech Understanding and Multitask Costs in Ecologically Relevant Listening Situations

Annelies Devesse, Astrid van Wieringen, and Jan Wouters

KU Leuven – University of Leuven, Department of Neurosciences, ExpORL, Herestraat 49
bus 721, B-3000 Leuven, Belgium (the same for each author)

Address correspondence to Annelies Devesse,
KU Leuven Department of Neurosciences, ExpORL
Herestraat 49 bus 721, B-3000 Leuven (Belgium)
Fax +32 16 330486 / Phone +32 16 32 14 68
E-mail: annelies.devesse@kuleuven.be

Source of Funding

This Research project was supported with grants from the Oticon Foundation (Oticon Fonden),
the Research Council of KU Leuven through project 0T/12/98 and a TBM-FWO Grant from
the Research Foundation-Flanders (T002216N).

Abstract

Objectives – There is a high need amongst clinicians and researchers for an ecologically valid measure of auditory functioning and listening effort. Therefore, we developed AVATAR: an ‘Audio-Visual True-to-life Assessment of Auditory Rehabilitation’ which takes important characteristics of real-life listening situations into account, such as multimodal speech presentation, spatial separation of sound sources and multitasking. As such, AVATAR aims to assess both auditory functioning and the amount of allocated processing resources during listening in a realistic yet controllable way. In the present study, we evaluated AVATAR and investigated whether speech understanding in noise and multitask costs during realistic listening environments changed with increasing task complexity.

Design – Thirty-five young normal hearing participants performed different task combinations of an auditory-visual speech-in-noise task and three secondary tasks on both auditory localization and visual short-term memory in a simulated restaurant environment. Tasks were combined in increasing complexity and multitask costs on the secondary tasks were investigated as an estimate of the amount of cognitive resources allocated during listening and multitasking. In addition to behavioral measures of auditory functioning and effort, working memory capacity and self-reported hearing difficulties were established using a reading span test and a questionnaire on daily hearing abilities.

Results – Whereas performance on the speech-in-noise task was not affected by task complexity, multitask costs on one of the secondary tasks became significantly larger with increasing task complexity. Working memory capacity correlated significantly with multitask costs, but no association was observed between behavioral outcome measures and self-reported hearing abilities or effort.

43 **Conclusions** – AVATAR proved to be a promising model to assess speech intelligibility and
44 auditory localization abilities and to gauge the amount of processing resources during effortful
45 listening in ecologically relevant multitasking situations by means of multitask costs. In contrast
46 to current clinical measures of auditory functioning, results showed that listening and
47 multitasking in challenging listening environments can require a considerable amount of
48 processing resources, even for young normal hearing adults. Furthermore, the allocation of
49 resources increased in more demanding listening situations. These findings open avenues for a
50 more realistic assessment of auditory functioning and individually tuned auditory rehabilitation
51 for individuals of different age and hearing profiles.

Introduction

A variety of methods is available to diagnose hearing impairment and, in current practice, behavioral measurements like pure-tone audiometry or speech-in-noise tests are commonly used. Although these methods are sensitive to detect hearing impairment, they often lack ecological validity to capture the hearing difficulties people typically experience in daily life. In general, traditional measures are a simplification of daily listening situations (Tun et al. 2012; Helfer et al. 2017) and might therefore underestimate the amount of cognitive resources allocated during listening, also referred to as *listening effort*. According to a recent consensus paper, listening effort refers to ‘*the deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a listening task*’ (Pichora-Fuller et al. 2016). Following this definition, challenging listening situations such as having a group conversation in a crowded restaurant, are more likely to make listening effortful as they impose more obstacles in goal pursuit and thus result in higher cognitive demands than favourable, silent environments do. Interestingly, this also implies that the same level of task performance may be achieved in situations that differ in demands, as long as the listener expends enough effort, i.e. allocates enough resources to overcome the obstacles. Noteworthy, listening will only be effortful when the listener deliberately allocates mental resources to overcome the encountered obstacles or, stated differently, when he *wants* to understand what is being said despite the challenges of the listening task. In contrast to daily listening situations, traditional measures such as unimodal, single-task speech-in-noise tests are conducted in easy and well-controlled environments and present few obstacles to understanding speech. As such, they might impose minimal cognitive demands and underestimate the listening effort encountered in daily life.

In the present article, we describe the evaluation of AVATAR (Audio-Visual True-to-Life Assessment of Auditory Rehabilitation), an ecologically relevant auditory assessment paradigm that complements traditional speech-in-noise tests by mimicking the cognitive

demands of daily listening situations in a quantitative and controllable way. Furthermore, AVATAR aims to estimate both auditory functioning and the amount of allocated cognitive resources during speech understanding in challenging situations by means of a behavioral test paradigm.

To date, many researchers in the field of cognitive hearing science have used behavioral measures to investigate effortful listening. A commonly used measure to estimate the amount of allocated mental resources is the dual-task paradigm (McGarrigle et al. 2014; Gagné et al. 2017; Strand et al. 2018), based on the assumption that an individual's cognitive resources to process information are limited in capacity and shared among concurrent tasks (Kahneman 1973). This capacity concept implies that, if two tasks have to be performed concurrently and if the total amount of required internal resources is higher than the available amount of resources, available resources will be redirected to the most important task and performance in terms of reaction time or accuracy on the least important task will drop compared to its single task performance. In general, the drop in performance is referred to as *dual-task cost* and is interpreted as a cost of divided attention between the two tasks that strive for the same limited capacity of mental resources. Hence, it could be used to make assumptions about the amount of effort spent during listening (Pichora-Fuller et al. 2016; Koch et al. 2018). In the field of cognitive hearing science, the primary task typically involves some measure of speech recognition, while the secondary task can vary in nature. Dual-task paradigms have proven sensitive to assess the effect of age (Degeest et al. 2015; Ward et al. 2017), hearing impairment (Tun et al. 2009; Xia et al. 2015) and hearing aid processing (Sarampalis et al. 2009; Desjardins & Doherty 2014; Neher et al. 2014) on dual-task costs. Multitask paradigms are especially suitable as an ecologically valid measure of listening effort, because individuals often have to engage in multiple tasks during real-world performance. However, multitasking in real-life often involves the processing of more than two tasks and it has been shown before that the

processing of more than two simultaneous cues may negatively influence performance on an alternate task. For example, Pitts and Sarter (2018) showed that performance on a driving task was impaired for both young and old participants when they simultaneously had to detect more than two nonredundant cues in different sensory modalities. Also, Dumas & Krampe (2013) found higher dual-task costs in a triple-task condition compared to a dual-task, when combining a working memory task together with a finger force and/or postural control task. To the best of our knowledge, the impact on speech understanding of having to focus on more than one secondary task has not been investigated before. Therefore, AVATAR combined a speech-in-noise task with not one, but three secondary tasks, which allowed to investigate the effect of multitasking on speech intelligibility and multitask costs as a proxy of listening effort.

The tasks and test set-up of AVATAR aimed to mimic the cognitive demands of everyday listening situations as closely as possible, in contrast to traditional behavioral speech-in-noise measurements. According to the Fifth Eriksholm Workshop on Hearing impairment and Cognitive Energy (Pichora-Fuller et al. 2016) and a special issue on Ecologically Valid Protocols for the Assessment of Hearing and Hearing Devices (Keidser 2016), ecologically relevant paradigms should at least include 1) non-aural information, including movements and visual cues and 2) auditory spatial complexity, by means of binaural rather than monaural listening conditions via sound recording and reproduction systems that include more than two channels. Both categories of demands have been incorporated in various test paradigms before, although only partially, separately from each other or not combined with behavioral measures of multitask costs.

First, with regard to visual speech cues, more and more auditory-visual speech tests have been developed that include virtual talking heads or video recordings of real humans (e.g. Schierholz et al. 2017; Schreitmüller et al. 2018). Also, both virtual and real visual cues have successfully been included in dual-task studies investigating effortful listening (Fraser 2010,

Gosselin & Gagné 2011, Stevens et al. 2013; Picou et al. 2014; Desjardins 2016). Nevertheless, if the aim is to mimic realistic listening situations, visual characteristics of the surrounding environment should also be included. Unfortunately, dual-task paradigms that present both a visual talker and a visual real-life listening environment are scarce. For example, Wu et al. (2014) combined an auditory-only speech recognition test with a virtual driving task in a dual-task paradigm to investigate the effect of hearing aid technologies on dual-task costs. Whereas the visual driving task was highly ecologically relevant, speech was presented through headphones and no visual speech cues were provided. Similarly, Lau et al (2016) developed a dual-task paradigm in which participants had to simultaneously walk on a treadmill in a virtual street crossing scenario and recognize spoken words presented auditory-only at different locations via a loudspeaker array. More recently, Hendrikse et al. (2018) investigated not only auditory-visual speech intelligibility, but also auditory localization performance, listening effort and the influence of visual cues on gaze and head movements. A horizontal 16-loudspeaker ring was used to present speech coming from four talkers at different locations, visually presented on a big screen by means of video recordings or virtual humans. Nevertheless, a neutral coloured background scene was used and listening effort was not inferred from behavioral measures but asked for in a questionnaire.

Interestingly, both the work from Lau et al. and Hendrikse et al. mentioned above addressed auditory spatial complexity, i.e. the second important shortcoming of test paradigms that lack ecological validity, by using loudspeaker arrays of more than two speakers. Other researchers have successfully incorporated arrays of five up to 10 speakers to study for example the effect of non-speech visual cues on sound localization (Best et al. 2007) or the benefit of lip reading and head rotations on speech intelligibility in multitalker situations (van Hoesel 2015; Grange & Culling 2016). Nevertheless, these paradigms did not include (behavioral) measures of listening effort. When it comes to auditory spatial complexity, loudspeaker arrays have

several benefits compared to a headphone set-up to mimic real-life acoustics, especially if the aim is to perform hearing aid research (Oreinos & Buchholz 2016). Loudspeaker arrays do not only allow the use of an individual's hearing aid, but also create binaural listening situations and ease to investigate the effect of head movements. Furthermore, they can simulate elementary everyday acoustics, such as different spatial sound sources or moving auditory objects, without having to calculate head-related transfer functions. Recently, extensive spherical loudspeaker arrays of up to 100 speakers have been created to mimic acoustic scenes in a virtual and highly controllable way, either model based or by means of microphone recordings (e.g. Best et al. 2015; Cubick & Dau 2015; Oreinos & Buchholz 2015; Grimm et al. 2016; Weller et al. 2016, Fuglsang et al. 2017). While these methods provide interesting research opportunities, they are less suitable for clinical practice due to high computational demands and equipment costs.

In AVATAR, we focussed on elementary spatial hearing abilities by means of a rather limited, modular and easily implementable horizontal speaker array that allowed to present auditory stimuli from different locations in the horizontal field. We included both speech stimuli and ecologically valid non-speech sounds (e.g. a telephone ringing), because it has been suggested that both types of auditory stimuli together with general spatial auditory awareness may influence the cognitive processes during listening (Edwards 2016). Next to demands related to auditory spatial hearing abilities, we answered the need for visual cues by incorporating virtual humans uttering speech auditory-visually in a virtual restaurant environment.

To summarize, we aimed to reconstruct real-life multitasking listening conditions in a well-controlled laboratory setting by mimicking more cognitive demands than classic speech-in-noise measurements do, to assess auditory-visual speech intelligibility, spatial hearing abilities and multitask costs. Compared to other ecologically relevant auditory assessment

paradigms, AVATAR is the first that allows to mimic and vary the effort during listening by means of a multitasking paradigm with more than one secondary task that mirrors real-life demands including both visual (speech) cues and auditory spatial cues.

The main aim of this study was to investigate the effect of different task combinations on both speech understanding and multitask costs in a realistic environment. To this end, three secondary tasks were added in increasing complexity to an auditory-visual speech-in-noise task. Sensitivity of all three tasks to provide multitask costs was investigated, both in single task conditions or multiple task combinations. Considering the literature on dual-task costs and the definition of listening effort, we hypothesized that the listening environment would require a higher amount of processing resources with increasing task complexity, resulting in higher multitask costs on the secondary tasks. We did not expect the performance on the main listening task to change, unless individuals failed to engage sufficiently in this task. As a secondary aim, we investigated whether multitask costs were related to outcome measures of a reading span test, based on the hypothesis that working memory capacity relates to the cognitive processes employed during speech understanding in adverse conditions (Akeroyd 2008; Rönnberg et al. 2013; Pichora-Fuller et al. 2016). Finally, we asked our participants about their daily hearing difficulties to investigate whether our behavioral outcome measures reflected self-reported hearing abilities, including listening effort.

Materials and Methods

Participants

Thirty-five young normal hearing adults (13 male, 22 female) age 18 to 30 (Mean = 23 years, SD = 2 years) participated in the study. All were native Dutch speakers and reported normal or corrected-to-normal vision. Bilateral audiometric pure-tone thresholds were measured in a sound-attenuating booth with a Madsen OB922 audiometer and TDH-39 earphones and were

less than or equal to 20 dB HL at 250, 500, 1000, 2000, 4000 and 8000 Hz. Written consent was obtained from each participant and the Medical Ethical Committee of the University Hospitals and University of Leuven approved the experiment and materials (approval number B322201526677).

Equipment

Measurements were conducted in a quiet and darkened room of about 3.6 m wide and 4 m deep. A virtual restaurant scenario including five female virtual humans was created in a 3D engine software (Unity 2016, Version 5.3.4) and projected on an acoustically transparent screen (1.73 x 3 m) by means of a short-throw beamer (Optoma EH200ST) (Figs. 1 and 2). A restaurant scenario was used, because it is representative of a typical daily challenging listening situation. Virtual humans were shown at eye-level and participants sat at a distance of 1 m to the projection screen, which reflects the distance of a typical two-side table conversation. Auditory stimuli were presented via an RME Fireface UCX soundcard through five FOSTEX 6301B3 loudspeakers positioned on stands in a straight line 0.11 m behind the screen and 1.20 m above the ground. The loudspeakers ranged from -45 to +45 degree-angle relative to the participant, with a spacing of 22.5 degree-angle. All loudspeakers were positioned towards the participant and every loudspeaker corresponded with one of the virtual humans. A sixth loudspeaker was placed at the zenith, right above the participant's head, and was used for noise stimuli only. The loudspeakers were calibrated at the participant's head position by means of a stationary speech-weighted noise. Testing was administered using tailored software programs on the experimenter's computer. The code to present and trigger auditory and visual stimuli, to control the experimental procedure and to store results was written in and administered through Processing (Processing 2016, Version 3.0.2). Auditory stimuli were presented via Pure Data (Pure Data 2015, Version 0.43.4), whereas the 3D software application Unity (Unity 2016,

Version 5.3.4) controlled the custom-made visual restaurant scenario, including the virtual humans.

The virtual humans used in this study were previously developed at our lab with the specific aim to present speech auditory-visually with intelligible mouth movements (Devesse et al. 2018). Speech sounds were mapped onto 17 different mouth shapes by means of a semi-automated text to speech synthesis software tool (Annosoft Lipsync Tool 2014, Version 4.0260). In a previous evaluation study, speech reading experiments demonstrated that the virtual humans were sufficiently intelligible; when speech reading mono- and disyllabic words, participants correctly identified 85% of the speech stimuli that were also correctly identified based on real video images (Devesse et al. 2018). We opted for virtual humans instead of video recordings because virtual talking heads generate high controllability of the visual speech information, can easily be integrated in virtual listening scenarios and allow to reuse existing speech test materials. The virtual humans in AVATAR differ from most existing talking heads (e.g. Schreitmüller et al. 2018) in that they represent full-body virtual humans and were specially tailored to be implemented in 3D listening scenarios.

Tasks and Stimuli

During a single lab visit, all participants performed single task measures of one main task and three secondary tasks followed by all possible combinations of the main task and one, two or three secondary tasks (see Table 1). While conditions were administered in increasing task complexity, i.e. first all single tasks, followed by 2-, 3- and 4-tasks respectively, the order of conditions within each block of single, 2-, 3- and 4-tasks was randomized between individuals. Within the single task combinations, participants always started with the main task, followed by a random order of the three secondary tasks. All tasks were conducted in multitalker babble noise presented at 60 dB A from the zenith loudspeaker, consisting of auditory-only recordings

of 20 English young adults (eight male, 12 female) simultaneously reading different passages (Francart et al. 2011). Participants could freely move their head throughout the entire experiment. All conditions were grouped in two to three test sessions up to one hour each, with a break of at least 15 minutes between each session. More or longer breaks were provided when participants required them.

Speech-in-noise task and secondary tasks: single tasks - The main task consisted of an adaptive auditory-visual speech-in-noise test (SPIN) including high-context sentences from the LIST-material (Leuven Intelligibility Sentences Test) (van Wieringen & Wouters 2008). These sentences were preferred to isolated words or non-meaningful matrix-sentences because they provide access to meaningful syntactic structures and grammatical and semantic relations within a sentence. Hence, LIST sentences activate semantic memory and urge individuals to engage in top-down language processing, as would happen in real-life (Lemke & Besser 2016). The LIST is one of the most commonly used speech materials in Flanders for clinical evaluation of hearing impairment. The high degree of test-retest reliability (1.2 dB within-subject reliability) makes it suitable for efficient experimental testing of speech intelligibility. The material includes 35 lists of 10 auditory-only sentences spoken by a woman, with two to three keywords per sentence (van Wieringen & Wouters 2008). In AVATAR, sentences were uttered auditory-visually by the middle virtual woman only, corresponding to the loudspeaker at 0 degree-angle. Participants were instructed to verbally repeat every part of the sentence and sentences were only scored correct if all keywords were repeated correctly. Starting at a signal-to-noise ratio (SNR) of -12 dB SNR, the level of the first sentence was increased with a step size of 2 dB until all keywords were correctly identified. From the second sentence on, sentence intensity was adjusted adaptively to the answer of the participant. After a correct answer, the level of the next sentence decreased with 2 dB. Conversely, a wrong answer increased the level with 2 dB. At the end of a list of 10 sentences, the speech reception threshold (SRT), i.e., the

SNR at which 50% of the keywords were correctly repeated, was calculated as the average SNR of the last five sentences plus the fictive 11th sentence.

In addition to the main SPIN-task, three secondary tasks were administered. In two of them, localization abilities in the horizontal plane were assessed by means of non-speech broadband stimuli spanning frequencies up to at least 4 kHz. In the localization task (LOC), participants heard a 1 s phone sound (Van den Bogaert et al. 2006) coming from one of the five loudspeakers in the horizontal field. Five phone stimuli, every direction once, were presented at 60 dB A with a roving level of ± 5 dB, in a random order. Participants had to point out the correct direction by verbally naming the letter on the shirt of the corresponding virtual human. Root mean square (RMS) errors were calculated as an outcome measure with n = number of presented stimuli and *stimulus* and *response* = the location of the corresponding loudspeaker in degree-angle:

$$RMS (^{\circ}) = \sqrt{\frac{\sum_{i=1}^n (stimulus_i - response_i)^2}{n}}$$

Next to LOC, detection of dynamic sounds (DYN) was administered by presenting the sound of a 3 s buzzing fly coming from the right or left at 60 dB A, through panning of the five loudspeakers in the horizontal plane. Participants were asked to verbally report the direction of the sound and percentages correct were computed as an outcome measure. Every direction was presented once, resulting in two DYN-stimuli per list. Third, in a visual one-back memory task (VIS), participants watched a series of 10 different two-digit numbers between 11 and 99. The number was part of a menu card pictured on the table of the restaurant, at the left bottom corner of virtual human C (Fig. 2). At five random moments, the experimenter asked the participants to verbally recall the preceding number. Percentages correct were calculated as an outcome measure. Test and retest were administered of the main task and each secondary task, except

for the VIS-task. An overview of all tasks with their corresponding outcome measure is listed in Table 1.

Speech-in-noise task and secondary tasks: task combinations - Once single task measures were gathered for all tasks separately (i.e., SPIN, LOC, DYN and VIS), the main SPIN-task was combined with all secondary tasks to investigate multitask costs (Table 1 and Fig. 3). All task combinations consisted out of 10 trials with the first trial including a sentence plus the default number of the VIS-task (i.e. 10). From the second trial on, sentences could be combined with a phone-, fly- and/or changing visual number, depending on the task condition. For example, for the 4-tasks condition, the nine remaining sentences were combined with five and respectively four phone- (LOC) and fly-stimuli (DYN), presented at random trials right before the start of a sentence. If a single trial contained both a phone- and fly stimulus, both stimuli were presented sequentially in a random order. After participants repeated the sentence, they were asked to verbally report the phone/and or fly of that same trial. The number of the VIS-task changed with every trial, within a time span of two up to five seconds starting from the beginning of the trial, and was visible during the entire trial. After five random trials participants had to report the number of the preceding trial. The same procedure was applied to the 2- and 3-tasks combinations, with one, or respectively two, secondary task(s) left out. Each condition was administered twice and participants were instructed to always prioritize the SPIN-task when having to perform multiple tasks at the same time.

Importantly, task interference is only expected on tasks that rely on the same pool of cognitive resources. Even though stimuli were presented sequentially during the AVATAR task combinations, we argue that all secondary tasks occurred concurrently with the main task because participants always had to remember the response to the stimuli of the secondary task while listening to the sentences of the SPIN-task. Stated differently, the cognitive processes necessary to remember the presented stimuli overlapped in time (Koch et al. 2018). As a result,

we assumed that all tasks shared at least cognitive memory resources to remember chunks of information from both the secondary tasks and the main SPIN-sentence.

Reading span test – In addition to behavioral measures of auditory functioning and multitask costs, working memory capacity was assessed in a non-auditory reading span test. We opted for an abbreviated, computerized version of the Dutch standard test by van den Noort et al. (2008) developed in the study of Vercammen et al. (2017), that taps into both processing and storage elements of working memory. Participants had to read out loud sentences presented one by one on a computer screen and remember the last word of each sentence. Twenty sentences were grouped into series of two up to six sentences. Series were presented in a random order and at the end of each series, participants had to verbally recall all memorized words of that series. The order in which they recalled the words was not important and a score of 1 was assigned to each correct word (maximum score of 20). Additionally, after a block of 20 sentences, questions were asked on the content of all previous sentences to make sure participants did not exclusively focus on reading the last word but actively processed the entire sentence. In total, three blocks of 20 sentences were administered, but only the total number of correctly recalled words determined the participant's final score (60 points maximum).

Questionnaire on daily hearing – Finally, as a measure of self-reported hearing abilities, the short form of the Speech, Spatial and Qualities of Hearing scale was administered (SSQ12) (Noble et al. 2013). In this pencil and paper questionnaire, 12 questions are assigned to three different subscales: five items refer to general speech understanding (Speech), three to localization abilities (Spatial) and four to the overall quality of perceived sounds (Quality). One item of Quality focusses on how hard individuals have to concentrate during listening, which is close to the notion of effortful listening. Participants had to indicate on a 10-point visual analogue scale how they would rate their hearing in a certain situation (0 = very bad to 10 =

perfect) as well as their concentration level during listening (0 = no concentration to 10 = extreme concentration).

Data analysis

Data were analyzed in the R software interface (R Core Team, 2017). Alpha-levels were set at 0.05 and p -values were corrected for multiple comparisons applying the Holm-Bonferroni method. Paired t tests (or a robust equivalent) did not reveal significant differences between test and retest conditions, thus conditions were averaged for further analysis. Because assumptions were not always met, non-parametric statistics were used to compute correlations and differences between conditions. Where possible, outcome measures were compared to literature.

Results

In what follows, we describe outcome measures for the main speech-in-noise task and the three secondary tasks for different task combinations. We investigated how speech understanding and multitask costs differed with increasing task complexity. Furthermore, we looked into relations between multitask costs and working memory capacity as well as self-reported hearing abilities including listening effort.

Speech understanding in noise

Performance on the main SPIN-task was stable across all task combinations, Friedman's ANOVA $\chi^2(7) = 3.36$, $p = 0.85$ (Fig. 4). As for the single task SPIN, the within-subject standard deviation of the SRT was 1.25 dB for auditory-visual sentences in noise. This measure was in the same order of magnitude as for the original auditory-only speech material, i.e., 1.17 dB (van Wieringen & Wouters 2008) and suggested good test-retest reliability. As expected, the

SRT of the single task condition (Mean = -10.3 dB SNR, SD = 2.2 dB) was better than the auditory-only norm SRT of -8.0 dB SNR (van Wieringen & Wouters 2008) due to differences in presentation mode and noise material. In our experiment, stimuli were presented auditorily-visually and through loudspeakers instead of auditory-only via headphones. Both visual cues and spatial separation of stimuli are known to improve speech intelligibility. The use of babble noise instead of speech-weighted noise attenuated this benefit (Francart et al. 2011).

Multitask costs

Three secondary tasks were added to the SPIN-task to increase task complexity and investigate multitask costs for different task combinations. With regard to the secondary LOC-task, performance did not differ between conditions, indicating participants could equally well localize the telephone signal in all task combinations, Friedman's ANOVA $\chi^2(4) = 1.93$, $p = 0.75$. Single task localization performance (Mean = 7, SD = 5 degree-angle) was in line with previous research (Mean = 7, SD = 3 degree-angle) (Van den Bogaert et al. 2006). Similar to the LOC-task, descriptive analysis of performance on the DYN-task showed no effect of multitasking. Having to detect the direction of the fly was particularly easy: scores were 100% in all conditions, except for the SPIN+DYN+VIS, SPIN+LOC+DYN and 4 tasks conditions, in which, in total, four times an individual score of 88% was obtained. Unlike localizing a telephone or fly signal, remembering a two-digit number resulted in significant multitask costs that increased with increasing task complexity (Fig. 5), as revealed by a Friedman's ANOVA $\chi^2(4) = 78.09$, $p < 0.01$ and post-hoc Wilcoxon signed rank comparisons. More specifically, performance on the single task condition was significantly better than on any of the task combinations, with p -values all < 0.05 and large effect sizes ($r > 0.50$) (Cohen 1988). Furthermore, participants had significantly lower multitask costs on the dual-task combination compared to the SPIN+LOC+VIS-condition and the 4-tasks combination, $p < 0.05$, medium

effect sizes ($0.30 < r$). Finally, multitask costs of the SPIN+DYN+VIS-condition were significantly lower than for the SPIN+LOC+VIS condition, $p < 0.05$, $0.30 < r$. In what follows, performance on the VIS-task was preferred over the LOC- and DYN-task to compute correlations between behavioral outcome measures and working memory capacity or self-reported listening effort, because significant multitask costs were found on this task only.

Working memory capacity

Results on the reading span test (Mean = 43.1/60, SD = 5.6) were comparable to those of Vercammen and colleagues (Vercammen et al. 2017), who found a mean score of 46.0/60 and a standard deviation of 5.9 for 56 normal hearing participants between 20 and 30 years of age. Kendall's τ correlations between results on the reading span test and results on the secondary VIS-task were computed to investigate whether working memory capacity related to multitask costs. The reading span test correlated significantly with performance on the VIS-task for both the SPIN+VIS- and SPIN+VIS+LOC-combination: $\tau = 0.26$ and $\tau = 0.30$ respectively, $p < 0.05$.

Self-reported hearing

Finally, responses on the SSQ12 questionnaire on daily hearing abilities fell all within the expected range set by Banh et al. (2012) (Table 2). None of the different subscales of the SSQ12 correlated significantly with their corresponding behavioral auditory outcome measure ('Speech' & single task SPIN, $\tau = 0.14$, $p = 0.24$ and 'Spatial' & single task LOC, $\tau = -0.19$, $p = 0.15$), nor did the SSQ12 question on concentration relate to multitask costs on the VIS-task, $-0.05 < \tau < 0.11$, $p > 0.40$.

Discussion

The presented work described the evaluation of an ecologically relevant speech-in-noise test, the AVATAR paradigm, to comprehensively assess daily hearing abilities as well as the amount of cognitive resources spent during listening, complementary to a traditional speech-in-noise test. Complex and dynamic real-life listening situations were mimicked in a controllable way by incorporating spatial hearing abilities into an auditory-visual speech-in-noise test. Furthermore, the speech-in-noise task was combined with three secondary tasks on either visual short-term memory or auditory localization of static and dynamic sounds. Turning a classic dual-task paradigm into a more realistic and challenging multitask situation provided a novel way of investigating multitask costs during everyday listening situations as a proxy of listening effort. Finally, the relation between multitask costs and working memory capacity was investigated, as well as how well behavioral outcome measures related to self-reported hearing abilities.

Speech understanding in noise and multitask costs

Our results showed that speech understanding in noise was not affected by task complexity. No matter how many secondary tasks were added to the main speech-in-noise task, speech reception thresholds remained the same. Stable performance on the main task in multitask conditions was in line with previous research on dual-task studies and suggested our participants followed the instruction to prioritize the SPIN-task (Gagné et al. 2017). Furthermore, it implied that the number of resources needed to understand speech did not exceed the total cognitive capacity of our young, normal hearing individuals. Nevertheless, significant performance decrements on the secondary tasks indicated that listening and multitasking required at least some effort. More particularly, significant multitask costs on the visual memory task in the dual-task condition showed that listening to sentences in noise

required a considerable amount of resources, because too few resources were available to perform the secondary task with the same accuracy as in its single task condition. Additionally, even higher performance decrements on the visual memory task in the triple- and quadruple-conditions suggested that addition of the auditory localization tasks further increased the effort during the AVATAR listening situations. As such, results argued in favor of the theory of a shared, limited capacity of cognitive resources (Kahneman 1973).

Figure 6 provides a schematic and hypothetical distribution of the amount of cognitive resources required for each task combination. First, we assumed that the amount of resources needed to achieve a maximum score on a certain task did not alter depending on the task combination, because the level of difficulty of the individual tasks did not change. Second, based on our results we hypothesized that different tasks required a different amount of resources for optimal task performance. To start with, the VIS-task seemed to require more cognitive resources than the LOC- and DYN-tasks respectively, as significant performance decrements in the dual-task combination were found on the visual memory task only. Next, localization of a static signal (LOC) did seem to require at least some cognitive processing, because the addition of the LOC-task to both the SPIN+VIS- and SPIN+DYN+VIS-conditions resulted in a decreased performance on the VIS-task during the corresponding triple- and quadruple-task combinations. On the other hand, adding the DYN-task never changed performance on any other secondary task, suggesting detection of a dynamic sound required only minimal allocation of resources. Interestingly, the fact that resources for the *auditory* LOC- and *visual* VIS-task seemed to compete during task performance, supports the idea that these resources were at least partially shared at a central, amodal stage (Picou & Ricketts 2014).

Significant multitask costs were measured for the VIS-task only, suggesting that participants preferred performing the LOC- and/or DYN-task over the VIS-task. While participants received explicit instructions on giving priority to the main SPIN-task, they were

free to choose which secondary task to prioritize. Hence, it is possible that after completion of the main SPIN-task, listeners flexibly divided their remaining resources over the secondary tasks in disadvantage of the VIS-task. Voluntary prioritization of one task over another is in line with research showing individuals can flexibly tune their resource allocation during multitasking (Koch et al. 2018). For example, Doumas and Krampe (2013) showed that individuals gave priority to tasks with high ecological relevance in triple-task conditions without explicit priority instructions. On the other hand, Ruiz Fernández et al. (2011) demonstrated that individuals selected processing order based on task difficulty.

According to Wickens et al. (2015), four different task attributes generally aid individuals in deciding whether or not to quit an ongoing task and switch to another one: 1) salience, 2) priority, 3) interest and 4) task difficulty. In the AVATAR paradigm, prioritization of one secondary task over another was most likely to be steered by differences in task difficulty. While priority instructions did not differ between tasks and the effects of salience and interest were hard to quantify, all participants indicated that they preferred to focus on the auditory localization tasks because they considered the VIS-task more challenging. Differences in perceived difficulty might have been related to the number of response alternatives of each secondary task. While the DYN-task contained two alternatives only, the LOC-task had five and the VIS-task 89. The higher the number of possible responses, the more difficult the participants rated the secondary task. This finding is in line with research demonstrating word recognition performance is more challenging in 1) an open-set task compared to a closed-set task and 2) a closed-set task with a high number of response alternatives compared to a low number of alternatives (Miller et al. 1951; Buss et al. 2016). Next to response alternatives, the timing and duration of the stimulus presentation differed between tasks. While the LOC- and DYN-stimuli were presented sequentially with the speech stimuli of the main task, the VIS-stimuli overlapped in time with the sentence. This possibly urged the listeners to shift attention

between the sentence and the number during the SPIN-task, forcing them to use more cognitive processes than only memory functions (Gagné et al. 2017). On the contrary, the number stayed visible between two trials, giving the participant the freedom to decide when to look at it. Indeed, 25 out of 35 participants reported that they tried to look at the number before the start of a new trial or sentence. Only two people reported they looked during the sentence while the remaining eight people had no clear or fixed looking strategy. Consequently, most participants had to remember the visual number for a longer period of time compared to the auditory stimuli, making the VIS-stimuli harder to remember.

To summarize, our results suggested that speech understanding in noise was not affected by task complexity, but did require a notable amount of processing resources as reflected by significant dual-task costs. Furthermore, the amount of allocated resources during multitasking in ecologically relevant listening situations increased with increasing task complexity. Of all secondary tasks, the visual memory task was the most sensitive measure of multitask costs. Nevertheless, both auditory localization tasks were relevant to implement in the AVATAR paradigm, because they provided additional information on the localization abilities of individuals and allowed to increase the task complexity.

Working memory capacity

In a second part of our study, we investigated whether behavioral outcome measures of the AVATAR paradigm related to working memory capacity. More specifically, we were interested in a possible relation between multitask costs and results on the reading span test. Based on the assumption that working memory capacity relates to the cognitive processes employed during speech understanding in adverse conditions, individuals with a high working memory capacity are expected to experience less effort during listening (Akeroyd 2008; Rönnberg et al. 2013; Pichora-Fuller et al. 2016). Although some studies have reported significant correlations

between working memory capacity measured through a reading span test and measures of effortful listening (e.g., Tun et al. 1991; Desjardins & Doherty 2013), a clear relation between dual-task costs and cognitive measures has not been established yet (Gagné et al. 2017). Our results suggest a possible link between multitask costs and working memory capacity, because significant correlations were found between results on the reading span test and performance on the VIS-task for two out of four multitasking conditions.

Self-reported hearing

Lastly, we investigated whether behavioral outcome measures on the SPIN-task and secondary tasks were related to self-reported hearing abilities for speech understanding, auditory localization and concentration. No significant associations were found for any category. Bearing in mind that our participants were young, normal hearing participants who all obtained high scores on both the behavioral tasks and self-reported hearing abilities, the lack of significant correlations is not surprising. With regard to the measures of concentration or effort, it is also possible that multitask costs on the secondary VIS-task and the SSQ12 question on concentration tapped into different characteristics of effortful listening (Gagné et al. 2017). This is in line with previous research that failed to demonstrate a clear relation between behaviorally measured and self-reported effort (McGarrigle et al. 2014). As suggested by Lemke and Besser (2016), effortful listening should be considered as a composition of two distinct types of effort: 'perceived effort' on the one hand and 'processing load' on the other. While 'perceived effort' refers to the subjective estimation of effort and is therefore linked to self-report, 'processing load' focusses on the amount of processing resources allocated during listening and is related to behavioral measures of effort (Lemke & Besser 2016). Additionally, interindividual differences in the subjective boundaries of task difficulty and quantification of effort may explain the absence of a significant correlation. It might be worth investigating whether

instantaneous questioning of the amount of perceived effort during the experiment might relate better to behavioral outcome measures, in contrast to investigating the general effort experienced in daily life circumstances. Similarly, we did not include measures of motivation or willingness to perform a task, although recent findings have pointed out that similar questions might increase our understanding of the perceived effort (Pichora-Fuller et al. 2016).

Conclusions and future work

Our findings demonstrate that AVATAR is an effective model to investigate both auditory functioning and the amount of cognitive resources allocated during listening in ecologically relevant conditions. Although speech understanding in noise was not affected by task complexity for a group of young normal hearing adults, listening to speech in challenging, noisy environments seemed to require a considerable amount of cognitive resources that increased with increasing task complexity. In future work, we would like to optimize the AVATAR model so it can be used as a complement to current auditory measures in clinical practice. As a first step, we aim to investigate whether the model is sensitive to differentiate between individuals of different ages and hearing profiles. This will open avenues for a more realistic assessment of auditory functioning and individually tuned auditory rehabilitation.

Acknowledgments

We thank prof. Ralf Krampe for sharing interesting insights and Alexander Dudek for his technical support. This research project was supported with grants from the Oticon Foundation (Oticon Fonden), the Research Council of KU Leuven through project 0T/12/98 and a TBM-FWO Grant from the Research Foundation-Flanders (T002216N).

References

- Akeroyd, M. A. (2008). Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *Int J Audiol*, 47(sup2), S53-S71.
- Annosoft Lipsync Tool [Computer software]. (2014). Retrieved from <http://www.annosoft.com/>
- Best, V., Keidser, G., Buchholz, J. M., et al. (2015). An examination of speech reception thresholds measured in a simulated reverberant cafeteria environment. *Int J Audiol*, 54(10), 682-690.
- Best, V., Ozmeral, E. J., Shinn-Cunningham, B. G. (2007). Visually-guided attention enhances target identification in a complex auditory scene. *J Assoc Res Otolaryngol*, 8(2), 294-304.
- Buss, E., Leibold, L. J., Hall, J. W. (2016). Effect of response context and masker type on word recognition in school-age children and adults. *J Acoust Soc Am*, 140(2), 968-977.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). New York: Academic Press.
- Cubick, J., & Dau, T. (2016). Validation of a Virtual Sound Environment System for Testing Hearing Aids. *Acta Acustica united with Acustica*, 102(3), 547-557.
- Degeest, S., Keppler, H., Corthals, P. (2015). The effect of age on listening effort. *J Speech Lang Hear Res*, 58(5), 1592-1600.
- Desjardins, J. L. (2016). The Effects of Hearing Aid Directional Microphone and Noise Reduction Processing on Listening Effort in Older Adults with Hearing Loss. *J Am Acad Audiol*, 27(1).
- Desjardins, J. L., & Doherty, K. A. (2013). Age-related changes in listening effort for various types of masker noises. *Ear Hear*, 34(3), 261-272.
- Desjardins, J. L., & Doherty, K. A. (2014). The effect of hearing aid noise reduction on listening effort in hearing-impaired adults. *Ear Hear*, 35(6), 600-610.

- 579 Devesse, A., Dudek, A., van Wieringen, A., et al. (2018). Speech intelligibility of virtual humans. *Int J*
580 *Audiol*, 57(12), 908-916.
- 581 Doumas, M., & Krampe, R. T. (2013). Ecological relevance determines task priority in older adults'
582 multitasking. *J Gerontol B Psychol Sci Soc Sci*, 70(3), 377-385.
- 583 Edwards, B. (2016). A model of auditory-cognitive processing and relevance to clinical applicability.
584 *Ear Hear*, 37, 85S-91S.
- 585 Francart, T., van Wieringen, A., Wouters, J. (2011). Comparison of fluctuating maskers for speech
586 recognition tests. *Int J Audiol*, 50(1), 2-13.
- 587 Fraser, S., Gagné, J.-P., Alepins, M., et al. (2010). Evaluating the effort expended to understand speech
588 in noise using a dual-task paradigm: The effects of providing visual speech cues. *J Speech Hear Res*,
589 53(1), 18-33.
- 590 Fuglsang, S. A., Dau, T., Hjortkjær, J. (2017). Noise-robust cortical tracking of attended speech in real-
591 world acoustic scenes. *Neuroimage*, 156, 435-444.
- 592 Gagné, J.-P., Besser, J., Lemke, U. (2017). Behavioral assessment of listening effort using a dual-task
593 paradigm: A review. *Trends Hear*, 21, 1-25.
- 594 Gosselin, P. A., & Gagné, J.-P. (2011). Older adults expend more listening effort than young adults
595 recognizing audiovisual speech in noise. *Int J Audiol*, 50(11), 786-792.
- 596 Grange, J. A., & Culling, J. F. (2016). Head orientation benefit to speech intelligibility in noise for
597 cochlear implant users and in realistic listening conditions. *J Acoust Soc Am*, 140(6), 4061-4072.
- 598 Grimm, G., Kollmeier, B., & Hohmann, V. (2016). Spatial acoustic scenarios in multichannel
599 loudspeaker systems for hearing aid evaluation. *J Am Acad Audiol*, 27(7), 557-566.
- 600 Helfer, K. S., Merchant, G. R., Wasiuk, P. A. (2017). Age-related changes in objective and subjective
601 speech perception in complex listening environments. *J Speech Lang Hear Res*, 60(10), 3009-3018.

- 602 Hendrikse, M. M., Llorach, G., Grimm, G., et al. (2018). Influence of visual cues on head and eye
603 movements during listening tasks in multi-talker audiovisual environments with animated
604 characters. *Speech Commun*, 101, 70-84.
- 605 Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall.
- 606 Keidser, G. (2016). Introduction to special issue: towards ecologically valid protocols for the assessment
607 of hearing and hearing devices. *J Am Acad Audiol*, 27(7), 502-503.
- 608 Koch, I., Poljac, E., Müller, H., et al. (2018). Cognitive structure, flexibility, and plasticity in human
609 multitasking — An integrative review of dual-task and task-switching research. *Psychol Bull*, 144(6),
610 557.
- 611 Lau, S. T., Pichora-Fuller, M. K., Li, K. Z., et al. (2016). Effects of Hearing Loss on Dual-Task
612 Performance in an Audiovisual Virtual Reality Simulation of Listening While Walking. *J Am Acad*
613 *Audiol*, 27(7), 567-587.
- 614 Lemke, U., & Besser, J. (2016). Cognitive load and listening effort: Concepts and age-related
615 considerations. *Ear Hear*, 37, 77S-84S.
- 616 McGarrigle, R., Munro, K. J., Dawes, P., et al. (2014). Listening effort and fatigue: What exactly are
617 we measuring? A British society of audiology cognition in hearing special interest group ‘white paper’.
618 *Int J Audiol*, 53(7), 433-445.
- 619 Miller, G. A., Heise, G. A., Lichten, W. (1951). The intelligibility of speech as a function of the context
620 of the test materials. *J Exp Psychol*, 41(5), 329-335.
- 621 Neher, T., Grimm, G., Hohmann, V., et al. (2014). Do hearing loss and cognitive function modulate
622 benefit from different binaural noise-reduction settings? *Ear Hear*, 35(3), e52-e62.
- 623 Noble, W., Jensen, N. S., Naylor, et al. (2013). A short form of the speech, spatial and qualities of
624 hearing scale suitable for clinical use: The SSQ12. *Int J Audiol*, 52(6), 409-412.

- 625 Oreinos, C., & Buchholz, J. M. (2015). Objective analysis of ambisonics for hearing aid applications:
626 Effect of listener's head, room reverberation, and directional microphones. *J Acoust Soc Am*, 137(6),
627 3447-3465.
- 628 Oreinos, C., & Buchholz, J. M. (2016). Evaluation of loudspeaker-based virtual sound environments for
629 testing directional hearing aids. *J Am Acad Audiol*, 27(7), 541-556.
- 630 Pichora-Fuller, M. K., Kramer, S. E., Eckert, et al. (2016). Hearing impairment and cognitive energy:
631 The framework for understanding effortful listening (FUEL). *Ear Hear*, 37, 5S-27S.
- 632 Picou, E. M., Aspell, E., & Ricketts, T. A. (2014). Potential benefits and limitations of three types of
633 directional processing in hearing aids. *Ear Hear*, 35(3), 339-352.
- 634 Picou, E. M., & Ricketts, T. A. (2014). The effect of changing the secondary task in dual-task paradigms
635 for measuring listening effort. *Ear Hear*, 35(6), 611-622.
- 636 Pitts, B. J., & Sarter, N. (2018). What You Don't Notice Can Harm You: Age-Related Differences in
637 Detecting Concurrent Visual, Auditory, and Tactile Cues. *Human factors*, 60(4), 445-464.
- 638 Processing [Computer software]. (2016). Retrieved from <https://processing.org/>
- 639 Pure data [Computer software]. (2015). Retrieved from <https://puredata.info/>
- 640 Rönnberg, J., Lunner, T., Zekveld, A., et al. (2013). The ease of language understanding (ELU) model:
641 theoretical, empirical, and clinical advances. *Front Syst Neurosci*, 7(31), 1-17.
- 642 Ruiz Fernández, S., Leonhard, T., Rolke, B., et al. (2011). Processing two tasks with varying task order:
643 Central stage duration influences central processing order. *Acta Psychol*, 137, 10–17.
- 644 Sarampalis, A., Kalluri, S., Edwards, B., et al. (2009). Objective measures of listening effort: Effects of
645 background noise and noise reduction. *J Speech Lang Hear Res*, 52(5), 1230-1240.
- 646 Schierholz, I., Finke, M., Kral, A., et al. (2017). Auditory and audio–visual processing in patients with
647 cochlear, auditory brainstem, and auditory midbrain implants: An EEG study. *Hum Brain Mapp*, 38(4),
648 2206-2225.

- 649 Schreitmüller, S., Frenken, M., Bentz, L., et al. (2018). Validating a Method to Assess Lipreading,
650 Audiovisual Gain, and Integration During Speech Reception With Cochlear-Implanted and Normal-
651 Hearing Subjects Using a Talking Head. *Ear Hear*, 39(3), 503-516.
- 652 Stevens, C. J., Gibert, G., Leung, Y., et al. (2013). Evaluating a synthetic talking head using a dual task:
653 Modality effects on speech understanding and cognitive load. *Int J Hum Comput Stud*, 71(4), 440-454.
- 654 Strand, J. F., Brown, V. A., Merchant, M. B., et al. (2018). Measuring listening effort: Convergent
655 validity, sensitivity, and links with cognitive and personality measures. *J Speech Lang Hear Res*, 1-24.
- 656 Tun, P. A., McCoy, S., Wingfield, A. (2009). Aging, hearing acuity, and the attentional costs of effortful
657 listening. *Psychol Aging*, 24(3), 761-766.
- 658 Tun, P. A., Williams, V. A., Small, B. J., et al. (2012). The effects of aging on auditory processing and
659 cognition. *Am J Audiol*, 21(2), 344-350.
- 660 Tun, P. A., Wingfield, A., Stine, E. A. (1991). Speech-processing capacity in young and older adults: A
661 dual-task study. *Psychol Aging*, 6(1), 3-9.
- 662 Unity [Computer software]. (2016). Retrieved from <https://unity3d.com/>
- 663 Van den Bogaert, T., Klasen, T. J., Moonen, M., et al. (2006). Horizontal localization with bilateral
664 hearing aids: Without is better than with. *J Acoust Soc Am*, 119(1), 515-526.
- 665 van den Noort, M., Bosch, P., Haverkort, M., et al.. (2008). A standard computerized version of the
666 reading span test in different languages. *Eur J Psychol Assess*, 24(1), 35-42.
- 667 van Hoesel, R. J. (2015). Audio-visual speech intelligibility benefits with bilateral cochlear implants
668 when talker location varies. *J Assoc Res Otolaryngol*, 16(2), 309-315.
- 669 van Wieringen, A., & Wouters, J. (2008). LIST and LINT: Sentences and numbers for quantifying
670 speech understanding in severely impaired listeners for Flanders and the Netherlands. *Int J Audiol*,
671 47(6), 348-355.

- 672 Vercammen, C., Goossens, T., Wouters, J., et al. (2017). How age affects memory task performance in
673 clinically normal hearing persons. *Neuropsychol Dev Cogn B Aging Neuropsychol Cogn*, 24(3), 264-
674 280.
- 675 Ward, K. M., Shen, J., Souza, P. E., et al. (2017). Age-related differences in listening effort during
676 degraded speech recognition. *Ear Hear*, 38(1), 74.
- 677 Weller, T., Best, V., Buchholz, J. M., et al. (2016). A method for assessing auditory spatial analysis in
678 reverberant multitalker environments. *J Am Acad Audiol*, 27(7), 601-611.
- 679 Wickens, C. D., Gutzwiller, R. S., & Santamaria, A. (2015). Discrete task switching in overload: A
680 meta-analyses and a model. *Int J Hum Comput Stud*, 79, 79-84.
- 681 Wu, Y. H., Aksan, N., Rizzo, M., et al. (2014). Measuring listening effort: Driving simulator vs. simple
682 dual-task paradigm. *Ear Hear*, 35(6), 623.
- 683 Xia, J., Nooraei, N., Kalluri, S., et al. (2015). Spatial release of cognitive load measured in a dual-task
684 paradigm in normal-hearing and hearing-impaired listeners. *J Acoust Soc Am*, 137(4), 1888-1898.

Figures

Figure 1. Test set-up including six loudspeakers, a short-throw beamer and an acoustically transparent projection screen.

Figure 2. Still frame of the restaurant scenario presented during the behavioral multitasking paradigm.

Figure 3. Overview of all possible trial contents and schematic time course of stimulus presentation (S) and required answer (A) within each trial, for the dual- and quadruple-task combinations. The time course of the triple-task conditions can easily be inferred from both. The nature of S differs for each task (e.g. S for SPIN = sentence, for LOC = telephone signal). Dotted arrows represent the time lapse during which stimuli of the secondary tasks needed to be remembered.

Figure 4. Individual results (points) on the main SPIN-task (N = 35). No difference in performance was observed between task combinations. For greater visual clarity, scores on the different dual- and triple-task combinations were averaged, but analysis was conducted on all eight conditions separately (see Table 1).

Figure 5. Individual results (points) on the secondary VIS-task (N = 35). Performance significantly decreased with increasing task complexity (3 tasks DYN = SPIN+DYN+VIS and 3 tasks LOC = SPIN+LOC+VIS).

Figure 6. Schematic representation of the amount of cognitive resources required for optimal task performance for each individual task. Every task combination that exceeded an individuals' total available capacity resulted in significant multitask costs.

707

Abbreviations

708 AVATAR = Audio-Visual True-to-life Assessment of Auditory Rehabilitation

709 DYN = dynamic sound detection task of AVATAR

710 LIST = Leuven Intelligibility Sentences Test

711 LOC = localization task of AVATAR

712 RMS = Root Mean Square

713 SNR = Signal-to-Noise Ratio

714 SPIN = speech-in-noise task of AVATAR

715 SRT = Speech Reception Threshold

716 SSQ = the Speech, Spatial, and Qualities of Hearing Scale

717 VIS = visual memory task of AVATAR

Figure 1

Figure 2

Figure 3

	Task combination	Content	Time lapse trial	Content	Time lapse trial
Trial 1	All	1x SPIN + VIS (no A needed)	<div>S</div> <div>A</div> S = '10'		
Trials 2 to 10	SPIN + LOC	4x SPIN	<div>S</div> <div>A</div>	5x SPIN + LOC	<div>S</div> <div>A</div> S -----> A
	SPIN + DYN	5x SPIN	<div>S</div> <div>A</div>	4x SPIN + DYN	<div>S</div> <div>A</div> S -----> A
	SPIN + VIS	4x SPIN + VIS (no A needed)	<div>S</div> <div>A</div> <div>S(n-1)</div> <div>Sn</div> ----->	5x SPIN + VIS (A needed)	<div>S</div> <div>A</div> <div>S(n-1)</div> <div>Sn</div> -----> <div>A(n-1)</div>
	SPIN + LOC + DYN + VIS	9x SPIN + VIS (5x A needed) + LOC (5x) and/or DYN (4x)	<div>S</div> <div>A</div> <div>S(n-1)</div> <div>Sn</div> -----> <div>A(n-1)</div> <div>S</div> -----> <div>A</div>		

Figure 4

Figure 5

Figure 6

Table 1: Overview of all task combinations and outcome measures of the behavioral multitasking paradigm. Per condition, one outcome measure was calculated for each task.

	No. of tasks	Name of condition	Trials + content	Task	Outcome measure
Single task	1 task	SPIN	10 sentences	Repeat sentence	SRT (dB SNR)
		LOC	5 phone signals, every location once	Report location	RMS error (°)
		DYN	2 fly signals, every direction once	Report direction	(% correct)
		VISs	10 numbers	Report previous number (1-back)	(% correct)
Multitasking	2 tasks	SPIN + LOC	<i>Example of 4-task combination</i>	Repeat sentence + if present:	SRT (dB SNR)
		SPIN + DYN			
	3 tasks	SPIN + VIS		report location (LOC) and/or direction (DYN)	RMS error (°)
		SPIN + LOC + DYN			% correct DYN
		SPIN + LOC + VIS		+ if present and asked for: report previous number (VIS)	% correct VIS
		SPIN + DYN + VIS			
	4 tasks	SPIN + LOC + DYN + VIS			

SPIN refers to the main speech-in-noise task and LOC (localization), DYN (dynamic sound) and VIS (visual memory) to the secondary tasks.

Table 2: Answers (Mean and SD) on the separate questions of the SSQ12, with average scores per category. Question 14 refers to the level of concentration during listening. Scores were in line with the scores collected by Banh et al. (2012).

Speech	Questionno.	1	10	11	4	12	All
	Score	8.8 ± 1.2	6.1 ± 2.1	8.4 ± 1.5	7.9 ± 1.5	8.6 ± 1.4	8.0 ± 1.1
Spatial	Questionno.	6	9	13			All
	Score	8.1 ± 1.6	7.9 ± 1.5	8.5 ± 1.1			8.2 ± 1.1
Qualities	Questionno.	2	7	9	14		All
	Score	8.3 ± 1.4	8.4 ± 1.5	9.2 ± 0.9	8.0 ± 1.8		8.0 ± 1.9

Questionnumbers refer to the original questions of the SSQ49.