DOI: 10.61782/fa.2023.0134

forum acusticum 2023

SPEECH-PERCEPTION-IN-NOISE AND COGNITION IN A REALISTIC LISTENING SCENARIO

Lyan Porto[1*] **Jan Wouters**[1] **Astrid van Wieringen**[1]
[1] Department of Neurosciences, Research Group ExpORL, KU Leuven, Belgium

ABSTRACT

Understanding speech in noise is an everyday task for adults and children alike. Many factors are known to affect how well one can understand speech in the presence of background noise, such as sound levels and spatial separation of speech and noise sources. Cognitive factors such as attention and working memory are also understood to play a role, but how these factors' effect on speech understanding in noise develops in children is not well understood, particularly in the case of children with hearing loss. As a first step towards shedding light on these questions, we developed a paradigm that aims to recruit attention and working memory in a speech-in-noise task by requiring participants to switch or maintain attention to different speakers in a realistic scene. Here, we present the first set of data as a validation of the paradigm and discuss the implications of its results.

Keywords: *speech perception in noise, working memory, attention, ecological validity*

1. INTRODUCTION

The environments where everyday communication takes place are often rife with noises and other conversations in the background, which people learn to filter out in order to successfully engage their conversational partners. Thus, cognitive factors such as attention are undoubtedly involved in speech understanding in noisy situations. Yet, many hearing tests do not take this fact into account and simplify the processes required for successful hearing. To address this, we developed AVATAR (Audiovisual True-to-life Assessment of Auditory Rehabilitation), a system aiming to replicate in a controlled test environment the many circumstances that both aid and hinder speech understanding in realistic situations. In this paper, we will describe the functionalities and latest developments in this system, some of the challenges encountered and the solutions therefor, as well as present some pilot results from our upcoming studies. Lastly, we discuss the implications of such a system and the possibilities it grants us.

2. AVATAR

AVATAR was developed by Devesse and colleagues [1-4] to investigate the impact of multiple tasks on speech perception in noise (SPIN). It consists of a program which displays a computer-graphics environment with virtual human figures (see Fig. 1). The scene is projected on a large screen so that the humans appear almost life-sized. The speech material is presented by speakers positioned behind each figure on the screen, so that the sound source is spatially aligned with the visual of each human figure as they speak. The virtual humans can be made to animate their mouths realistically in synchrony with audio material being presented by a separate program. This paradigm has proven successful in eliciting audiovisual benefit [2]. In addition to the SPIN task, three other tasks can be added: a localization task, wherein a phone rings in one of the five possible speaker locations, and the participant is asked to indicate which direction the ringing had come from; a dynamic task, wherein the sound of a fly passing from left to right or from right to left is presented, and the participant is asked to indicate which direction it had gone; and a visual task, wherein participants were required to remember a number present on the scene and which changes between trials. Adding these secondary tasks in isolation or combined has been shown to cause a detrimental effect on

[*] ***Corresponding author***: *lyan.porto@kuleuven.be*

performance on the main SPIN task, even in young adults with typical hearing [1].

In order to investigate specifically how the presence of simultaneous speakers affects performance on a SPIN task, new functionalities had to be added to the original system. First, all studies originally used only female figures and voices, so changes were made to allow male and female models and corresponding speech materials to be presented at once. Second, the original system displayed several people on screen, but only one could speak at a time. Thus, changes were made to allow two talkers to produce speech simultaneously. And finally, the system was coupled with an eye-tracker (Pupil Core from Pupil Labs) to provide real-time insight into looking behavior and pupil size as the participant completes the task. This was done to provide measures of attention and listening effort, respectively [5, 6]. Below we present pilot data from an upcoming study using this updated version of the system to validate its suitability for answering the sorts of questions outlined here.

Figure 1. AVATAR – restaurant environment with five female models.

3. METHODS

3.1 Participants and set-up

Participants were three young adults with self-reported typical hearing recruited from local university students. The participant sits on a chair in front of a large screen where a projection shows a living room scene with two virtual humans, one male and one female, each shown sitting on a sofa at roughly ±37° azimuth. Behind the screen there are loudspeakers lined up with the mouths of the virtual humans on screen, about 140cm from the floor. The chair's height is adjusted so that the participant's eyes and ears are level with these loudspeakers. Noise made up of 20 overlapping recordings of men and women reading passages is presented from a speaker positioned directly above the participant's head at 65 dB A. Participants wear an eye-tracker during the entire procedure but are able to move freely in the chair; for the sake of brevity, eye-tracking data will not be discussed in this article.

3.2 Procedure

Testing consists of two phases. In the first, SRT50s are calculated for both male and female materials when each is presented from both the left side of the screen (-37° azimuth) and the right side of the screen (+37° azimuth). This is done twice for each gender × side combination, for a total of 8 blocks. Each block consists of 10 sentences from the LIST corpus [7-8], which participants are asked to repeat as closely as possible. Trials are marked as correct only if all keywords are repeated correctly. Blocks start presentation at -8 dB SNR and are adjusted in a one-up, one-down procedure depending on the participant's response, with correct responses increasing the SNR and incorrect responses lowering it. The average of the last 5 trials plus the 11th fictive trial for a block is taken as that block's resulting SRT50.

In the next phase, the two-talker phase, the same scene with the same virtual humans is presented on screen. The participant is instructed again to repeat the sentence heard as closely as possible, but is informed that the two speakers might speak partially over one another. There are four types of trials: target-only (TO), when the woman speaks alone, as in the previous phase; distractor-only (DO), when the man speaks alone, as in the previous phase; target-first (TF), when the woman speaks first but is partially covered by the man speaking; and distractor-first (DF), when the man goes first but is partially covered by the woman speaking. In all but distractor-only trials, the participant is asked to repeat what the woman has said; in distractor-only trials, they are instructed to repeat what the man has said. This is done to prevent the participant from ignoring the distractor completely; since the participant does not know at the beginning of the trial what type it will be, when the distractor begins speaking, the participant is hypothesized to allocate their attention to him, as it may be a distractor-only trial. If then the woman begins speaking, we expect the participant to switch their attention to her. The timing of the onset of the second speaker is adjusted depending on the duration of the female sentence for that trial, such that the target sentence is always half covered by the distractor sentences. To ensure that participants understand the task, a short familiarization block is completed with only four trials (one of each type) and no noise. In this phase, the levels do not vary adaptively and are set separately for each gender at

the best SRT values obtained in the previous phase. This is done to ensure that the task is challenging enough that participants are expected to shift their visual attention to look at the virtual human speaking. Two blocks of 20 trials are completed in this phase, each trial containing one target sentence from the LIST. Each block contains five trials of each type.

We expected TO and DO trials to yield performance of around 50%, given that they are presented at the level calculated to be SRT50% for each respective participant. In TF trials, requiring attention maintenance and inhibition of the distractor only, we expect participants to perform worse overall; in DF trials, which requires attention switching from distractor to target and then maintenance and inhibition of the distractor, should yield the worst performance across all conditions.

4. RESULTS

Each bar in Figure 2 below represents the average number of correct trials of each respective type out of 10 completed by each participant.

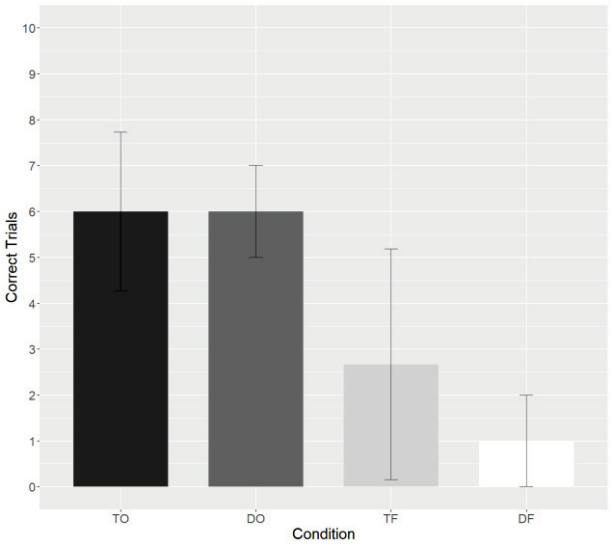

Figure 2. Means and SDs of the correct number of trials in each condition.

5. DISCUSSION

In conformity with our predictions, TO and DO trials yield the best performance, at 60% correct trials; seeing as the SNR used was that calculated as the SRT50% for each individual participant, it is somewhat surprising that these means are above 50%. Also as expected, both simultaneous conditions, TF and DF, have yielded lower performance. Moreover, the difference between TF and DF trials conform to our prediction that TF trials are "easier" than DF trials.

The larger variability in TF trials relative to DF trials may suggest differences in strategies between participants. As in these trials, participants are already attending the target sentence from the beginning, some participants opted not to wait for the distractor sentence to finish to provide their response, while others did not want to speak until the distractor was finished. Participants who chose the latter had to hold the target sentence in memory for a few seconds more while preventing the distractor sentence from interfering, which may have caused a decrease in performance. This is not seen in DF trials as the stimulus presentation finished with the target sentence being produced alone, so participants can begin producing their response immediately.

6. CONCLUSION

In spite of its small sample size, the latest set of data collected using this protocol provides an optimistic perspective on the usability of this method to determine the impact of attention switching and maintenance typical of real-life conversations on speech understanding in noise. Coupled with eye-tracking data and scores on standardized cognitive measures, we expect future data collected will provide an even greater insight into how cognition plays a role in everyday listening scenarios. Moreover, we believe the paradigm described here can be adapted to answer other questions that require greatly ecological validity in measurements. Specifically, our future studies will focus on how children and adults differ in their use of cognition to overcome the difficulties imposed by simultaneous talkers and background noise, as well as individuals with hearing loss.

7. ACKNOWLEDGMENTS

This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklowdoska-Curie Grant Agreement No. 860755 (Comm4CHILD project).

8. REFERENCES

[1] A. Devesse, A. van Wieringen and J. Wouters, "AVATAR Assesses Speech Understanding and Multitask Costs in Ecologically Relevant Listening

Situations," *Ear & Hearing,* vol. 41, no. 3, pp. 521-531, 2020.

[2] A. Devesse, A. van Wieringen and J. Wouters, "Age Affects Speech Understanding and Multitask Costs," *Ear & Hearing,* vol. 41, no. 5, pp. 1412-1415, 2020.

[3] M. I. Posner, "Orienting of Attention: Then and Now," *Q J Exp Psychol (Hove),* vol. 69, no. 10, pp. 1864-1875, 2016.

[4] A. Micula, J. Rönnberg, L. Fiedler, D. Wendt, M. C. Jørgensen, D. K. Larsen and E. H. N. Ng, "The Effects of Task Difficulty Predictability and Noise Reduction on Recall Performance and Pupil Dilation Responses," *Ear & Hearing,* vol. 42, pp. 1668-1679, 2021.

[5] A. Devesse, A. Dudek, A. van Wieringen and J. Wouters, "Speech intelligibility of virtual humans," *International Journal of Audiology,* vol. 57, no. 12, pp. 914-922, 2018.

[6] A. Devesse, A. van Wieringen and J. Wouters, "The Cost of Intrinsic and Extrinsic Cognitive Demands on Auditory Functioning in Older Adults With Normal Hearing or Using Hearing Aids," *Ear & Hearing,* vol. 42, no. 3, pp. 615-628, 2021.

[7] A. van Wieringen and J. Wouters, "LIST and LINT: Sentences and numbers for quantifying speech understanding in severely impaired listeners for Flanders and the Netherlands," *International Journal of Audiology,* pp. 348-355, 2008.

[8] S. Jansen, K. Raphael, J. Wouters and A. van Wieringen, "Development and validation of the Leuven Intelligibility Sentence Test with male speaker (LIST-m)," *International Journal of Audiology,* vol. 53, no. 1, pp. 55-9, 2014.