PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

The ability to allocate attentional resources to a memory task predicts speech-on-speech masking for older listeners

Frederick J. GALLUN[1,2]; Kasey M. JAKIEN[2]

[1] VA Portland Health Care System, Portland, OR, USA

[2] Oregon Health and Science University, Portland, OR, USA

ABSTRACT

To examine the degree to which attention and working memory tasks predict performance in complex auditory environments, 51 participants (21-77 yrs) with a range of pure-tone hearing thresholds completed both an auditory/visual working memory (WM) experiment and a competing speech task. WM tested auditory and visual memory both in a single-modality list recall task and in dual-modality selective and divided attention span tasks. The competing speech task used three closed-set sentences presented simultaneously via earphones using a Virtual Spatial Array, such that the target sentence was always at 0° azimuth angle and the maskers were either colocated or positioned at +-45°. Each condition was tested four times. WM performance under conditions of selective and divided attention was correlated with speech performance independently of age and hearing loss. These results suggest that the capacity to allocate attentional resources to sensory stimuli can help explain some of the variance in the ability to understand speech in complex acoustical environments.

Keywords: Attention, Auditory, Spatial

1. INTRODUCTION

In a situation where a listener must attend to one talker and ignore others, the question of where to put one's auditory and visual attention (and how that sensory information is stored and drawn from memory) is relevant to how well speech in noise is processed and perceived. The rise of Auditory Virtual Reality (AVR) has allowed researchers to simulate complex listening environments over headphones (1), which in turn has opened the possibility of testing many more listeners than could be done using only loudspeaker arrays. One consequence of this is that it is now possible to do clinical research on those individual and environmental variables that relate to the ability to process speech in complex listening environments, such as the "cocktail party" scenario (2). The focus of the study reported here is to better understand the ways in which the attention system underlies the ability to understand speech in complex environments.

Successfully reporting the information contained in a target sound presented in the presence of competition depends upon a series of operations, beginning with the encoding of the acoustical energy impacting the eardrum and ending with successful recall of the target information. When listening in environments with multiple sound sources, the ability to conduct all of the necessary operations can be impaired by a combination of internal and external factors. This study is a further exploration of the work that has previously indicated that working memory can explain some of the variability in speech understanding in competition observed with older and hearing impaired listeners (3,4). The methods are drawn from previously published work on attention and working memory (5) as well as methods developed to study speech on speech masking (6). Here the

operations of the working memory system are modeled as depending on the selective attention system, as proposed by Cowan (7).

The role of attention is important to consider because it is the "entry point" into the working memory system, through which stimuli are encoded for later analysis. It is hypothesized that some portion of the difficulties older and hearing impaired listeners report with communicating in noisy environments (8) are evidence of dysfunction in the attentional selection capacity of the working memory system. In order to ensure that working memory abilities and speech performance would vary among our participants, we tested individuals from 21-77 years of age and we allowed hearing ability to vary along with age. We hypothesized that if the cognitive operations tested in our working memory tasks are important predictors of performance, then we should be able to use those measures to improve upon the degree to which age and hearing loss are known to predict variations in speech understanding in the presence of competing stimuli.

2. METHODS

2.1 Participants

Fifty-one individuals participated (age range 21-77 yrs., average age 47.75 yrs.). All listeners had standard bilateral pure tone average (PTA standard) values (500 Hz, 1kHz, 2 kHz) from 0.83 dB HL to 40.83 dB HL (mean of 13.63 dB HL; standard deviation of 9.35 dB HL). High frequency PTA (PTA high) values (4kHz, 6kHz, 8kHz) ranged from -0.83 dB HL to 72.50 dB HL (mean of 23.97 dB HL; standard deviation of 20.43 dB HL). All listeners had fairly symmetrical hearing at 2 kHz and below (most had differences between the ears of less than 10 dB at all frequencies, and none had differences exceeding 20 dB). Participants were tested in a sound-attenuated booth at the National Center for Rehabilitative Auditory Research (NCRAR) in Portland, Oregon. Completion of testing took approximately 1 hour per participant. All participants were monetarily compensated for their time. All procedures were approved by the VA Portland Health Care System Institutional Review Board. The speech data were taken from a larger data set of 100 participants that was collected between 2012 and 2016 and subsets of which have been reported previously, but only with regard to the effects of age and hearing loss (6,9,10).

2.2 Working Memory Test

Participants were asked to report (in the order presented) a series of digits presented aurally via insert earphones and/or letters presented visually on a computer touchscreen located on a table in front of them. The auditory stimuli were recordings of the digits 1-9 spoken by a female talker. All were time-compressed or expanded in order to have a duration of exactly 500 ms and were presented at an RMS level 40 dB above the Speech Reception Threshold (SRT) measured for spondaic words presented in quiet. This level never exceeded 80 dB SPL. Letters were drawn from the set (ACEFHILOR) and were presented in 16 point font at a distance of approximately 45 centimeters in front of the participant.

Each trial involved the presentation of a sequence of letters and/or digits at a rate of one element per second. The visual stimuli appeared in white in the center of a black background and followed an orienting stimulus that was present for three seconds before the trial began. For the dual-modality conditions, the onset of each visual element was synchronized to the time at which each of the auditory elements began to play and were presented on the screen for the full one second duration. After the sequence was complete the participants were then asked to recall the series of numbers/digits and input their responses with the touchscreen. Responding was via a graphical-user interface in which keypads showing letters and numbers were displayed and each element selected appeared

at the top of the touchscreen. All stimuli were designed and implemented using MATLAB (Mathworks, Natick, MA).

Tasks were divided into single- and dual-modality conditions. In the auditory-only condition (AS), participants heard and then recalled a sequence of *digits* presented over headphones. In the visual-only condition (VS), participants recalled a sequence of *letters* presented visually. First sequences of three elements were presented and over time the sequence length increased, finally presenting as many as nine elements. Three trials were tested at each sequence length, and the run was stopped when the participant failed to correctly report all of the elements of at least one of three of the sequences in the correct order.

After completing the single-modality conditions (AS and VS), four dual-modality conditions were tested in interleaved blocks. Each consisted of simultaneously presented auditory and visual stimuli and the test modality was either cued (C) or uncued (U). Dual-modality conditions were thus divided into those in which participants knew in advance that they would report the auditory (CA) or visual (CV) stimuli and those in which participants were NOT informed in advance which modality they would be asked to report (UA and UV). The four conditions were presented in a randomly interleaved fashion, with sequences that started with three elements and progressed in steps of one to seven elements. No additional training was given. The cue that informed the observer which elements to report was a cartoon of an eye or an ear. This always appeared immediately after the sequences were finished and on C trials it also appeared as the fixation stimulus before the trial began.

Scoring was based on the total number of correctly identified elements, averaged across the number of conditions. Thus, if the total number of elements was the same regardless of the length of the sequence, then the total for each sequence length would equal the average. This method has the potential to obscure differences among participants due to combining performance for shorter and longer sequences. An alternative method commonly used is the maximum sequence length correctly reported. It has been shown, however, that composite scores such as this one are more reliable than scores that do essentially reflect performance on a single trial when performance was maximal (11). Future work should compare various metrics to determine if one more accurately predicts performance on the speech tasks with which this work is primarily concerned. All participants were able to accurately report six sequences of three practice elements before testing started, suggesting that all stimuli were audible and visible.

2.3 Spatial Release from Masking Test

Auditory Virtual Reality (AVR) based on binaural manikin Head-Related-Transfer-Functions (HRTFs) was utilized to present spatial conditions over headphones that accurately simulated the colocated and spatial separation between target and maskers that would be found in an anechoic chamber (12). AVR was created using the methods of Gallun et al. (6). Participants completed four runs of two spatial conditions in which a target sentence from the Coordinate Response Measure corpus (CRM) was at 0 degrees azimuth angle and two masking sentences from the CRM were either colocated with the target (at 0 degrees) or symmetrically separated from the target by +/- 45 degrees.

CRM sentences take the form of "Ready (CALLSIGN), go to (COLOR) (NUMBER) now" and consist of eight possible call signs: (Arrow, Baron, Charlie, Eagle, Hopper, Laker, Ringo, Tiger) and 12 keywords: four colors (red, green, white, and blue) as well as the numbers 1–8. For this test, three male talkers were used, speaking all 256 combinations

of the call signs, colors, and numbers. Participants were asked to identify the color-number combination after the callsign "Charlie".

Performance was estimated by testing two target sentences at each of ten target-to-masker ratios (TMRs) varying from 10 dB to -8 dB in 2 dB steps. Thresholds were estimated by an established method (6,9,10) in which the total number of sentences reported correctly, which can vary from 0 to 20, is subtracted from 10, and the result is the estimated threshold. This method has been shown to provide reliable thresholds that are highly repeatable across multiple repetitions of the test (9). The difference between thresholds for a given run has been termed spatial release from masking (SRM) and represents the improvement in threshold (in dB) associated with the addition of spatial separations among the three speech stimuli.

3. RESULTS

Data analysis was conducted using SPSS v25 (IBM, Armonk, NY). Average performance on all tests are shown in Table 1. TMR: Target-to-Masker Ratio; SRM: Spatial Release from Masking

Table 1 – Speech and Working Memory Task Performance

Task			Min	Max	Mean	SD
Speech		Colocated (TMR)	-1.25	4.00	1.57	1.20
		Separated (TMR)	-8.50	3.00	-4.36	2.75
		SRM (Colocated – Separated)	0.00	9.50	5.93	2.32
Working Memory						
Single Modality		Auditory Sequence Length	2.83	5.33	4.11	0.56
		Visual Sequence Length	2.75	5.11	3.73	0.59
Dual Modality	Cued	Auditory Sequence Length	2.27	5.00	4.27	0.63
		Visual Sequence Length	1.73	5.00	3.58	0.87
	Uncued	Auditory Sequence Length	2.20	4.87	3.55	0.68
		Visual Sequence Length	0.40	4.73	2.26	0.96

The correlations among these variables, as well as with age and hearing thresholds, are shown in Table 2. Values in bold type are significantly correlated at a level of $p < .01$ and the correlation accounted for a minimum of 10% of the shared variance.

Table 2 – Correlations Among Age, Hearing, Speech Performance, and Working Memory. Values in bold indicate p < 0.01

		PTA		Single Modality		Cued		Uncued	
	Age	Standard	High	Auditory	Visual	Auditory	Visual	Auditory	Visual
Age		**0.45**	**0.68**	**-0.42**	-0.37	**-0.41**	**-0.51**	-0.30	**-0.50**
PTA St	**0.50**		0.**76**	-0.34	-0.28	**-0.45**	**-0.46**	-0.30	-0.30
PTA High	**0.68**	**0.76**		**-0.29**	-0.19	**-0.38**	**-0.35**	-0.30	-0.29
Colocated	**0.53**	0.25	**0.37**	**-0.49**	**-0.44**	**-0.45**	**-0.52**	**-0.36**	**-0.57**
Separated	**0.52**	**0.65**	**0.62**	**-0.51**	**-0.45**	**-0.61**	**-0.68**	**-0.40**	**-0.52**
SRM	-0.34	**-0.63**	**-0.54**	0.35	0.30	**0.49**	**0.53**	0.28	0.32

To gain further perspective on the relative importance of these multiple correlations, three stepwise regressions were conducted in which each speech measure was entered as the dependent measure and all of the potential predictors were entered into the model on the basis of the strength of the correlation. Each successive predictor was only added if doing so resulted in a significant change in the predictive power of the regression model. The relationships between observed and predicted thresholds for the three speech measures are shown in Figure 1. In each case, the regression model had two significant factors. For the colocated condition (Figure 1, left panel), speech target-to-masker ratio (TMR) was predicted by a combination of the average sequence length in the uncued visual condition and age. For the separated condition (Figure 1, middle panel), TMR was predicted by the cued visual condition and PTA. Spatial release (the difference between TMRs in colocated and separated) is shown in right panel of Figure 1 and was predicted by PTA and average sequence length in the cued visual condition.

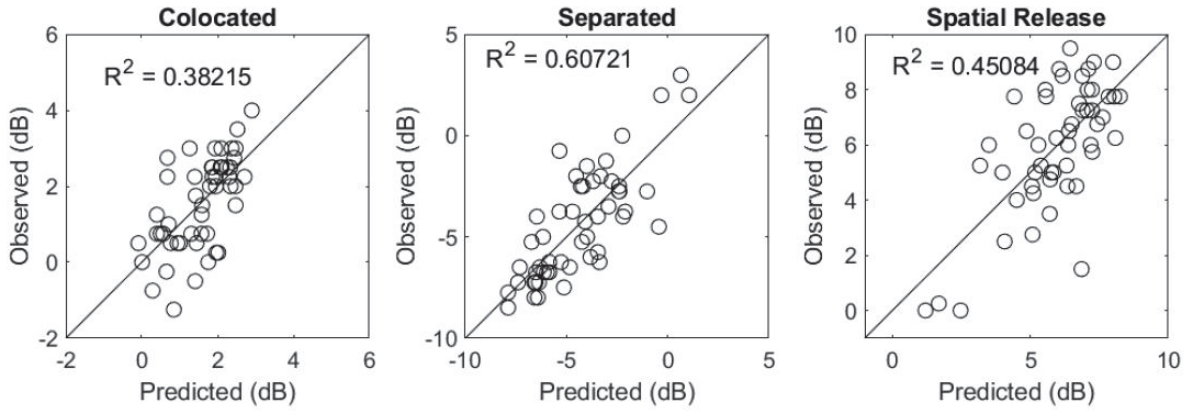

Figure 1 – Relationships between observed and predicted performance, based on the multiple regression models described in the text.

4. DISCUSSION

These results are consistent with the work of those who have suggested that speech understanding in the presence of competing speech depends not just on audiometric factors but also on cognition. It is interesting to note that the three speech conditions were

each predicted best by different factors, suggesting that there is different information to be gained from characterizing listening ability using all three metrics.

For the colocated condition, the finding that age and not audiogram predicts performance is consistent with the analysis of Jakien and Gallun (10), which was based on a slightly different subset of the same 100 listeners run in our laboratory between 2012 and 2016 (82 listeners who completed two runs of the speech tests rather than these 51 who completed four runs of the speech tests and also completed the WM tests). Including the WM performance reveals 1) that the WM task is even more important than age as a predictor, 2) that the dual-sensory conditions are more predictive than the single-sensory conditions, and 3) that the uncued visual condition was a better predictor than the uncued auditory or either of the cued dual-modality tasks.

The finding that the dual-modality task was a better predictor than the single-modality task suggests that the limitation on performance is not simply the ability to recall sequences of sensory input, but the ability to do so in the presence of competing information. Furthermore, the stronger relationship with the unknown modality test than the known modality test suggests that the task of selectively attending to a single talker in the presence of competing talkers when no spatial cues are present is dependent on the ability to store information in a form where it can be retrieved after a few seconds of interference rather than the ability to select one source of information in the presence of competition. It is important to note that age accounted for additional variance, suggesting that there are age effects unrelated to hearing loss or WM abilities that modulate performance on the colocated speech task.

Similar to previous analysis of this data set (9,10), the separated speech condition was predicted by hearing loss, but in this case PTA high was a better predictor than were the lower frequency PTAs. More importantly, the best predictor of performance was the dual-modality condition in which the visual stimulus was the target, which the participants knew in advance. This suggests that 1) WM in competition is important for performance in this task, as with the colocated condition, but that 2) it is the ability to select a sensory input in advance that limits performance, rather than the ability to store and recall sensory input. This is consistent with the increased reliance upon good peripheral hearing, which may provide cues to selecting one talker in the presence of competition at the lower TMR values at which the separated task can be performed (13).

Finally, the difference of the two thresholds, the spatial release measure, was predicted most strongly by PTA standard, with additional variance related to CV performance. The PTA result is similar to previous analyses of a different subset of this data set (6,9,10), but the finding that PTA is a stronger predictor than WM for spatial release has important implications. This suggests that by looking at the improvement in performance provided by adding spatial separation, the influence of WM has been reduced (but not eliminated). Furthermore, it is important to note the predictive ability of PTA standard in spatial release as the primary predictor when PTA high predicted performance in the separated condition. This difference suggests that it may be that although PTA high is important for the exact level of performance in the separated condition, that use of the spatial cues, which accounts for the difference between colocated and separated, is more strongly related to lower-frequency hearing ability. Future work should explore these and the other speculative explanations provided here prospectively now that the presence of strong relationships has been confirmed.

One finding that is important to mention is that average performance on the single-modality tasks was lower than the dual-modality tasks, although the maximum performance was higher for the single modality tasks. There is good reason to believe that the lower average performance on this single-modality tasks may reflect a learning or

familiarization effect as these conditions were tested first. For this reason, it would be useful to collect further data on tasks in which both the single and dual-modality tasks are tested in an interleaved fashion in order to verify the strength of the relationships between WM with and without interference and the speech tasks.

5. CONCLUSIONS

Across multiple studies, these spatial release tests have been found to be associated with age and hearing loss to various degrees (6,10,14,15). The importance of each has been found to vary with spatial separations tested, and, not surprisingly, the distribution of ages and hearing losses and their covariation have significant influences on the strengths of the relationships observed. This study represents a meaningful step forward in understanding the factors involved, and it is of substantial theoretical importance to note that age, which is strongly correlated with performance in the spatially separated condition, is no longer a factor when working memory and attention are taken into account. This suggests that at least some of the effects of age found in previous work are likely mediated by the cognitive processes tested in this study. It is also important to note, however, that in the colocated condition, age was still explaining additional variance when the cognitive factors had been considered. Future work should continue to explore these effects so that the factor of "age" can be replaced by a more precise (and hopefully actionable) descriptor of the mechanism(s) limiting performance.

ACKNOWLEDGMENTS

We are grateful to the many participants, both Veterans and non-Veterans, who volunteered their time to this study. Data collection was conducted by multiple members of the lab, including Sean Kampel, Meghan Stansell, Rachel Ellinger, and Nirmal Srinivasan. The work was supported by the Department of Veterans Affairs Rehabilitation Research and Development Service and the National Institutes of Health's National Institute for Deafness and Communication Disorders (R01 DC011828; R01 DC 015051). The work was supported with resources and the use of facilities at VA RR&D National Center for Rehabilitative Auditory Research, which is located at the VA Portland Health Care System. The contents of this article are the private views of the authors and should not be assumed to represent the views of the Department of Veterans Affairs or the United States Government.

REFERENCES

1. Wightman FL, Kistler DJ. Headphone simulation of free-field listening. I: Stimulus synthesis. J Acoust Soc Am. 1989 Feb 1;85(2):858–67.
2. Cherry EC. Some Experiments on the Recognition of Speech, with One and with Two Ears. J Acoust Soc Am. 1953 Sep 1;25(5):975–9.
3. Akeroyd MA. 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. 2008 Jan 1;47(sup2):S53–71.
4. Füllgrabe C, Moore BC, Stone MA. Age-group differences in speech identification despite matched audiometrically normal hearing: contributions from auditory temporal processing and cognition. Front Aging Neurosci. 2015;6:347.
5. Cowan N, Fristoe NM, Elliott EM, Brunner RP, Saults JS. Scope of attention, control of attention, and intelligence in children and adults. Mem Cognit. 2006 Dec 1;34(8):1754–68.
6. Gallun F, Diedesch A, Kampel S, Jakien K. Independent impacts of age and hearing loss on spatial release in a complex auditory environment. Front Neurosci. 2013;7:252.

7. Cowan N. Attention and Memory: An Integrated Framework. Oxford University Press; 1998. 342 p.
8. Souza P, Hoover E, Blackburn M, Gallun F. The Characteristics of Adults with Severe Hearing Loss. J Am Acad Audiol. 2018;
9. Jakien KM, Kampel SD, Stansell MM, Gallun FJ. Validating a Rapid, Automated Test of Spatial Release From Masking. Am J Audiol. 2017 Dec 12;26(4):507–18.
10. Jakien KM, Gallun FJ. Normative Data for a Rapid, Automated Test of Spatial Release From Masking. Am J Audiol. 2018 Sep 19;1–10.
11. Conway ARA, Kane MJ, Bunting MF, Hambrick DZ, Wilhelm O, Engle RW. Working memory span tasks: A methodological review and user's guide. Psychon Bull Rev. 2005 Oct 1;12(5):769–86.
12. Xie B. Head-related transfer function and virtual auditory display. J. Ross Publishing; 2013.
13. Best V, Thompson ER, Mason CR, Kidd G. An Energetic Limit on Spatial Release from Masking. J Assoc Res Otolaryngol. 2013 Aug 1;14(4):603–10.
14. Srinivasan NK, Jakien KM, Gallun FJ. Release from masking for small spatial separations: Effects of age and hearing loss. J Acoust Soc Am. 2016 Jul 1;140(1):EL73–8.
15. Jakien KM, Kampel SD, Stansell MM, Gallun FJ. Validating a Rapid, Automated Test of Spatial Release From Masking. Am J Audiol. 2017 Dec 12;26(4):507–18.