Auditory aspects of multisensory working memory are retrieved better than visual aspects

Tori Turpin[1], Işıl Uluç[1,2,*], Kaisu Lankinen[1,2], Fahimeh Mamashli[1,2], Jyrki Ahveninen[1,2]

[1] Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Charlestown, MA, USA
[2] Department of Radiology, Harvard Medical School, Boston, MA, USA

Corresponding author *:
Işıl Uluç, Ph.D.
CNY 149, 13th St.
A.A. Martinos Center for Biomedical Imaging,
Department of Radiology,
Massachusetts General Hospital,
Charlestown, MA 02129
iuluc@mgh.harvard.edu
Tel: 617 726 6584, Fax: (617) 726-7422

Abstract

Working memory (WM) reflects the transient maintenance of information in the absence of external input, which can be attained via multiple senses separately or simultaneously. Pertaining to WM, the prevailing literature suggests the dominance of vision over other sensory systems. However, this imbalance may be attributed to challenges of finding stimuli that are represented in comparable ways across modalities. Here, we addressed this methodological problem by using a balanced multisensory "retro-cue" WM design. The to-be-memorized stimuli consisted of combinations of auditory (ripple sounds) and visuospatial (Gabor patches) patterns, which have been shown to undergo similar transformations during WM encoding and retrieval. Using a staircase procedure, the auditory ripple velocities and spatial frequencies of Gabor patches were adjusted relative to each participant's just noticeable differences (JND) separately in each modality, before the main task. The task was to audiovisually compare the probes to the memorized items. In randomly ordered trials, the probe either fully matched or differed from the memory item auditorily, visually, or audiovisually. The participants correctly rejected a significantly larger number of auditory non-match probes than visual non-match probes. Our findings suggest that, in the case of inter-sensory competition during feature maintenance, auditory attributes of multisensory WM items can be retrieved more precisely than their visual counterparts when complexity of the content and task demands are bimodally equated.

Keywords: auditory working memory, visual working memory, audiovisual working memory, multisensory working memory, multimodal working memory, feature discrimination

Introduction

Working memory (WM) reflects our ability to encode input from the environment, store information in internal representations, and call upon representations to guide our behavior in future situations. In everyday settings, inputs accumulate through multiple sensory systems. Information gathered through different senses have been proposed to be maintained in modality-specific subsystems of WM (A. Baddeley, 1986; De Renzi & Nichelli, 1975; Quak, London, & Talsma, 2015). Of these subsystems, the visual WM has been widely considered to have the highest capacity and most precise performance.

Several lines of evidence support the view that human WM is dominated by the visual modality. For example, Cohen et al. (2009) examined whether the robust retention capabilities ascribed to visual memory are also found in the auditory domain. Using a range of complex, naturalistic sounds and images with equal memorability, they tested the recognition for sound clips, verbal descriptions, and matching pictures alone. Recognition accuracy in the experiments were consistently and substantially lower for sounds than visual scenes. Similarly, Bigelow and Poremba (Bigelow & Poremba, 2014) reported that human auditory WM performance was inferior to that of visual, specifically with retention periods longer than a few seconds. Comparable findings have been reported in behavioral studies in non-human primates (Colombo & D'Amato, 1986; Scott, Mishkin, & Yin, 2012), as well as in recent neurophysiological studies in humans (Wolff, Kandemir, Stokes, & Akyurek, 2020).

However, the marked modality differences in the precision and capacity could represent innate processing differences between auditory and visual WM (Li & Cowan, 2021; Noyce, Cestero, Shinn-Cunningham, & Somers, 2016; Penney, 1989). The traditionally strong emphasis on visual studies might have resulted in designs that favor memoranda in this modality over the others (Lehnert & Zimmer, 2008). For example, in the visual domain, one can encode multiple distinct objects in parallel (Luck & Vogel, 1997). In the auditory domain, WM objects need to be distributed across relatively long periods of time (Bizley & Cohen, 2013), which in itself could increase the memory load (A. D. Baddeley, Thomson, & Buchanan, 1975). Comparisons of performance using classic capacity measures, such as the number of briefly presented items that can be maintained in WM, could thus inherently favor the visual domain.

In the present study, we compared the fidelity of auditory and visual WM by using stimuli that reflect corresponding parameters in each modality, as motivated by previous attempts to directly compare parametric features of these modalities by utilizing simple, manipulable stimuli (Visscher, Kaplan, Kahana, & Sekuler, 2007) (**Fig. 1**). In the visual domain, we chose static Gaussian-windowed sinusoidal gratings, Gabor patches, a staple stimulus family used to probe visual WM. To guarantee comparable stimulus representations that reflect auditory WM, we utilized frequency-varied broadband sounds, referred to as dynamic ripple sounds, which are resistant to semantic contamination but possess spectrotemporal properties similar to human speech (Depireux, Simon, Klein, & Shamma, 2001). As the primary task, we used a retrospective cue (retro-cue) paradigm, which helps separate the account of actively maintained WM content from the stimulus history (Backer & Alain, 2012; Kumar et al., 2016; Lepsien & Nobre, 2006; Mamashli et al., 2021; Uluc, Schmidt, Wu, & Blankenburg, 2018). Most importantly, unlike the majority of studies that have utilized similar approaches, in our main Experiment 1 we compared the performance of auditory and visual WM in a multisensory design to allow modality comparisons

across the same trials. We hypothesized that memory for feature information in audition and vision would at least be equally reliable when systematically equated in content and complexity.

Methods

Experiment 1

Participants. The study protocol was performed in accordance with the Human Research Committee and the Institutional Review Board (IRB) of Massachusetts General Hospital. A total of 22 healthy adult volunteers (12 men, 10 women; self report metric) participated in the study for monetary compensation. Subjects self-reported no hearing deficits and normal or corrected-to-normal vision. Written informed consent was obtained from each participant prior to experiments. *Apparatus.* Stimuli were projected from a 24" Dell Model U2410 Monitor with 1920x1200 pixel resolution driven at 59 Hz. Viewing distance at eye-level was 50 cm in front of the subject. The display was affixed at the center of the visual field, subtending a 5.6° visual angle. All visual stimuli were presented in the center of a grey background. Head movements were stabilized with the aid of a chinrest. Sounds were presented binaurally through HD-280 supra-aural headphones (Sennheiser Electronic Corporation). Presentation of the experiment was controlled using Presentation Software (Neurobehavioral Systems, San Francisco, CA).

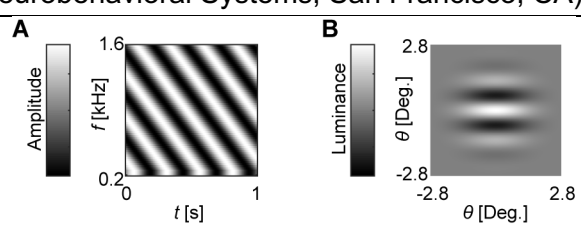

Figure 1. Auditory and visual WM items. (A) An Example for Auditory Stimulus. Dynamic ripple sounds refer to broadband (here, 0.2–1.6 kHz) sounds that are modulated across time (ripple velocity ω; measured in cycles per second) and frequency (spectral density Ω; measured in cycles/octave). (B) An Example for Visual Stimulus. Representations of a Gabor visual gratings plotted as a function of space measured in degrees of visual angle θ around fixation.

Auditory Stimuli. The dynamic ripple waveform sounds were created by layering 20 sinusoids per octave with frequencies ranging from 0.2 kHz to 1.6 kHz. The frequencies, *f*, upon which ripple velocities vary (at a rate of ω cycles per second) are represented by the equation: $s(g,t) = Do + D \cdot cos[2\pi(\omega t + \Omega g) + \Psi]$ where $g = \log(\frac{f}{f_0})$, t is time, Do is baseline intensity, D is modulation depth, ω is ripple velocity, Ω is spectral density, and Ψ is the phase of the ripple. At presentation, the auditory stimuli featured a fade in and out for 10 ms at the beginning and end of each stimulus. *Visual Stimuli.* The static Gaussian windowed, sinusoidal gratings (Gabor patch) visual stimuli were created by function $g = Acos(fz)$, where g is grating pattern, A is maximum contrast of the

image, f is spatial frequency (cycles per pixel), and z is rotated coordinates with angle θ, so that $z = xcos\theta + ysin\theta$, where x and y are the coordinates of the image. We used values $A = 0.2$, θ = 90 degrees and image resolution 257x257 (defining the range of x and y). The grating pattern was masked with a Gaussian function $m = Bexp - \left(\left(\frac{x}{2\sigma}\right)^2 + \left(\frac{y}{2\sigma}\right)^2\right)$, where B is a scaling factor and σ is standard deviation. We used values $B = 1$ and σ = 32 pixels (corresponding to 5.6° degrees around fixation). Furthermore, the image was scaled so that the maximum amplitude of the image was 1 and that the edge luminance was exactly 0.5 to blend the patch with the surroundings.

Preliminary Detection Thresholds. To ensure that the auditory and visual WM materials were identical in difficulty, stimulus-specific preliminary detection tasks controlling for individual baseline differences in sensory discrimination were administered. Subjects received separate tasks in the auditory and visual modalities prior to the memory experiment. Modality task order was counterbalanced according to parity of subject identification number. Just noticeable difference (JND) values were estimated using a 2-down, 1-up staircase algorithm separately in each individual participant. From these values, we generated modality-specific auditory and visual arrays of six stimuli, in which each stimulus was 1.5 JNDs away from its nearest neighbor. All main experiments' stimuli were drawn from these arrays. The classes 2–5 were used for the WM item stimuli, whereas all classes were used for the probes. Given that stimuli were individually tailored to each participant, baseline stimulus values (stimulus 1 in each array) were constant across the group, whereas subsequent stimulus values 2-6 were adapted for each participant. Three participants whose ripple stimulus classes included stimuli with ripple velocities above 500 cycles/s were excluded from further analysis (Supin et al., 2018).

Procedure. Behavioral experiments took place in a quiet, dimly lit room. The experimenter explained instructions prior to administering the task. During explanation, participants had the opportunity to ask for any clarifications they needed. Subjects participated in 10 practice trials and were left to complete the experiment. The retro-cue paradigm was split into six sequence blocks, each comprised of 48 trials, 288 trials in total.

Trials began with a “!” to denote a new sequence. Then, the two sets of concurrently audiovisual stimuli were presented in succession for 1 s each with a 350 ms offset-to-onset interstimulus interval (ISI) in between. This was followed by a 500-ms visual retro-cue (“1” or “2”) that instructed which item to retain in memory over a pseudorandomized subsequent maintenance period (either 2 or 10 s). Then, a probe stimulus was presented for 1 s, at which time participants were to respond to the probe. In the case of non-match probes, the probes' auditory and/or visual attributes were separated by one step along the feature axis, that is, 1.5 JNDs from the relevant item.

Audiovisual probe and memory items that were conceptually perceived as identical items were defined as a match (same or match trials). Conversely, non-match items were specified by three categories in which; sound features differed (auditory non-match or ANM trials), visual features differ (visual non-match or VNM trials), or both modality features differed (audiovisual non-match or NM trials). Unimodal non-match trials differed such that items were nonidentical in feature representation of auditory components or visual components. ANMs were represented by nonidentical ripple velocities, whereas VNMs were represented by nonidentical Gabor spatial frequencies. Audiovisual NM trials were bimodally nonidentical in stimulus features. Notably,

participants were naïve to partial unimodal match conditions as to promote conceptual congruency within the item. An equal number of match and non-match trials (with even distribution between the three non-match conditions) were presented in random order throughout the six sequence blocks. Four stimulus values were drawn from individualized pools of classes among each modality, resulting in four auditory and visual test features culminating in 16 possible audiovisual multifeatured combinations.

Response mapping was counterbalanced according to parity of subject number. Half of the participant pool indicated a same response with the right index finger press and a different response with the right middle finger press, and the other half of the subjects vice versa. If no response was given within a 1.5 s time window the monitor reminded the participant to respond until a response was received. A feedback was provided by three crosses "+++" to signify a correct response or three dashes "---" for an incorrect response. New trials commenced following an intertrial interval (ITI) of 1 s.

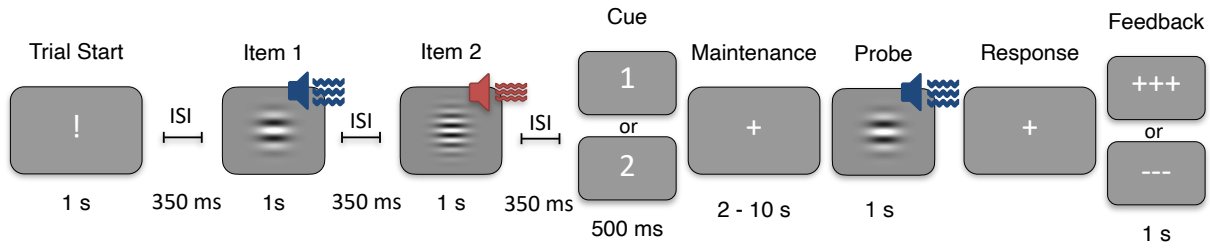

Figure 2. Multimodal Trial timeline. Experiment 1 sequences contained two audiovisual items, presented in a row, with parametrically modulated features. A brief cue indicated the to-be-memorized item (1 or 2), followed by a delay interval of 2 or 10 s. Then, a new bimodal probe was presented. Participants were to determine whether the cued item was the same or different from the probe. The match, ANM, VNM and NM trials as well as 2- and 10-s trials were randomized across the experiment. Retro-cue '1' and '2' trials were also randomzied and the retro-cue indicated the to-be-remembered item.

Quantification of Behavioral Performance. The study implemented a within-subjects, repeated measures experimental design. The primary dependent variable of interest was the percent of correct responses (PCR). The simple PCR measure was utilized because our primary purpose was to determine differences in correct rejection of ANM vs. VNM trials. It is worth noting that in our randomized design, all non-match conditions have a "common reference", the Match condition. Comparing more advanced sensitivity (e.g., d') and bias measures (Green & Swets, 1966) between ANM and VNM conditions would thus not provide any additional information, because due to this "common reference" these comparisons would be driven by differences in PCR only.

Statistical Analysis. All statistical analyses were conducted in R Statistical Software and RStudio (R. Team, 2020; R. D. C. Team, 2009). To test the effect of condition and standardized

discrimination thresholds on performance, we used a linear mixed effects model (LMEM) (*lmer* in R) for PCR and non-match modality conditions (Bates & Maechler, 2009). The LMEM predicted PCR by the fixed-effect of trial condition (ANM, VNM, NM, or Match) and the random effect of subject identity. More detailed comparisons between conditions were conducted by paired *t*-tests. To correct for multiple comparisons, we adjusted the *p* values of these *t*-tests by assessing the false discovery rate (FDR) (Benjamini & Hochberg, 1995).

Experiment 2

As control, we performed separate unimodal WM experiments for auditory and visual sensory modalities. Components of the methods and analyses are adapted from Experiment 1. The only adjustments are explained in this section.

Participants. A total of 7 healthy adult volunteers (2 men, 5 women; self report metric) participated in the study. Subjects self-reported no hearing deficits and normal or corrected-to-normal vision. Written informed consent was obtained from each participant prior to experiments. All procedures were approved by the Institutional Review Board (IRB) of Massachusetts General Hospital.

Apparatus, stimuli, and procedure. Experiment 2 is closely in accordance with the procedures in Experiment 1 with slight changes (see Fig. 3). Multisensory items were separated by modality into visual and auditory domains, thus memory for either Gabor or ripple stimuli were tested rather than combined audiovisual stimuli. Each trial began with two sample stimuli in a row, which was followed by a cue to denote the relevant stimulus according to serial position. A randomized short or long retention interval ensued, after which a probe was either identical (match trial) or different (non-match trials) to the relevant, cued stimulus. The order in which task types were administered (auditory, visual; or visual, auditory) was counterbalanced across participants. The unimodal retro-cue task featured 24 trials across six blocks amounting to 144 trials in total.

Statistical analysis. Statistical analysis was conducted analogously to Experiment 1.

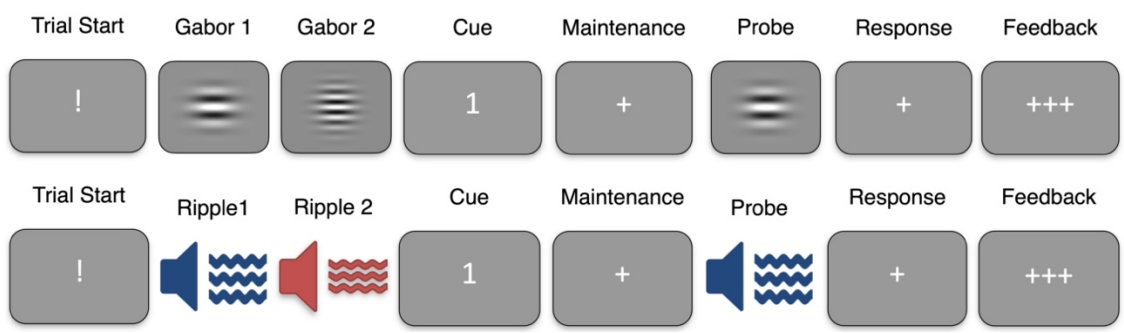

Figure 3. Unimodal Trial timelines, Experiment 2. The sequences reflect sample retro-cue trials that differ from Experiment 1 only by domain separation. Visual and auditory presentations in

unimodal tasks are displayed in the first and second rows, respectively. Visual trials tested memory for visual Gabors while auditory trials tested memory for ripple sounds. Experiment timing and stimulus durations are matched exactly to Experiment 1.

Results

Experiment 1

We examined the behavioral performance of 22 healthy neurotypical participants with a reto-cue working memory paradigm with multisensory auditory ripple velocity and visual spatial frequency content. In this study, we refer to the two sets of audiovisual stimuli presented prior to the delay as “items”, and the subsequent audiovisual reference stimuli to test for a match as “probes”. Results suggest that subjects’ ability to detect auditory non-matches (ANM) outperforms their detection of comparable visual spatial non-matches (VNM) between audiovisual probes and maintained audiovisual WM items.

Main analysis. **Figure 4A** shows the means and standard errors of means (SEM) of PCR in all four conditions. Our LMEM showed a highly significant main effect of task condition on PCR ($F_{3,63}$ = 40.8; p < 0.001). According to the subsequent FDR-corrected comparisons across individual conditions, PCR was significantly smaller in VNM trials than ANM trials in this comparison ($p_{\text{Corrected}}$ < 0.001; for detailed parameters, see **Table 1**). In addition, we also found that when compared to the congruent NM trials, PCR was significantly reduced in both ANM ($p_{\text{Corrected}}$ < 0.01) and VNM trials ($p_{\text{Corrected}}$ < 0.001; **Table 1**). Not surprisingly in the light of these results, PCRs in both ANM and VNM trials were also significantly smaller than in the Match trials ($p_{\text{Corrected}}$ < 0.001 in both cases; **Table 1**). However, the PCR differences in NM trials vs. Match trials did not reach statistical significance.

Effect of retention delay. Group mean PCR values across the two delay conditions are depicted in **Fig. 4B**. According to our FDR-corrected *t*-test, PCR was significantly lower in the 10-s than in the 2-s condition ($p_{\text{Corrected}}$ < 0.001; **Table 1**). More detailed information of the means and standard deviations of PCR values in each task and delay condition are itemized in **Table 2**.

Effect of cue set position. The means and SEMs of PCR for the two different reto-cue conditions are shown in **Figure 4C**. There was a trend toward reduction of PCR across all trials where the subject was cued to maintain item 1, but this effect did not reach statistical significance (**Figure 4C**).

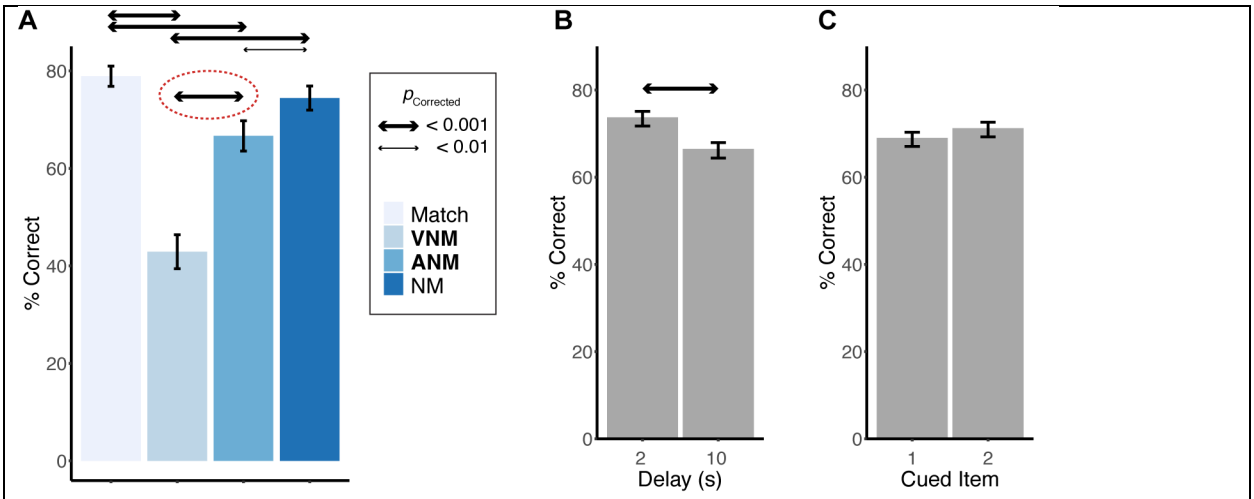

Figure 4. Percent of correct responses (PCR) in the bimodal WM Experiment 1. (A) Comparison of PCR across task conditions. Most interestingly, PCR is significantly smaller in the VNM than ANM condition (two middle bars). The details of statistical parameters are shown in Table 1. (B) PCR as a function of retention delay length. PCR was significantly smaller in the longer than shorter delay conditions. (C) PCR as a function of retro-cue type.

Experiment 2

As a control analysis, we applied the retro-cue procedure to modalities in isolation to test direct comparisons of unimodal working memory.

Main analysis. **Figure 5A** demonstrates the comparison of the group means and SEMs of PCRs across all unimodal auditory and visual trials. No significant group differences were observed in the comparisons of PCRs of auditory and visual conditions when analyzed across all delays and cue conditions (**Table 3**).

Effect of Retention Interval. Group means and SEMs of auditory and visual PCR values at the two retention delays are shown in **Figure 5B**. The results of the respective FDR-corrected *t*-tests comparing PCR values across the two retention delays and unimodal task conditions are shown in **Table 3**. A significant reduction of PCR was observed across the 2- and 10-s retention conditions in the unimodal auditory experiment ($p_{Corrected} < 0.05$; detailed parameters in **Table 3**), but in the visual experiment, this effect remained statistically non-significant. However, as shown by the data in **Figure 5B** and the statistical parameters reported in **Table 3**, there were no significant differences between auditory and visual PCRs at either retention delays.

Effect of cue position. Group means and SEMs of auditory and visual PCR values as a function of the cued item are shown in **Figure 5C**. There were no significant differences in PCR values across trials where the participant was cued to memorize the first or second item in either auditory or visual control experiments.

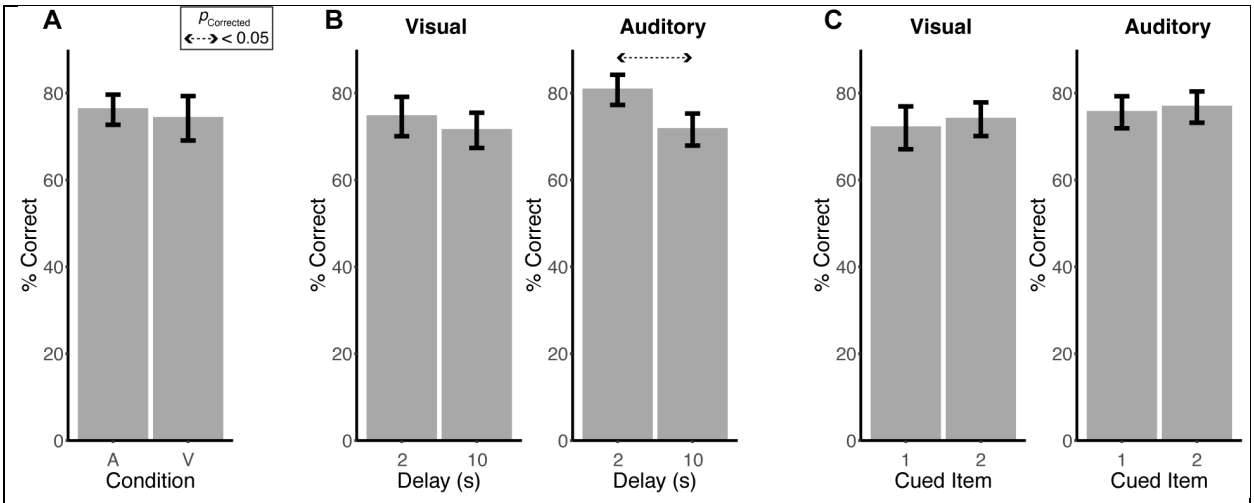

Figure 5. Percent of correct responses (PCR) in the unimodal WM Experiment 2. (A) Comparison of PCR across task conditions. No significant differences were observed across auditory and visual unimodal task performance overall. (B) PCR as a function of retention delay length, separately for the visual and auditory conditions. (C) PCR as a function of the cued item in the auditory and visual control task.

Discussion

We examined the working memory performance on sets of non-conceptual audiovisual stimuli. Our main purpose was to compare trials in which the memorized item's 1) auditory, 2) visual, and 3) audiovisual properties differed from the probe. To compare the different modalities on an equal footing, the auditory and visual stimulus sets were homogenized in discriminability by using comparable discrimination threshold tests (Mamashli, et al., 2021; Visscher, et al., 2007). Using this balanced design, we show that subjects' WM for auditory ripple velocity content outperforms visual frequency content, as quantified by the participants' ability to detect partial non-matches in the probes vs. maintained items. These modality differences were absent in our second experiment, which compared WM performance for the same auditory and visual stimuli in a unimodal design.

The retro-cue audiovisual paradigm of our main experiment required the participants to encode bi-sensory information into WM, actively maintain two intermodally competing memories, and then retrieve the relevant WM features for comparison with a bimodal probe stimulus. Adjacent cognitive processes such as divided attention and change detection are intrinsically recruited and utilized during these processing stages (Cowan, 1988; Lavie, 2005). One explanation for the subjects' better performance in rejecting ANM than VNM probes could thus be attributed to the multisensory load of maintaining multiple sensory items at one point in time. Given the limited capacity nature of WM, it is possible that intersensory competition for limited storage elicited cross-modal interference effects and led to disadvantage with visual items, which have been reported to be susceptible to interruption in memory and attention (Li & Cowan, 2021).

In a similar vein, the results of classic psychophysical studies, which compared the behavioral effects of attention cues, suggest that in a multimodal design auditory stimuli tend to be more alerting than visual stimuli (Harvey, 1980). It is thus possible that the present ANM probes resulted in stronger feedforward-type activation of the decision-making circuits needed for responding to the retrieval task than their VNM counterparts (Harvey, 1980; Posner, Nissen, & Klein, 1976; Turatto, Benso, Galfano, & Umilta, 2002). This line of thinking can be supported by recent psychophysical studies, which suggest that auditory change detection is less reliant on sustained selective attention than visual change detection (Demany, Semal, Cazalets, & Pressnitzer, 2010). Such an auditory superiority of bottom-up type change detection has been attributed to an implicit (sensory) memory buffer, which might be considerably weaker in the visual domain (Demany, et al., 2010). Although the present retro-cueing task is believed to control for effects explainable by sensory buffers alone (Lepsien & Nobre, 2006), these attention-independent change detection systems could have facilitated rejection of ANM probes under inter-sensory competition, with the focus of selective attention being shifted back and forth between auditory and visual VM representations.

Modality differences in WM performance could arise from methodological challenges in matching the relative difficulty of auditory vs. visual materials (Ward, Avons, & Melling, 2005). Here, the matching of intermodal difficulty was therefore enabled via a threshold test that was similar for auditory and visual stimuli. As discussed in previous studies from which the present stimuli were adapted, such JND estimates could still have become easier for the auditory than visual stimuli (Visscher, et al., 2007). Such a bias could follow from that the participants are, inherently, more familiar in comparing stimuli in one modality vs. another. To control for these kinds of biases, we

therefore implemented unimodal paradigm (Experiment 2) in a way that eliminates the attentional costs of cross-modal competition in Experiment 1. In the unimodal comparisons in Experiment 2, no significant differences were found in the PCR in auditory and visual WM performance overall. This result is, in principle, consistent with previous intermodal comparisons using unimodal designs, which suggest that with matched feature content and difficulty, performance of auditory and visual WM follow relatively similar functional principles (Visscher, et al., 2007; Ward, et al., 2005).

As expected, with all bimodal probe types considered, PCR of WM performance decreased significantly as a function of maintenance delay duration in Experiment 1. In the control unimodal Experiment 2, this effect was significant only in the auditory runs, but a slight trend towards a similar effect was also observable in the visual runs. This could be seen as supporting the trend observed in many other studies where the retention of auditory items has decreased more rapidly as a function of increasing delay than visual retention (Bigelow & Poremba, 2014; Jensen, 1971). However, it is worth noting that in the comparison at individual delay conditions, there were no significant differences in PCR between the two modalities at either the shorter or longer delay conditions in our unimodal Experiment 2. Finally, no statistically significant differences were observed in PCR between trials where the participant was retrospectively cued to maintain the first vs. second item in either Experiment. This supports the interpretation that the participants' performance was based on WM instead of passive sensory aftereffects in all task conditions.

Conclusion

Our results demonstrate that auditory attributes of multisensory WM items can be retrieved more precisely than their visual counterparts during inter-sensory competition. No modality differences were observed between auditory and visual WM performance in our unimodal control task. These results could provide a new perspective to the ongoing debate on modality differences in human WM performance.

Acknowledgments

This work was supported by the NIH grants R01DC016915, R01DC016765, R01DC017991, and P41EB015896.

Table 1. Statistical parameters of FDR-corrected t-test results in Experiment 1 (*** $p < 0.001$, ** $p < 0.01$).

Comparison	*t*	*d.f.*	FDR-corrected *p*		Diff. means	CI 95% Low	CI 95% High
ANM vs. VNM	5.2	21	0.000	***	23.8	14.3	33.2
ANM vs. NM	-3.1	21	0.008	**	-7.8	-13.0	-2.5
ANM vs. Match	-4.0	21	0.001	***	-12.2	-18.5	-6.0
VNM vs. NM	-8.2	21	0.000	***	-31.5	-39.5	-23.5
VNM vs. Match	-8.3	21	0.000	***	-36.0	-45.0	-27.0
NM vs. Match	-1.9	21	0.074		-4.5	-9.4	0.5
Delay 2 s vs. 10 s	4.8	21	0.000	***	0.07	0.04	0.10
Cued Item 1 vs. 2	-2.1	21	0.057		-0.02	-0.04	0.00

Table 2. Group Mean PCR (left) and SD (right) by condition in Experiment 1.

	Overall		Retention Interval			
			2 s		10 s	
	Mean	SD	Mean	SD	Mean	SD
MA	78.91	9.45	89.65	7.80	66.86	14.68
VNM	42.9	15.95	34.09	14.31	51.70	20.82
ANM	66.67	14.21	64.20	17.35	69.13	15.42
NM	74.43	11.37	73.30	10.78	75.57	14.45

Table 3. Statistical parameters of FDR-corrected t-test results in Experiment 2 (* $p < 0.05$).

Comparison	*t*	*d.f.*	FDR-corrected *p*		Diff. means	CI 95% Low	CI 95% High
A vs. V	1.0	6	0.611		0.03	-0.03	0.09
A Delay 2 vs. 10 s	5.0	6	0.023	*	0.09	0.05	0.14
V Delay 2 vs. 10 s	2.1	6	0.359		0.03	-0.01	0.07
A vs. V, 2 s	1.7	6	0.400		0.06	-0.03	0.15
A vs. V, 10 s	0.1	6	0.952		0.00	-0.08	0.08
A Cue 1 vs. 2	-0.5	6	0.707		-0.01	-0.07	0.05
V Cue 1 vs. 2	-0.7	6	0.632		-0.02	-0.09	0.05
A vs. V, Cue 1	0.9	6	0.611		0.04	-0.06	0.13
A vs. V, Cue 2	1.0	6	0.611		0.03	-0.04	0.10

References

Backer, K. C., & Alain, C. (2012). Orienting attention to sound object representations attenuates change deafness. *J Exp Psychol Hum Percept Perform, 38*(6), 1554-1566. doi: 10.1037/a0027858

Baddeley, A. (1986). *Working memory*. New York, NY, US: Clarendon Press/Oxford University Press.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior, 14*(6), 575-589. doi: https://doi.org/10.1016/S0022-5371(75)80045-4

Bates, D. M., & Maechler, M. (2009). lme4: Linear mixed-effects models using S4 classes.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, 57*, 289-300.

Bigelow, J., & Poremba, A. (2014). Achilles' Ear? Inferior Human Short-Term and Recognition Memory in the Auditory Modality. *PLoS One, 9*(2), e89914. doi: 10.1371/journal.pone.0089914

Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nat Rev Neurosci, 14*(10), 693-707. doi: 10.1038/nrn3565

Cohen, M. A., Horowitz, T. S., & Wolfe, J. M. (2009). Auditory recognition memory is inferior to visual recognition memory. *Proc Natl Acad Sci U S A, 106*(14), 6008-6010. doi: 10.1073/pnas.0811884106

Colombo, M., & D'Amato, M. R. (1986). A comparison of visual and auditory short-term memory in monkeys (Cebus apella). *Q J Exp Psychol B, 38*(4), 425-448.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychol Bull, 104*(2), 163-191.

De Renzi, E., & Nichelli, P. (1975). Verbal and non-verbal short-term memory impairment following hemispheric damage. *Cortex, 11*(4), 341-354. doi: S0010-9452(75)80026-8 [pii]

10.1016/s0010-9452(75)80026-8

Demany, L., Semal, C., Cazalets, J. R., & Pressnitzer, D. (2010). Fundamental differences in change detection between vision and audition. *Exp Brain Res, 203*(2), 261-270. doi: 10.1007/s00221-010-2226-2

Depireux, D. A., Simon, J. Z., Klein, D. J., & Shamma, S. A. (2001). Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *Journal of Neurophysiology, 85*(3), 1220-1234.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory annd psychophysics*. New York: Wiley.

Harvey, N. (1980). Non-informative effects of stimuli functioning as cues. *Quarterly Journal of Experimental Psychology, 32*(3), 413-425. doi: 10.1080/14640748008401835

Jensen, A. R. (1971). Individual differences in visual and auditory memory. *Journal of Educational Psychology, 62*(2), 123-131.

Kumar, S., Joseph, S., Gander, P. E., Barascud, N., Halpern, A. R., & Griffiths, T. D. (2016). A Brain System for Auditory Working Memory. *J Neurosci, 36*(16), 4492-4505. doi: 10.1523/JNEUROSCI.4341-14.2016

Lavie, N. (2005). Distracted and confused?: selective attention under load. *Trends Cogn Sci, 9*(2), 75-82. doi: 10.1016/j.tics.2004.12.004

Lehnert, G., & Zimmer, H. D. (2008). Modality and domain specific components in auditory and visual working memory tasks. *Cogn Process, 9*(1), 53-61. doi: 10.1007/s10339-007-0187-6

Lepsien, J., & Nobre, A. C. (2006). Cognitive control of attention in the human brain: insights from orienting attention to mental representations. *Brain Res, 1105*(1), 20-31. doi: 10.1016/j.brainres.2006.03.033

Li, Y., & Cowan, N. (2021). Attention effects in working memory that are asymmetric across sensory modalities. *Mem Cognit, 49*(5), 1050-1065. doi: 10.3758/s13421-021-01142-9

10.3758/s13421-021-01142-9 [pii]

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*(6657), 279-281. doi: 10.1038/36846

Mamashli, F., Khan, S., Hamalainen, M., Jas, M., Raij, T., Stufflebeam, S. M., . . . Ahveninen, J. (2021). Synchronization patterns reveal neuronal coding of working memory content. *Cell Rep, 36*(8), 109566. doi: 10.1016/j.celrep.2021.109566

Noyce, A. L., Cestero, N., Shinn-Cunningham, B. G., & Somers, D. C. (2016). Short-term memory stores organized by information domain. *Atten Percept Psychophys, 78*(3), 960-970. doi: 10.3758/s13414-015-1056-5

10.3758/s13414-015-1056-5 [pii]

Penney, C. G. (1989). Modality effects and the structure of short-term verbal memory. *Memory & Cognition, 17*(4), 398-422. doi: 10.3758/BF03202613

Posner, M. I., Nissen, M. J., & Klein, R. M. (1976). *Psychological Review, 83*(null), 157.

Quak, M., London, R. E., & Talsma, D. (2015). A multisensory perspective of working memory. *Front Hum Neurosci, 9*, 197. doi: 10.3389/fnhum.2015.00197

Scott, B. H., Mishkin, M., & Yin, P. (2012). Monkeys have a limited form of short-term memory in audition. *Proc Natl Acad Sci U S A, 109*(30), 12237-12241. doi: 10.1073/pnas.1209685109

Team, R. (2020). RStudio: Integrated Development for R. Boston, MA: RStudio, PBC. Retrieved from http://www.rstudio.com/

Team, R. D. C. (2009). R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Retrieved from www.R-project.org

Turatto, M., Benso, F., Galfano, G., & Umilta, C. (2002). Nonspatial attentional shifts between audition and vision. *J Exp Psychol Hum Percept Perform, 28*(3), 628-639. doi: 10.1037//0096-1523.28.3.628

Uluc, I., Schmidt, T. T., Wu, Y. H., & Blankenburg, F. (2018). Content-specific codes of parametric auditory working memory in humans. *Neuroimage, 183*, 254-262. doi: 10.1016/j.neuroimage.2018.08.024

Visscher, K. M., Kaplan, E., Kahana, M. J., & Sekuler, R. (2007). Auditory short-term memory behaves like visual short-term memory. *PLoS Biol, 5*(3), e56. doi: 10.1371/journal.pbio.0050056

Ward, G., Avons, S. E., & Melling, L. (2005). Serial position curves in short-term memory: functional equivalence across modalities. *Memory, 13*(3-4), 308-317. doi: 10.1080/09658210344000279

Wolff, M. J., Kandemir, G., Stokes, M. G., & Akyurek, E. G. (2020). Unimodal and Bimodal Access to Sensory Working Memories by Auditory and Visual Impulses. *J Neurosci, 40*(3), 671-681. doi: 10.1523/JNEUROSCI.1194-19.2019