Improving Visual Memory with Auditory Input

Scott R. Schroeder (schroeder@u.northwestern.edu)
Department of Communication Sciences and Disorders, 2240 Campus Drive
Evanston, IL 60208 USA

Viorica Marian (v-marian@northwestern.edu)
Department of Communication Sciences and Disorders, 2240 Campus Drive
Evanston, IL 60208 USA

Abstract

Can input in one sensory modality strengthen memory in a different sensory modality? To address this question, we asked participants to encode images presented in various locations (e.g., a depicted dog in the top left corner of the screen) while they heard spatially *un*informative sounds. Some of these sounds matched the image (e.g., the word "dog" or a barking sound) while others did not. In a subsequent memory test, participants were better at remembering the locations of images that were encoded with a matching sound, even though these sounds were spatially uninformative – an effect that was mediated by whether the sounds were verbal or non-verbal. Because the sounds did not provide any relevant location information, better spatial memory cannot be attributed to auditory memory; rather, it is attributed to visual memory being strengthened by the matching auditory input. These findings provide the first behavioral evidence for cross-modal interactions in memory.

Keywords: Audio-Visual Integration; Memory; Multisensory Processing; Visual Spatial Memory

Introduction

We live in a multi-sensory world, where auditory, visual, and other sensory inputs merge together to form our experiences, many of which we later try to remember. For example, consider the experience of parking your car in an unfamiliar place and then later trying to find it. After parking your car, you will likely try to commit the location of the car to memory. At the same time as you are memorizing the car's location and preparing to walk away, you might lock the car's doors, and the car might make a sound to indicate that the doors have been successfully locked. Does hearing this auditory input (the sound of the car) help you encode and then later remember the important visual input (the car's location in space)? More generally, can input in one sensory modality strengthen episodic memory in a different sensory modality?

The answer to this question has basic-science implications for our understanding of how memory works in real-world, ecologically-valid contexts, as our everyday experiences are largely multi-sensory in nature. The answer may also have applied-science implications for educational programs, cognitive therapies, and human factors designs, as multi-sensory input might improve memory performance.

Despite its relevance to basic- and applied-science and despite evidence for cross-modal interactions in other cognitive domains, there is, to our knowledge, no persuasive behavioral evidence for cross-modal interactions in episodic memory. In several previous studies, hearing congruent auditory input during the encoding of a visual image (for example, hearing a dog's bark while encoding a picture of a dog) was found to aid later recognition of the previously-viewed visual image (the picture of the dog) (Lehmann & Murray, 2005). However, the finding that congruent auditory input improves visual item memory (also known as visual *what* memory) might not reflect auditory input changing visual memory. When participants were presented with a visual image (a picture of a dog) on the recognition memory test, they may have remembered hearing the image's congruent sound (the barking sound of a dog), which could be used to correctly indicate that they had previously seen the image, even if the participant forgot the image. Thus, the use of the helpful auditory memory trace may have led to better visual memory performance, even if hearing the auditory input did not actually strengthen visual memory.

To determine if auditory input can truly strengthen visual memory, we created a multi-sensory audio-visual memory task in which memory for the auditory input itself could not help participants perform the visual memory test. Specifically, we presented to-be-encoded visual objects in various spatial locations on the screen (for example, an image of a dog placed in the top left corner of the screen), along with spatially *un*informative auditory cues (for example, a barking sound played to both ears and thus not linked in any way to the location of the image). Then, in a later memory test, we assessed visual spatial memory (also known as visual *where* memory) for the previously-seen images. If *where* memory performance for a visual image is improved when the image is encoded with a spatially uninformative but congruent auditory cue (relative to a control condition), it would have to reflect better visual memory per se. It could not reflect the use of auditory memory to help participants perform the *where* memory test because the auditory memory trace does not contain any relevant location information and therefore would not help spatial memory performance. Thus, better visual *where* memory performance in this task would provide evidence that input in one modality can strengthen episodic memory in a different modality.

A cross-modal effect of auditory input on visual memory might depend on the type of auditory input. Verbal sounds

(i.e., spoken words like "dog") and non-verbal sounds (i.e., environmental sounds like a dog barking) are known to affect visual processing differently (Chen & Spence, 2011; Edmiston & Lupyan, 2015). Environmental sounds and spoken words differ in the location information they provide during visual spatial processing. Environmental sounds (such as a barking dog) provide helpful information about the location of the relevant object (the dog). In contrast, spoken labels (such as "dog") provide location information about the speaker but not about the relevant object (i.e., the dog), as the word "dog" can be uttered irrespective of the specific location (or even presence) of the dog (a feature of language known as displacement). Because environmental sounds and spoken words differ in their spatial informativeness for relevant objects, they may have different effects on visual *where* memory.

In the current study, we examined the effects of environmental sounds and spoken words on visual *where* memory (as well as visual *what* memory). In an environmental sounds experiment and a spoken words experiment, participants encoded a series of visual objects (for example, a dog) located in one of the four corners of the screen while hearing task-irrelevant, spatially uninformative auditory cues. The auditory cues were either *congruent* with respect to the visual object (the sound of a dog barking while seeing a dog), *incongruent* with respect to the object (the sound of a motorcycle's exhaust while seeing a trumpet), or *neutral* with respect to the object (a semantically-meaningless beep sound while seeing a helicopter). Participants then performed an item (*what*) and spatial (*where*) memory task to test memory for what pictures they saw and where they saw them. If hearing a congruent auditory cue were shown to help *what* memory performance relative to the neutral control condition, this finding would replicate previous research and demonstrate the benefits of having two sensory memory traces (or dual-codes) for memory performance; however, it would not provide evidence for cross-modal effects. The crucial test for cross-modal effects is *where* memory performance. A finding of better *where* memory for the congruent condition relative to the neutral control condition would provide evidence for cross-modal effects in memory.

Methods

Participants

Forty English-speaking young adults (median age = 21.5 years; 32 females, 8 males) were included in the study. Participants received monetary compensation or course credit for their participation. The experiment was approved by the Northwestern University Institutional Review Board.

Materials and Procedure

Participants completed the environmental sounds experiment and the spoken words experiment in a counterbalanced order, such that 20 participants completed the environmental sounds experiment first while the other 20 participants completed the spoken words experiment first. None of the auditory or visual stimuli that appeared in the first experiment also appeared in the second experiment.

Environmental Sounds and Visual Memory Experiment

Encoding. Participants viewed 60 pictures during the encoding task. In the set of 60 pictures, 20 were presented with their *congruent* environmental sound (i.e., the sound associated with that object), 20 with an *incongruent* environmental sound (i.e., the sound associated with a different object), and 20 with a *neutral* control sound (i.e., one of twenty tonal beep sounds). A semantically neutral sound (a tonal beep) was used for the control condition instead of no sound. The reason for using a semantically neutral sound as the control rather than no sound is that, in no-sound trials there would be a matching context between encoding and retrieval (i.e., both contexts would be silent), whereas in the congruent and incongruent trials, there would be a mismatching context between encoding and retrieval (i.e., there would sound at encoding but not at retrieval). The degree of match between encoding and retrieval context affects memory performance (Smith & Vela, 2001). By using a neutral sound (a tonal beep), the change in context from encoding to retrieval (a sound at encoding and no sound at retrieval) is the same for all trials, be they congruent, neutral, or incongruent.

To ensure that potential differences in the memorability of the pictures across congruent, neutral, and incongruent conditions did not affect the results, we created the stimuli in the following way. Four lists of 20 picture-sound pairs were compiled. Each of the four lists served in one of four positions – (1) as the 20 picture-sound pairs in the congruent trials, (2) as the 20 pictures in the neutral trials, which were paired with a tonal beep sound, (3) as the 20 pictures in the incongruent trials, or (4) as the 20 sounds in the incongruent trials. The four lists rotated, serving in all four positions an equal number of times across participants. The benefit of creating four lists and having them rotate positions is that if one list of pictures is easier or harder to remember than others, the results will not be affected because each list appears in each condition an equal number of times across participants and thus each condition is equally influenced by any discrepancies between lists. Nevertheless, care was taken to ensure that the lists were equivalent. The words associated with the picture-sound pairs (e.g., the word "cat" for a picture of a cat and the cat's meow sound) were matched across all four lists on English frequency, concreteness, familiarity, and imageability (MRC Psycholinguistic Database).

All pictures were selected to be similar in saturation and line thickness. Each sound was edited to be 1000 milliseconds in duration, so that congruent, neutral, and incongruent auditory cues were matched in duration. The 20 tonal beeps were sine waveforms ranging from 300 Hz to 2200 Hz with each tone being 100 Hz different from its nearest two tones. All sounds were peak-amplitude

normalized using Audacity. The sounds were presented using monophonic sound reproduction and played to both ears through headphones, so as to ensure that they did not provide any relevant spatial information.

Of the 60 pictures, 15 were presented in the upper left corner, 15 were presented in the upper right corner, 15 were presented in the lower left corner, and 15 were presented in the lower right corner.

Each trial started with a 200-millisecond fixation cross in the center of the screen. Following the fixation cross, a picture was displayed for 1000 milliseconds. Simultaneous with the onset of the picture, a sound was played for 1000 milliseconds. Figure 1 provides a visual representation of the encoding task.

In the instructions of the encoding task, participants were asked to try to remember the pictures for a later memory test but not to be concerned with remembering the sounds. After the encoding task, participants completed a five-minute filler task in which they performed a simple math test (participants determined which of two values is larger). The purpose of the filler task was to prevent recency effects (Cohen, 1989), which might lead participants to remember predominantly the last sequence of pictures, regardless of condition.

Retrieval. In the retrieval task, participants viewed 120 pictures: the 60 pictures they had seen in the encoding task ('old' pictures), plus 60 foil pictures that they had not seen before ('new' pictures). The retrieval task had two components – an item or *what* memory component and a spatial or *where* memory component. In each trial, a picture was displayed and participants had to click 'new' (indicating that they did not recognize the picture from the encoding task) or 'old' (indicating that they did recognize the picture from the encoding task). If participants clicked 'old' (indicating that they had seen the picture before), they then made a judgment about its spatial location. They did so by clicking in one of four boxes located in the four corners of the screen. A visual representation of the retrieval task is shown in Figure 1.

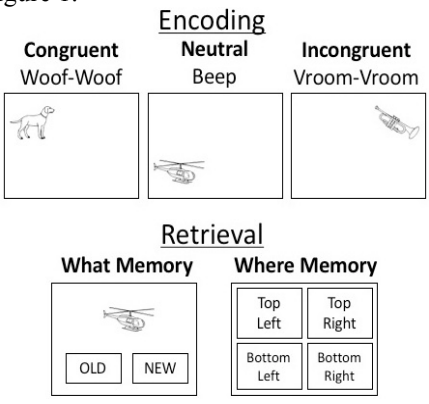

Figure 1: Top row depicts a congruent, neutral, and incongruent trial in the encoding phase of the *Environmental Sounds and Visual Memory* experiment. Bottom row depicts the *what* and *where* retrieval trials.

Spoken Words and Visual Memory Experiment

The spoken words experiment had the same methodology as the environmental sounds experiment unless noted below.

Encoding. In the spoken words experiment, participants were shown 64 pictures, of which 16 were presented with a *congruent* auditory cue (the English word for the visual object), 16 with an *incongruent* auditory cue (the English word for a different object), 16 with a *neutral non-linguistic* control auditory cue (a tonal beep sound), and 16 with a *neutral linguistic* control auditory cue (a pseudoword). The neutral *non-linguistic* control was included as in the environmental sounds experiment. In addition to the neutral *non-linguistic* control, a neutral *linguistic* control was also included for the purpose of having a linguistic control against which to compare the congruent linguistic condition. (The spoken words experiment had more pictures than the environmental sounds experiment in order to accommodate the additional condition in the spoken words experiment; note, however, that a pilot study that included an equal number of pictures and conditions in both experiments yielded the same results.)

Five lists of 16 picture-word pairs were compiled. The five lists served in one of the five positions – (1) as the 16 picture-word pairs in the congruent condition, (2) as the 16 pictures in the neutral non-linguistic condition, which were paired with a tonal beep sound, (3) as the 16 pictures in the neutral linguistic condition, which were paired with a pseudoword, (4) as the 16 pictures in the incongruent condition, or (5) as the 16 words in the incongruent condition. The pairings in the incongruent condition were created by matching a picture from one list (e.g., a trumpet) with a word from another list (e.g., the word "dog"). The five lists rotated, serving in every position an equal number of times across participants.

The words in the picture-word pairs were matched across all five lists on English frequency, English phonological neighborhood size, English biphone frequencies, number of English phonemes, concreteness, familiarity, and imageability (MRC Psycholinguistic Database; CLEARPOND; Marian, Bartolotti, Chabal, & Shook, 2012).

The pseudowords came from Colbertian, an artificial language (Bartolotti & Marian, 2012). Colbertian pseudowords were designed to conform to phonotactic rules of English and did not differ from the five lists of picture-word pairs in number of phonemes or in English biphone frequencies (CLEARPOND)

The spoken word stimuli were recorded at 44100 Hz by a female native English speaker. All words and pseudowords were equal to or shorter than 1000 milliseconds in duration. The tonal beep sounds were 1000 milliseconds in duration and ranged from 250 Hz to 1750 Hz, with each tone being 100 Hz different from its nearest two tones. None of the tones had the same frequency as the tones in the environmental sounds experiment.

After the encoding task, participants completed the five-minute filler math test, as in the environmental sounds experiment, but with different numbers.

Retrieval. In the retrieval task, participants viewed 128 pictures: the 64 pictures they had seen in the encoding task ('old' pictures), plus 64 foil pictures that they had not seen before ('new' pictures). As in the environmental sounds experiment, the retrieval task had two components – an item or *what* memory component and a spatial or *where* memory component. A visual representation of the encoding and retrieval phases is presented in Figure 2.

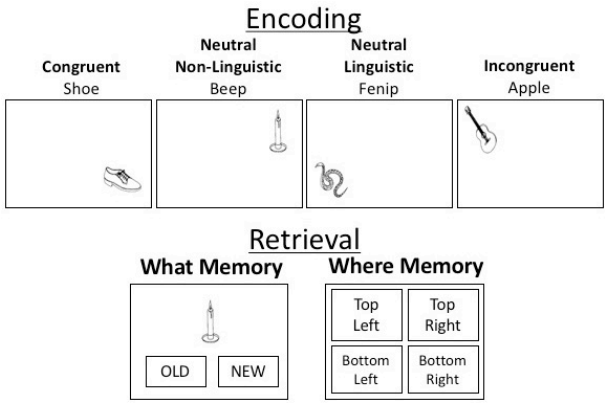

Figure 2: Top row depicts a congruent, neutral non-linguistic control, neutral linguistic control, and incongruent trial in the *Spoken Words and Visual Memory* experiment. Bottom row depicts the *what* and *where* retrieval trials.

Results

Where Memory

Environmental Sounds. The effects of environmental sounds on visual spatial or *where* memory were analyzed using a repeated-measures ANOVA with condition (Congruent, Incongruent, Neutral) as the independent variable and accuracy on the spatial memory task as the dependent variable. Accuracy rates by condition are displayed in Figure 3. The ANOVA yielded a significant main effect of condition, F (2, 78) = 11.72, $p < .001$, $\eta_p^2 = .23$. The significant main effect was followed up with contrasts between the experimental conditions (congruent and incongruent) and control condition (neutral). The contrasts indicated that the locations of pictures in the congruent condition were remembered significantly better than the locations of pictures in the neutral control condition (68.7% versus 56.9%), F (1, 39) = 14.74, $p < .001$, $\eta_p^2 = .27$. Conversely, the locations of pictures in the incongruent condition were not remembered significantly differently than the locations of pictures in the neutral control condition (57.8% versus 56.9%), F (1, 39) = 0.11, $p > .1$, $\eta_p^2 = .003$. These results suggest that visual spatial (*where*) memory was improved by hearing a congruent environmental sound. To determine whether this group-level effect was consistent across individuals, we computed the number of participants who remembered more locations in the congruent condition than in the neutral condition (and vice versa). Twenty-eight participants remembered more locations in the congruent condition, whereas only 12 remembered more locations in the neutral condition.

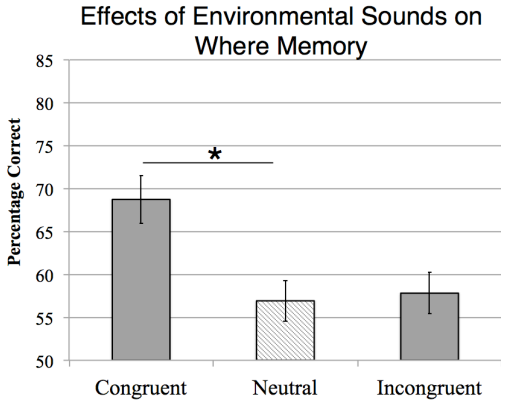

Figure 3: Memory accuracy on the spatial memory trials in the *Environmental Sounds and Visual Memory* experiment.

Spoken Words. To analyze the effects of spoken words on visual spatial or *where* memory, a repeated-measures ANOVA was conducted with condition (Congruent, Incongruent, Neutral Non-Linguistic, Neutral Linguistic) as the independent variable and spatial memory accuracy as the dependent variable. Accuracy by condition is presented in Figure 4. The ANOVA revealed no significant main effect of condition, F (3, 111) = 0.24, $p > .1$, $\eta_p^2 = .01$. These results indicated that visual spatial (*where*) memory was not improved by hearing a congruent spoken word. Consistent with these group-level results, individual-level results indicated that 18 participants remembered more locations in the congruent condition than in the neutral non-linguistic condition and 20 participants remembered more locations in the neutral non-linguistic condition than in the congruent condition.

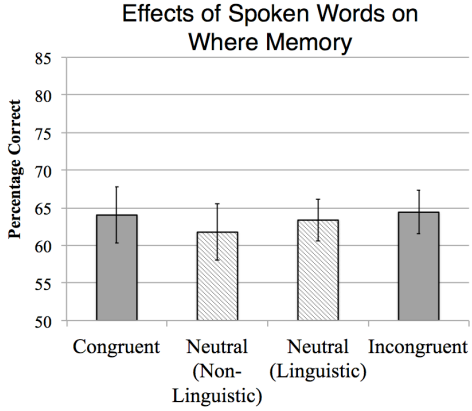

Figure 4: Memory accuracy on the spatial memory trials in the *Spoken Words and Visual Memory* experiment.

What Memory

Environmental Sounds. To analyze the effects of environmental sounds on visual item or *what* memory, a repeated-measures ANOVA was conducted with condition (Congruent, Incongruent, Neutral) as the independent variable and accuracy on item memory trials as the dependent variable. The accuracy rates for the three conditions are displayed in Figure 5. The ANOVA yielded a significant main effect of condition, F (2, 78) = 4.21, p < .05, η_p^2 = .10. Follow-up contrasts revealed that pictures in the congruent condition were recognized at a significantly higher rate than pictures in the neutral control condition (79.3% versus 73%), F (1, 39) = 5.47, p < .05, η_p^2 = .12. Similarly, pictures in the incongruent condition were recognized with significantly higher accuracy than pictures in the neutral control condition (77.9% versus 73%), F (1, 39) = 5.32, p < .05, η_p^2 = .12. These results indicate that hearing a congruent (or incongruent) environmental sound helped visual item (*what*) memory performance. These results at the group level were consistent with the results at the individual level, as 24 participants remembered more congruent pictures than neutral pictures, 13 participants remembered more neutral pictures than congruent pictures, and 3 participants remembered the same number of congruent and neutral pictures.

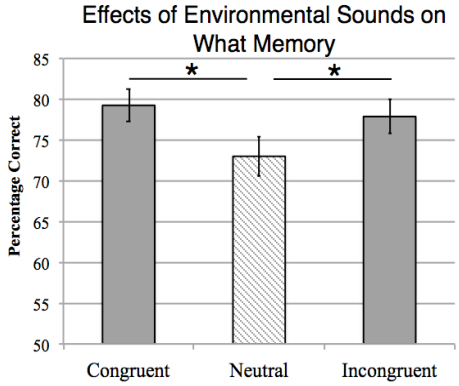

Figure 5: Memory accuracy on the item memory trials in the *Environmental Sounds and Visual Memory* experiment.

Spoken Words. The effects of spoken words on item or *what* memory were analyzed using a repeated measures ANOVA with condition (Congruent, Incongruent, Neutral Non-Linguistic, Neutral Linguistic) as the independent variable and accuracy on item memory trials as the dependent variable. Accuracy rates by condition are presented in Figure 6. The ANOVA yielded a significant main effect of condition, F (3, 108) = 4.56, p < .01, η_p^2 = .11. Follow-up contrasts revealed that pictures in both the congruent (77.5%) and incongruent (78%) conditions were recognized significantly better than pictures in the neutral non-linguistic control condition (71.2%) (F (1, 37) = 5.13, p < .05, η_p^2 = .12 and F (1, 37) = 10.31, p < .01, η_p^2 = .22, respectively). Moreover, congruent pictures were remembered marginally better and incongruent pictures were remembered significantly better than pictures in the neutral linguistic control condition (73.8%) (F (1, 37) = 3.00, p = .09, η_p^2 = .08 and F (1, 39) = 7.37, p < .05, η_p^2 = .17, respectively). These results provide evidence that hearing a congruent (or incongruent) spoken word improved visual item (*what*) memory performance. These group-level results were also reflected in the individual data, with 19 participants remembering more congruent pictures than neutral non-linguistic pictures, 13 participants remembering more neutral non-linguistic pictures than congruent pictures, and 6 participants remembering the same number of both.

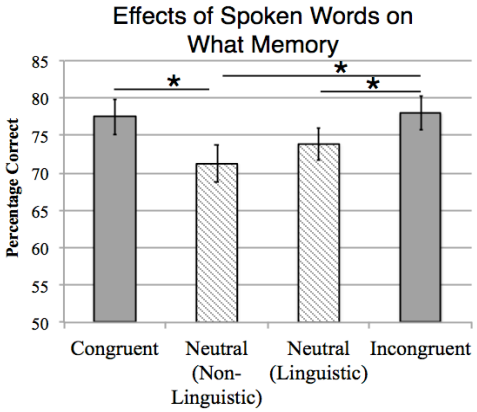

Figure 6: Memory accuracy on the item memory trials in the *Spoken Words and Visual Memory* experiment.

Discussion

We examined whether auditory input can impact visual memory. The results showed that hearing a congruent yet spatially uninformative environmental sound (for example, a barking sound played to both ears) improved memory for where a visual object was located (for example, a dog located in the top left corner of the screen). Because improved spatial memory in this case cannot be attributed to better auditory memory (memory for the environmental sound had no valid location information and therefore would not lead to a correct answer on the visual spatial memory test), the results are attributable to visual memory being improved by the environmental sound. To our knowledge, these results provide the first behavioral evidence for a cross-modal interaction in memory.

The effects of auditory input on *where* memory depended on the type of sound. While environmental sounds strengthened visual spatial memory, spoken words did not. These differences can be explained by theories positing that people unconsciously generate expectations on the basis of learned regularities (for example, predictive coding and schema theory). According to these theories, our cognitive system would expect an environmental sound, such as a dog's bark, to carry helpful location information because of the learned regularity that environmental sounds nearly always correlate with the location of the visual object (that

is, the dog's bark comes from the location of the dog). Because of this expectation, we may become more responsive to the relevant visual spatial information upon hearing an environmental sound. In contrast, our cognitive system would be unlikely to expect a spoken label, such as the word "dog", to carry helpful location information about the dog because spoken labels rarely correlate with the location of the visual object to which they refer. With a weaker expectation for valid spatial information about the referent, we might not be especially responsive to the relevant visual spatial information upon hearing a spoken label.

A second possible explanation for why congruent environmental sounds (but not congruent spoken words) enhanced *where* memory is that congruent environmental sounds may have elicited deeper processing, heightened arousal, or increased attention to the stimuli (relative to congruent spoken words). However, if congruent environmental sounds prompted deeper processing, heightened arousal, or increased attention to the stimuli, then congruent environmental sounds should have also enhanced *what* memory to a larger degree than congruent spoken words because deeper processing, heightened arousal, and increased attention to the stimuli are all known to produce stronger *what* memory (Craik & Tulving, 1975; Dolcos, LaBar, & Cabeza, 2004; Naveh-Benjamin, Guez, & Marom, 2003). Yet, relative to a neutral control sound, congruent environmental sounds did not enhance *what* memory more than congruent spoken words.

Still another explanation for why congruent environmental sounds (but not congruent spoken words) enhanced *where* memory relates to the ventriloquism effect. According to the ventriloquism effect (Slutzky & Recanzone, 2004), simultaneously hearing a sound and seeing an image can sometimes lead to the illusion that the sound is coming from the image. If the ventriloquism effect occurred for congruent environmental sounds (but not for congruent spoken words), it may have yielded helpful spatial encoding of the congruent environmental sounds. However, this explanation assumes that the ventriloquism effect depends on semantic congruence and on type of auditory cue, is precise enough to distinguish spatial locations within centimeters, and is reliable with headphones. At present, none of these assumptions has strong empirical support.

The finding that congruent environmental sounds and spoken words increased *what* memory replicates several previous studies (e.g., Lehmann & Murray, 2005). It is possible that these results are due in part to cross-modal interactions in memory, but they can also be attributed to dual-coding, where a memory is encoded in both the auditory modality and visual modality, and later visual memory performance is helped by remembering the encoded auditory cue.

In conclusion, we found that auditory input can strengthen visual episodic memory. The current results provide evidence for cross-modal interactions in memory and extend multi-sensory research on perception and attention by showing that audio-visual interactions are not just short-lived perceptual and attentional effects; they have longer-term consequences that persist in memory. Because most of our experiences are multi-sensory, these results capture how memory works in everyday situations. These findings may also carry practical implications, as cross-modal enhancements may be applied to increase memory in educational programs, cognitive therapies, and human factors designs.

Acknowledgments

This research was funded by grant NICHD R01 HD059858 to Viorica Marian.

References

Bartolotti, J., & Marian, V. (2012). Language learning and control in monolinguals and bilinguals. *Cognitive Science*, *36*(6), 1129-1147.

Chen, Y. C., & Spence, C. (2011). Crossmodal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(5), 1554-1568.

Cohen, R. L. (1989). The effects of interference tasks on recency in the free recall of action events. *Psychological Research*, *51*(4), 176-180.

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268-294.

Dolcos, F., LaBar, K. S., & Cabeza, R. (2004). Dissociable effects of arousal and valence on prefrontal activity indexing emotional evaluation and subsequent memory: An event-related fMRI study. *Neuroimage*, *23*(1), 64-74.

Edmiston, P., & Lupyan, G. (2015). What makes words special? Words as unmotivated cues. *Cognition*, *143*, 93-100.

Lehmann, S., & Murray, M. M. (2005). The role of multisensory memories in unisensory object discrimination. *Cognitive Brain Research*, *24*(2), 326-334.

Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). CLEARPOND: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PloS one*, *7*(8), e43230.

Naveh-Benjamin, M., Guez, J., & Marom, M. (2003). The effects of divided attention at encoding on item and associative memory. *Memory & Cognition*, *31*(7), 1021-1035.

Slutsky, D. A., & Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect. *Neuroreport*, *12*(1), 7-10.

Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin & Review*, *8*(2), 203-220.