Auditory recognition memory is inferior to visual recognition memory

Michael A. Cohen[a], Todd S. Horowitz[a,b], and Jeremy M. Wolfe[a,b,1]

[a]Brigham and Women's Hospital, [b]Harvard Medical School, Boston, MA 02115

Edited by Anne Treisman, Princeton University, Princeton, NJ, and approved February 24, 2009 (received for review November 24, 2008)

Visual memory for scenes is surprisingly robust. We wished to examine whether an analogous ability exists in the auditory domain. Participants listened to a variety of sound clips and were tested on their ability to distinguish old from new clips. Stimuli ranged from complex auditory scenes (e.g., talking in a pool hall) to isolated auditory objects (e.g., a dog barking) to music. In some conditions, additional information was provided to help participants with encoding. In every situation, however, auditory memory proved to be systematically inferior to visual memory. This suggests that there exists either a fundamental difference between auditory and visual stimuli, or, more plausibly, an asymmetry between auditory and visual processing.

For several decades, we have known that visual memory for scenes is very robust (1, 2). In the most dramatic demonstration, Standing (3) showed observers up to 10,000 images for a few seconds each and reported that they could subsequently identify which images they had seen before with 83% accuracy. This memory is far superior to verbal memory (4) and can persist for a week (5). Recent research has extended these findings to show that we have a massive memory for the details of thousands of objects (6). Here, we ask whether the same is true for auditory memory and find that it is not.

Results

For Experiment 1, we recorded or acquired 96 distinctive 5-s sound clips from a variety of sources: birds chirping, a coffee shop, motorcycles, a pool hall, etc. Twelve participants listened to 64 sound clips during the study phase. Immediately following the study phase, we tested participants on another series of 64 clips, half from the study phase and half new. Participants were asked to indicate whether each clip was old or new. Memory was fairly poor for these stimuli: the hit rate was 78% and the false alarm rate 20%, yielding a d' score* of 1.68 (s.e.m. 0.14). To put this performance for a mere 64 sound clips in perspective, in Shepard's original study with 600 pictures, he reported a hit rate of 98%, whereas Standing reported a hit rate of 96% for 1,100 images.

There are several possible explanations for the poor performance on this auditory memory task. It could be that the remarkable ability to rapidly encode and remember meaningful stimuli is a feature of visual processing. Alternatively, these might have been the wrong sounds. A particular stimulus set might yield poor performance for a variety of reasons. Perhaps the perceptual quality was poor; for example, many of our stimuli were recorded monaurally but played over headphones. It is also possible that the sound clips were too closely clustered in the stimulus space for observers to distinguish between them. Or the stimuli might simply be the wrong sort of auditory stimuli for reasons unknown. To distinguish between the poor memory and poor stimuli hypotheses, we replicated the experiments with a second set of stimuli that were professionally recorded (e.g., binaurally) and designed to be as unique as possible (e.g., the sound of a tea kettle, the sound of bowling pins falling). Each sound was assigned a brief description (e.g., "small dog barking"). In a separate experiment, 12 participants were asked to choose the correct name for each sound clip from a list of 111 descriptions (chance = 0.90%), and they succeeded exactly with 64% of the sounds. Two-thirds of the remaining errors being "near misses" (e.g., "Big dog" for the sound of a small dog barking would be considered a near miss; "tea-kettle" for the sound of bowling pins falling would not). Thus, with this second set of sound clips, participants were able to identify the sound clips relatively well. For each sound clip in this new set, we also obtained a picture that matched the description.

There were 5 conditions in Experiment 2. In each condition, 12 new participants were tested using the same testing protocol as Experiment 1. The study phase contained 64 stimuli. In the test phase, participants labeled 64 stimuli as old or new. We measured memory for the sound clips alone, the verbal descriptions alone, and the matching pictures alone. We also added 2 conditions intended to improve encoding of the sound clips. In 1 condition, the sound clips were paired with the pictures during the study phase. In the other, the sound clips were paired with their verbal descriptions during study. In both of these conditions, participants were tested for recognition of the sound clips alone.

The results, shown in Fig. 1, were unambiguous. According to Tukey's WSD test, memory for pictures was significantly better than for all other stimuli, while the remaining conditions did not differ from one another. Recall for sound clips was slightly higher than in the first experiment, but still quite low (d' = 1.83; s.e.m. = 0.21) and far inferior to recall for pictures (d' = 3.57; s.e.m. = 0.24). Supplying the participants with descriptions together in the study phase did not significantly improve recall for sound clips (d' = 2.23; s.e.m. = 0.17). This may not be surprising, because recall for the verbal descriptions by themselves was also relatively poor (d' = 2.39; s.e.m. = 0.15). However, even pairing sound clips with pictures of the objects at the time of encoding did not improve subsequent testing with sound clips alone (d' = 1.83; s.e.m. = 0.16). Note that these were the same pictures that, by themselves, produced a d' of 3.57.

Again, it is still possible that these were the wrong stimuli. In terms of information load, the auditory stimuli we used may simply be more impoverished than pictures. Thus, poor memory performance with sounds may be due solely to the nature of the particular stimulus we used. Perhaps richer stimuli would lead to more efficient encoding and storage in memory. To explore this possibility, in Experiment 3 we replicated the testing procedures from Experiments 1 and 2 using 2 new types of stimuli: spoken language and music. Both classes of stimuli might contain more information than the natural auditory sounds used in Experiments 1 and 2. Spoken language conveys information about the speaker's age, gender, and nationality, in addition to a wealth of

Author contributions: M.A.C., T.S.H., and J.M.W. designed research; M.A.C. performed research; M.A.C., T.S.H., and J.M.W. analyzed data; and M.A.C., T.S.H., and J.M.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: jmwolfe@rics.bwh.harvard.edu.

*d', a standard index of detectability derived from signal detection theory (7), is computed from hit and false alarm rates. Because false alarm rates are not available for all of the early picture memory studies, we also report hit rates.

 www.pnas.org/cgi/doi/10.1073/pnas.0811884106

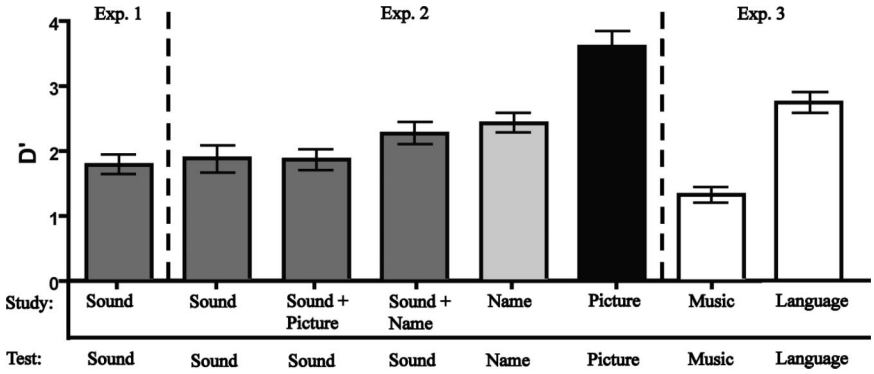

Fig. 1. Memory performance in units of d'. Error bars denote the standard error of the mean. The leftmost part shows the results from Experiment 1, the center part shows the results from Experiment 2, and the rightmost part shows the results from Experiment 3.

semantic information about the topic being discussed. Music, when there is a vocalist, can convey much the same information as spoken language, in addition to information about rhythm, harmony, and instrumentation.

Experiment 3 consisted of 2 groups of 12 participants, all native English speakers. In the spoken language condition, participants were tested using 90 unique speech clips (7–15 s) on a variety of topics (e.g., politics, sports, current affairs, sections from novels). Participants were debriefed afterward to confirm that they had no problem understanding what was being said, in terms of both content and speaker's pronunciation. Performance in this condition ($d' = 2.7$; s.e.m. = 0.16) was better than every other sound condition, but was still worse than the picture only condition of Experiment 2 [t (11) = 3.31, $P < 0.01$]. In the music condition participants were tested using 90 novel popular music clips (5–15 s). Each participant was debriefed after the experiment, and none reported having ever heard any of these specific clips before. Performance in this experiment ($d' = 1.28$; s.e.m. = 0.11) was actually worse than in the sound only condition of Experiment 2 [t (11) = 2.509, $P < 0.05$], and far worse than the picture only condition [t (11) = 14.14, $P < 0.001$]. Thus, memory for a variety of auditory stimulus classes, some of which potentially carry more information than natural auditory sounds, is inferior to visual memory for scenes and objects.

Experiment 3 suggests that poor auditory memory is not simply the product of impoverished stimuli. However, it would be more satisfying to directly measure the quality of visual and auditory stimulus sets in the same units. Here, we used the classification task previously used to calibrate the auditory stimuli in Experiment 2, asking participants to assign each stimulus a label from a prespecified list of labels. Recall that for the auditory stimuli, participants were able to perform at 64% on this 111-alternative choice task, using a conservative scoring criterion. For comparison, we obtained a set of images that had been created by taking 256 × 256 pixel images, reducing them to 16 × 16 pixel resolution, then upsampling to create 256 × 256 pixel images for display. This resulted in very degraded, blurred versions of the originals (8). Previous work with these same images demonstrated that this procedure leads to a decrease in performance on a broad categorization task as compared to higher resolution images (8).

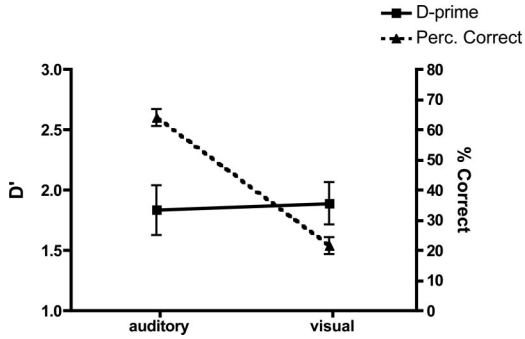

Fig. 2. Auditory stimuli *vs* degraded visual images. Memory performance (squares, solid line) is plotted against the left *y*-axis in units of d'. Percent correct for the naming experiment is plotted against the right *y*-axis. Error bars denote standard error of the mean.

For the first part of Experiment 4, we tested 12 participants in the same memory protocol as in the previous experiments using 102 upsampled images. As Fig. 2 shows, performance on this condition ($d' = 1.89$; s.e.m. = 0.17) was not significantly different from performance with the auditory stimuli from Experiment 2 [t (11) = 0.21, $P > 0.8$]. In the second condition, we then asked 12 participants[†] to choose the correct name for each degraded image from a list of 102 descriptions (chance = 0.98%). Participants successfully matched an image with its description just 21% of the time, significantly worse than the 64% classification performance for the auditory stimuli reported earlier [t (11) = 21.22, $P < 0.001$]. Using the more liberal scoring criterion that corrects for "near misses" (e.g., "highway" for the image of a forest road would be considered a near miss; "bedroom" for the image of a "beach" would not), performance was still only 24% against 83% for the auditory stimuli [t (11) = 30.277, $P < 0.001$].

Fig. 2 makes our point graphically. To equate the memorability of visual and auditory stimuli, we needed to render the visual stimuli almost unrecognizable. Participants were much better at classifying/identifying the auditory stimuli than the degraded visual stimuli (triangles, right *y*-axis). This is consistent with an asymmetry between visual and auditory processing. Stimuli of equal memorability are not equally identifiable. Highly identifiable auditory stimuli are not remembered well.

Discussion

It is clear from these results that auditory recognition memory performance is markedly inferior to visual recognition memory on this task. Note that we do not claim that long-term auditory memory, in general, is impoverished. Clearly, some form of auditory long-term memory allowed our participants to identify the stimuli as tea kettles, dogs, and so forth. Moreover, with practice, people can commit large bodies of auditory material (e.g., music) to memory. The striking aspects of the original picture memory experiments are the speed and ease with which

[†]Note that 5 participants participated in both conditions of experiment 4, but were only allowed to complete the classification condition after having completed the memory condition.

complex visual stimuli seem to slide into long-term memory. Hundreds or thousands of images, seen for a few seconds at a time, are available for subsequent recognition. It is this aspect of memory that seems to be markedly less impressive in audition. Two explanations suggest themselves. Auditory objects might be fundamentally different from visual objects. In their physics or psychophysics, they may actually be less memorable than their visual counterparts. Alternatively, auditory memory might be fundamentally different/smaller than visual memory. We might simply lack the capacity to remember more than a few auditory objects, however memorable, when they are presented one after another in rapid succession. In either case, it is unlikely that anyone will find 1000 sounds that can be remembered with anything like the accuracy of their visual counterparts.

Materials and Methods

Participants. One hundred thirteen total participants (aged 18–54) participated in the experiments. For each condition there were 12 participants, with a total of 11 conditions/experiments. Each participant passed the Ishihara test for color blindness and had normal or corrected to normal vision. All participants gave informed consent, as approved by the Partners Healthcare Corporation IRB, and were compensated $10/h for their time.

Stimuli. In Experiment 1, stimuli were gathered using a handheld recording device (Panasonic PV-GS180) or were obtained from a commercially available database (SoundSnap). In Experiment 2, stimuli were gathered from SoundSnap.com. In Experiment 3, music clips came from the collections of members of the laboratory. Songs were uploaded into WavePad and 7- to 15-s clips were extracted. Speech clips used came from various podcasts obtained online and were also uploaded into WavePad to obtain 5- to 15-s clips. Degraded visual images used in Experiment 4 were obtained from A. Torralba (Massachusetts Institute of Technology, Cambridge, MA). A list of the stimuli used is provided on our website: search.bwh.harvard.edu.

Experimental Blocks. The memory experiments consisted of a study block and a test block. In the study block, participants listened to or viewed a set of sound clips or sound clips and their correlating images/names (60–66 clips) for approximately 10 min. Their instructions were simply to carefully study to the clips and try to commit them to memory as best they could. In the test block, participants were presented with another set of clips (60–64 clips), half that were repeated from the study block (old) and half that had never been presented before (new). Participants were asked to make an ''old/new'' discrimination after every trial. Note that on 1 condition of the memory experiments the basic paradigm remained the same, but participants were presented with only visual images (picture only). The naming/classification experiments comprised a single block lasting approximately 20 min. Participants were shown each stimulus for 5 s and would then type in the name of what they had heard/seen from a list provided (102–110 names).

Apparatus. Every experiment was conducted on a Macintosh computer running MacOS 9.2, controlled by Matlab 7.5.0 and the Psychophysics Toolbox, version 3.

ACKNOWLEDGMENTS. We thank Christina Chang, Karla Evans, Yair Pinto, Aude Oliva, and Barbara Shinn-Cunningham for helpful comments and suggestions on the project, and Antonio Torralba for providing the degraded images used in Experiment 4. This work was funded in part by NIMH-775561 and AFOSR-887783.

1. Shepard RN (1967) Recognition memory for words, sentences, and pictures. *J Verb Learn Verb Behav* 6:156–163.
2. Pezdek K, Whetstone T, Reynolds K, Askari N, Dougherty T (1989) Memory for real-world scenes: The role of consistency with schema expectation. *J Exp Pscyhol Learn Mem Cog* 15:587–595.
3. Standing L (1973) Learning 10,00 pictures. *Q J Exp Psychol* 25:207–222.
4. Standing L, Conezi J, Haber RN (1970) Perception and memory for pictures: Single-trial learning of 2500 visual stimuli. *Psychon Sci* 19:73.
5. Dallet K, Wilcox SG, D'Andrea L (1968) Picture memory experiments. *J Exp Psychol* 76:312–320.
6. Brady TF, Konkle T, Alvarez GA, Oliva A (2008) Visual long-term memory has a massive storage capacity for object details. *Proc Natl Acad Sci US* 105:14325–14329.
7. Macmillan NA, Creelman CD (2005) in *Detection Theory*: *A User's Guide* 2nd ed. (Lawrence Erlbaum Assoc, Mahwah, NJ) 2nd Ed.
8. Torralba A (2009) How many pixels make an image? *Visual Neurosci*, epub ahead of print.