Psychon Bull Rev (2015) 22:1777–1783
DOI 10.3758/s13423-015-0854-z

BRIEF REPORT

Maintenance of auditory-nonverbal information in working memory

Alexander Soemer[1] · **Satoru Saito**[1]

Published online: 12 May 2015

Abstract According to the multicomponent view of working memory, both auditory-nonverbal information and auditory-verbal information are stored in a phonological code and are maintained by an articulation-based rehearsal mechanism (Baddeley, 2012). Two experiments have been carried out to investigate this hypothesis using sound materials that are difficult to label verbally and difficult to articulate. Participants were required to maintain 2 to 4 sounds differing in timbre over a delay of up to 12 seconds while performing different secondary tasks. While there was no convincing evidence for articulatory rehearsal as a main maintenance mechanism for auditory-nonverbal information, the results suggest that processes similar or identical to auditory imagery might contribute to maintenance. We discuss the implications of these results for multicomponent models of working memory.

Keywords Auditory short-term memory · Working memory · Phonological loop · Timbre · Auditory imagery

Decades of research have provided a great deal of support for the multicomponent model of working memory that distinguishes between domain-specific resources for verbal and visuospatial information (see Baddeley, 2012, for a recent version). Within this framework, evidence has been accumulated for a verbal short-term store based on a phonological code and for an articulation-based maintenance mechanism operating on the store's contents. Furthermore, it has been suggested that this "phonological loop" component also handles auditory-nonverbal information, even types of auditory-nonverbal information that are difficult to encode phonologically and are difficult or impossible to articulate. This report investigates one such type, namely, timbre information (the quality or uniqueness of a sound). We asked were whether maintenance for such auditory-nonverbal information is accomplished by the articulatory rehearsal, as assumed in the multicomponent model, or by other potential mechanisms.

Although there is a rich body of literature on auditory short-term memory predating the rise of the multicomponent model, this early research was mainly concerned with various forms of passive auditory storage (Cowan, 1984). In contrast, *active* maintenance of auditory-nonverbal information is a rather new research area (e.g., Williamson, Baddeley, & Hitch, 2010) and, to date, not much is known about potential mechanisms that prolong the survival of auditory features over a delay of several seconds.

Within the framework of the multicomponent model, associating auditory information with the phonological loop was not without reason. For example, it is well known that irrelevant sound presented during a delay in a short-term memory task disrupts verbal maintenance (Colle & Welsh, 1976), which has been explained by direct access of auditory information into the phonological store. However, a challenge to the notion that auditory information is maintained in the loop arises from the assumption that its representational code is phonological. While it is true that phonemes are being distinguished by articulatory features underlying their production, these features are normally thought to be abstracted away from the acoustic level (Gaskell, Quinlan, Tamminen, & Cleland, 2008). This poses the question of how auditory-nonverbal materials lacking phonemic properties could be converted into a phonological code (in what "phonological" feature would the sound of a violin and the sound of a cello differ?). Even if conversion is unnecessary and the loop can

✉ Alexander Soemer
alexander@soemurai.de

[1] Graduate School of Education, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan

handle nonphonological materials as well, how can the timbres of these instruments be maintained via articulation[1]?

In fact, previous research by McKeown, Mills, and Mercer (2011) using a discrimination task suggests that timbre information may not be stored in the loop, as shown in a lack of interference from reading aloud between standard and probe presentations. McKeown and Mercer (2012), furthermore, showed that timbre memory is prone to an inevitable and slow decay over time, which seems to suggest that active maintenance might not be possible at all. However, these experiments were concerned with fine-grained differences in timbre for which verbal labeling might be extremely difficult. Golubock and Janata (2013) recently estimated working memory capacity for sounds differing in timbre on broader acoustic feature dimensions (Experiment 2) and found a similar decline in capacity estimates over time. This finding, however, does not rule out the possibility that active maintenance mechanisms can slow down the decay.

If one assumes that active maintenance of timbre is possible, one will have to consider both the phonological loop and other logical possibilities of memory maintenance. For example, timbre information might be maintained by central executive mechanisms, or, alternatively, there might be a functionally specialized and yet-to-be-explored mechanism. We conducted two experiments to investigate these alternatives.

Experiment 1

In Experiment 1, we first investigated the standard hypothesis stating that categorically different auditory-nonverbal information is maintained in the phonological loop (Baddeley, 2012). Participants were required to remember two to four artificial sounds, differing in spectral and dynamic timbre, played on the same fundamental frequency (131 Hz) over a delay of either 3 seconds (short-delay condition) or 12 seconds (long-delay condition). After the delay, either one sound of the set or a distractor sound was presented as a probe. Participants had to judge whether or not the probe was part of the set. During the delay in one half of the trials, one group of participants performed articulatory suppression, while a second group of participants performed key tapping as a standard dual-task control condition known to be minimally demanding in terms of attention. In the other half of the trials, the participants were not required to perform any secondary task.

[1] In fact, the literature on neural correlates, which is beyond the scope of this article, provides some support for an involvement of motor areas in timbre imagery (see Hubbard, 2013, for a review).

Materials

About 1,000 artificial sounds taken from different synthesizers were selected fulfilling the following requirements: The sounds (1) were perceptually dissimilar to existing instruments or environmental sounds and perceptually dissimilar to other already selected sounds, (2) evoked a clear pitch percept, and (3) were at least 500 milliseconds long. The remaining sounds were trimmed to 500 milliseconds and normalized by their root-mean-square amplitude. Unsatisfying results of this procedure for some items were adjusted manually.

We computed acoustic measures that have previously been identified as reliably indicating perceptual similarity (McAdams, 1999): attack time (the time between onset and reaching 90% of the maximum amplitude), spectral centroid (the location in the spectrum with the highest average amplitude), and spectral flux (the degree of temporal variation in the spectrum). Pairwise similarities between these sounds were computed as the Euclidean distance on these three dimensions. To obtain a set of sounds in which all items are reasonably well spaced to each other in similar space, we computed for each sound the average distance and the standard deviation to all other selected sounds. The distribution of these pairwise similarity averages showed a reasonably bell-shaped form. We excluded each 25% of sounds on the left side (sounds that were similar to many other sounds) and the right side (very unique sounds) of the curve. Furthermore, we excluded 10% of sounds with the highest pairwise similarity standard deviations (sounds that were very similar to some items but very dissimilar to others). The final set contained 88 stimuli.

Participants

Sixty undergraduate students, aged between 18 and 25 years (median age = 21), served as participants for Experiment 1 (male = 39, female = 21). Participants were evenly split between the suppression condition and the tapping condition. All participants reported normal hearing and were rewarded with a 1,000 JPY book coupon after the completion of the experiment.

Procedure

The experiment was conducted in a soundproof room with external stereo speakers placed approximately 40 cm away from the participants. In a trial, participants were first prompted with a fixation cross for 500 milliseconds. After a blank of 500 milliseconds, 2, 3, or 4 sounds were played for each 500 milliseconds and separated from each other by 500 milliseconds of silence. After the disappearance of the last sound, white noise was played for 200 milliseconds to eliminate sensory traces. Between mask offset and probe presentation, either 5 (short-delay) or 20 (long-delay) unfilled circles

appeared in the center of the screen. Each 600 milliseconds a circle was color-filled from left to right to indicate the pace with which participants were required to either articulate the syllable "da" (suppression condition) or to press a specified key (tapping condition). Participants were told to ignore the circles on the control trials.

Prior to the experiment, participants were given opportunity to practice eight control and eight treatment trials with different memory loads. In the main session, participants performed four trials for each possible combination of task condition (control/treatment), set size (2/3/4), delay (3 s/12 s), and probe type (same/different). In total, each participant performed 96 main trials. Each participant started with an item load of two and ended with an item load of four. Control and treatment conditions were switched every four trials. Half of the participants started with the control condition and the other half with the treatment condition. Within each condition, delays and probe types were intermixed. Participants could rest at anytime between trials.

Results

Overall accuracy was analyzed with mixed-effect logistic regression models (hits and correct rejections were coded as 1, misses and false alarms as 0). Regression coefficients and measures of uncertainty were estimated running MCMC simulations in the Gibbs sampler JAGS (mcmc-jags.sourceforge.net). Model selection was based on the Deviance Information Criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002), a measure of fit with a penalty for complexity common in Bayesian model comparison.

The optimal random effects structure was determined first, starting with a model that included subject and item intercepts, as well as subject-by-delay, subject-by-set size, and subject-by-task slopes. In all cases, models containing only subject and item intercepts were selected. Subsequently, the fixed effects structure was determined using the same procedure. The reliability of the remaining fixed effects in the selected model was indicated by 95% Bayesian credible intervals (below abbreviated as CRI) around the regression coefficients *b*. If intervals were entirely below or above 0, the effect was considered reliable with the interval itself giving the direction of the effect and an estimation of effect size.

Data of one participant from the suppression group were lost due to a recording error. Modeling of the remaining data points was carried out separately for the two groups (see Fig. 1). The selected model for the tapping group contained reliable main effects of delay, b = -0.14 units/sec, CRI = (-0.21, -0.06) and set size, b = -0.44 units/item, CRI = (-0.68, -0.22), suggesting a performance decrease with longer delays and more memory items. The selected model for the suppression group contained reliable main effects of treatment, b = -1.32, CRI = (-2.06, -0.60); delay, b = -0.05 units/sec, CRI = (-0.07, -0.03); and set size, b = -0.51 units/item, CRI = (-0.67; -0.34) as well as a reliable interaction between treatment and set size, b = 0.38 units/item, CRI = (0.15, 0.60), suggesting lower performance in the treatment condition and with longer delays and more items.

The interaction indicated that the disruptive effect of suppression was only present with small set sizes. Separate analyses carried out for each set size confirmed that there was a reliable suppression effect only for two items, b = -0.65, CRI = (-1.00, -0.31). In all set size conditions there were main effects of delay, two items: b = -0.08 units/sec, CRI = (-0.12, -0.04); three items: b = -0.06 units/sec, CRI = (-0.10, -0.02); and four items: b = -0.03 units/sec, CRI = (-0.06, -0.01).

Discussion

The most important result of Experiment 1 is that maintenance of timbre information remained surprisingly robust under articulatory suppression. Only in the two-item condition, did suppression cause a reliable decline in performance of 9 percentage points. This suggests that the effect was potentially not caused by blocked articulatory rehearsal but by a factor that is sensitive to practice effects (recall that participants started with two-item blocks and finished with four-item blocks). In addition, participants might have, indeed, used a verbal labeling strategy that broke down when dealing with more than two items.

Such a result is not convincingly supporting the view that auditory-nonverbal information is maintained in the phonological loop. This leaves one wondering what other mechanisms could be behind the reasonably high performance of our participants. One possibility within the existing working memory framework is that the materials are maintained by central executive or attentional mechanisms. This idea would be consistent with the small disruption of articulatory suppression, as this task is assumed to place minimal demands on the central executive.

On the other hand, introspection suggests that auditory information could be maintained by a specialized mechanism experienced as auditory imagery (Hubbard, 2010, 2013). Auditory imagery can be conceived as an auditory analogue to visual imagery or as imaging a sound before the mind's ear. Previous research has shown that certain aspects of timbre information are contained in auditory images. Crowder (1989) presented participants with a sine wave tone played on a certain pitch, and they then had to imagine what the tone would sound like if played on an instrument (e.g., trumpet). After a short delay, the same or a different tone was played on either an instrument matching the imagery instruction or a different instrument. When the instruments matched, participants were faster to make correct "same tone" judgments than when the instruments differed. Using a similar paradigm, Pitt

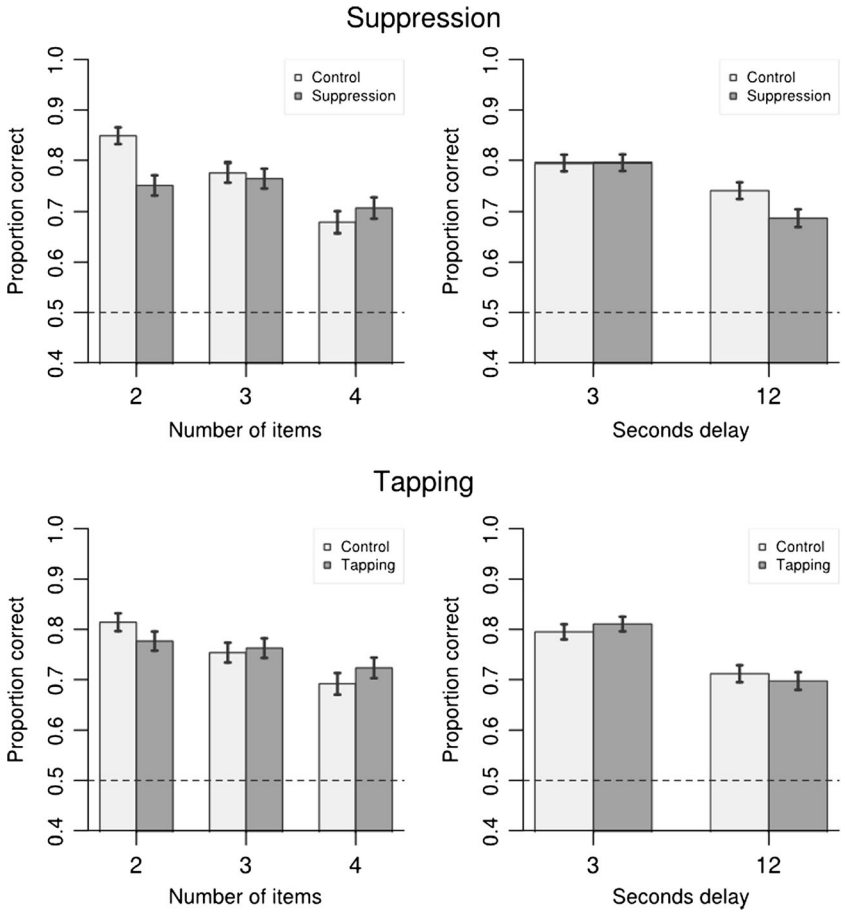

Fig. 1 Proportion correct per set size (left panels) and delay (right panels) in Experiment 1. Error bars depict standard errors

and Crowder (1992) found that a match of spectral but not dynamic properties led to faster "same tone" responses, suggesting that auditory images may contain spectral, but perhaps not dynamic, timbre information. However, in these experiments, since memory for timbre was not required, it is yet to be investigated whether auditory imagery can act as a maintenance mechanism for timbre information.

Experiment 2

In Experiment 2, participants were again required to maintain two to four sounds over a delay of several seconds. In half of the trials, one group of participants was to perform a secondary auditory imagery task, while a second group of participants was to perform a secondary visual imagery task. The auditory imagery task was designed to interfere with a potential maintenance mechanism based on auditory imagery, while the visual imagery task served as a control task. We set up both tasks as comparable as possible in terms of difficulty, presentation, and response mode in order to distinguish between potential disruptions specific to audition as opposed to potential disruptions involving the central executive.

Materials

We reused the timbre stimuli from Experiment 1. For the imagery tasks, we had to make sure that participants really rely on imagery. This was accomplished by using tasks that can be made objective to some extent—here, a pitch comparison task and a brightness comparison task. Rating studies were conducted prior to the experiment in which one group of subjects not participating in the main experiment judged the pitch of auditory images evoked by a set of words, while a different group of subjects judged the brightness of visual images evoked by a different set of words. These ratings were then used to define groups of words referring to sounds of low pitch versus high pitch, and images of low brightness versus high brightness, respectively.

For the auditory imagery task, 20 participants (female = 9, median age = 21) rated the imagined pitch of auditory images evoked by 58 words. To ensure robust auditory images, all words were onomatopoeic expressions related to sound ("giongo" in Japanese) and consisted of two repeating two-mora[2] combinations (e.g., *chi-ku-chi-ku*). Participants rated

[2] A *mora* in Japanese phonology consists of single vowels or combinations of consonant/glides and vowels.

the pitch of each item on a scale from 1 (*very low*) to 7 (*very high*). These ratings were reasonably distributed between 2 and 7, allowing for the creation of two groups of words, one of which contained 21 words with pitch ratings below 3.5 (low-pitch group: $M = 2.92$, $SD = .47$) and another 21 words that contained pitch ratings above 4.5 (high-pitch group: $M = 5.48$, $SD = .81$). The remaining 16 words were removed from the set.

An analogous brightness rating study with 68 onomatopoeic words related to appearance or movement ("gitaigo" in Japanese) was conducted with another 14 participants (male = 12, female = 2, median age = 23). The items had the same two-mora structure and were rated on a scale from 1 (*very dark*) to 7 (*very bright*). Again, two groups were created containing each 19 words (low-brightness group: $M = 2.8$, $SD = .39$; high-brightness group: $M = 5.1$, $SD = .91$). The remaining 30 words were removed from the set.

Participants

Thirty-six undergraduate and graduate students, aged between 18 and 32 years (median age = 21), served as participants for Experiment 2 (female = 15). Participants were evenly split between the auditory and visual imagery conditions. All participants reported normal hearing, and they were rewarded with a 1,000 JPY book coupon after the completion of the experiment.

Procedure

The procedure was largely the same as in Experiment 1, with the main exception being that the secondary tasks were unpaced (no dots appearing on the screen during the delay). In the imagery tasks, one word from the low-pitch or low-brightness group and one word from the high-pitch or high-brightness group were randomly assigned to the left and the right side of the screen. Participants had to judge which of the words evoked an auditory/visual image of higher pitch/brightness. Responses were made via pressing keys corresponding to the words on the left and the right side of the screen, respectively. After each decision, the next pair of words appeared on the screen. Thus, the whole delay was filled with pitch or brightness decisions. In comparison to Experiment 1, we increased the delay from 3 to 6 seconds in the short-delay condition, as we expected imagery processing to take some time.

Results

We first checked the "accuracy" of the imagery tasks, defined as the ratio of *decisions in line with the ratings* (collected from the rating studies) to the *number of decisions* made in correct (hit or correct rejection) trials. This accuracy was between 95.7% and 97.5% (overall 96.8%) for the auditory group, and between 89.6% and 94.5 (overall 92.8%) for the visual group, depending on set size and delay. To analyze the reliability of this difference, a linear regression model was fitted to arcsine-transformed accuracy data. This analysis revealed a reliable accuracy disadvantage for the visual task, $b = -0.11$; CRI = (-0.18, -0.04). Performing a similar analysis on the average number of decisions made during the retention intervals did only reveal a reliable effect of delay, $b = 0.81$ items/sec; CRI = (0.72, 0.90), which amounts to around five more decisions made in the long-delay conditions.

Modeling of correct responses was carried out separately for the two groups. The selected model for the auditory imagery group contained reliable main effects of treatment, $b = -0.35$, CRI = (-0.58, -0.11); delay $b = -0.04$ units/sec, CRI = (-0.08, -0.01); and set size, $b = -0.30$ units/item, CRI = (-0.44, -0.15). The coefficients show that performance was lower in the auditory imagery condition and decreased with the delay and with the number of items. In contrast, the selected model for the visual imagery group contained main effects of delay, $b = -0.17$ units/sec, CRI = (-0.32, -0.02) and set size, $b = -0.64$ units/item, CRI = (-1.10, -0.16), indicating that only the length of the delay and number of items decreased performance reliably (see Fig. 2).

One explanation for the treatment effect might be a shift in response strategy rather than poorer memory for the items in the imagery condition. To test this, we calculated C as a measure of response bias ($C_{Imagery} = 0.13$, $C_{Control} = -0.02$) and d' as a bias-free sensitivity measure ($d'_{Imagery} = 1.70$, $d'_{Control} = 1.32$). The difference in C between the conditions, however, failed to reach reliability, $\Delta_C = 0.16$, CRI = (-0.02, 0.33), while d' did, $\Delta_{d'} = 0.38$, CRI = (0.03, 0.74).

Discussion

Overall, we found a disruptive effect of an auditory imagery task on the maintenance of timbre materials. In contrast, visual imagery did not reliably disrupt participants' maintenance activities, despite the fact that the visual imagery task seemed to be more difficult (as indicated by the reliably lower performance in this condition). Because of the similarity in the two tasks, one can be confident that specific cognitive activities carried out during the auditory but not the visual imagery task interfered with maintenance.

General discussion

We conducted two experiments exploring the maintenance mechanism for auditory-nonverbal materials. Participants had to maintain a number of sounds over a delay of several

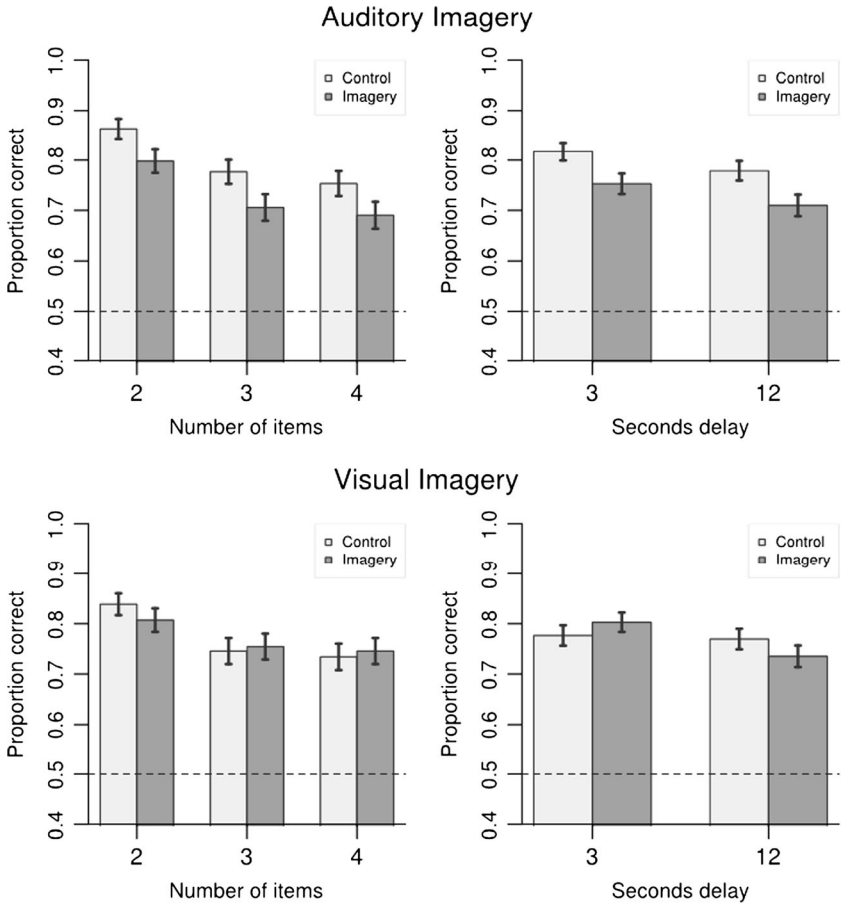

Fig. 2 Proportion correct per set size (left panels) and delay (right panels) in Experiment 2. Error bars depict standard errors

seconds and had to perform concurrent secondary tasks in some conditions. Experiment 1 demonstrated only small disruptions by articulatory suppression when two items had to be maintained. In Experiment 2, on the other hand, we found a reliable disruption of maintenance by a secondary auditory imagery task but not by a comparable visual imagery task.

From these results one can derive several important conclusions. First, we have shown that timbre information can be maintained over the short term even with higher item loads, essentially replicating a previous study of Golubock and Janata (2013). Second, we have shown that this maintenance is *active*; in other words, despite being subject to a very slow decline, performance in an auditory short-term memory task can additionally be disrupted by auditory imagery. Third, disruption by articulatory suppression was surprisingly small and inconsistent, which implies that articulatory rehearsal is likely insufficient for, or even not involved in, maintenance.

We conclude that within the framework of the multicomponent working memory model there is no existing mechanism available to explain the results of the two experiments. On the other hand, our results are consistent with the view that cognitive processes related to auditory imagery act as a maintenance mechanism for auditory information. One might even contend a stronger hypothesis, stating that auditory information is maintained in working memory separately (but not necessarily independently) from abstract-phonological information by means of processes similar to or identical with auditory imagery. This would solve the apparent contradiction in current versions of the phonological loop model that assume that nonphonological and unutterable materials are stored in a phonological code and rehearsed by an articulation-based mechanism.

In fact, a separate store for pitch information has previously been suggested by Williamson et al. (2010). However, Williamson et al. also proposed that information in this store is maintained by articulatory rehearsal. This might be a reasonable assumption in their case, because pitch it is an acoustic property that can be articulated ("sung"). On the other hand, it is questionable whether this assumption can be extended to timbre, which is arguably difficult to articulate. Because this difference might be crucial for the effectiveness of articulatory rehearsal in maintaining auditory information, pitch might be considered a special, though important, case of auditory-nonverbal information.

Finally, although it seems that humans are able to actively maintain auditory information, both experiments also demonstrated a small but reliable decline in performance with and without any intervening task. This phenomenon is well known in the literature on auditory short-term memory (McKeown & Mercer, 2012) and points to the possibility that active maintenance of auditory information has certain limits—a possibility future research should explore.

References

Baddeley, A. D. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology, 63,* 1–29.

Colle, H. A., & Welsh, A. (1976). Acoustic masking in primary memory. *Journal of Verbal Learning and Verbal Behavior, 15,* 17–32.

Cowan, N. (1984). On short and long auditory stores. *Psychological Bulletin, 96,* 341–370.

Crowder, R. G. (1989). Imagery for musical timbre. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 472–478.

Gaskell, M. G., Quinlan, P. T., Tamminen, J., & Cleland, A. A. (2008). The nature of phoneme representation in spoken word recognition. *Journal of Experimental Psychology: General, 137,* 282–302.

Golubock, J. L., & Janata, P. (2013). Keeping timbre in mind: Working memory for complex sounds that can't be verbalized. *Journal of Experimental Psychology: Human Perception and Performance, 39,* 399–412.

Hubbard, T. L. (2010). Auditory imagery: Empirical findings. *Psychological Bulletin, 136,* 302–329.

Hubbard, T. L. (2013). Auditory imagery contains more than audition. In S. Lacey & R. Lawson (Eds.), *Multisensory imagery* (pp. 221–247). New York, NY: Springer.

McAdams, S. (1999). Perspectives on the contribution of timbre to musical structure. *Computer Music Journal, 23,* 85–102.

McKeown, D., Mills, R., & Mercer, T. (2011). Comparisons of complex sounds across extended retention intervals survives reading aloud. *Perception, 40,* 1193–1205.

McKeown, D., & Mercer, T. (2012). Short-term forgetting without interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 1057–1068.

Pitt, M. A., & Crowder, R. G. (1992). The role of spectral and dynamic cues in imagery for musical timbre. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 728–738.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology), 64,* 532–563.

Williamson, V. J., Baddeley, A. D., & Hitch, G. J. (2010). Musicians' and nonmusicians' short-term memory for verbal and musical sequences: Comparing phonological similarity and pitch proximity. *Memory & Cognition, 38,* 163–175.