PsyArXiv preprint 10.31234/osf.io/vs48d

Auditory tracking of multiple naturalistic moving objects

Lauri Oksama[a] (Corresponding author) (email: loksama@utu.fi)

Timo Heikkilä[b] (email: timo.t.heikkila@utu.fi)

Lauri Nummenmaa[b, c] (email: latanu@utu.fi)

Mikko Sams[d] (email:mikko.sams@aalto.fi)

Jukka Hyönä[b] (email: hyona@utu.fi)

a Finnish Defence Research Agency, Human Performance Division
P.O. Box 5 (Rantatie 66, Tuusula)
FI-04401 Järvenpää, Finland

b University of Turku, Department of Psychology
Assistentinkatu 7
20014 Turku, Finland

c Turku PET Centre, Turku University Hospital,
P.O. Box 52, FIN-20521, Turku, Finland

d Departments of Neuroscience and Biomedical Engineering and Department of Computer Science,
Aalto University, P.O. Box 11000 (Otakaari 1B)
FIN-00076 AALTO, Finland

Declarations of interest: none

Word count: 3000 words

Abstract

Thirty participants tracked auditorily moving sound sources to estimate the capacity for multiple identity tracking by hearing. The participants sat blindfolded in a gym hall. Four assistants moved about semi-randomly in a circular area around the participant and constantly repeated a proper name. Two to four of the assistants were designated as the targets. The participants were to keep track of the designated targets during the 10-sec movement phase. After the movement stopped, one target was probed and the participant provided the name of the probed target. Auditory tracking capacity was estimated to be 1.5 items, which is half the size of the visual tracking capacity. It is suggested that the limited capacity for auditory tracking is related to the difficulty in refreshing what-where -bindings in the auditory modality.

Keywords

Auditory tracking - Multiple object tracking - Multiple identity tracking - Visual tracking - Dynamic auditory attention

1. Introduction

When walking on a busy street or in a park, vehicles or birds move about around you. You may not see many of them, but you can hear them moving. How well can you track moving auditory objects by relying just on your hearing? Can you point to the position of individual objects? Surprisingly, no previous research exists on the human capacity for auditorily tracking distinct moving targets through space.

Here we determine the human capacity for auditorily tracking moving sound sources. Visual tracking of moving objects has been extensively studied. Most studies have focused on position tracking using the MOT (multiple object tracking) paradigm, where identical objects are tracked (Pylyshyn & Storm, 1988). However, in real life it is imperative to also know the identity of objects among other moving objects. For instance, in traffic it is critical to be aware that a pedestrian is currently on one's left side and a car is on the right, and not vice versa. Oksama and Hyönä (2004, 2008) devised a multiple identity tracking (MIT) paradigm to examine how distinct visual targets are tracked, that is, to estimate the capacity of maintaining what-where -information in a dynamically changing visual scene. Their model of MIT proposes that to keep track of moving identities, each target identity needs to be bound to its correct location and the constructed identity-location bindings need to be constantly refreshed, as the targets continuously move about. This refresh mechanism is assumed to be based on overt or covert attention shifts between targets (Oksama & Hyönä, 2016).

The capacity of visually tracking distinct identities is estimated to be 3-4 items, and it varies as a function of the type of identity (e.g., Horowitz et al., 2007; Li et al., 2019; Oksama & Hyönä, 2004). Is the auditory tracking capacity similar to the visual one? One possibility is that it is smaller, because sound-source localization is not as precise as object localization in the visual modality. In the visual system, retinotopic maps allow accurate localization of objects (Golomb & Kanwisher, 2012). In the auditory system, sound-source localization is based on two types of cues on the horizontal plane (reviewed in Middlebrooks, 1991; Risoud et al., 2018): interaural time difference (i.e., the sound propagation time between the two ears) and interaural level difference (i.e., the intensity difference between the two ears for the same sound). Listeners are prone to make back-front errors in sound localization. Yet, Zhong and Yost (2017) showed that listeners are able to identify and locate at least four static sound sources. They also found that the separation of different sound sources is better for broad spectrum speech sounds (names of countries played from 12 different loudspeakers) than tonal sounds. As listeners are capable of perceiving four simultaneous sound sources (see also Eramudugolla et al., 2005), localization accuracy does not set a lower limit to the auditory tracking capacity when compared to visual tracking. Thus, it is possible that stimulus modality does not play a crucial role in determining the tracking capacity, but instead, common higher-order cognitive functions based on supramodal representation of space may be behind the capacity limits (e.g., Andersen et al., 1997; Farah et al., 1989). If so, the auditory tracking capacity should be similar to the visual tracking capacity.

We created a naturalistic environment to study auditory tracking (Figure 1). It has been argued that laboratory experiments may fundamentally misrepresent how cognitive processes (including attention) operate in situations mimicking real-world environments (Adolphs et al., 2016; Risko et al., 2016). Our blindfolded participants sat on a chair in the center of a gym hall. Four assistants moved quasi-randomly and silently (wearing wool socks) around the participant each reciting a male name. The participant's task was to track 2–4 moving targets. After 10 seconds, the movement stopped and one of the targets was probed by the chosen assistant sounding a beep from a handheld loudspeaker. All assistants then moved to a predefined area and the participant removed the blindfold. She was then given a name list from which to choose the probed name. Maximum target set-size was set to four, the number of simultaneously active static sound sources listeners are capable of identifying (Zhong & Yost, 2017). As sound localization is better for wideband sounds, we used speech sounds rather than tones as the targets. Response accuracy and tracking capacity were used as indices of performance success.

2. Methods

2.1 Participants

Thirty participants (university students, four men; mean age 24.0 years, SD=3.8) were recruited. They gave an informed consent for participation and received study credit in return. The sample size was determined to be similar to corresponding studies on visual tracking. Set-size effects in tracking are robust and can be reliably measured even with relatively small samples.

2.2 Hearing

Hearing ability was assessed using pure tone audiometry screening at 25 dB hearing level (0.5 kHz, 1 kHz, 2 kHz and 4 kHz). All participants passed the screening test.

2.3 Auditory tracking

The experiment was conducted in a gym hall, where a circle with a diameter of 14 meters was marked as the outer boundary of the experimental area. Another circle with a diameter of 4 meters was marked in the center of the area (see Figure 1 for the setup). The area between the inner and outer circles marked the boundaries where the assistants, i.e., the to-be-tracked targets, were allowed to move. The participants were tested in pairs so that they were seated blindfolded back-to-back on chairs in the middle of the inner circle. Four assistants were moving about within the designated area at a walking pace (average speed of 12.3 degrees/s) and repeating aloud a common male name at a steady rate and a normal speaking voice. The participants' task was to keep track of the locations and identities of the designated targets. At the end of each trial, one assistant (assigned as the target at the trial beginning) produced a probe sound (3 kHz) with a handheld electronic loudspeaker. The participant then chose from a name list the probed target's name (see supplementary materials for a video of a trial; https://osf.io/bxhpq/?view_only=1b3dcad5bf3e468da226d07dab0426db).

The experiment comprised three blocks, each containing 11 trials lasting for 10 seconds each. In the first block, two assistants were assigned as the targets, while the other two acted as distractors. In the second block, the number of targets was three (one distractor), and in the final block all four assistants were tracked (no distractors). Before each trial, the target assistants repeated aloud twice their name, while standing on randomly designated starting positions. The order in which the target identities were announced was randomized. The experimenter then signaled the trial to commence and timed the trial with a stopwatch, while the assistants were moving and repeating their names. After 10 seconds, the experimenter raised his hand to mark the trial end, after which one target was probed. The assistants then moved away from the experimental area to a predesignated area, after which the experimenter gave permission for the participants to remove their blindfold and mark down their answer. The answers were given by circulating the name from a printed list constituting all the names used in the trial. The mean latency between hearing the probe sound and being given the permission to answer was approximately 7 seconds.

The selected names were chosen from among the most common two-syllable Finnish male names provided by the Finnish Population Register Centre (2018). These names were randomized between the assistants, blocks and trials. Within a trial, all names were unique. Four lists were created for the assistants' use that, in addition to the trial-specific identities, contained information about whether they were one of the tracked targets. These lists also provided the assistants with their randomized starting positions and movement direction for each trial.

The starting positions were determined by dividing the experimental area into 8 sectors (similar to cardinal and intercardinal compass directions), and each sector was divided into 5 locations that determined the distance from the participants. The closest starting point was at the boundary of the inner circle and the farthest point at the boundary of the outer circle, while the three remaining points were evenly distributed between the two. Movement was quasi-random; the assistants were instructed to move towards a randomized trial-specific direction until they would collide either with a boundary or with another assistant. After this, they were to change direction as they pleased, while keeping at least one-meter distance from other assistants. Their task was also to ensure that their mutual positions from the point of view of the participants always changed from their starting positions. Prior to the experiment, a training session was held for the assistants to practice constant movement speed without making unnecessary noise (e.g., woolen socks were used to dampen the sound of the steps), to train in the usage of a similar voice volume and in the simultaneous termination of the name repetition at the trial end.

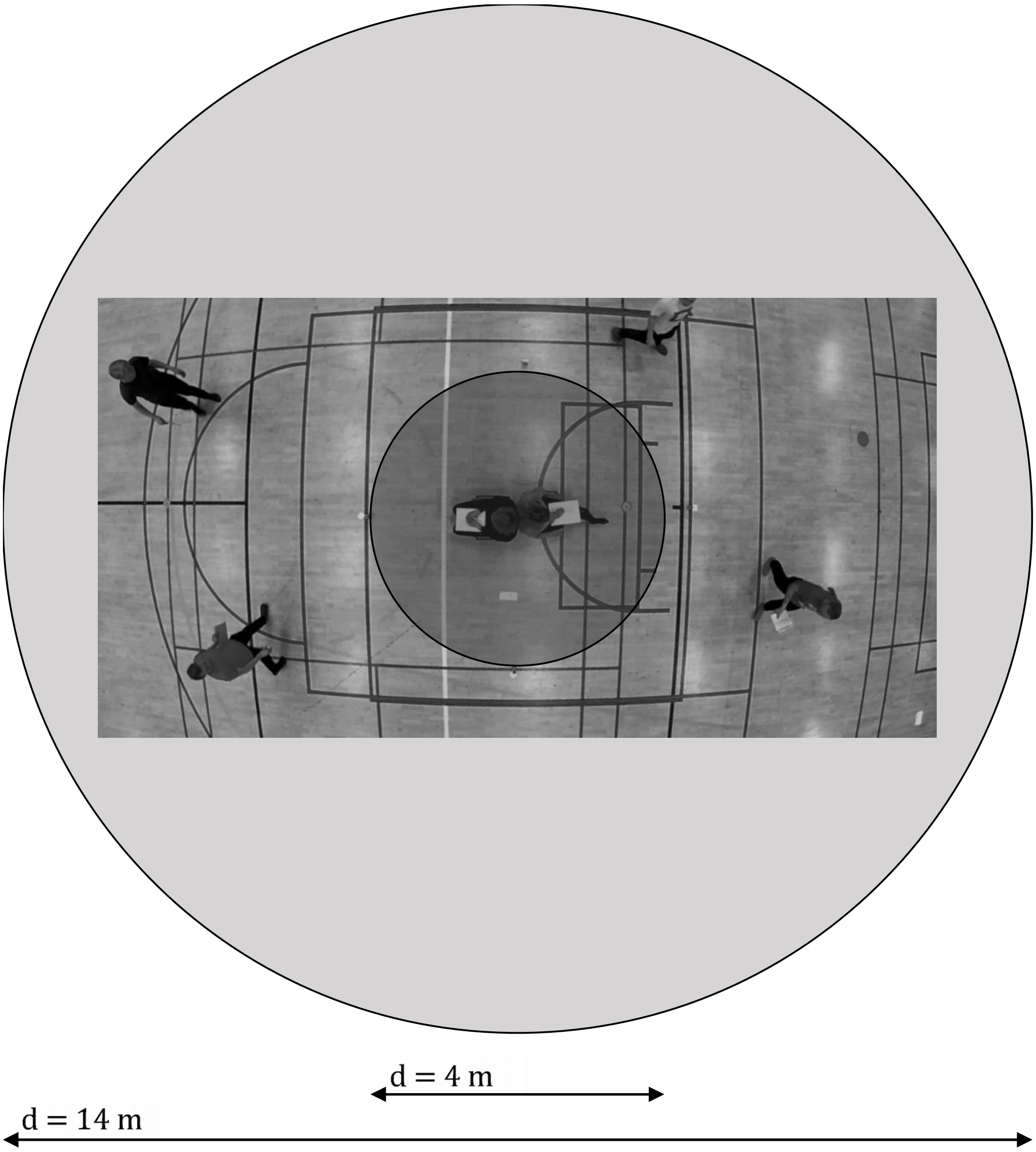

Figure 1. A screenshot of the experimental trial. The assistants were allowed to move only in the area between the inner and outer circle. The diameters of the inner and outer circles were 4 meters and 14 meters, respectively. Two participants were seated in the center back-to-back with blindfolds. See a supplementary video of two experimental trials https://osf.io/bxhpq/?view_only=1b3dcad5bf3e468da226d07dab0426db.

3. Results

3.1 Accuracy and capacity in auditory tracking

A repeated-measures analysis of variance (ANOVA) was performed on the performance accuracy (number of correct responses) in auditory tracking with target set-size (2–4) as the independent variable. The main effect of set-size was significant, $F(2,58)=47.58$, $\eta_p^2=.622$, $p<.001$. Performance deteriorated linearly as a function of set-size (see Table 1).

Tracking capacity was estimated using the following formula proposed by Horowitz et al. (2007). It assumes that participants are good at guessing.

$$k=\frac{a+pt-1-\sqrt{(1-a-pt)^{2}-4(apt-1)}}{2}$$, where

k =tracking capacity
a = number of possible response options
p = observed tracking accuracy
t = number of targets

An ANOVA revealed a significant effect of set-size in tracking capacity, *F*(2,58)=4.59, η_p^2=.137, *p*=.014. As apparent from Table 1, the capacity estimate was identical for set-size 2 and 3 but deteriorated for set-size 4.

Table 1. Performance accuracy and capacity in auditory tracking

Measure	Set-size 2	Set-size 3	Set-size 4
Performance accuracy (%)	86.29 (13)	71.72 (13)	57.18 (16)
Capacity estimate	1.57 (.38)	1.58 (.47)	1.29 .65)

3.2 Comparing auditory and visual tracking capacity

To compare the performance accuracy in auditory tracking to that of visual tracking, we conducted an ANOVA, where auditory tracking was compared to that of tracking line drawings of common objects (e.g., shoe, coat, watch; Oksama & Hyönä, 2004) and printed words (Hyönä et al., 2020). The speed of motion and target set-size were comparable to those of auditory targets. Set-sizes 3 and 4 were used in the analyses, as printed word tracking did not contain set-size 2. Fifty-six participants completed the object tracking experiment and 30 participants took part in the word tracking experiment. Tracking mode was entered as a between-participants factor. Performance accuracy was markedly lower in auditory than visual tracking (see Figure 2, left panel), as revealed by a highly significant main effect of tracking mode, *F*(2,113)=60.52, η_p^2 =.517, *p*<.001. The main effect of set-size was also significant, *F*(1,113)=93.64, η_p^2 =.453, *p*<.001; tracking was poorer with 4 than 3 targets. Their interaction was non-significant, *F*<1.2.

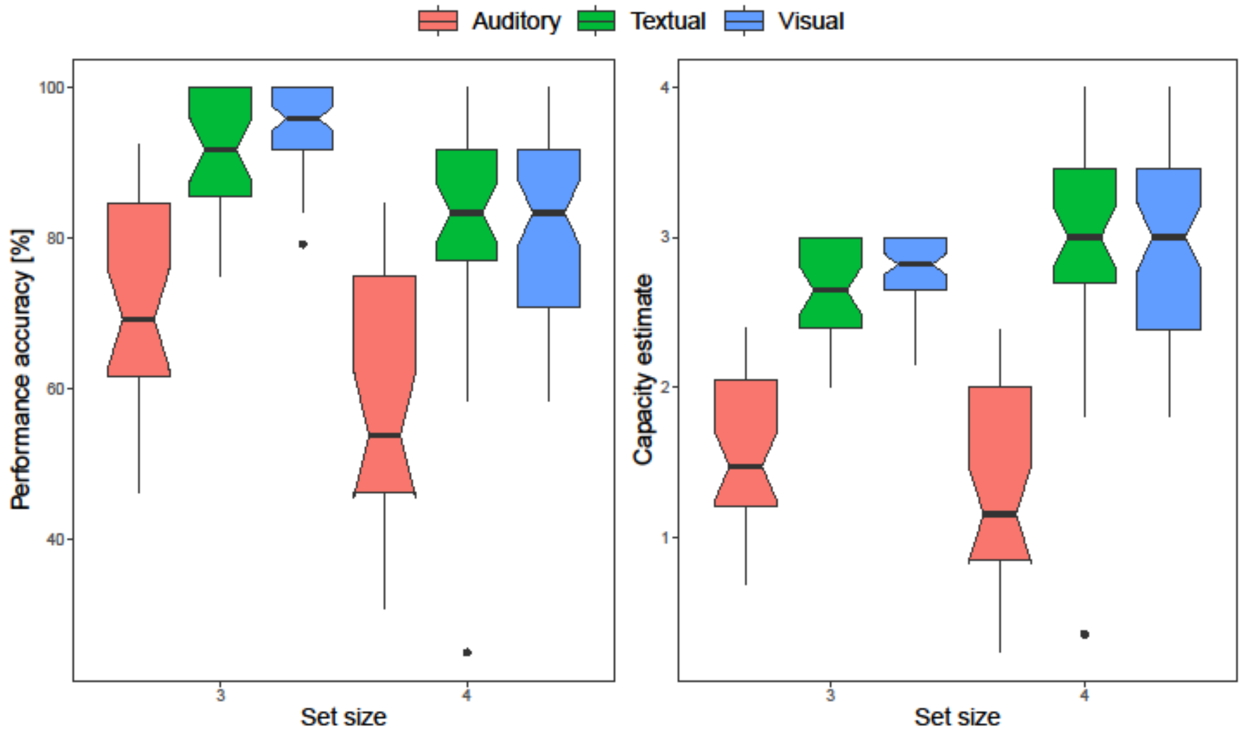

Figure 2. Performance accuracy (left panel) and tracking capacity (right panel) in auditory, visual object and printed word tracking.

In tracking capacity, ANOVA revealed a significant main effect of tracking mode, *F*(2,113)=111.59, η_p^2 =.664, *p*<.001. Capacity was much lower in auditory tracking, about half the size of that in visual tracking (see Figure 2, right panel). Tracking mode also interacted with set-size, *F*(1,113)=9.24, η_p^2 =.141, *p*<.001. In auditory tracking, the estimated tracking capacity was poorer for set-size 4 than 3, whereas the opposite was the case for visual tracking.

4. Discussion

We established, for the first time, the capacity limits for tracking moving targets in the auditory modality. All previous studies have investigated visual tracking (but see Woods & McDermott, 2015, for auditory tracking in static feature space). Our results show that auditory tracking of moving targets is remarkably difficult. In this task, the visual modality outperforms the auditory modality by a large margin. This is the case even when only two targets are tracked; the superiority of the visual modality becomes particularly noticeable with three and four targets (see Figure 2). We estimated the auditory tracking capacity to be about 1.5 items, while the visual tracking capacity is twice that size. As the auditory tracking capacity is only one or two items, it is no wonder why the performance breaks down with set-size 4, for which the capacity estimate is even smaller than that for set-size 3.

4.1 Why auditory tracking capacity is so limited?

According to the MOMIT model of Oksama and Hyönä (2008; see also Li et al., 2019), an effective tracking system temporarily stores in the short-term memory (1) identity, (2) location and (3) the identity-location –binding information of the targets. In addition, (4) a mechanism is needed which constantly updates the what-where –bindings. In the visual modality, this refresh mechanism is assumed to be based on overt or covert attention shifts between targets (Oksama & Hyönä, 2016).

According to MOMIT, identity tracking does not operate automatically but requires continuous and effortful attention. If what-where -bindings are not refreshed, situation awareness deteriorates and targets are lost from short-term memory/awareness. Thus, an active refresh mechanism is assumed to be an essential part of visual tracking. In principle, it may be possible to refresh bindings via a preattentative early-vision mechanism (Pylyshyn & Storm, 1988). However, all recent evidence points to attentional and oculomotor activity during tracking of visually distinct objects (for a review, see Hyönä et al., 2019).

In what follows, we apply the above theoretical framework to auditory tracking to discern possible strengths and weaknesses of the auditory system. (1) The temporary storage of identity information should not be a major problem for the auditory system. The capacity of the phonological store (Baddeley, 1986) is sufficient for maintaining via the rehearsal mechanism the 2-4 target identities (proper names) needed in the present task. (2) Sound localization of several simultaneous sound sources is not a problem for the auditory system either, as listeners are able to localize four separate static sound sources (Zhong & Yost, 2017). (3) Analogously to the visual system, what –where – bindings of the tracked sound objects may be stored in the amodal episodic buffer (Baddeley, 2000).

However, (4) the active refresh mechanism of what-where bindings appears as the most vulnerable part of the auditory tracking mechanism. As argued above, the effective tracking system responsible for the maintenance of dynamic information, be it visual or auditory, updates what-where bindings by continuously switching the attentional focus between targets. It is unlikely that the auditory system would possess any effective automatic (early perceptual) updating mechanism for what-where bindings. Thus, in order to refresh which auditory object is where, the auditory system has to either overtly turn the head (and the ears) toward the sound sources or covertly switch the focus of attention between the moving targets represented by the sensory auditory system. Overt attentional shifts in terms of head movements may provide some help (e.g., Perrett & Noble, 1998) in the updating process (and the additional directional ear movements in some animals). However, overt shifting of auditory attention by head movements between moving targets is less accurate and much slower than overt or covert attention shifts in the visual system. Thus, temporarily lost targets may be difficult to recover by overt head movements. Covert auditory attentional shifts between targets may also be more difficult. The auditory scene consists of overlapping sound sources spreading out across the frequency map of the cochlea, while visual objects typically occupy local regions in the retina (McDermott, 2009). Thus, the sensory analysis of the complex auditory signal would need time, and therefore, slows down the updating process causing situation awareness to lag behind the present moment, at least with several targets. Our results are consistent with this explanation.

Another possibility is that auditory tracking is based on crossmodal orienting of attention. Spence et al. (2017) review evidence suggesting that “vision naturally guides the (re-)calibration of spatial hearing” (p. 1140). This may take the form of code conversion from the auditory modality to the visual one. In this alternative, auditory information is converted to visuospatial representations with the help of visual imagery. However, such a mechanism would be laborious and demanding, as it requires continuous and probably very difficult code conversion and updating of visual imagery held in the visuospatial short-term memory. Yet, our results cannot refute this explanation. Spence et al. also review evidence suggesting that audio-motor or tactile spatial information can recalibrate spatial hearing. However, this possibility is not readily applicable to the present results, as our participants had not access to such information.

Finally, our results are inconsistent with the idea of a higher-order tracking system with supramodal spatial representations common to both the auditory and visual system. If such a higher-order system existed, similar tracking capacities should have been observed in both modalities. Instead, marked differences in tracking capacity point to modality-specific tracking systems.

5. Conclusions

We started this paper by asking how many moving birds or vehicles we can track just by relying on hearing. The present results suggest a surprising and unintuitive answer: we can auditorily track only about one bird at the time, while visually we can track more than twice as many targets. We argue that the limited capacity for auditory tracking is related to the difficulty in refreshing what-where - bindings in the auditory modality.

Acknowledgments. Portions of these findings were presented in a poster session at the 2019 Vision Sciences Society meeting.

Funding. This work was supported by the Academy of Finland [grant #273413] to Lauri Oksama

Supplemental Material. All data have been made publicly available via OSF and can be accessed at https://osf.io/bxhpq/?view_only=1b3dcad5bf3e468da226d07dab0426db. Additional video material can be found at https://osf.io/bxhpq/?view_only=1b3dcad5bf3e468da226d07dab0426db

References

Adolphs, R., Nummenmaa, L., Todorov, A., & Haxby, J.V. (2016). Data-driven approaches in the investigation of social perception. *Philosphical Transactions of The Royal Society of London Series B, 371*

Andersen, R. A., Snyder, L. H., Bradley, D. C., & Xing, J. (1997). Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience*, *20*, 303–330. https://doi.org/10.1146/annurev.neuro.20.1.303

Baddeley, A. (1986). *Working memory*. Oxford: Oxford University Press.

Baddeley. A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*, 417-423. https://doi.org/10.1016/S1364-6613(00)01538-2

Eramudugolla, R., Irvine, D. R. F., McAnally, K. I., Martin, R. L., & Mattingley, J. B. (2005). Directed attention eliminates 'change deafness' in complex auditory scenes. *Current Biology*, *15*, 1108-1113. https://doi.org/10.1016/j.cub.2005.05.051

Farah, M. J., Wong, A. B., Monheit, M. A., & Morrow, L. A. (1989). Parietal lobe mechanisms of spatial attention: Modality-specific or supramodal? *Neuropsychologia*, *27*(4), 461–470. https://doi.org/10.1016/0028-3932(89)90051-1

Golomb, J. D., & Kanwisher, N. (2012). Retinotopic memory is more precise than spatiotopic memory. *Proceedings of the National Academy of Sciences*, *109*(5), 1796-1801.

Horowitz, T. S., Klieger, S. B., Fencsik, D. E., Yang, K. K., Alvarez, G. A., & Wolfe, J. M. (2007). Tracking unique objects. *Perception & Psychophysics*, *69*(2), 172–184. https://doi.org/10.3758/BF03193740

Hyönä, J., Li, J., & Oksama, L. (2019). Eye behavior during multiple object tracking and multiple identity tracking. *Vision*, 3(3), 37: https://doi.org/10.3390/vision3030037

Hyönä, J., Oksama, L., & Rantanen, E. (2020). Tracking the identity of moving words: Stimulus complexity and familiarity affects tracking accuracy. *Applied Cognitive Psychology*, 34(1), 64-77.

Li, J., Oksama, L., & Hyönä, J. (2019). Model of Multiple Identity Tracking (MOMIT) 2.0: Resolving the serial vs. parallel controversy in tracking. *Cognition*, *182*, 260-274. https://doi.org/10.1016/j.cognition.2018.10.016

McDermott, J.H. (2009). The cocktail party problem. *Current Biology*, 19, R1024-R1027. https://doi.org/10.1016/j.cub.2009.09.005

Middlebrooks, J.C. (1991). Sound localization by human listeners. *Annual Review of Psychology*, *42*, 139-152.

Oksama, L., & Hyönä, J. (2004). Is multiple object tracking carried out automatically by an early vision mechanism independent of higher-order cognition? An individual difference approach. *Visual Cognition*, *11*, 631-671.

Oksama, L., & Hyönä, J. (2008). Dynamic binding of identity and location information: A serial model of multiple identity tracking. *Cognitive Psychology*, *56*, 237-283.

Oksama, L., & Hyönä, J. (2016). Position tracking and identity tracking are separate systems: Evidence from eye movements. *Cognition*, *146*, 393-409.

Perrett, S. & Noble, W. (1997) The contribution of head motion cues to localization of low-pass noise. *Attention, Perception & Psychophysics*, *59*, 1018–1026. DOI: 10.3758/bf03205517

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, *3*(3), 179–197. https://doi.org/10.1163/156856888X00122

Risko, E. F., Richardson, D. C., & Kingstone, A. (2016). Breaking the fourth wall of cognitive science: Real-world social attention and the dual function of gaze. *Current Directions in Psychological Science*, *25*(1), 70–74. https://doi.org/10.1177/0963721415617806

Risoud, M, Hanson, J.-N., Gauvrit, F., Renard, C., Lemesre, P.-E., Bonne, N.-X., & Vincent. C. (2018). Sound source localization. *European Annals of Otorhinolaryngology*, *Head and Neck Diseases*, *135*, 259-264. https://doi.org/10.1016/j.anorl.2018.04.009

Spence, C., Lee, J., & van der Stoep, N. (2017). Responding to sounds from unseen locations: Crossmodal attentional orienting in response to sounds presented from the rear. *European Journal of Neuroscience*, *51*, 1137-1150. doi:org.10.1111/ejn.13733.

Woods, K.J.P., & McDermott, J.H. (2015). Attentive tracking of sound sources. *Current Biology*, *25*, 2238-2246. https://doi.org/10.1016/j.cub.2015.07.043

Zhong, X., & Yost, W.A. (2017). How many images are in an auditory scene? *The Journal of the Acoustical Society of America*, *141*, 2882-2892. https://doi.org/10.1121/1.4981118