Fall 2019

The Top-Down Influences of Characteristic Sounds on Visual Search Performance in Realistic Scenes

Ghazaleh Mahzouni
San Jose State University

THE TOP-DOWN INFLUENCES OF CHARACTERISTIC SOUNDS ON VISUAL SEARCH PERFORMANCE IN REALISTIC SCENES

A Thesis

Presented to

The faculty of the Department of Psychology

San José State University

In Partial Fulfillment

of the Requirement for the Degree

Master of Arts

by

Ghazaleh Mahzouni

December 2019

The Designated Thesis Committee Approves of the Thesis Titled

THE TOP-DOWN INFLUENCES OF CHARACTERISTIC SOUNDS ON VISUAL SEARCH PERFORMANCE IN REALISTIC SCENES

by

Ghazaleh Mahzouni

APPROVED FOR THE DEPARTMENT OF PSYCHOLOGY

SAN JOSÉ STATE UNIVERSITY

December 2019

Cary Feria, Ph.D.	Department of Psychology
Evan Palmer, Ph.D.	Department of Psychology
Mark Van Selst, Ph.D.	Department of Psychology

ABSTRACT

THE TOP-DOWN INFLUENCES OF CHARACTERISTIC SOUNDS ON VISUAL SEARCH PERFORMANCE IN REALISTIC SCENES

by Ghazaleh Mahzouni

The purpose of this experiment was to investigate whether meaningful sounds can facilitate visual search performance in the context of realistic scenes. It also aimed to determine whether the stimulus onset asynchrony (SOA) of sound and picture is a significant factor in enhancing performance. A 3 X 4 X 2 within subject design was used with independent factors sound congruency (congruent, incongruent and white noise), SOA (-1000, -500, 0, 300 ms), and target presence (present and absent). Participants were 55 (34 female and 21 male) college aged students at San Jose State University. On each trial participants were presented with a word cue indicating the target object, then depending on the condition they either 1) heard a sound and saw a picture simultaneously (SOA 0), 2) heard a sound followed by a scene (negative SOA), or 3) viewed a scene followed by a sound (positive SOA). The results indicated a congruency effect only at the negative SOAs, when the sound preceded the picture by 1000 or 500 ms. However, we did not observe a significant advantage of -1000 SOA over -500 SOA. Moreover, performance was significantly degraded at the positive SOA 300. Overall, these results suggest that congruent characteristic sounds can enhance visual search performance in realistic scenes, provided that they are presented at least 500 ms before the picture.

AKNOWLEDGMENTS

I am immensely grateful to my mentor and advisor Dr. Cary Feria. This project would not have been possible without her endless help and support. Dr. Feria, thank you for guiding me every step of the way and encouraging me to pursue what I'm passionate about. Your enthusiasm, hard work, and dedication have been a source of inspiration throughout my academic journey. I would also like to thank Dr. Evan Palmer and Dr. Mark Van Selst for their helpful insights, patience, and willingness to serve as my committee members.

I would like to thank my family who continuously encouraged and motivated me to finish this project. To my lovely mom Vida, thank you for your unconditional love, support and words of encouragement. To my kind sister, Sahar, thank you for always making me laugh and believing that I could do this. Last but not least, a special thank you to my cat Mr. Meow Meow, whose constant purrs have made this arduous process a little more pleasant.

TABLE OF CONTENTS

List of Tables viii

List of Figures ix

Introduction 1
- Research Problem 1
- Literature Review 1
- Bottom-up Processing in Multisensory Integration 3
- Top-Down Processing in Multisensory Integration 4
- The Effect of Meaningful Stimuli on Perception and Behavior 6
  - Characteristic sounds. 6
  - Real-world scenes. 9
- Conceptual Short-Term Memory (CSTM) 10
- The Role of the SOA in Multisensory Integration 12
- Deficiencies in the Literature 13
- Significance of the Study 14
- Research Question and Hypothesis 14

Method 18
- Participants 18
- Research Design 18
- Apparatus and Stimuli 19
  - Realistic scenes. 19
  - Characteristic sounds. 22
- Procedure 22

Results 25
- Three-Way ANOVA 25
- Two-Way ANOVA 26
- Simple Main Effect of Congruency in Each SOA Condition 27
- Simple Main Effect of SOA in Each Congruency Condition 29
- SOA -1000 vs. SOA -500 30
- Sound vs. No Sound 31
- Three-Way ANOVA 32
- Two-Way ANOVA 33
- Accuracy with No Sound vs Sound 33

Discussion 35
- Hypothesis 1 35
- Hypothesis 2a 35
- Hypothesis 2b 38
- Hypothesis 2c 39

Hypothesis 3 40
Implications 40
Limitations 41
Future Direction 42
References 44

Appendices 48
Appendix A : Consent Form 48
Appendix B : Questionaire 50

LIST OF TABLES

Table1. A list of target objects and their characteristic sounds that were used in the main experiment…… 20

LIST OF FIGURES

Figure 1.	A diagram representing the CSTM theory……...	11
Figure 2.	Structure of a single trial……...	24
Figure 3.	A schematic diagram representing the manipulation of the SOAs between realistic scenes and characteristic sounds……..	26
Figure 4.	A graph comparing RTs in the incongruent and congruent trials at each SOA condition…….	31
Figure 5.	Results of the two-way ANOVA…….	34
Figure 6.	Accuracy graph that shows the results in the two-way ANOVA with factors SOA and congruency……...	36

Introduction

Research Problem

Traditional research has primarily focused on how unimodal sensory signals influence perception and behavior in isolation. However, we live in a multisensory world where different sensory signals from the environment are combined to form a coherent representation of the world. Sound, in particular, is a rich source of information that can convey object identity, meaning, or location in space. Research based on visual search performance suggests that presenting congruent characteristic sounds that provide no spatial information can facilitate finding the target object faster (Iordanescu, Guzman-Martinez, Grabowecky & Suzuki, 2008). However, this phenomenon has not been investigated in visual search performance in realistic scenes. In addition, the optimal stimulus onset asynchrony (SOA) for presenting the characteristic sounds and picture of realistic scenes has not been addressed yet.

Literature Review

While there is a large body of psychological research dedicated to visual search performance, very few studies have looked at the multisensory nature of meaningful sounds on visual search in the context of realistic scenes. Consequently, this lack of investigation has resulted in limited understanding of the effects of multisensory integration in real life situations. For this reason, the current study aimed to extend the generalizability of prior findings by utilizing realistic pictures of real-world scenes and meaningful sounds.

Multisensory integration refers to the process of combining inputs from different sensory modalities. For a long time, it was believed that visual and auditory inputs from the environment are first processed independently within their primary cortices and only combined at later stages of cortical processes. However, Falchier, Clavagnier, Barone and Kennedy (2002) provided anatomical evidence for the connectivity of the primary visual cortex (v1 or area 17) and the auditory cortex of Cynomolgus monkeys. They injected retrograde tracers to the visual cortex of the monkeys and found that this area receives direct projections from the auditory cortex.

Further neuroimaging studies have demonstrated the same connectivity in the human brain. For instance, Romei, Murray, Merabet and Thut (2007) applied transcranial magnetic stimulation (TMS) over the occipital pole of the human brain and measured behavioral responses to visual or auditory and visual stimuli. They found that TMS slowed reaction times to the visual stimuli, confirming that activity within the visual cortex was inhibited by the TMS pulse. However, reaction times to the visual stimuli were faster when the visual stimuli were paired with a simple auditory tone. This was taken as evidence that the activity within the primary visual cortex was enhanced in the presence of the auditory stimuli than vision alone. In other words, the brain areas that were inhibited by the TMS pulse were "disinhibited" when an auditory tone was present.

The activation of visual cortex as a result of auditory input has a direct influence on perception. For example, Franssinetti, Bologinin and Ladvas (2002) presented a faint flash of green light at one of the several possible locations on the screen. The flash was either presented alone or with a simultaneous tone that was irrelevant to the task. They

found that visual sensitivity (d') for detecting the flash was increased when the tone was present. Overall, these studies suggest that the visual cortex is excited in the presence of auditory information which can consequently enhance perception and cognitive functions such as attention.

Given that our cognitive system is capacity limited, attention is required to select relevant information. Attention can be directed in two ways: 1. Bottom-up processing in which attention is captured automatically based on the properties of the stimuli. 2. Top-down processing which is the voluntary allocation of attentional resources based on the observer's goals, intentions and relevance to the goals. (Theeuwes, 1991; Wolfe, Butcher, Lee & Hyle, 2003). The following two sections reviews evidence of bottom-up and top-down processing in multisensory settings.

Bottom-up Processing in Multisensory Integration

Most research on multisensory integration has focused on rudimentary auditory stimuli and their effect on perception. Research using simple tones suggests that auditory information can enhance visual perception in a bottom-up manner. For instance, Stein and London (1996) demonstrated that a brief auditory tone can significantly enhance the perceived intensity of LED lights. Similarly, Van der Burg, Oliver, Bronkhorst and Theeuwes (2008) demonstrated that a simple auditory "pip" can significantly reduce the search time for finding an otherwise hard-to-find target. They presented participants with a visual search array that contained several vertical and horizontal line segments. On each trial, the target and distractors changed colors continuously, making the target hard to find. Participants were required to make a speeded response to the orientation of the

target line. On some trials a simple auditory "pip" accompanied color change. The results indicated that reaction times for finding the target was significantly lower when the auditory "pip" was present. They argued that the auditory stimulation automatically makes the target object pop out in a bottom-up manner. In another study, Matusz and Elmer (2011) showed that multisensory integration enhances the saliency of sensory input which in turn enhances attentional capture to facilitate performance. They found that visual search performance was faster when the target cue (color change) was accompanied by a brief tone.

Top-Down Processing in Multisensory Integration

In contrast, some studies show that enhancement of visual and auditory processing is a result of top-down cognitive processes rather than automatic processes. For example, Talsma and Woldorff (2005) presented participants with random auditory, visual or audiovisual stimuli in two lateral spatial positions. The participants were instructed to attend to only one of the spatial locations. The event-related potential (ERP) analysis showed that when participants were instructed to pay attention to the unisenosry visual target, the N1 and P1 components, which are generally known to be related to attentional enhancement, were enhanced compared to when they were not paying attention. In addition, when they were paying attention to the unisensory audio target, the N1 component was enhanced compared to when they were not paying attention. In the multisensory condition, when the stimuli were presented at the attended location, the audiovisual condition showed a greater amplitude for both the N1 and P1 component compared to either of the unimodal stimuli. The researchers also found that the

unattended location elicited considerably smaller N1 and P1 responses. This provides evidence that directing attention in a voluntarily manner can modulate the multisensory process.

In a subsequent study Talsma, Doty and Woldorff (2007) presented participants with a rapid serial visual presentation (RSVP) task which included a stream of letters appearing above a fixation cross. Every 1-10 seconds the letter was replaced by a random digit that served as the target. Directly below the fixation cross there was either a visual stimulus, which was a horizontal square wave grating, auditory stimuli, a tone pip, or audiovisual stimuli which was a simultaneous presentation of the wave grating and the tone. The results showed that when participants were instructed to attended to both stimuli, the early component of the ERP (P50) was enhanced compared to attending to only the visual or auditory stimuli. This finding provides further evidence for the influence of top-down directed attention on responding to bimodal stimuli.

Additionally, Lippert, Logothetis and Kayser (2007) showed that participants only showed improvement in performance when the sound carried information about the visual target that was not obtainable from the visual display. Furthermore, Laurienti, Kraft, Maldjian, Burdette and Wallace (2004), demonstrated that pairing visual objects with semantically congruent auditory stimuli can facilitate object recognition. For instance, recognition of a red or blue circle was improved only when the circle was presented simultaneously with the verbalization of the target congruent color. Moreover, object recognition was degraded in the presence of incongruent auditory stimuli. Overall,

these studies demonstrate that observers can guide their attention voluntarily based on their intentions and knowledge of stimuli in a multisensory stetting.

The Effect of Meaningful Stimuli on Perception and Behavior

Characteristic sounds. Although research has clearly established the effects of simple sounds on visual perception, little is known about characteristic sounds. Characteristic sounds are sounds that inform us about the identity of an object. For example, barking, meowing and keys jingling clearly identify specific objects: dog, cat and keys. Several behavioral and neuroimaging studies have shown that characteristic sounds can significantly enhance performance in cognitive tasks such as object recognition, categorization tasks and visual search performance. The idea is that when the meaningful sounds matches the target object (congruent), the semantic information from the characteristic sound excites the visual cortex, which then crossmodally enhances visual processes. Below we review neuroimaging and behavioral evidence of this process.

Molholm (2004) measured high density event related potentials (ERPs) while participants performed an object recognition task. The task either paired pictures of animals with their characteristic sounds or the picture of the animal was shown alone. Participants made a speeded response to the appearance of the target animal. The ERP result indicated the activation of occipito-temporal cortices, which is part of the ventral stream, known for processing object recognition. Behavioral results showed that reaction times were significantly faster and more accurate at identifying the target animal when the picture and sound were congruent, than when the target was presented without sound.

This indicates that even if the target is easily identifiable through one sensory input (vision) input from another sense (sound) can interact to enhance object recognition. This shows that visual and auditory input interact to enhance ventral stream processing.

Other behavioral studies have also shown that semantically meaningful sounds enhance object identification. Chen and Spence (2010) used an identification task to investigate audiovisual semantic congruency. Participants were briefly shown a line drawing picture of an animal, which was immediately masked by a scrambled picture. They then had to make an un-speeded response identifying the animal by typing the name of the animal on a keyboard. The stimulus onset asynchrony (SOA) of sound and picture were varied such that the characteristic sound could occur simultaneously with the picture (SOA 0) or 300 or 522 ms after the picture (positive SOA). They did not manipulate the negative SOAs (when sound occurred before picture) in this study. The results showed that picture identification was improved when the congruent characteristic sound and the line drawing were presented simultaneously (SOA 0) and when the onset of sound was delayed by 300 ms (SOA +300). Interestingly, no semantic congruency effect was observed when the onset of sound was delayed by 522 ms (positive SOA) This finding suggests that that the semantic congruency effect of characteristic sound is present at SOA 0 and can be extended with an auditory delay of 300 ms.

In a subsequent study Chen and Spence (2011) used a picture detection task to investigate the influence of characteristic sounds when they occurred before the picture (negative SOAs). In this experiment, participants first saw a blank frame for 600 ms, then followed by either a frame of picture (line drawing) or another blank frame. The

congruent characteristic sound could occur simultaneously (SOA 0) or 345 ms before the picture (negative SOA). Positive SOAs were not manipulated in this study. The visual stimulus was then masked by scrambled drawing. Participant had to indicate whether they had seen a picture (regardless of its identity) right before the mask. Results showed that only characteristic sounds enhanced sensitivity to semantically congruent pictures when sound occurred 345 ms before the picture (negative SOA).

Using a categorization task (living vs. non-living), Chen and Spence (2018) tested the congruency effect of characteristic sounds with 7 different SOAs (-1000, -500, -250, -100, 0, 100 and 250). They found that responses were faster and more accurate only when the auditory cue was presented 250, 500, or 1000 ms before the visual pictures (negative SOAs). The congruency effect was not present with the simultaneous presentation of sound and picture (SOA 0). The authors suggested that a when the sound precedes the picture a short-term buffer for semantic processing is necessary to allow for each sensory modality to access their meaning and integrate to enhance performance.

Prior studies have shown that characteristic sounds that provide non-spatial information facilitate finding the location of a target object. Iordanescu et al. (2008) presented four pictures of common objects in quadrants and the search display was accompanied by the simultaneous presentation of characteristic sounds. The characteristic sounds were either consistent with the target, consistent with the distractor or a sound unrelated to anything on the display. The results showed that search was more efficient when the target was paired with target-consistent sound than a distractor-consistent or unrelated sound. Interestingly, while target consistent sounds facilitated search

performance, the target inconsistent sound did not impair performance relative to unrelated sounds. One criticism of this study is that random and unrelated pictures were used without any context. In real life however, objects do not appear randomly in their surroundings. Since there was no scene information, the ability to generalize these results to real-world situation is limited.

While the above studies have provided compelling evidence for the enhancement of visual perception in the presence of sounds, two questions still remain unresolved: can characteristic sounds influence more realistic tasks such as visual search performance in real world scenes? Is temporal disparity between auditory and visual stimuli a significant factor in obtaining congruency effect in realistic scenes?

Real-world scenes. Real-world scenes are complex, and attention is required for scene perception (Wolfe, Alvarez, Rosenholtz, Kuzmova & Sherman, 2011). Studies have shown that humans are particularly good in scene perception. In fact, significant information about a scene can be extracted in a very brief glimpse (Biederman, Rabinowiz, Glass & Stacy, 1974). For example, participants are able to accurately identify the scene type (outdoor vs. indoor) as quickly as 45-135 ms, by an analysis of global features (Henderson & Hollingworth, 1999). In addition, one can understand the meaning of a complex novel scene when the image is blurred (Schyns & Oliva, 1994). Moreover, the gist of a scene can be acquired in a single glance (Biederman, 1981; potter, 1999). To illustrate, in a go/no-go task in which participants had to decide whether a photograph that was flashed for only 20 ms contained an animal, ERP measures revealed that this could be achieved in less than 150 ms (Thorpe, Fize & Marlot, 1996).

Understanding the gist of a scene allows one to access the schema representation of object identity and spatial location (Henderson, 2003).

However, finding the location of objects within the context of a scene is generally more difficult and require more time than 150 ms (De Graef, 1990). De Graef (1990) argues that there are several stages in finding an object in a scene. The first stages of scene perception are for the scene specific information in which scene schemas are activated from memory and later stages are based on the object information. The cognitive guidance theory states that the most important factor that guides attention within a complex scene is meaning (Henderson & Hayes, 2018). This view holds that attention is directed by the cognitive system to the specific scene regions that are semantically informative and relevant to the observer's goal. For instance, Eckstein, Koehler, Welbourne and Akbas (2017) mis-scaled some objects within realistic scenes such that they were significantly larger than other items in surroundings. They found that the mis-scaled target (ex. A giant parking meter) was missed because it was not part of the goal of the observer. This finding shows that the top-down cognitive factors such as the goal of the observer, are more important than bottom-up saliency in scene perception.

Conceptual Short-Term Memory (CSTM)

CSTM is a form of working memory in selective attention that helps access conceptual representation of stimuli in the environment (Potter, 1999). According to the CSTM theory when people observe a scene or hear a sound, a series of conceptual information regarding those stimuli are quickly activated and held in CSTM. This leads to the retrieval of additional relevant information from the long-term memory (LTM).

The relevant information is then dynamically structured together to achieve the goal of the observer. If information is not incorporated or selected it will be forgotten immediately. Potter (1993,1999) argues that the entire process is very quick and takes less than 1 second. Figure 1 shows a diagram representing the structure of this theory.

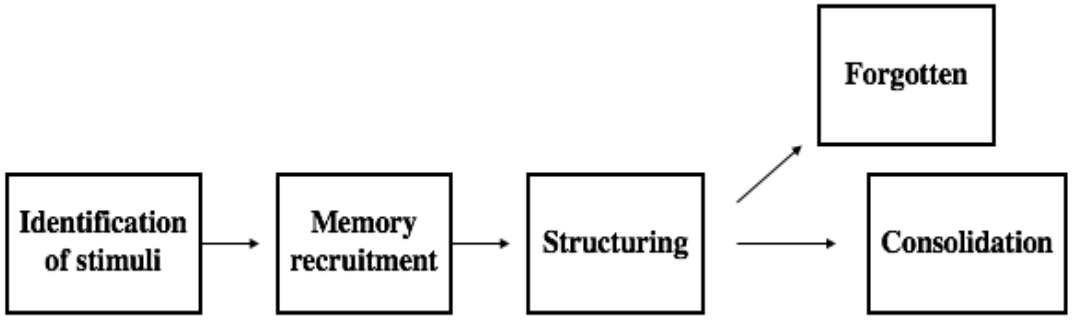

Figure 1. A diagram representing the CSTM theory. Once the stimuli are identified a semantic representation of the stimuli is quickly activated. Relevant information is recruitment from long-term memory and linked to achieve the observer's goal. Linked information can be consolidated into long-term memory and unused information is rapidly forgotten.

In the present experiment the goal of the observer was to locate a known target object in a complex scene while hearing a characteristic sound (congruent or incongruent to the target) or white noise. The CSTM theory would suggest that the congruent characteristic sound provides coherent information regarding the identity and the location of the target object which enhance the consolidation of the scene stimuli. On the other hand, the incongruent sound activates a conceptual representation that is not helpful the goal of the observer and it will not be structured with the scene stimuli (Chen & Spence, 2010;

Potter,1999). Similarly, the white noise does not provide any relevant information, so it will not be linked with the scene stimuli.

The Role of the SOA in Multisensory Integration

The above studies show that the combination of sensory information from different senses about a common source enhances object identification, discrimination or localization. King (2005) argues that a necessary step for this process is for the sensory signals to bind together. He further argues that the co-occurrence of sensory stimuli is a powerful cue in this process. This is based on the idea that there are multisensory neurons in the brain (such as those in the superior colliculus) that are activated when sound and picture are presented simultaneously (Meredith, Nemitz & Stein, 1987). In this view, response to a visual stimulus is enhanced only if it is similarly accompanied by a sound. However other studies have shown that response to visual stimuli can be enhanced when the onset of sound is varied. The SOA refers to the amount of time between the start of one stimulus and the start of a second stimulus. Previous studies have shown contradicting results regarding the influence of the SOA on tasks involving visual and auditory stimuli. Some studies indicated that performance is enhanced when the characteristic sound precedes the picture, i.e., negative SOA (Chen & Spence, 2011; Chen and Spence 2018), while others have found that performance is also enhanced when the sound is lagged by 350 ms, i.e., positive SOAs (Chen, 2011; Meredith, Nemitz & Stein, 1987).

These discrepancies can be attributed to the type of task used in each experiment. For instance, in the picture detection and picture categorization tasks where negative SOAs

have shown improvements, participants made a speeded response to a quickly shown line drawing. In this case when the characteristic sounds were presented before the picture, participants had enough time to access the meaning of the sound before completing the task, therefore, the sound could significantly enhance perception. However, in the same task at SOA 0, the sound did not have adequate time to access its meaning before the completion of task and could not influence perception. In contrast, in the picture identification tasks, where positive and 0 SOA have shown improvements, participants made un-speeded responses to visual stimuli by simply typing the identity of the target on the keyboard. It can be argued that in this task participants were able to wait until the meaning of the sound was achieved before task competition, thus at positive or 0 SOA the sound could enhance performance. Overall, these studies seem to suggest that the time frame for characteristic sounds to influence visual perception is flexible and depends on the task at hand. It is not clear how the temporal dynamics of characteristic sounds will affect performance in a visual search task with more complex stimuli such as realistic scenes.

Deficiencies in the Literature

Although research on visual search is rich, studies have not looked at the multisensory nature of visual search performance in the context of realistic scenes. Thus far, characteristic sounds have only been shown to facilitate visual search among isolated pictures of unrelated objects. Currently, no published studies have investigated whether characteristic sounds that are spatially uninformative can facilitate finding an object in a

complex scene. In addition, no research has investigated the temporal dynamics of characteristic sound in a complex scene.

Significance of the Study

The current study utilized a cognitive task that is relevant to everyday life activity to investigate the top-down influences of sounds on performance. More specifically, we used characteristic sounds that provide information about an objects' identity, but not its location in space, on finding that object in a complex and realistic scene. The results of this study enhance the ecological validity of prior findings and a better understanding of how attention is distributed among realistic stimuli.

Research Question and Hypothesis

In this study we were interested in answering the following questions:

(1) Can congruent semantic information from auditory stimuli enhance visual search performance in realistic scene in a top-down manner?

(2) Is there an advantage of presenting the auditory information first to allow proper sensory integration?

(3) Is timing between the presentation of stimuli crucial for multisensory integration during complex scene searches.

To answer these questions, we presented participants with a visual search task containing realistic scenes and characteristic sounds that can be either congruent or incongruent to the target objects or hear white noise or no sound. We also systematically varied the SOA to gain a better understanding of the influence of temporal disparity between each stimulus.

Hypothesis 1: Participants will be faster in finding the target objects when the scene is paired with a target-congruent sound compared to incongruent and white noise.

Hypothesis 2:

(a) The congruency effect will be different at different levels of the SOA. Specifically, the congruency effect will only be present when the congruent sound is presented before the scene (negative SOAs) compared to when the congruent sound is presented after the scene (positive SOA).

(b) In the congruent condition, RTs will be significantly different at each level of the SOA (SOA -1000 < -500 < 0 < 300). Whereas, in the incongruent condition and white noise condition, RTs will not be significantly different at each level of the SOA.

(c) There will be a trend where longer and negative SOA will lead to stronger congruency effect. The longer negative SOA -1000 will lead to a stronger congruency effect than the shorter negative -500 SOA.

Hypothesis 3:

Performance in the no sound condition will be higher than RTs in all congruent sound conditions.

The rationale for the first hypothesis is that hearing a characteristic sound that matches the target object will guide one's attention to the likely location that the target would appear in the scene, thus, making that object easier to find. To illustrate, in a study Iordanescu, Grabowecky, Franconeri, Theeuwes and Suzuki (2010) measured eye movements during visual search performance with congruent and incongruent characteristic sounds and found that the time it took for participants to saccade to the

target object was significantly reduced with the characteristic sound was congruent to the target object in comparison to when it was incongruent.

The rationale for hypothesis 2a is that performance seems to be enhanced when sound precedes the picture by a certain amount of time. For instance, Chen and Spence (2011) found that the characteristic sounds give rise to semantic congruency effect in picture identification task only when the sound lead the picture by more than 346 ms. They argue that this amount of time is enough for the meaning of the sound to be fully accessed.

The SOAs for the present experiment are -1000, -500, 0 and 300 ms. These specific SOAs were chosen based on the results of prior studies that used characteristic sounds and visual stimuli. For instance, the SOA -1000, and -500 were chosen because Chen and Spence (2018) demonstrated congruency effect in a picture categorization task when the characteristic sound lead the pictures at SOA -1000 and -500. SOA 0 was chosen because Iordanescu, et al. 2008, found that search performance was enhanced when congruent characteristic sounds were presented simultaneously with the search display. Moreover, the SOA 300 was chosen because some studies have shown characteristic sounds still enhance object identification even if auditory stimuli briefly lagged visual targets. For example, Chen and Spence (2010) found that identification of masked picture was facilitated at SOA +300 but not +500. However, we argue that it is unlikely that our experiment would show a congruency effect at SOA 300 due to complexity of our visual stimuli. The realistic scenes used in our experiment is more complex, thus, the meaning of the scene cannot be held in the CSTM and bind to the characteristic sound when the scene appears before the sound.

The rationale for hypothesis 2c is that longer negative SOA (SOA -1000) would allow more time for the CSTM to access the conceptual representation of the characteristic sound and retrieve relevant information from long term memory to guide attention to the likely location of the target object.

The rationale for hypothesis 3 is that evidence form behavioral and neuroimaging studies show that performance is facilitate when two congruent stimuli from different senses are combined. For this reason, we expect that the no sound condition will have higher reaction times than all the congruent sound conditions.

Method

Participants

Participants for this study were college aged students who self-reported normal or corrected-to-normal vision and hearing. All participants were recruited from the San José State University (SJSU) psychology pool and received course credit for participation. Written informed consent were obtained from all participants prior to participation in the study. This study was approved by the Institutional Review Board (IRB) at SJSU. The number of participants was determined prior to the study to achieve adequate statistical power. G*power suggested that a minimum of 48 participants were needed to conduct a within-subject analysis of variance (ANOVA) with an effect size of alpha = .25 and power = .85. A larger sample of 60 was included in order to account for technical difficulties and attrition rate. Six participants were excluded from the experiment due to technical difficulties with the computer; thus, the total sample size for this experiment was 55 (34 female and 21 male) students.

Research Design

This experiment was a 3 X 4 X 2 factorial design. The independent variables were sound (congruent, incongruent, white noise), SOA (-1000, -500, 0, 300 ms) and target presence (present and absent). The dependent variables were reaction time (RT) in milliseconds (ms) and proportion of correct responses. We also included a no sound condition which only included the target present and absent variables.

In the congruent condition, the target object was always presented with its characteristic sound (e.g., picture of a dog in a park was matched with the sound of

barking). In the incongruent condition the target object was presented with a sound that was not relevant to anything in the scene (e.g., a picture of a cat sitting on a bed was paired with the sound of toaster popping). In the control condition, the scene was paired with white noise. Additionally, in the no sound condition, the scene was presented alone, without any auditory input. The four SOAs and the no sound condition were divided into five separate blocks. Each SOA block contained six conditions of congruent/present, incongruent/present, white noise/present, as well as congruent/absent and incongruent/absent and white noise/absent. There were 20 trials in each condition, resulting in a total of 120 trials in each block. The no-sound block also contained 120 trials but only included the target present and target absent conditions (60 target- present trials and 60 target-absent trials). The order of trials within each block was randomized.

Apparatus and Stimuli

The visual stimuli were presented on a 20-inch monitor at a resolution of 1920 x 1080 with a 120 Hz refresh rate. Participants were seated approximately 57 cm away from the monitor. The auditory stimuli were presented via two loud speakers that were placed on each side of the computer monitor. The experiment was generated and presented via Opensesame (Mahot, Schreij & Theeuwes, 2012).

Realistic scenes. The visual stimuli for this experiment consisted of 620 color images of realistic scenes that included various scene types such as bedroom, bathroom, living room, kitchen, street, office, and farm. Target objects within the scene were chosen such that they were placed in a reasonable location, easily recognized, and were not occluded by other objects. The target objects could be household objects, musical instruments,

animals, etc. There were 20 target objects in the experiment; a list of all target objects is included in Table 1.

Table 1

A list of target objects and their characteristic sounds that were used in the main experiment

Target Objects	**Scene Types**	**Characteristic Sounds**
Alarm Clock	Bedroom	Alarm ring
Car	Street	Car engine
Cat	Indoor	Meow
Clock	Living space	Clock ticking
Computer Mouse	Office	Mouse clicking
Cow	Farm	Moo
Dog	Outdoor	Barking
Duck	Lake	Quack
Egg	Kitchen	Egg cracking
Faucet	Bathroom	Water
Guitar	Living space	Strumming
Keyboard	Office	Typing on keyboard
Keys	Living Space	Jingling
Motorcycle	Street	Motorcycle engine
Phone	Living Space	ringing
Piano	Living Space	Piano note
Rooster	Farm	Crowing
Shoes	Living Space	footsteps
Toaster	Kitchen	Toaster popping
Wine glass	Kitchen	Tapping on wine glass

Pictures were selected from Google images and were presented only once in the experiment. In target present conditions, the target object was in a reasonable location within a relevant scene (e.g., an alarm clock was on a bedside table in a bedroom). In target absent condition, the scene was relevant to the target object that was indicated at the beginning of the trial, but it was not in the scene (e.g. picture of a bathroom with no

faucet). There were 30 different scenes for each of the target object; 15 unique scenes for target present and 15 unique scenes for target absent condition. For instance, the target object, toaster, appeared in 15 different kitchen setting; there were also 15 kitchen scenes that did not include a toaster. The target object within each scene could look different, however they were always the same type. For example, a phone was always a landline phone, never a smart phone, or the toaster was always a pop-up toaster and not a toaster oven.

In order to control for the complexity of the pictures, participants were randomly assigned to three picture groups. Participants in group one saw the same set of scenes for the congruent, incongruent and white noise conditions. To illustrate, for one third of participants, the scene containing the target object keys on a blue lanyard was always paired with the congruent sound of jingling, regardless of the order of the SOAs. Additionally, the scene containing keys on a shelf was always paired with a random incongruent sound and the scene with keys on a nightstand was paired with white noise. In contrast, for participants in group two, the keys on the blue lanyard was paired with a random incongruent sound, while the keys on shelf was paired with white noise, and keys on nightstand was paired with the congruent jingling sound. For the last third of participants, the keys with blue lanyard was paired with white noise, keys on shelf with jingling sound and keys on night stand was paired with an incongruent sound. This allowed us to make sure that each scene was presented in all the congruency conditions regardless of the order of the SOAs, thereby reducing the influence of the picture complexity on performance. It is important to note that participants saw 15 unique scenes

for all the 20 target objects in each group. The order of the blocks was counterbalanced using Latin square.

Characteristic sounds. The auditory stimuli for this experiment were 26 characteristic sounds that represented the target objects. All sounds were downloaded from www.freesound.org. Only one sound was chosen to represent a target object. All sounds were trimmed to 850 ms. To ensure that the sounds were a good representation of the target objects, a pilot study was done with seven naive participants. The participants listened to each sound once and indicated the object that they thought the sound represented. All sounds that were chosen for this study reached 100% accuracy rate among the participants. In the congruent conditions, only the sound that matched the target object was used from the list. For the incongruent conditions, a sound that did not matched the target object was randomly assigned from the list of 26 sounds.

Procedure

Participants were tested individually in a dark and quite room. The researcher read all instructions out loud to the participants. Participants completed 15 practice trials to familiarize themselves with the experiment. The target objects for the practice trials were the following: bird, blackboard, camera, paper, pig. Data from the practice trials were excluded from the analysis. Following the practice session, participants began the main experiment which included 4 blocks of SOA and 1 block of no sound condition, each containing 120 trials. Participants were given a mandatory 3-mintures breaks in between blocks to avoid eye strain or fatigue. During the breaks they were allowed to walk outside

of the lab, drink water or use the bathroom. The entire experiment lasted approximately one hour.

Each trial began with a fixation cross at the center of the screen. After 500 ms a written word indicating the target object was presented at the center of the screen for 650 ms. Then, depending on the condition, either a realistic scene was presented followed by the characteristic sound or a characteristic sound was heard followed by a realistic scene at varying SOAs. The characteristic sounds were presented for 850 ms and the scenes were presented for 500 ms (see Figure 2 for the structural of the trials).

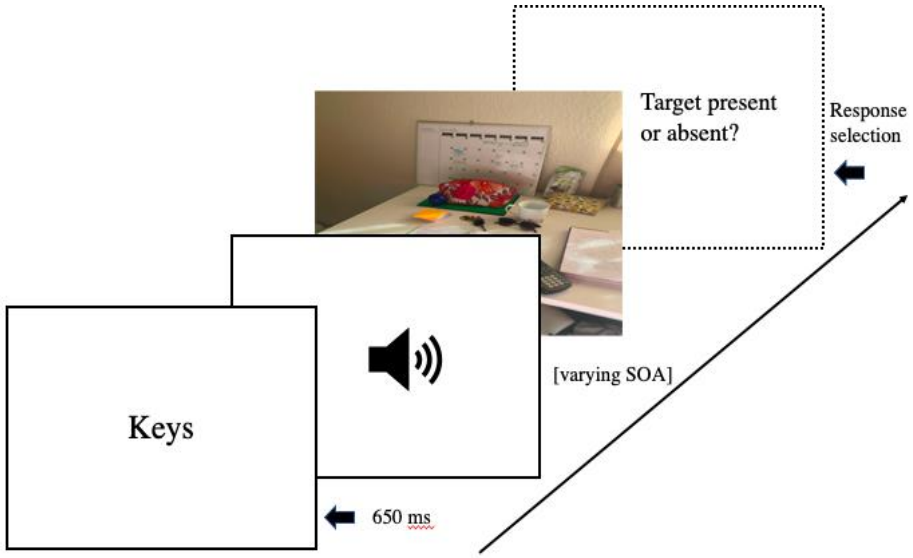

Figure 2. Structure of a single trial. Participants first saw a written cue. After 650 ms, depending on the condition, either a realistic scene or a sound was presented. This example shows a negative SOA where the sound was presented first. In the positive SOA the picture preceded the sound and in the 0 SOA condition the sound and picture were presented at the same time.

The characteristic sounds could be either congruent or incongruent with the target object. The incongruent conditions did not have objects that could be associated with the incongruent sound. The possible SOAs for this experiment were -1000, -500, 0, 300 ms.

These SOAs represented the amount of time between the start of picture and the start of sound. A positive SOA indicated that the picture precedes the sound, while a negative SOA indicated that the sound precedes the scene. An SOA of 0 indicates the simultaneous presentation of the scene and the sound together (Figure 3). The target object could be present in 50% of trials or absent from the scene in the other 50%. Participants were asked to press the right arrow on the keyboard as soon as they found the target object and press the left arrow if the target was not present in the scene. They were given feedback of their performance at the end of each trial. Additionally, at the end of each block accuracy rate and RTs were presented on the screen. Once participants finished the experiment, they provided verbal answers to a brief questionnaire.

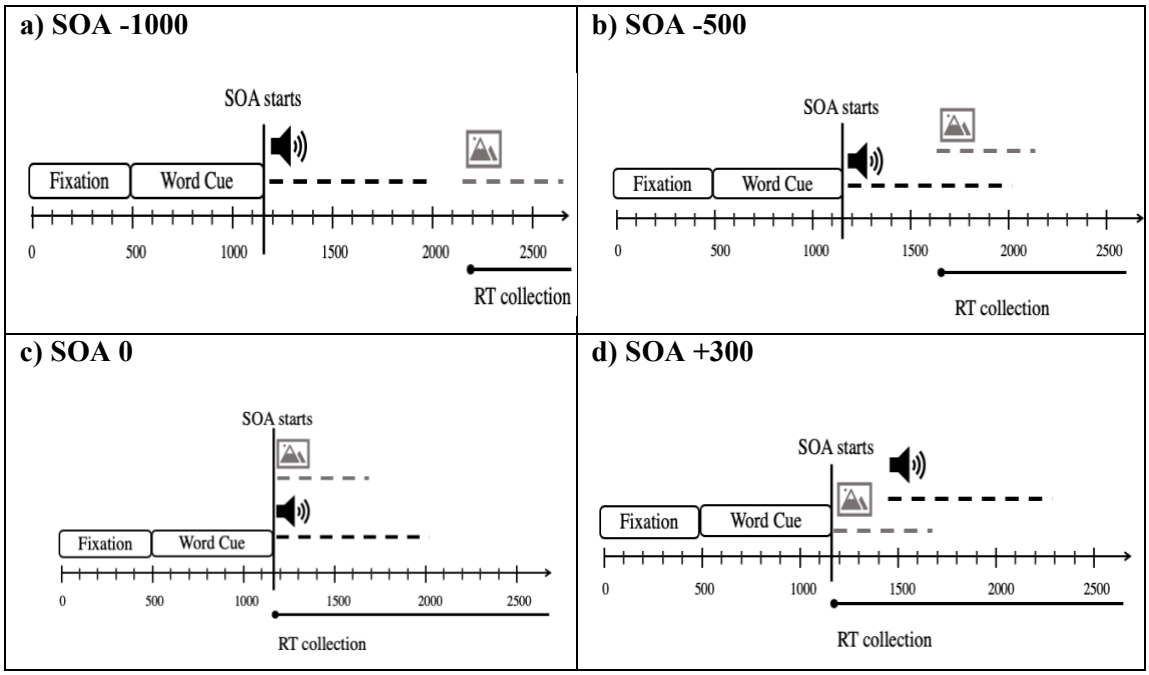

Figure 3. A schematic diagram representing the manipulation of the SOAs between realistic scenes and characteristic sounds.

Results

Data from the 15 practice trials were excluded from the analysis. The average RT in ms were collected from all conditions. RTs less than 200 and greater than 3000 ms were considered outliers and removed from all analysis. This exclusion criteria removed .02% of all data. The number of correct trials was divided by the total number of trials to calculate accuracy rates in all conditions. Once accuracy rate was established, all the incorrect trials were deleted from the RT analysis. This removed 3.05% of the overall data. Thus, the main analysis only included RTs in correct trials and target present conditions. Greenhouse-Geiser correction was used to account for the sphericity assumption in all conditions with more than 2-levels.

Three-Way ANOVA

First, RT data were submitted to a three-way repeated measures ANOVA with factors SOA (-1000, -500, 0, 300 ms), congruency (congruent, incongruent, white noise) and presence (present and absent) to assess whether there is a main effect of target presence. All ANOVA analyses were conducted using SPSS Statistics Version 25.

Results showed that there was a significant main effect of presence, $F(1, 54) = 212.334$, $p < .001$. Reaction times were significantly faster in the present condition ($M = 797.39$, $SD = 156.72$) than the absent condition ($M = 1068.75$, $SD = 259.83$). After this, all the absent trials were removed from the analysis as they are not helpful to our hypotheses.

Two-Way ANOVA

Hypothesis 1: To test the hypothesis that hearing a target congruent sound will facilitate search performance, a separate repeated measures two-way ANOVA was conducted with variables SOA (-1000, -500, 0, 300 ms), congruency (congruent, incongruent, white noise), using only the RTs from the present and correct trials. In addition, we employed Bayes' Factor (BF) calculations with default priors using JASP statistical software (JASP Team, 2019) in our post hoc comparisons. This test was done because we are making the assumptions that in some conditions RTs will be the same rather than different. BF calculations test how likely our data are under the alternative hypothesis (RTs in some conditions are different) compared to the null hypothesis (RTs are the same). Additionally, unlike traditional frequentist statistics, Bayesian statistics can provide evidence in support of the null hypothesis. As such, we included the BF calculation for all post hoc tests (Kass & Raftery, 1995). According to Kass and Raftery (1995) a BF value less than one provides evidence for accepting the null hypothesis. BF values between 1 and 3 are considered not worth more than a mention, BF 3 to 10 provide moderate evidence, and BF > 10 provides strong rejecting the null hypothesis.

Results showed that there was a Significant main effect of the SOA, $F(1.8, 94.9) = 15.24$, $p < .001$, $\eta_{p2} = .220$. Least significance difference (LSD) post hoc test was used to test the difference between the SOAs. It showed that RTs were significantly longer at SOA 300 ($M = 887.27$, SD = 250.31) than at SOA 0 ($M = 765.17$, $SD = 164.82$), SOA-500 ($M = 766.89$, $SD = 154.98$), and SOA -1000, ($M = 770.24$, $SD = 152.71$), $p < .001$, all BF > 10, indicating "strong" evidence against the null hypothesis. The other comparisons

did not reach statistical significance, $p > .05$, BF < 1, which provides positive evidence for accepting the null hypothesis.

Additionally, there was a significant main effect of congruency, $F(2, 107.9) = 11.857$, $p < .001$, $\eta_{p2} = .180$. LSD post hoc showed that RTs in the congruent trials were significantly faster ($M = 778.43$, $SD = 154.70$) than incongruent sounds ($M = 813.45$, $SD = 153.68$), $p < .001$, BF > 10, and white noise ($M = 800.30$, $SD = 170.41$), $p < .01$, BF = 2.56. The difference between incongruent and white noise was not statistically significant, $p > .05$, BF = .23. This BF value provides positive evidence that the white noise and incongruent conditions are the same. The interaction between SOA and congruency did not reach statistical significance, $F(4.9, 262.8) = 1.955$, $p = .088$, $\eta_{p2} = .035$.

Simple Main Effect of Congruency in Each SOA Condition

Hypothesis (2a): To understand whether the congruency effect was different at each SOA we looked at the simple main effect of congruency at each SOA. There was a significant simple main effect of congruency at SOA -1000, $F(1.9, 101.1) = 7.560$, $p < .01$, $\eta_{p2} = .123$. LSD post hoc analysis showed that at SOA -1000, RTs were significantly faster in the congruent condition ($M = 736.80$, SD = 143.11) than in the incongruent ($M = 785.34$, $SD = 164.43$), $p < .01$, BF = 18.01 and white noise condition ($M = 788.57$, $SD = 186.23$), $p <. 01$, BF = 10.49. The difference between incongruent and white noise did not reach statistical significance, $p > .05$, BF = .15. The BF value provides evidence in favor of the null hypothesis. This allows us to conclude that at SOA -1000, the RTs in incongruent and white noise conditions are the same.

The simple main effect of congruency at SOA -500 was significant, $F(1.9, 102.5) = 6.438$, $p < .01$, $\eta_{p2} = .107$. LSD post hoc analysis showed that at SOA -500, RTs were significantly faster in the congruent condition ($M = 739.77$, $SD = 141.55$) than RTs in the incongruent condition *(*$M = 783.48$, $SD = 158.40$), $p < .001$, BF = 54.91 and white noise condition ($M = 777.43$, $SD = 191.09$), $p < .01$, BF = 5.18. The difference between incongruent and white noise conditions was not statistically significant, $p > .05$, BF = .16. This BF value provides evidence in favor of the null hypothesis, which allows us to conclude that at SOA -500 the RTs in white nose and incongruent conditions are the same.

The simple main effect of congruency at SOA 0 was significant, $F(2.0, 105.8) = 3.870$, $p < .05$, $\eta_{p2} = .067$. LSD post hoc analysis showed that the at SOA 0 the RTs in the congruent condition were significantly faster ($M = 749.09$, $SD = 174.11$) than incongruent ($M = 790.38$, $SD = 190.05$) condition, $p < .05$, BF = 2.99. The difference between congruent and white noise was not significant, $p > .05$, BF = .16. This BF factor provides positive evidence that at SOA 0 the congruent and white noise are the same. The white noise condition ($M = 756.04$, $SD = 170.18$) was significantly faster than incongruent, $p < .05$, BF = 1.00.

The simple main effect of congruency at SOA 300 was not statistically significant, $F(1.7, 89.4) = .435$, $p > .05$. All BF comparisons were <1 supports the idea that the at SOA 300, the RTs in congruent, incongruent and white noise were the same. Figure 4 depicts a graph comparing RTs in the congruent and incongruent conditions at each SOA condition. This effect was calculatd by subtracting the average RTs in the

incongruent/present conditiosn from congruent/present conditions in each SOA condition. The graph clearly shows significant congruency efffect at SOA -1000, -500, 0 but not SOA 300.

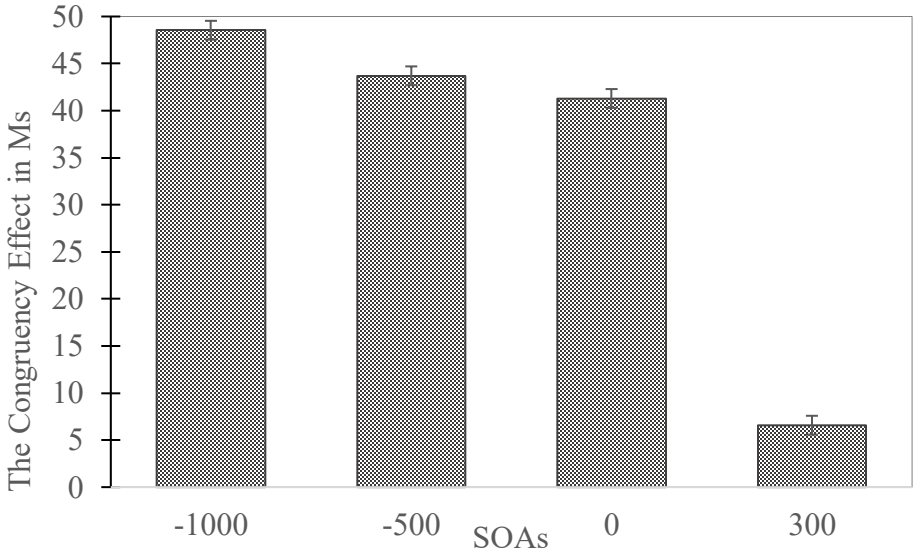

Figure 4. A graph comparing RTs in the incongruent and congruent trials at each SOA condition.

Simple Main Effect of SOA in Each Congruency Condition

Hypothesis (2b): To understand the interaction better, we looked at the simple main effect of SOAs in each of the congruency condition. The results indicated that the simple main effect of SOA in the congruent condition was significant, $F(1.6, 89.0) = 20.145$, $p < .001$, $\eta_{p2} = .272$. LSD post hoc showed that in the congruent condition, at SOA 300 ($M = 888.04$, $SD = 262.41$) RTs were significantly longer than the RTs at SOA 0 ($M = 749.09$, $SD = 174.11$), $p < .001$, SOA -500 ($M = 739.77$, $SD = 141.55$) and SOA-1000 ($M = 736.80$, $SD = 143.11$), $p < .001$; all BF values for these comparisons were > 10. The

difference between other SOAs did not reach statistical significance, $p > .05$, all BF values were < 1.

The simple main effect of SOA in the incongruent condition was also significant, $F(2.4, 131.0) = 9.396$, $p < .001$, $\eta_{p2} = .148$. Similarly, LSD post hoc revealed that in the incongruent condition, RTs at SOA 300 were significantly longer ($M = 894.62$, $SD = 240.40$) than SOA 0 ($M = 790.38$, $SD = 190.05$), $p < .01$, SOA -500 ($M = 783.48$, $SD = 158.40$), $p < .001$ and -1000 ($M = 785.34$, $SD = 164.43$), $p < .001$, BF factors for these comparisons were > 10. The other comparisons did not reach statistical significance, $p > .05$, BF values < 1.

The simple main effect of SOA in the white noise condition was significant, $F(2.1, 111.9) = 8.028$, $p < .001$, $\eta_{p2} = .129$. Similar to congruent and incongruent conditions the LSD post hoc showed that in the white noise condition the RTs were significantly longer at SOA 300 ($M = 879.15$, $SD = 276.66$) than SOA 0 ($M = 756.04$, $SD = 170.18$), $p <.001$, BF > 10, SOA -500 ($M = 777.43$, $SD = 191.09$), BF = 6.31 SOA -1000 ($M = 788.57$, $SD = 186.23$), $p < .01$, BF = 7.51. All the other comparisons did not reach statistical significance, p > .05, BF < .01.

SOA -1000 vs. SOA -500

Hypothesis (2c): To test whether there is trend where the longer and negative SOA leads to faster RT, we conducted a separate two-way ANOVA with SOA (-1000, -500) and congruency (congruent, incongruent, white noise). The results revealed that the main effect of SOA was not significant, $F(1, 54) = .073$, $p > .05$. However, the main effect congruency was significant, $F(2.0, 107.6) = 13.635$, $p < .001$, $\eta_{p2} = .202$. LSD post hoc

showed that the congruent condition (M = 738.28, SD = 134.55) was significantly faster than incongruent (M = 784.41, SD = 147.86) and white noise (M = 783.00, SD = 173.89), $p < .001$, BF > 10. The difference between white noise and incongruent was not statistically significant, $p > .05$, BF < 1. The interaction between the SOAs and congruency was not statistically significant, $F(1.9, 101.6) = .263$, $p > .05$.

Sound vs. No Sound

It is also of interest to find out whether search performance was different in the no sound condition compared to the sound conditions. To answer this question, 12 dependent samples t-tests were used to compare the RTs in the no sound/target present conditions to the sound/target present conditions (congruent, incongruent, white noise) in all SOAs (-1000, -500, 0, 300).

The results showed that SOA -1000/congruent condition (M = 736.80, SD = 143. 11) was significantly lower than the no sound/target present condition (M = 785.96, SD = 184.79), $t(54) = 2.371$, $p < .05$, BF = 1.07. Similarly, at SOA -500 congruent (M = 739.77, SD = 141.55) RTs were significantly lower than the no sound condition, $t(54) = 2.241$, $p < .05$, BF = 1.07. All sound conditions at SOA 300 were significantly different that the no sound condition. The congruent (M = 888.04, SD = 262.41), $t(54) = -3.655$, $p < .01$, BF = .79, incongruent (M = 894.62, SD = 240.39), t(54) = - 4.759, $p < .001$, BF = 1.28 and white noise (M = 879.15, SD = 276.66), $t(54) = -3.124$, $p < .01$, BF = .51, were all significantly higher than the no sound condition. None of the other comparisons reach statistical significance, p > .05, BF < 1 (Figure 5).

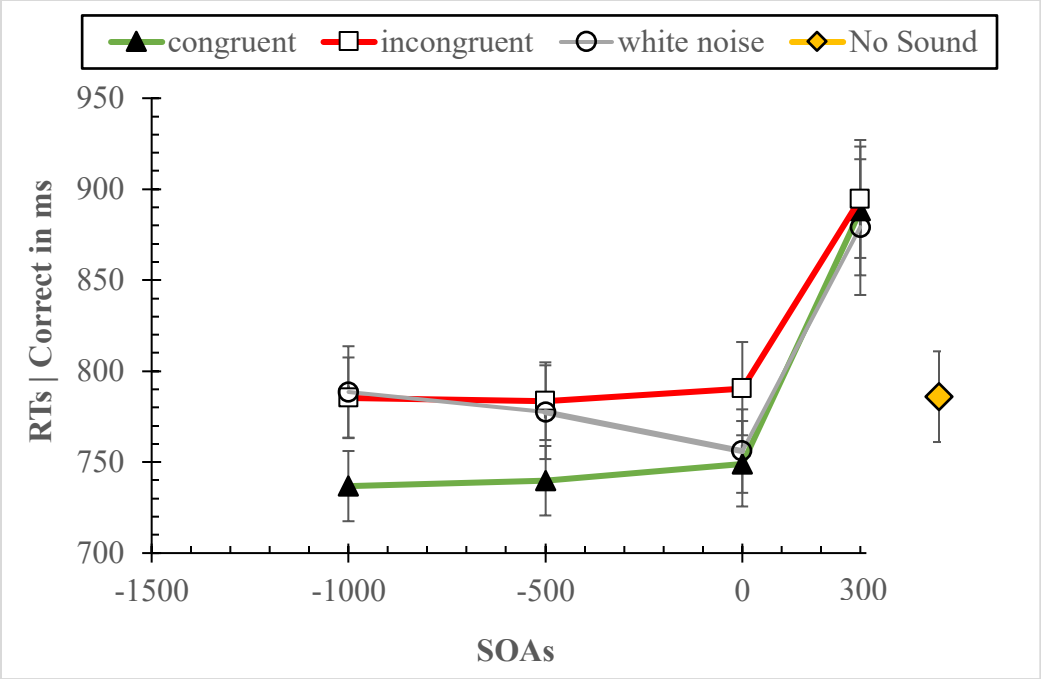

Figure 5. Results of the two-way ANOVA that shows the interaction between congruency (congruent, incongruent, white noise) and the SOA (-1000, -500, 0, 300) conditions. A congruency effect was observed at SOA -1000, -500 and 0, but not at SOA 300. Additionally, RTs in the no sound condition were significantly higher than SOA -1000/congruent and -500/congruent, but significantly lower than all the sound conditions of SOA 300.

Accuracy Results

Three-Way ANOVA

To identify whether there was a speed accuracy trade off, the proportion of the correct responses were submitted into a three-way ANOVA with factors SOA (-1000, -500, 0, 300) and congruency (congruent, incongruent, white noise) and presence (present and absent). The results revealed a significant main effect of presence, $F(1, 53) = 71.698$, $p < .001$, $\eta_{p2} = .575$. Accuracy rate was significantly higher in absent trials ($M = .98$, $SD = .01$) than present trials ($M = .96$, $SD = .02$). This shows that while participants were faster

in target present conditions, they were less accurate in finding the target objects. Likewise, participants were slower in target absent trials, but they were more accurate in finding the target object.

Two-Way ANOVA

Then, the data from the absent trials were excluded and proportion of correct responses were submitted into a two-way ANOVA with factors SOA (-1000, -500, 0, 300) and congruency (congruent, incongruent and white noise). The results revealed that the main effect of SOA was not significant, $F(2.80, 151.47) = 2.128$, $p > .05$. The main effect of congruency was also not significant, $F(2, 106.6) = .334$, $p > .05$. The interaction between SOA and congruency was not significant, $F(5.4, 292.8) = 1.483$, $p > .05$

Accuracy with No Sound vs Sound

To test whether accuracy rates were different in no sound and sound conditions, we conducted 12 dependent samples t-tests using the proportion of correct responses in all conditions. The difference between sound and no sound did not reach statistical significance, $p > .05$ (Figure 6).

.

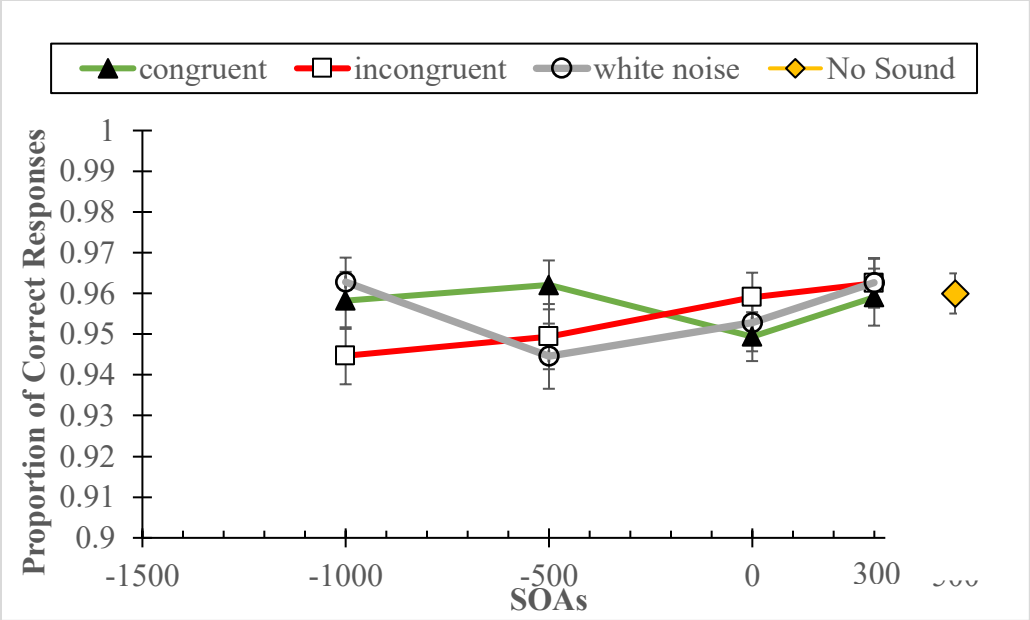

Figure 6. Accuracy graph that shows the results of the two-way ANOVA with factors SOAs and congruency. Accuracy rates were not statistically different in the different conditions. The three-way ANOVA that included the target presence condition indicated a speed accuracy trade-off between present and absent conditions such that target present trials were faster but less accurate than target absent trials. However, once the absent trials were removed for the two-way analysis, the speed accuracy trade off disappeared indicating that accuracy rates were similar across all conditions.

Discussion

Hypothesis 1

The purpose of this experiment was to investigate the multisensory nature of visual search performance in realistic scenes while systematically varying the SOA. First, we looked whether the effect of congruency was present in the visual search performance. Our results indicated a congruency effect in visual search performance; i.e., participants were faster in finding an object in a scene when it was matched with its characteristic sound. This finding supported hypothesis 1 that congruent sounds facilitates visual search performance compared to incongruent sounds or white noise. This finding is consistent with prior research that found the congruent semantic information from auditory stimuli can significantly enhance behavioral performance in visual tasks (Laurenti et al, 2004 Chen, 2010, Chen, 2011). consistent with our hypothesis, our results also indicated that performance was similar in the incongruent and white noise conditions. This shows that the incongruent sound which provided misleading information about the target object did not impair performance compared to white noise. Iordenscu et al. (2008) showed similar results that congruent characteristic sound significantly enhanced visual search performance, yet the incongruent sound did not impair performance compared to white noise.

Hypothesis 2a

Second, we examined the influence of timing between the start of visual stimuli and auditory stimuli on the congruency effect. Hypothesis 2a predicted that the congruency effect would only be present at SOA -1000 and -500 and not SOA 0 or 300. Accordingly,

our results indicated that the congruency effect was dependent on when the auditory information was presented. More specifically, the congruency effect occurred not only at SOA -1000, -500 ms but also at SOA 0. This finding is in line with prior studies that show enhancement of performance when sound precedes the picture by 1000 or 500 ms. (Chen, 2018, Chen and Spence, 2011). However, our results directly contrast those of Chen and Spence (2010) in which they show a congruency effect at the positive SOA. Moreover, Chen and Spence (2018) reported an absence of congruency effect at SOA 0. Chen and Spence (2010) argue that the meaning of the picture is kept in the CSTM for 300 ms, therefore it can still bind with the sound and improve performance. However, we argue that the realistic scenes used in our experiment were more complex and required more time to retrieve relevant information from memory. For this reason, the meaning of the scene cannot be held in the CSTM and bind to the characteristic sound when the scene appears before the sound (positive SOA). On the other hand, the meaning of the characteristic sounds can be accessed quickly and easily. Therefore the characteristic sound is able to be maintained in CSTM and guide attention to the location of target in a subsequent scene (negative SOA). This view would explain the faster performance in the negative SOAs shown in our experiment.

Additionally, neuroimaging and behavioral studies have found enhancement of visual processing with the co-occurrence of congruent auditory information. For instance, Molhom (2004) found that participants are faster in responding to a picture when it's simultaneously presented with its characteristic sounds. Moreover, they noted a

modulation of ERP component N1, which is associated with object processing, as evidence for enhanced visual processing.

It is important to note that our results indicated that at SOA 0 the congruent condition produced similar RTs as the white noise condition. Since white noise does not provide any information regarding target's identity or its location, we conclude that the meaning of the sound in the congruent condition was not responsible for enhancing the performance at SOA 0. Therefore, it is reasonable to assume that the incongruent sound significantly impaired performance when it was presented simultaneously with the picture.

Performance was equally bad in all congruency conditions of SOA 300. In this condition, the participant first saw the scene containing a target object and 300 ms later heard a sound. This does not support Chen and Spence (2010). One possible explanation for the lack of congruency effect when the sound is presented after the picture is that information from the two senses are processed separately. Whereas in negative SOAs and SOA 0, the visual and auditory information are combined to enhance perception, in the positive SOA the visual and auditory information are processed separately without influencing perception.

According to the *Unity Assumption Theory*, a multisensory event can either be perceived as a single multisensory event or two (or more) separate events. The observer makes this assumption based on the consistency of the available information from the sensory inputs (Welch,1999; Spence, 2007). Accordingly, when the two sensory information appear consistent or "go together", the observer is more likely to assume that

they share a common spatiotemporal origin and will bind them together as a single event as oppose to separate events (Bedford, 2001). Based on our results we argue that when participants heard a congruent sound 500 ms or 1000 ms before the scene, they made the assumptions that the sound and the scene "go together" and therefore were able to bind the sensory inputs into a coherent single multisensory event. Conversely when the sound was presented after the scene, binding of stimuli did not occur because they were perceived separately. Processing these two sources of information separately would take longer than if they were combined, which would explain the sharp increase in RTs at SOA 300. It is also possible that when the sound and picture were presented simultaneously (SOA 0) the congruent and white noise were perceived as single event, whereas the incongruent sound was perceived as a separate event.

Hypothesis 2b

Next, in order understand the interaction of SOA and congruency effect better, we took a closer look at performance in each congruency condition. In Hypothesis 2b we stated that in the congruent conditions there will be an incremental increase in RTs as SOAs got shorter (-1000 < -500 < 0 < 300), while RTs in the incongruent and white noise will be similar at each SOA (-1000 = -500 = 0 = 300). Our results partially supported this hypothesis. While we found that RTs were significantly longer at SOA 300, performance at SOA -1000, -500 and 0 were similar in the congruent condition. This finding contrasts our idea that more time is needed to access the meaning of both stimuli in order to enhance performance. As noted above, the congruency effect at SOA 0 cannot be attributed to the congruent semantic of the characteristic sound, because performance

with white noise was the same in that condition. Therefore, we conclude that the congruency effect occurs to the same extend at SOA -1000 and -500. Additionally, in contrast to our hypothesis RTs in the incongruent and white noise conditions were not similar across the SOA; SOA 300 produced significantly longer RTs than -1000, -500 and 0 SOAs.

Hypothesis 2c

Hypothesis 2c stated that longer negative SOA -1000 will lead to a stronger congruency effect than shorter negative SOA -500. Our results did not support this hypothesis. We found that in the both negative SOAs the congruency effect occurred to similar extent. One possible interpretation of this is that in negative SOAs the incongruent sound is suppressed before the presentation of the scene. In other words, it may be that the participant sees the written cue then hear the sound and quickly realize that the sound is not relevant to what they are supposed to look for, therefore they suppress the sound. In contrast, in the congruent conditions the matching sound has adequate time to activate a schema of the of the target objects or its likely spatial location, before the scene is presented. It can also be argued that at SOA 0, the incongruent sound is ignored in a similar fashion because it is not relevant to the picture.

The realistic scenes encompass a large amount of information. The human cognitive system is limited in its ability to process all the information at the same time. Therefore, attention is needed to properly select the information that is relevant to goal of the observer. According to the Cognitive Guidance Theory, the most important factor that guides attention within a complex scene is meaning (Henderson, 2007, 2009). In this

view, attention is directed to the regions of the scene that are semantically informative and relevant to the observer's goal. Our results show that hearing a characteristic sound a few milliseconds before the scene can be guiding factor for directing attention to the likely location of the target object in a realistic complex scene.

Hypothesis 3

We also hypothesized that performing the visual search with no sound will produce longer reaction times than the congruent sounds. In line with our hypothesis, our results showed that finding the target object was significantly faster in the SOA -1000/ congruent, and SOA -500/congruent condition than no sound. However, Performance in the no sound condition was not different from the SOA 0/congruent condition. Moreover, the no sound condition produced significantly faster RTs than the SOA 300/congruent sounds. These results indicate an advantage of bimodal stimuli (congruent sound and picture) over unimodal (vision only), provided that the congruent auditory stimuli is presented -1000 or 500 ms in advance. This finding adds to the existing literature on multisensory integration where it is shown that two sources of relevant information from different modalities enhance perception compared to unisensory information (Murry 2016, Chen and Spence 2018, Iordenscue 2008, 2010).

Implications

Overall, the result of our experiment provides evidence for the existence of top-down control of characteristic sound on realistic visual search task. We showed that hearing a characteristic sound that matches the target will facilitate finding that target in complex and realistic scenes. We further showed that the timing of the onset of the characteristic

sound is a crucial factor in enhancing performance. The implication of the current study is that we used stimuli and tasks that are realistic and relevant to everyday life activities. Previous studies have mostly focused on using basic stimuli such as line drawings, pictures of objects in isolation and simple tones to investigate multisensory integration. While basic stimuli allow researchers to conduct well-controlled experiments, they limit the extent of generalizability of results to real-world experiences. In our experiment we used real-world scenes comprising of background information and multiple discrete objects that were meaningfully arranged. In addition, the visual search task resembled everyday life activities such as looking for your keys, finding your car in a parking lot or searching for ingredients to make yourself breakfast. We combined this task with meaningful characteristic sounds that would normally be present in the environment. For these reasons, our results extend the ecological validity of prior findings. This study demonstrates that our perception of the world is influenced by the semantic component of sounds and the timing of sensory information.

Limitations

One limitation of the current study is that we did not control for the size of target objects that appeared in scenes. It is possible that some objects were easier or harder to find depending on the amount of space they occupied in the scene. For instance, the target object piano generally occupied more space within the scene, hence might have been easier to locate than a smaller object. In comparison, the target object alarm clock was significantly smaller in size and might have been more challenging to locate. In this view, it is possible that the larger target objects were found automatically due to bottom-

up factors (saliency) and not cognitive top-down factors as we expected. A second limitation of our study is that we only used 20 target objects that were repeated in different scenes. The target object was usually placed in the same location within the scenes. For example, a toaster was always on the countertop in a kitchen, or an alarm clock was always next to a bed in a bedroom. Thus, it is possible that after repeatedly searching for the same target object in various scenes, participants already built an expectation of where to look for in the scene and did not rely on the congruent sound as much. The effect of repeated search in similar scenes on the congruency effect may have been less pronounced in the negative SOAs where participants heard the characteristic sound before the scene. However, when the sound was presented after or simultaneously with the scene, the expectation of where to look for might have influence the congruency effect to a greater degree. For instance, when participants saw the kitchen they knew that the toaster would be on the countertop (since that's where it's always been), therefore they did not need to use the subsequent characteristic sound to guide their attention.

Future Direction

Future research should look at the pattern of eye movements in a similar experimental design to confirm that the characteristic sounds influence the first saccade to the target object in realistic scenes. This would provide direct evidence for our finding that a congruent sound directs attention to the location of a target object in a complex scene. Moreover, future studies should look at the interactive nature of realistic scenes and sounds. It would be interesting to see if finding an object is enhanced by hearing its characteristic sound in an interactive situation, where the participants is allowed to walk

and move around; this would further enhance the generalizability of our laboratory results to real world situations. Another factor that is yet to be investigated is individual difference in performance. Our study focused on neurotypical participant. It has been demonstrated that individuals with autism spectrum disorder (ASD) show hypersensitivity to sensory stimuli which can lead to enhanced perceptual processing (Foss-Feig et al, 2010). Some studies have shown that ASD individuals show superior visual and auditory perceptual discrimination. For instance, O'Riordon and Passetti (2006) used simple auditory tones in a pitch discrimination task and found that ASD individuals show better performance in discriminating two tones with similar frequencies relative to the neurotypical individuals. Is it not known whether this enhanced perceptual processing is only limited to basic stimuli or can be extended to more complex stimuli that resembles those in everyday life. Can individuals on the autism spectrum show the same or better congruency effect than we found in our experiment? Extending our study to "neurodivergent" population such as ASD will enhance our fundamental understanding of how multisensory integration occur in different individuals, thus providing an avenue for better treatments for them.

References

Bedford, F. L. (2001). Towards a general law of numerical/object identity. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, *20*(3–4), 113–175.

Biederman, I. (1981). Do background depth gradients facilitate object identification? *Perception*, *10*(5), 573–578

Biederman, I., Rabinowitz, J. C., Glass, A. L., & Stacy, E. W. (1974). On the information extracted from a glance at a scene. *Journal of Experimental Psychology*, *103*(3), 597–600.

Chen, Y., & Spence, C. (2010). When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. *Cognition, 114*(3), 389-404.

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

Chen, Y., & Spence, C. (2018). Audiovisual semantic interactions between linguistic and nonlinguistic stimuli: The time-courses and categorical specificity. *Journal of Experimental Psychology, 44*(10), 1488-1507.

de Graef, P., de Troy, A., & d'Ydewalle, G. (1992). Local and global contextual constraints on the identification of objects in scenes. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *46*(3), 489–508.

Eckstein, P., Koehler, K., Welbourne, L., & Akbas, E. (2017). Humans, but Not Deep Neural Networks, Often Miss Giant Targets in Scenes. *Current Biology, 27*(18), 2827-2832.e3.

Fabre-Thorpe, M. (2011). The characteristics and limits of rapid visual categorization. *Frontiers in Psychology*, *2*(243).

Falchier, A., Clavagnier, S., Barone, P., & Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. *The Journal of Neuroscience, 22*(13), 5749-5759.

Frassinetti, F., Bolognini, N., & Làdavas, E. (2002). Enhancement of visual perception by crossmodal visuo-auditory interaction. *Experimental Brain Research, 147*(3), 332-343.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, *7*(11), 498–504.

Henderson, J. M., & Hayes, T. R. (2018). Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps. *Journal of Vision*, *18*(6), 10.

Henderson, J., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology, 50*(1), 243-271.

Iordanescu, L., Grabowecky, Franconeri, Theeuwes, & Suzuki. (2010). Characteristic sounds make you look at target objects more quickly. *Attention, Perception & Psychophysics, 72*(7), 1736-1741.

Iordanescu, L., Guzman-Martinez, E., Grabowecky, M., & Suzuki, S. (2008). Characteristic sounds facilitate visual search. *Psychonomic Bulletin & Review, 15*(3), 548-554.

Kass, Robert E, & Raftery, Adrian E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773-795.

King, A. (2005). Multisensory integration: Strategies for synchronization. *Current Biology, 15*(9), R339-R341.

Laurienti, P., Kraft, R., Maldjian, J., Burdette, J., & Wallace, M. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research, 158*(4), 405.

Lippert, M., Logothetis, N., & Kayser, C. (2007). Improvement of visual contrast detection by a simultaneous sound. *Brain Research, 1173*, 102-109.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314-324.

Matusz, P. J., & Eimer, M. (2011). Multisensory enhancement of attentional capture in visual search. *Psychonomic Bulletin & Review*, *18*(5), 904–909.

Meredith, M., Nemitz, J., & Stein, B. (1987). Determinants of multisensory integration in superior colliculus neurons. I. temporal factors. *The Journal of Neuroscience, 7*(10), 3215-3229.

Molholm, S. (2004). Multisensory visual-auditory object recognition in humans: A high-density electrical mapping study. *Cerebral Cortex, 14*(4), 452-465.

Murray, M., & Spierer, L. (2009). Auditory spatio-temporal brain dynamics and their consequences for multisensory interactions in humans. *Hearing Research*, *258*(1-2), 121-133.

Potter, M. (1975). Meaning in visual search. *Science, 187*(4180), 965-966.

Potter, M. C (1999) Understanding sentences and scenes: the role of conceptual short-term memory. *In Fleeting Memories* (Coltheart, V., ed.) p. 13-46, MIT Press

Romei, V., Murray, M., Cappe, C., & Thut, G. (2009). Preperceptual and Stimulus-Selective Enhancement of Low-Level Human Visual Cortex Excitability by Sounds. *Current Biology, 19*(21), 1799-1805.

Romei, V., Murray, M., Merabet, L., & Thut, G. (2007). Occipital transcranial magnetic stimulation has opposing effects on visual and auditory stimulus detection: Implications for multisensory interactions. *J Neurosci, 27*(43), 11465-11472.

Schyns, P., & Oliva, A. (1994). From Blobs to Boundary Edges: Evidence for Time- and Spatial-Scale-Dependent Scene Recognition. *Psychological Science, 5*(4), 195-200.

Spence, C. (2007). Audiovisual multisensory integration. *Acoustical Science and Technology, 28*(2), 61-70.

Stein, B. E., & London, N. (1996). Enhancement of perceived visual intensity by auditory stimuli: A psychophysical analysis. *Journal of Cognitive Neuroscience*, *8*(6), 497.

Talsma, D., & Woldorff, M. G. (2005). Selective Attention and Multisensory Integration: Multiple Phases of Effects on the Evoked Brain Activity. *Journal of Cognitive Neuroscience*, *17*(7), 1098–1114.

Talsma, D., Doty, T. J., & Woldorff, M. G. (2007). Selective attention and audiovisual integration: Is attending to both modalities a prerequisite for early integration? *Cerebral Cortex*, *17*(3), 679–690.

Theeuwes, J. (1991). Cross-dimensional perceptual selectivity. *Perception & Psychophysics*, *50*(2), 184–193.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520-522.

Van der Burg, E., Olivers, C., Bronkhorst, A., & Theeuwes, J. (2008). Pip and pop: Nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology*, 34(5), 1053-1065.

Welch, R. B. (1999). Meaning, attention, and the “unity assumption” in the intersensory bias of spatial and temporal perceptions. *Advances in Psychology, 129*(15), 371-387.

Wolfe, J. M., Butcher, S. J., Lee, C., & Hyle, M. (2003). Changing your mind: On the contributions of top-down and bottom-up guidance in visual search for feature singletons. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(2), 483–502.

Wolfe, J., Alvarez, G., Rosenholtz, R., Kuzmova, Y., & Sherman, A. (2011). Visual search for arbitrary objects in real scenes. *Attention, Perception, & Psychophysics, 73*(6), 1650-1671.

Appendix A
Consent Form

Request for Your Participation in Research

Title of Study: Visual Search Performance in Realistic Scenes

Name of the Researcher: Ghazaleh Mahzouni, SJSU graduate student and Dr. Cary Feria, Faculty advisor

Purpose: You have been asked to participate in a research study investigating finding objects within realistic scenes. Not all aspects about the purpose of the study are being shared at the outset but will be provided to you afterwards.

Procedures: You will be asked to view computer displays showing pictures of realistic scenes and find a target object in that scene. You will answer a short questionnaire afterward. The study will last approximately one hour and will be done in Hugh Gillis Hall 246. Not all aspect about the study are being shared upfront but will be provided afterwards during a debriefing. This study will be done during Spring 2019.

Potential Risks: This study presents no more than minimal risks of fatigue and eye strain. Participants will be allowed to take a break during the study.

Potential Benefits: You will receive no direct benefits from participating in this study. It is possible that you may indirectly benefit by furthering the general knowledge of visual perception.

Compensation: As a student in the psychology research subject pool, you will receive partial credit towards your Psychology class research requirements. Credit will be granted even if you decide to withdraw from this study at any time. No other compensation is provided for participation in this study.

Confidentiality: Although the results of this study may be published, no information that could identify you will be included. The researchers are required to report cases of abuse, neglect and intent to harm self or others, when applicable.

Participant Rights: Your participation in this study is completely voluntary. You can refuse to participate in the entire study or any part of the study without any negative effect on your relations with San Jose State University. You also have the right to skip any question you do not wish to answer. This consent form is not a contract. It is a written explanation of what will happen during the study if you decide to participate. You will not waive any rights if you choose not to participate, and there is no penalty for stopping your participation in the study.

Questions or Problems: You are encouraged to ask questions at any time during this study. For further information about the study, please contact Ghazaleh Mahzouni) at Ghazaleh.mahzouni@sjsu.edu. Complaints about the research may be presented to Dr. Clifton Oyamot (Chair, Department of Psychology, SJSU) at (408) 924-5600. For questions about participants' rights or if you feel you have been harmed in any way by your participation in this study, please contact Dr. Pamela Stacks, Associate Vice President of the Office of Research, San Jose State University, at 408-924-2479.

Signatures:
Your signature indicates that you voluntarily agree to be a part of the study, that the details of the study have been explained to you, that you have been given time to read this document, and that your questions have been answered. You will receive a copy of this consent form for your records.

Participant's Name (printed) Participant's Signature Date

Researcher Statement:
I certify that the participant has been given adequate time to learn about the study and ask questions. It is my opinion that the participant understands his/her rights and the purpose, risk, benefits, and procedures of the research and has voluntarily agreed to participate.

Signature of Person Obtaining Informed Consent Date

Appendix B
Questionnaire

subject # _____

Questionnaire

The researcher asks questions #1-6 aloud to the subject and will writes down the subject's responses.

1. Please describe, in detail, what you were doing.

2. Did you find any of the trials to be more difficult than other trials?

3. Some trials included sounds; did you find these sounds helpful to your performance or distracting?

4. During the experiment did you find yourself focusing more on the sound or the picture?

5. Did any problems occur with the computer during the experiment?

6. Is there any other aspect of the experiment you would like to comment on?

For Question #7, the subject will write the answer themselves.

Please circle your sex: MALE FEMALE