Effects of Sound and Visual Congruency on Product Selection and Preference

Brock Bass
Clemson University
brockb@clemson.edu

Felipe Fernandez
Clemson University
ffernan@clemson.edu

Drew Link
Clemson University
dlink@clemson.edu

Andy R. Schmitz
Clemson University
andy@arschmitz.com

ABSTRACT

Visual attention in a shopping context is a necessary part of applied research. Eye-tracking methodologies help to understand where attention is allocated in shopping settings, however there are few examples of applied research that studies the effects of sound and visual congruency in product selection and preference. The current study applies concepts of the effects of congruent sound stimuli on visual attention to a shopping context. Results are presented from this study that indicates there is no effect of a natural sound congruent with a corresponding visual stimulus. Recommendations for further research are suggested to determine if there is some difference between visual attention and product selection when presented with a cross-modal representation of audio and visual stimuli.

Author Keywords

Tobii eye tracker, cross-modality, visual and auditory congruency, packaging

INTRODUCTION

Auditory and visual stimuli rarely operate independently of one another in real-world scenarios. This is apparent when people are driving, working, shopping, or completing a wide range of activities on a day-to-day basis. In these scenarios, visual information is presented in conjunction with auditory information. Often, audio and visual information can coexist to create cross-modal functionality. This can improve information processing [11] and sometimes the interaction of the two modalities can help produce automatic attention capture in an individual [1, 7]. However, audio and visual information can conflict causing distractions and adversely affect attention [10]. This bottom-up attentional shift due to environmental sounds produces interesting questions about the role audition plays in attention, object identification, and object preference.

Eye tracking measures can be incorporated into studies of cross-modal integration to gain novel dependent variables. Previous studies have taken advantage of eye tracking measures and gained the dependent variables of eye movements, saccades, fixations, and blinks [3]. In one specific study to test cross-modal integration, non-spatial auditory information was used as a stimulus while participants watched a video [3]. While the stimuli and procedure of this study are not comparable to our proposed study, there are relevant findings worth discussing. One of these being the influence that sound has over time. It was found that the influence of sound on eye movements was delayed until approximately 500 ms after the stimulus was presented [3]. This finding is relevant because a continuous auditory stimulus might not affect eye movements in the same manner as a sudden noise burst, which has been used in other studies [6]. Examining the effect of a specific auditory stimulus on a specific visual stimulus has been suggested for future research [3]. This suggestion describes the nature of research we are interested in conducting.

We propose a study that focuses on the role of audio stimuli on guiding attention and decision-making when identifying and selecting an object in a visual field. Just as audio can inhibit or enhance our selective attention, it is our goal to study how audio can affect preference and ultimately selection of an object in an environment. The proposed study will create a scenario in which a natural audio cue (congruent or incongruent) will be played as the participants search for the target visual stimulus in an array of stimuli. We will measure the fixations and scan-paths using a Tobii eye-tracker, and compare that data with behavioral responses involving the selection of an object. Our hypotheses are as follows: (1) When a natural auditory stimulus is presented along with a series of visual stimuli, fixations will increase on the target object of highest congruence, and (2) when a natural auditory stimulus is presented along with a series of visual stimuli, more selections of the object of highest congruence will be made. To test these hypotheses, we will create a simulated shopping experience in which natural audio will be played while the participants shop and select items in the store. The justification for this scenario is that this is a natural, every day task in which behaviors are susceptible to influence involving bottom-up attention capture in the form of advertisements, announcements, and other stimuli.

Cross-modality of visual and auditory stimuli

When auditory and visual features are paired in an environment, cross-modal correspondences can influence how we direct our attentions. Whether it is a visual stimulus that sensitizes our selective attention for audio, or a sound that helps us visually locate an object in the environment, cross-modality allows us to increase the speed

and accuracy with which we respond to targets and events. For instance, previous research observed how cross-modal congruence between the pitch and frequency of a sound and the location of an object in the environment allows for faster detection and location of that object [2]. Cross-modality was shown in such a way that lower pitched sounds resulted in an automatic visual search for an object lower in a visual field. The opposite was also true, where participants have visually searched for objects higher in a visual field when a higher pitched sound was played. Similarly, cross-modality and a high correspondence have been observed in the pitch of a sound and the contrast of an object. Object selection and preference for lighter surfaces were observed when paired with a higher pitch sound, while responses were faster in selecting a darker image when that image corresponded with a lower pitch sound [8]. Individually, these high correspondences are interesting for building the theory for cross-modality and its impact on automatic attentional shifts, speed of processing, and object/target selection. However, highly specific sounds or tones do not occur naturally in the environment often, nor do they have a contextual significance to them.

In a comprehensive study, visual modalities of different kinds were paired with auditory sounds that varied in pitch [4]. The purpose of the study was to see if there were any interactions in addition to the main effects observed by previous studies when visual features were combined. In essence, rather than simply comparing contrast, size, or location, all of these features were varied to see if a sound could increase the speed and/or accuracy of the detection of the target objects. The findings indicated that there was a strong relationship between audio pitch and visual stimuli size, position, and spatial frequency. They were unable to replicate previous findings in the effect of audio pitch on selection of objects based on contrast, however. Despite this, this lends strong evidence to the potentially complex nature of the effects of audio on automatic, sensory-level processing and selection of objects in a visual field. While the above studies have focused on audio and visual stimuli that exhibit an automatic response, there have been very few studies that investigate the effect of natural sounds as a cue for selection of congruent natural imagery.

Cross-modality of spatial and temporal proximity

The modalities of audition and vision are used daily by humans to sense and perceive their environment. The concept of combining or integrating these two modalities is not a novel idea. Previous research has delved into the area known as cross-modal integration to study this concept. An aspect of cross-modal integration is concerned with the spatial and temporal positioning of the modalities. Understanding how the spatial positioning and temporal presentation of auditory and visual stimuli effect human perception of those stimuli is of particular interest. Potentially, the presentation of stimuli in a cross-modal context could enhance the human's perception of one of the stimuli. In order to determine the effects of spatial and temporal positioning on perception, it is necessary to design an experiment that creates permutations of the visual and auditory stimuli. A previous experiment was conducted that investigated the effects that cross-modal (visual and auditory) integration had on human visual perception [6]. Two major issues were scrutinized. The first issue concerned the spatial location of the auditory stimuli and the visual stimuli (i.e., the proximity of one stimulus to the other). The second issue dealt with temporal proximity by creating two conditions; a simultaneous stimuli presentation condition (auditory stimulus was played simultaneously with visual stimulus presentation) and an asynchronous stimuli presentation condition (auditory stimulus preceded visual stimulus). The study found that the spatial positioning of the two stimuli did affect the perceptual sensitivity of the participant. Essentially, when auditory cues where placed in close proximity to the visual stimuli the participants' ability to detect the visual stimulus improved. In terms of temporal proximity, results showed that participants experienced improvement in detecting the visual stimulus only in the simultaneous presentation condition. These findings support the theory that humans have the ability to capitalize on cross-modal integration through means of appropriate spatial positioning and temporal presentation [6].

Spatial location or congruity has been found to be of varied importance based on what data collection measures are used. For instance, a previous study found that spatial congruity was of little importance in relation to participants' detection of targets and their reaction times when behavioral data (i.e., percentage of hits and mean RT) was recorded [9]. However, there was a cross-modal effect for participants in spite of this lack of spatial congruity effect. This means that the presence of two stimuli (auditory and visual) was more beneficial for participants' performance than a uni-modal stimulus. Spatial congruity was found to be of significant importance when neural response patterns were measured. This means that when participants were in spatially congruent conditions their neural response patterns differed from when the participants were in spatially incongruent conditions. These findings show that understanding the effect of spatial congruity may take several data collection measures to converge on what is actually occurring with participants' performance.

Cross-modality of natural sounds and contextually congruent imagery

The nature of the relationship between more natural sounds and contextually congruent imagery has not been as rigorously studied. This may be in part due to the fact that these stimuli involve much more complicated systems than simple visual and auditory perceptions, and may involve a higher-level cognitive process to guide attention, but some studies do suggest there is a potential relationship between the two. For instance, visual location of objects have been

shown to be enhanced by positional and dynamic audio in a virtual environment, such that a non-static high or low pitched sound can help facilitate the location of an object in a virtual space [5]. Additionally, neurological studies have shown that complex and dynamic sounds have increased the activity in ventral-occipital stream when finding objects in a visual environment, providing neurological evidence that complex sounds activate a neurological reaction and subsequent deliberate, guided search behavior [9]. However, the complex sounds also activate the dorsal stream in processing visual information, suggesting there is also an element of automatic processing involved. No studies have been done to assess the role of natural sounds on cross-modal mapping using eye-tracking technology. As such, it is of particular interest in this study to combine eye-tracking data with behavior in the form of object selection and identification.

METHODOLOGY

Figure 1: CUshop stocked shelves

Apparatus

Eye movements were gathered using the Tobii suite of technology.

Tobii Glasses
The Tobii Glasses are a head-mounted camera worn as a pair of glasses. One participant facing camera, records the right-eye's retinal movement, while an opposing, scene-facing camera records the participant's field of vision. According to the manufacturer, the glasses record video at 30Hz, have a visual angle accuracy of 0.5 degrees, and have an active range of 60-250cm.

Recording Assistant
The Recording Assistant was a wearable, walkman-sized computer that stores both video and audio. It correlates the recorded retinal movements to the recorded field of vision in order to determine the scanpath for a participant.

Infrared (IR) Markers
The Infrared Markers, attached to the shelving units, serve as a visual point-of-reference for later analysis.

Tobii Studio Software
Within the Tobii Studio software we can use those points-of-reference to create areas of analysis (AOA) and areas of interest (AOI). The software can then count the number, and duration, of fixations within those areas in order to create the data we need.

Audio playback
The different sounds for congruent and congruent sounds were ambient mp3s of a city sound (incongruent) and an ocean (congruent). The mp3s were looped and played from an application for an iPad and streamed via Bluetooth to an LG 280W sound system. Volume was kept constant for both conditions, and the sound system was positioned such that the audio was equally perceived in all areas of the store.

Calibration

Calibration of the Tobii Glasses was accomplished by having participants stand 1 m from a wall that displayed a 3x3 grid of points. The experimenter asks the participant to follow an IR marker from point to point with their eyes. The participant is also instructed to keep their head stationary while tracking the target. Once each point has been followed, and is confirmed by the eye tracker equipment, calibration is considered to be successful and the experimenter concludes the calibration.

Stimulus

The visual stimuli in this study were a series of soapboxes. These soapboxes were displayed on a shelf much like they would be in a realistic shopping scenario. Boxes were displayed in the manner that is shown in Figure 2. Adobe Illustrator CS5 and CS6 were used to design the boxes and edit the images on the boxes. The boxes were printed on a Roland LEC 330 printing machine. Konsberg XP is the cutting machine used to cut the soapboxes.

Figure 2: Design of the visual stimulus and how it appears on the shelf.

The auditory stimuli in this study were two ambient sounds generated via an ambient sound generator on an Apple Ipad, and displayed on a compatible speaker system. The sound files are 30 minutes long and looped. The sounds chosen were an ambient ocean and city sounds. For the third level of this variable there was no sound played.

Participants

The participants for this study were 31 (18 female) people from the Clemson area between 18 to 64 years. Recruitment was done through the university human participant pool (SONA Systems) as well as verbal recruiting on campus. All participants had normal or corrected to normal vision, however two participants could not be calibrated properly with their glasses so their data was dropped from analysis. We concealed the purpose of the study from the participants, and no participants had prior knowledge about the purpose of the study.

Experiment design

A one-way ANOVA was conducted to evaluate the three levels on the categorical independent variable of audio stimulus (No sound, incongruent sound, congruent sound) on the dependent variables of fixation and visit duration. Additionally, we conducted a chi-square analysis of the target object in the experimental condition with the control and incongruent conditions (i.e., two discrete/categorical variables, with the dependent variable being a binary "Yes/No" choice for selecting an object). This assessment investigated the behavioral effect of the levels of the independent variable on decision-making, while the ANOVA aimed to assess the effect of the independent variable on eye-movement behavior and visual attention.

Procedure

The participants who volunteered for our study were led through a 4 step process:

1. Introduction

The participant was brought in and told general information about the study. Then the participant was given an informational letter discussing the benefits, risks, costs, and legal information. As this study has no more than minimal risk, a signed consent form was not required.

2. Calibration

We calibrated the glasses to the participant by using a single IR Marker. The participant followed the IR Marker with their eyes as we moved it around the participant's field of vision.

3. Shopping

The participants used a shopping list and selected the specific products they preferred by speaking the product's shelving number aloud. The participants also wrote the corresponding number on a tablet computer. This was recorded by a researcher for later use. The shopping list was presented in a random order to account for any potential order effects associated with the product selection portion of the study.

4. Survey

After the study, participants completed a demographics survey, and then were debriefed on the full details of the experiment.

RESULTS

All data collected with the eye tracker was exported and converted to be analyzed by the PASW statistical package. Data collected via eye tracking included fixation count and duration, and visit to an area of interest (AOI) count and duration.

The between-subjects one-way ANOVA of the sound condition on total duration of fixations on the target soap box was not significant ($F(1,30) = 0.49$, $p = .67$), and the duration of each individual fixation on the target soap box was not significant ($F(1,30) = 2.25$, $p = .12$). Additionally, the between-subjects one-way ANOVA on the total duration of visits to an AOI group for the target stimulus was not significant ($F(1,30) = .46$, $p = .63$), and the duration of each individual visit to the target stimulus was not significant ($F(1,30) = .03$, $p = .95$). See Table 2(a) for descriptive statistics associated with the fixation measure, and 2(b) for descriptive statistics associated with the AOI visit measure.

Condition	**Individual duration (s)**		**Total duration (s)**	
	M	SD	M	SD
No sound	.39	.25	2.72	1.26
Incongruent	.37	.28	2.00	1.66
Congruent	.43	.26	2.99	3.56

(a)

Condition	**Individual visit duration (s)**		**Total visit duration (s)**	
	M	SD	M	SD
No sound	.55	.43	2.77	1.29
Incongruent	.58	.45	2.06	1.71
Congruent	.59	.37	3.02	3.57

(b)

Table 1: Mean durations in seconds for individual and total fixations within an AOI group (a) and mean durations in seconds for individual and total visits within an AOI group (b).

For the analysis of the behavioral response data, the chi-square test conducted was also not significant, therefore the percentage differences between participants did not differ significantly on selection of the target stimulus $\chi 2(1, N = 30) = 3.76$, $p = .15$.

DISCUSSION

Results indicate that the ambient sound played throughout the store did not have a significant effect on either eye-tracking responses or behavioral responses when the sound matched the visual target stimulus contextually. Therefore, neither of our hypotheses were supported. However, both fixation durations and AOI visit durations were highly variable from condition to condition (see Figure 4). Though mean durations of eye tracking responses on the target stimulus were higher than the no sound condition, they were also comparable to the incongruent sound condition. One possible explanation for this lack of difference from the congruent to incongruent conditions could potentially be that the sounds of the ocean ambient noise were similar in frequency, tone, and volume. During a few pilot trials of the study, we asked pilots to identify the sounds they were hearing. The sounds were identified correctly. However, no manipulation check was conducted during the actual study to determine if participants were capable of identifying the sound or even perceiving it.

Another interesting aspect of the results is the fact that each stimulus was attended to more or less equally. Figure 3 shows the heatmap for all participants in the study. It was

Figure 3: Heatmap (all participants) across the three conditions. The center area of the second shelf from the top is the target stimulus.

generally observed that people looked at each box of soap before making a selection. Results indicated that there was no significant difference between groups for fixations on the target box (the center box in Figure 3). This may be due to this fact that people looked at all boxes equally before making a decision. Interestingly, the chi-square analysis showed that 70% of participants in the congruent sound condition selected the target stimulus, while the other conditions reported a 30% choice of the target (no sound condition) and a 34.6% choice of the target (incongruent condition) (see Figure 5). Typically, in order to achieve sufficient power with a chi-square analysis, many more participants are required. A limitation of this study was that only 31 participants were tested. An argument could still be made that, given more participants, frequency of selecting the target may prove to be significant compared to the incongruent or no sound conditions. This would suggest that people may attend to all stimuli equally disregarding the condition they are assigned to, and yet still behave in a way that seems to support the hypothesis that a natural contextually similar sound may lead to more selections of an object that matches that sound. Of course, this is conjecture at this point, as our data showed no significance in the behavioral data collected.

The effectiveness of cross-modality when manipulating spatial and temporal proximity was indirectly investigated in the current study. Previous research found that the spatial and temporal proximity of an auditory stimulus was influential in the detection of a visual stimulus. Using an auditory stimulus that was close to the visual stimulus in terms of spatial and temporal proximity yielded improvements in the detectability of the visual stimulus [6]. The current study yielded data that was not in concert with this previous finding. There was no significant effect for the manipulation of the visual stimulus and auditory stimulus. However, there are a few potential methodological issues that may have led to the findings of the current study.

The main hypothesis of the current study stated that a natural auditory stimulus would be most effective (i.e., increase fixations on the target object) when the visual target object and natural auditory stimulus were congruent. A potential methodological issue comes from the spatial proximity of the auditory stimulus in relation to the visual stimulus in the current study. The auditory stimulus was placed behind the participants as they completed the study. This was done pragmatically to hide the stereo from the participants and keep any electrical cords out of the participants' path of walking. While initially this did not seem to be an issue, perhaps the spatial proximity of the auditory stimulus was a detrimental factor in the participants' eye fixations and eventual product choice. In light of previous research, our experimental environment may not have capitalized on the spatial proximity effect. Our experiment may have unintentionally disengaged the spatial proximity effect by not placing the auditory stimulus closer to the visual stimulus of congruence.

Another issue that may have detrimentally affected the current study's findings is the temporal proximity of the auditory stimulus. Previous research found that a simultaneously played auditory stimulus led to better detectability of the visual stimulus. This simultaneous auditory stimulus was described as a "noise burst" [6]. The potential issue is that the current study used an auditory stimulus that was continuously in the environment for the duration of the experiment. The participants were not exposed to a "noise burst" as the previous study describes. The continuous auditory stimulus was chosen to resemble a naturalistic shopping experience, but this may have led to a decrease in the effect of temporal proximity on fixations and product choice. It may be worth investigating whether manipulations of auditory stimulus length (i.e., noise burst

versus continuous) are a significant factor in participants' eye movements and selection of items. Given the experimental environment that the current study used to test participants, it is possible that the known cross-modal integration effects of spatial and temporal proximity were unintentionally negated.

CONCLUSION

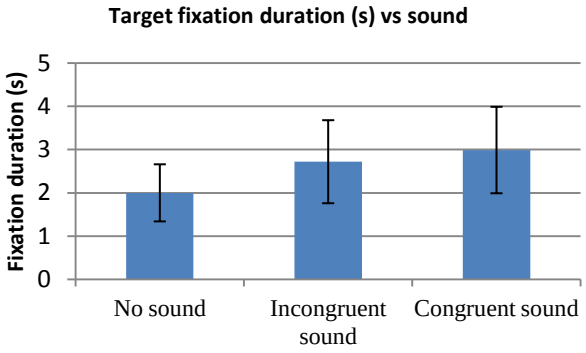
The objective of this study was to establish a potential effect of a natural sound on the behaviors and attention of consumers in a shopping context. The method for doing this was to apply the concepts of congruent sounds, cross-modality, and contextual cues to potentially observe any behavioral changes that may have occurred as a result. This study showed no conclusive evidence that a natural, contextual sound resulted in preference being given to a congruent visual stimuli or more visual attention allocated to the congruent stimuli.

Despite the lack of significant findings, there seems to be a trend in behavioral data that suggests a future study may indicate a tendency to select an object that visually matches a similar audio stimulus. A limitation of this study, in this sense, is the dearth of participants. The number of participants per cell is generally adequate for establishing power for an experimental factorial analysis, however the sample size is insufficient for achieving power for a chi-square analysis. If a discrepancy between the lack of significant differences when attending to visual stimuli from incongruent to congruent sound conditions and significant differences in selection of a product can be established, this would prove to be useful for both applied and theoretical research. It would provide evidence that attention to package design and subsequent behavioral responses may not always complement each other.

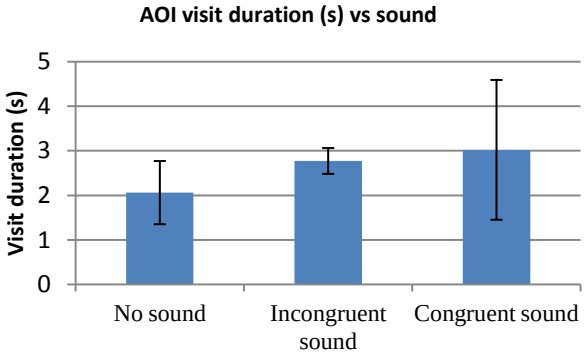
Future research could be conducted in a way to account for the limitations in our design mentioned in the discussion section, however further research could supplement these findings by conducting a behavioral study that simply looks at the effects of congruent audio and visual stimuli on selection of a product. This study would allow for the higher number of participants needed to find an effect. Shopping is becoming an increasingly interactive experience, with different methods being established to draw consumers' attention to their designs. Continuing this research would allow for a deeper understanding of the effects of congruent stimuli in a cross-modal fashion and the effects on consumer behavior.

Acknowledgments

We would like to thank Andrew Duchowski for providing access and support for using the eye-tracking system. We would also like to thank the staff of the Harris A. Smith Building for allowing us access to the lab for conducting this research.



(a)



(b)

Figure 4: Results: Eye tracking metrics for (a) fixations and (b) AOI visits

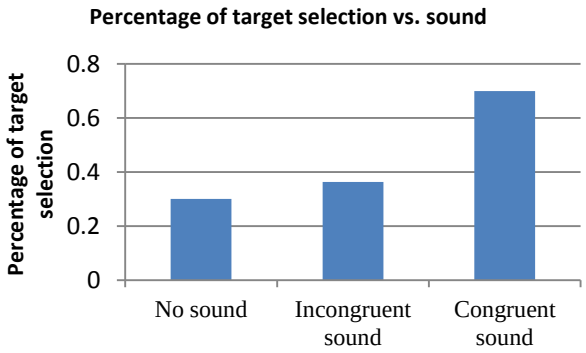


Figure 5: Results: Percentage selection of target

REFERENCES

1. Begault, D. R., and Pittman, M. T. Three-dimensional audio versus head-down traffic alert and collision avoidance system displays. *The International Journal of Aviation Psychology, 6, 1* (1996), 79-93.
2. Bernstein, I. H., and Edelstein, B. A. Effects of some variations in auditory input upon visual choice reaction time. *Journal of Experimental Psychology, 87* (1971), 241-247.
3. Coutrot, A., Guyader, N., Ionescu, G., & Caplier, A. Influence of soundtrack on eye movements during video exploration. *Journal of Eye Movement Research, 5, 4* (2012), 1-10.
4. Evans, K. K., and Treisman, A. Natural cross-modal mappings between visual and auditory features. *Journal of Vision, 10, 1, 6* (2010), 1-12.
5. Flanagan, P., McAnally, K. I., Martin, R. L., Meehan, J. W., Oldfield, S. R. Aurally and visually guided search in a virtual environment. *Human Factors, 40, 3* (1998), 461-468.
6. Frassinetti, F., Bolognini, N., & Ladavas, E. Enhancement of visual perception by crossmodal visuo-auditory interaction. *Experimental Brain Research, 147, 3* (2002), 332-343.
7. Klatzky, R. L., Marston, J. R., Giudice, N. A., Golledge, R. G., and Loomis, J. M. Cognitive load of navigating without vision when guided by virtual sound versus spatial language. *Journal of Experimental Psychology, 12, 4* (2006), 223-232.
8. Marks, L. E. On cross-modal similarity: Auditory-visual interactions in speeded discrimination. *Journal of Experimental Psychology and Human Perception Performance, 13,* (1987), 384-394.
9. Teder-Salejarvi, W. A., Di Russo, F., McDonald, J. J., and Hillyard, S. A. Effects of spatial congruity on audio-visual multimodal integration. *Journal of Cognitive Neuroscience, 17, 9* (2006), 1396-1409.
10. Tiippana, K., Andersen, T. S., and Sams, M. Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology, 16, 3* (2004), 457-472.
11. Wickens, C. D. Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science 3, 2* (2002), 158-177.