Exploring the Influence of Audio in Directing Visual Attention During Dynamic Content

Brooke E. Wooley*
David S. March§
Media Research Labs

Abstract

The mechanisms underlying the allocation of visual attention toward dynamic content are still largely unexplored. Due to the number of variables present during dynamic content, it is often difficult to confidently determine what components direct visual attention. In this study, we manipulated the presence of audio in an attempt to explore the contribution of audio in driving visual attention during dynamic content. Participants viewed a reel of non-global commercials while their eye movements were recorded. Participants were either exposed to content containing the original audio track or content in which the audio track was edited out. Dynamic heat maps were created for each ad in order to identify areas of high visual attention between the conditions. Fixation durations and fixation counts for each area of interest were then computed. Analyses showed that the presence of audio has an influence on the allocation of visual attention during dynamic content, most notably in regard to on-screen text. Understanding the influence of audio in directing visual attention may help future researchers control for the extraneous influence of audio in eye-tracking methodologies.

CR Categories: H.1.2 [Information Systems]: User/Machine Systems – Human factors

Keywords: Eye movements and cognition, AOI methods, Heat maps

1 Introduction

Audio is an important factor to consider when attempting to understand visual attention in dynamic content. Audio content has been shown to have a direct influence on the allocation of visual attention during non-dynamic content. Frens, Van Opstal, and Van Der Willigen (1995) reported that saccades toward a visual target were shortened by the presence of an auditory cue (whose source was near the target) that played slightly before the visual target appeared. Wiebe and Annetta (2008) found that narration of the text on a slide resulted in longer gaze durations on the slide, however, when narration was absent and the textual information on the slide was dense, participants allocated more visual attention to the text at the expense of the slide's graphic. Also, participants spent more time looking at an animated graphic when narration was present than when narration was absent, suggesting that auditory narration in a stimulus impacts the allocation of visual attention. Vilaró et al. (2012) describe findings where viewing patterns of a video clip varied due to the experimenters' manipulation of audio during content.

Given the novelty of eye-tracking research, what elements guide viewer allocation of visual attention during dynamic content is still relatively unexplored. There is a strong tendency towards the center of the screen (Brasel & Gips, 2008; Goldstein, Woods & Peli, 2006); though, data indicate that gaze dispersion is highly variable when the dynamic content is analyzed frame by frame (Tatler, 2007). Additionally, Yantis and Hillstrom (1994) found that the sudden onset of an object captured attention during a search task, possibly suggesting that, in dynamic content, the introduction of new objects in the scene might demand attention. However, the processes underlying allocation of visual attention in scenes depend on a multitude of variables, including the salience and informativeness of objects, visual context, and viewer goals (Buswell, 1935; Loftus & Mackworth, 1967; Wedel & Pieters, 2006). Wedel and Pieters (2006) argue that we give visual attention to objects based on their "informativeness," or relevancy to the viewer. Thus, we tend to fixate on regions within a scene, such as people and faces, because they are generally considered highly relevant to the viewer. Therefore, it is possible that during a viewing task, viewers' goals affect where they look during content. How the goals of the content might impact visual attention must also be considered.

Although top-down factors such as viewer goals and motivation are important in understanding how eye movements explain attention, top-down processes in visual attention are effortful and are generally a slower process than bottom-up allocation of attention (Pieters & Wedel, 2007). Though, it is not always feasible to lend effortful, planned attention to stimuli within our visual field. In instances where cognitive load cannot accommodate top-down processes of visual attention, it has been suggested that bottom-up processes will have a greater influence on visual attention (Pieters & Wedel, 2007). These situations might call for the use of scanpaths (a series of fixations and saccades that are systematic in nature and used upon repeated viewing of a stimulus) to help explore visual scenes in a previously learned way. A number of studies have shown that visual strategies exist for different types of viewing (e.g., Noton & Stark, 1971; Pieters, Rosbergen, & Wedel, 1999), and that these strategies interact to create the most efficient viewing experience (Egeth & Yantis, 1997). The allocation of our visual attention is very much impacted by both top down and bottom up processes. While viewing static content may encourage the use of learned scanpaths, this scanpath might be altered or non-existent in the presence of dynamic stimuli and audio. Together, the informa-

*e-mail: b.wooley@themediapanel.com
§e-mail: d.march@themediapanel.com

ETRA 2014, March 26 – 28, 2014, Safety Harbor, Florida, USA.

ACM 978-1-4503-2751-0/14/03 $15.00

tiveness of objects, as well as viewer goals will impact viewers' visual attention during dynamic content.

The purpose of the present study was to examine how the presence or absence of audio affects the allocation of visual attention during dynamic content (e.g., Coutrot, Guyader, Ionescu, & Caplier, 2012). We wanted to explore the differences in viewer's reliance on highly relevant areas of interest (AOI) when audio is present or absent. Consistent with previous research on the presence of audio, we expected that viewers would lend more visual attention to textual items during dynamic content without audio than viewers viewing the same content with audio.

2 Methodology

Stimuli. A reel of seven 30-second Australian produced commercials for Australian products was created in order to reduce familiarity effects. The seven commercials were: Affordable Living (financial services), Bunnings (home improvement store), Telstra (cell phone provider), RAC (car insurance), ANZ (bank), Aussie Farmers Direct (grocery delivery), Lasoo (electronics store), Ambipur (air freshener). One reel contained ads with their original audio track intact, while the other reel had the audio edited out using Final Cut Pro 10.

Commercials were used as stimuli since each commercial provides a 30 second space in which an entire idea is conveyed. Furthermore, all commercials have common goals; normally, comprehension and retention are at the top of that list. Thus, commercials are created with certain components that appear in ads for wide ranging products. Because we wanted to examine visual attention across several different clips of dynamic stimuli, commercials provided us with a variety of dynamic situations and a consistent set of AOIs.

Sample. Our sample consisted of 64 participants (27 males, 37 females, age $M = 37.6$). Out of 35 audio condition recordings, there were 27 usable eye gaze recordings. Out of 29 no audio condition recordings, there were 25 usable files. All participants gave informed consent and were compensated with a $10 gift card for their participation.

Eye-tracker and software. Eye-tracking data were collected using a Tobii T60 tracker with a 60 Hz sampling rate and a screen resolution of 1024x768. The software that displayed the stimuli and recorded eye-tracking data was Attention Tool 5.0 by iMotions Global. This software was also used in analysis.

Procedure. Participants were randomly allocated to the audio or no audio ad reel. For each condition, commercials were displayed in a random order to each participant to control for order effects. Participants were seated at a desk approximately 60 cm from a Tobii T60 tracker. Participants were told that they would watch a reel of advertisements and that their ads may or may not have audio. Once calibrated to the tracker, participants watched a calibration validation clip and then the ad reel began. After their ad reel finished, participants were taken to another computer to answer a post exposure questionnaire.

Questionnaire. After participants watched the ad reel, they responded to an online questionnaire about their experience. All respondents answered questions about each ad they remembered. If they answered 'yes' to having remembered an ad, then further questions about the ad would be presented. Three screen shots of the commercial represented a visual recognition task for the participant to determine if they remembered the ad. Then, four Likert-type questions measured participants' emotional and evaluative responses to the ad. These questions served as an additional explanatory measure.

3 Results

3.1 Questionnaire results

Information provided by the questionnaire showed that, overall, participants felt more emotion (positive or negative) during the no audio condition ($M = 3.81$) than during the audio condition ($M = 3.31$; $F = 16.931$, $p < .001$). Participants also liked the pace (number of cuts) more during the no audio condition ($M = 4.49$) than during the audio condition ($M = 4.27$; $F = 3.95$, $p = .047$). No other significant effects were noted.

3.2 Eye-tracking results

Eye gaze recordings were used if the participant successfully fixated on at least 10 out of the 12 targets during the calibration validation phase.

Focusing on AOIs that appear in most commercials, as well as past research examining the effects of audio on visual attention, we created a set of AOIs to identify in each commercial. These AOIs included: *talking characters, non-talking characters, products, product illustrations* (i.e. pictures or graphical representations of the advertised product), *textual product information, small print, brand name, textual brand info,* and *non-branded textual items*. Additionally, we combined these AOIs into three groups: *people*, *text*, and *objects*. Past research has suggested that humans give preferential visual attention to faces (Itti, Koch, & Niebur, 1998; Wedel & Pieters, 2006), which is why we chose to include faces as an area of interest even if faces do not always appear across commercials. We also separated out textual items (both branded and non-branded) given that past research has shown that the absence of audio can direct visual attention to textual items in a scene (Wiebe & Annetta, 2008; Krejtz et al., 2012). Additionally, given that the product in a commercial is highly informative, we created the third group called *objects* that encompassed both the actual product and/or graphical representations of it.

Using iMotions Global Attention Tool 5.0 software, dynamic areas of interest were drawn over each of these AOIs within each commercial. This allowed us to compute fixation durations and fixation counts for each AOI as well as create dynamic heat maps of each commercial.

Grouped AOIs. The objective of our first pass analysis was to look at the effect of AOI type on visual attention between conditions and across commercials. Areas of interest that included *talking* and *non-talking characters* were grouped into an AOI type called *people*. AOIs that included textual items (*textual product information, small print, brand name, textual brand information and non-branded textual items*) were grouped into an AOI type called *text*. And finally, AOIs that included graphical or object type items (*products* and *product illustrations*) were grouped into an AOI type called *objects*. These groups functioned as grouping variables in our analysis between conditions and across commercials.

To explore the effects of the AOI type *people*, a one-way ANOVA was computed with fixation duration and fixation counts as our dependent variables and audio or no audio as our

a. b.

Figure 1: *Heat maps displaying visual attention to the AOIs "Product Illustration" and "Product Information" in the (a) audio and (b) no audio conditions. (Picture material taken from Bunnings TV commercial)*

independent variable. Results from this test revealed no significant differences in fixation durations and fixation counts between the audio and no audio conditions.

To explore the effects of the AOI type *text*, a one-way ANOVA was computed with fixation duration and fixation counts as our dependent variables and audio or no audio as our independent variable. Results from this test revealed that participants in the no audio condition fixated significantly more times on textual items across commercials (M = 5.22) than participants in the audio condition (M = 4.57; F = 4.95, p = .026).

To explore the effects of the AOI type *objects*, a one-way ANOVA was computed with fixation duration and fixation counts as our dependent variables and audio or no audio as our independent variable. Results from this test revealed no significant differences in fixation durations and fixation between the audio and no audio conditions.

Individual AOIs. The objective of our second pass analysis was to look at the effect of each AOI on visual attention between conditions and across commercials. We wanted to see if the results from our first pass analysis were being driven by individual AOIs as opposed to our grouping of AOIs. In this phase, an ANOVA was computed for each of the nine AOIs with fixation durations and fixation counts as our dependent variable and audio or no audio as our independent variable.

AOI type *text* was broken down into AOIs containing some form of text. Somewhat unexpectedly, the *small print, brand name, textual brand information,* and *non-branded textual item* AOIs did not produce any significant differences in fixation durations and fixation counts between our two conditions. The only textual AOI to reveal any significant differences was *textual product information*, which identified additional information about the product in a commercial (e.g., price, adjectives describing the product). This is not to be confused with text items, like the brand's website address or phone number; these items are considered part of the *textual brand information* category. Participants in the no audio condition made significantly more fixations on *textual product information* items (M = 5.78) than those in the audio condition (M = 4.19; F = 10.08, p = .002; see Figure 1 for example). The no audio participants appeared to be fixating longer on these items as well, though this did not reach significance.

The AOI type *objects* was broken down into AOIs *products* and *product illustrations*. For the AOI *product*, participants in the no audio condition fixated longer than participants in the audio condition (no audio M = 2097ms, audio M = 1364ms; F = 7.82, p = .006), while also fixating more times (no audio M = 10.14, audio M = 5.93; F = 19.095, p < .001). However, for the AOI *product illustrations,* participants in the audio condition fixated longer than those in the no audio condition (no audio M = 595ms, audio M = 1401ms; F = 19.433, p < .001), while also looking more times (no audio M = 2.64, audio M = 5.76; F = 22.091, p < .001; see Figure 1).

AOI type *people* was broken down into AOIs *talking character* and *ton-talking character*. Analyses revealed no significant effects in either of these conditions.

4 Discussion

Our analyses suggest that the presence of audio has an influence on the allocation of visual attention during dynamic content. Consistent with past research investigating the effect of audio on visual attention (e.g., Krejtz et al., 2012), participants in the no audio condition tended to look more at textual items in the scene. However, upon further analysis, this effect only occurred for textual information that provided extra information about the product being advertised, and did not include textual items such as the brand name and brand information (e.g., websites and phone numbers). This is not to say participants in the no audio condition failed to look at these other textual items, but that they looked at textual product information significantly more times than participants in the audio condition. Past research suggests that the number of fixations made in an AOI during a non-directed task might indicate that the AOI is of greater interest to the viewer (Jacob & Karn, 2003), and might also suggest that the participant is lacking an efficient means of viewing the scene (Goldberg & Kotval, 1999). In this study, more fixations on textual items might not only indicate that these areas are important to the viewer, but that perhaps if audio were present, the large number of fixations would not be necessary, as the audio could better guide visual attention through the scene.

Furthermore, the preference for *textual product information* from our no audio group may be influenced by the presence of other AOIs on screen. In several commercials, *product illustrations* (one of our nine AOIs) often appeared on screen with *textual product information*. This might offer a reason why participants in the audio condition spent more time looking at *product illustrations* than participants in the no audio condition. Participants with audio could more efficiently take in the scene by utilizing the audio information and lending visual attention to non-textual items (like *product illustrations*). The fact that our no audio participants looked so often at the *textual product in-*

formation suggests that they may have been visually confused as to what elements in the scene would offer the most information. *Product illustrations* were fixated on by no audio participants, as were *textual product information* items by audio participants. What is telling is that participants in the no audio condition looked significantly more times at the *textual product information,* indicating a possible visual confusion that was absent in the audio condition. It should be noted that the presence of multiple AOIs on screen was not controlled for in these analyses, potentially limiting interpretability.

Participants in the no audio condition spent more time looking at the *product* AOI than did participants in the audio condition. Out of the six commercials that were analyzed, four of them had physical products on screen at some point during the ad. In these commercials, when the actual product is on screen there was normally a character on screen interacting with the product. It may be that participants in the no audio condition spent more time looking at the *product* in these instances than the audio participants because lending visual attention elsewhere might not reveal what exactly the scene is about. When there is no audio track providing additional information, identifying the product and what it does provides valuable information about the scene. Future research should build off these findings by including a condition in which non-relevant audio is present.

Our survey results were mainly exploratory in nature. Participants in the no audio condition rated the commercials as more emotional than participants in the audio condition. However, our emotional scale did not account for positive versus negative emotions, but simply the presence of emotion. Participants in the no audio condition also found the pace of commercials more appealing, but further study is necessary to determine what this indicates.

This study build on previous research on the influence of audio on static content by showing that audio has an impact on visual attention during dynamic content. These results may support the notion that audio enhances a viewer's ability to efficiently navigate a dynamic scene, and without it, the viewer is left making more inefficient, and/or different visual allocations. In order to understand the complexities of visual attention in dynamic content, we must first take steps to isolate the effects of certain variables present. This study attempts to add to our understanding of audio's influence on directing attention in dynamic content so that future eye-tracking researchers can attempt to account for some of its extraneous effects.

References

Buswell, G. T. (1935). How people look at pictures: a study of the psychology and perception in art.

Coutrot, A., Guyader, N., Ionescu, G., & Caplier, A. (2012). Influence of soundtrack on eye movements during video exploration. *Journal of Eye Movement Research*, *5*(4).

Egeth, H. E., & Yantis, S. (1997). Visual attention: Control, representation, and time course. *Annual review of psychology*, *48*(1), 269-297.

Frens, M. A., Van Opstal, A. J. & Van Der Willigen, R. F. (1995). Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. *Perception & Psychophysics, 57 (6)*, 802-816.

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics*, *24*(6), 631-645.

Goldstein, R. B., Woods, R. L., Peli, E. (2006). Where people look when watching movies: Do all viewers look at the same place? *Computers in Biology and Medicine, 37 (7)*, 957-964.

Hillstrom, A. P., & Yantis, S. (1994). Visual motion and attentional capture. *Perception & Psychophysics*, *55*(4), 399-411.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *20*(11), 1254-1259.

Jacob, R. J., & Karn, K. S. (2003). Eye-tracking in human-computer interaction and usability research: Ready to deliver the promises. *Mind*, *2*(3), 4.

Krejtz, I., Szarkowska, A., Krejtz, K., Walczak, A., & Duchowski, A. (2012, March). Audio description as an aural guide of children's visual attention: evidence from an eye-tracking study. In *Proceedings of the Symposium on Eye-tracking Research and Applications* (pp. 99-106). ACM.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(4), 565.

Noton, D., & Stark, L. (1971). Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision research*, *11*(9), 929-IN8.

Pieters, R., Rosbergen, E., & Wedel, M. (1999). Visual attention to repeated print advertising: A test of scanpath theory. *Journal of Marketing Research*, 424-438.

Pieters, R., & Wedel, M. (2007). Goal control of attention to advertising: The Yarbus implication. *Journal of Consumer Research*, *34*(2), 224-233.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14).

Vilaró, A., Duchowski, A. T., Orero, P., Grindinger, T., Tetreault, S., & di Giovanni, E. (2012). How sound is the Pear Tree Story? Testing the effect of varying audio stimuli on visual attention distribution. *Perspectives*, *20*(1), 55-65.

Wedel, M., & Pieters, R. (2006). Eye tracking for visual marketing, Foundations and Trends in Marketing, Vol. 1. *Issue*, *4*, 231-320.

Wiebe, E., & Annetta, L. (2008). Influences on visual attentional distribution in multimedia instruction. *Journal of Educational Multimedia and Hypermedia*, *17*(2), 259-277.