Looking at sound: Sound Design and the audiovisual influences on gaze

Jonathan P. Batten & Tim J. Smith

Birkbeck, University of London

***Corresponding Authors contact details**
Department of Psychological Sciences
Birkbeck, University of London
Malet Street
LONDON
UK
WC1E 7HX
Tel: +44 207 631 6359
e-mail: jonobatten@gmail.com

Bio

Jonathan P. Batten

Jonathan P. Batten BSc. Hons, MSc (Gold.) is a PhD student within the CINE (Cognition in Naturalistic Environments) Lab, in the Department of Psychological Sciences, Birkbeck, University of London. Under the supervision of Dr. Tim Smith and Prof. Fred Dick he studies the influence of sound on when and where vision orients in complex scenes (including film and naturalistic environments), and how this affects perception and memory. He is experienced in quantifying active vision with eye-tracking, utilizing psychophysical and behavioral measures to address how sound (music, dialogue, sound-effects) can orient attention through time.

Tim J. Smith

Tim J. Smith BSc. Hons, PhD. (Edin.) is a Reader/Associate Professor in the Department of Psychological Sciences, Birkbeck, University of London. He is the head of the CINE (Cognition in Naturalistic Environments) Lab which studies audiovisual attention, perception and memory in real-world and mediated environments (including Film, TV and VR) as well as the impacts of such media on social and cognitive development. He is an expert in active vision and eye tracking and applies empirical Cognitive Psychology methods to questions of Film Cognition publishing his work on the subject in both Psychology and Film journals.

Abstract

From the earliest films to the blockbusters of today, film has rarely been silent. Live musical accompaniment of silent movies progressed into the synchronized sound of 'talkies' and today film sound is a highly developed craft in which sound-designers believe they have the power to represent and accentuate aspects of the scene, focusing the viewer's attention to specific events and conveying emotion (e.g. Bordwell & Thompson, 2013; Chion, 1994; Murch, 2001). This chapter will attempt to empirically validate some of these beliefs by exploring the separate and integrated influence of each of the primary auditory components of a film's sound design (musical score, dialogue and sound effects; known as "sound stems") on viewer behavior, specifically observing the role of sound in guiding a viewer's gaze through a film. This chapter will approach these issues from the perspective of experimental cognitive psychology. For a review of sound design practice see Sonnenschein, (2001); for a review of the psychological impacts of music and sound see Cohen (2014) and for reviews of film theory on classic and modern sound design see Gorbman (1980) and Donnelly (2009), respectively.

This chapter considers the influence of sound design in two "found" experimental case studies in which filmmakers claim to have manipulated viewer behavior through sound design. Firstly, a highly dynamic and edited scene from *How to Train Your Dragon* (DeBlois & Sanders 2010) was viewed with the three sound stems independently (dialogue, sound effects, music and a silent condition), the attentional synchrony and affective responses between the sound conditions will be compared. Secondly, gaze behavior during the famous single long opening shot from *The Conversation* (Coppola 1974) compared the sound influences (the presence and absence of sound) during discrete sound events within the sequence and viewer gaze behavior via quantitative analysis of heat-maps. We conclude that the influence of sound on viewer gaze during film viewing is not as pronounced as often

thought. Future studies are required to further our understanding of the nuanced influence of sound design and how it shapes our whole experience of a film including attention and affective responses.

Watching a film extends beyond simply viewing a visual sequence, it is an immersive audiovisual experience that engages both senses (and may invoke others; Sobchack, 2000) in order to entertain, inform and transport its audience to narrative worlds. The composer Virgil Thompson quoted in Copeland (1939: 158) conveyed this well: “The quickest way to a person’s brain is through his eye but even in the movies the quickest way to his heart and feelings is still through the ear”. In this chapter will investigate how the auditory and visual modalities interact; refining, placing and contextualizing each other in a continuous semantic interplay that conveys the narrative, the scene context, and the emotional nuances of the scene. Sound enhances the visual scene as an additive force, providing energy, dialogue, motion, warmth, and grounding the limited visual perspective in a 360-degree aural world that is believed to immerse and guide the viewer through the narrative (Gorbman, 1980; Chion, 1994; Sonnenschein, 2001). In this chapter we will explore the empirical evidence for how audio influences our experience of narrative film with a specific focus on whether sound design influences viewer gaze.

Although the early years of cinema did not have synchronized sound this is not to say that the percept of the viewer was absent of any sound. The cinematic world being viewed clearly had sound in which interacting actors could hear each other and the world around them. Rather the movie required the audiences’ imagination to ‘hear’ (Raynaud 2001). Additionally, early cinema screenings were commonly accompanied by an array of audio cues including narrators, live interpreters as well as live music (Elsaesser & Barker, 1990). These served a number of purposes, 1) creating continuity between the traditional use of sound design in theatrical performances which may have shared the bill with an early movie; 2) communicating narrative information; 3) drowning out the whirring mechanical projector; and 4) adding audio energy and emotion to the otherwise ghostly and unnatural looking silent actions (Gorbman, 1980). Since the introduction of the ‘talkies’ the role of sound in film has developed exponentially, with modern films utilizing complex soundscapes for Dolby 5.1 and 7.1 immersive surround sound that envelope the audience in 360-degree spatialized sound. The requirements for film sound are vast, so common practice in film production is to divide sound into three distinct stems: dialogue, music, and sound effects. The sound effect stem encompasses diegetic sounds (the sounds of the scene, including foley) and non-diegetic sound (sounds not attributable to the scene, for example sounds added for dramatic effect). Both the dialogue and diegetic sound effects are altered to conform to the intended phonic world (the phonic resonance of the visually projected space, for example adding reverb and compression). Music is usually non-diegetic and completely for the benefit of the audience (the characters do not generally hear or interact with it), as an emotive and narrative emphasis or counterpoint (Gorbman, 1980). This chapter will consider how each of these three stems individually and when integrated, influence where and when viewers attend to visual features in Hollywood, narrative film.

Prior to investigating how audiovisual influences may alter film viewing behavior, we must first consider the nature of the two perceptual systems. When comparing the perceptual attributes of the auditory and visual systems, two key features stand out. Firstly, the human field of view is limited to around 130° (where 360° is a full circle around the viewer's head; Henderson 2003) and our ability to perceive high-level detail and color is further limited to the visual information projected close to the center of the retina (known as the *fovea*), with image quality decreasing rapidly with eccentricity, further limiting the useful field of view. To perceive visual events in the world the eyes must continuously move so that the parts of the scene we are interested in are projected on to this high-resolution part of the retina (on average three times per second for scenes; Rayner 1998). Where the eyes move is subject to constant competition between visually salient image features and task/semantic relevance (Tatler et al. 2011), and this focus of visual information means that visual events that occur outside of this ‘spotlight’ are less likely to be processed sufficiently to make it into our conscious awareness (Jensen et al. 2011). As a result, the visual system suffers from severe sensory capacity limitation. In contrast, there is no “field of view” for audition as all audible information from our 360^{o} surroundings are received by the auditory system. However, for auditory information to be perceived neural processes are required that inhibit, isolate, and group sounds into attributable sources, a process known as auditory scene analysis (Bregman 1990).

The second feature that contrasts the two modalities is the dominance of vision in processing *where* information is (spatial), and in the auditory modality for *when* it occurs (temporal). Both senses have spatio-temporal components but the difference in emphasis is a direct product of how the sensory information is formed: sound is produced by changes in air pressure over time whereas visual information is largely a product of the difference in absorption of photons by adjacent parts of the human retina. To identify and attend to a sound source (for example a person speaking), requires the binding of a continuous stream of auditory features through time by temporally grouping sounds based on phonic similarities (Bregman 1990). Whereas perceiving a visual object involves processing the changes in brightness projected spatially across the retina in order to identify edges and bind these together to form an object (Marr, 1982). But real-world perception is rarely unimodal and both auditory and visual information are perceptually bound by their relative spatio-temporal features, what the Soviet filmmaker, Sergei Eisenstein termed ‘the synchronization of the senses’ (Eisenstein 1957: 69). In binding information, the perception systems utilize the relative strengths of the two modalities to form a coherent and efficient precept of the world. When there is perceptual ambiguity for one of the senses, information from the other is employed which can produce perceptual illusions. For example, the ‘ventriloquism effect’ (Thurlow & Jack 1973) where highly simplified spatially separated audiovisual stimuli are perceived as joined when their presentation is synchronized in time. Or the McGurk effect, whereby simultaneous mismatching mouth shapes and syllabic sounds form an integrated but illusory different auditory percept not present in either modality (McGurk & MacDonald 1976). A notable example of an illusory audiovisual percept from film identified in Chion (1994: 12), is the 'pssssht' door sound used in the early Star Wars films, which gives the viewer a percept of doors closing yet the doors are never seen in motion. The use of sound combined with the abrupt visual cut is fused to provide an illusory percept of visual motion that matches the temporal dynamics of the audio.

Beyond the ability to generate audiovisual illusions, the combination of audio and visual information is generally perceptually advantageous. In a psychophysical 'pip and pop' paradigm, Van der Berg et al., (2008) found that participants' identification time for detecting an ambiguous target line within a complex array of lines was significantly reduced (i.e. the line seemed to 'pop' out) if the visual presentation was accompanied by an auditory tone (a 'pip'). This effect provides evidence that temporal binding of both the audio and visual information is used to efficiently disambiguate visual information, and that this bound representation is perceptually enhanced (more salient) in a viewer's attention. A fundamental benefit of a bound audiovisual representation is that it can inform the temporal dynamics of attention (a limited resource) through time. An example of this was observed in a simple visual discrimination task with music by Escoffier et al. (2010). The authors presented visual images both in synchrony with the musical beat and randomly in time. They found that reaction times on-beat were significantly faster than off-beat presentation, suggesting an entrainment of visual attention to the music, i.e. the use of predictable auditory temporal events (musical pacing) enhanced the predictive dynamics of visual attention through time. These findings are compelling, but ultimately in contrived (somewhat reductive) scenarios, that have little auditory or visual complexity. To date there is little research that extends these psychophysical paradigms up to more complex naturalistic scenes, or applies them to film. Were these effects to scale up to the complexity of film, the temporal correspondence of sound to a visual event or object should enhance film viewers' attention to it, increasing the probability of gaze fixating it (as has been suggested by sound designers; Sonnenschein 2001). Secondly, the musical rhythm of film scores would influence attention to key visual elements introduced on-beat (and inversely be detrimental to off-beat moments), potentially altering and influencing memory and narrative understanding subject to these time points as has been proposed by theorists of classical narrative film scoring (e.g. Gorbman 1980).

A fundamental example of how sound designers believe they can influence viewer attention is the introduction of a sound corresponding to a visual object (Bordwell & Thomson 2013; Murch 2001). Chion (1994) believed the inclusion of sound has influenced how complex the visual content in film can be. He noted that silent cinema, demanded a simpler visual scene as without synchronized sound the visual complexity of a scene would overwhelm the viewer, fail to highlight the important details and lead to confusion. Such gaze guidance by sound was also predicted by Sergei Eisenstein (1957) in relation to a sequence in his 1938 film, *Alexander Nevsky*. Eisenstein believed that the score (composed by Sergei Prokofieff) directed the rise and fall of viewers' attention in synchrony with the rise and fall of the music. A recent empirical test of his predictions by Smith (2014), provided some limited correspondence between his predictions and viewer gaze allocation. However, the overarching musical influence of Prokofieff's score on where gaze was located was not supported as viewers' gaze was no different with the music than in silence. Rather the changes in the music complemented the existing changes in the visual scene across cuts, producing vertical gaze shifts in time with the rise and fall of the music but no significant association between music and gaze was found within shots. These findings potentially confirming Prokofieff's ability to see the visual patterns of the scene and feel them on his own gaze before expressing them in the musical score.

Eye movement evidence in support of auditory influences on where people look when watching films is limited. When watching edited sequences, the gaze of viewers often clusters around faces, hands, and points of motion in the scene, a phenomenon we have termed '*attentional synchrony'* (Smith, Levin & Cutting, 2012; Smith & Mital, 2013; Smith, 2013). The *attentional synchrony* of multiple viewer's eye movements is unsurprising when you consider the tendency in film to frame the salient action centrally (Cutting, 2015). A highly effective viewing strategy for watching a film is therefore to simply maintain gaze to the screen center (Tseng et al. 2009; Le Meur et al. 2007). The frequent central and close framing of action in narrative films combined with the general tendency for gaze to cluster around these centrally located salient visual features (faces, hands and points motion) limits the possibility for audio influences to draw attention away from the screen center and direct it to peripheral screen locations. In fact, the apparent dominance of visual features and shot composition on viewer attention has been empirically shown to be so robust that we have recently referred to it as '*the Tyranny of Film'* (Loschky et al. 2015). Despite these complexities there is some evidence that audio can influence dynamic scene viewing. A study by Vo et al. (2012) eye-tracked two groups of participants watching a series of ad-hoc interviews (pedestrians on the street) that were either accompanied by synchronized speech with background music or simply background music. They found that gaze was captured to the faces of people, and when they spoke people looked at the speaker's mouth. This mouth capture was notably reduced when watching the scene without the speech (music condition). Similar evidence for gaze differences with and without a film's original soundtrack has been presented by Rassel, Robinson and colleagues (2015; 2016). In two eye tracking studies examining viewer gaze behavior during the Omaha Beach sequence from *Saving Private*

Ryan (Spielberg 1998) and the climactic chase sequence from *Monsters Inc.* (Docter, Silverman & Unkrich 2001) they reported a qualitative trend towards greater gaze exploration of the screen periphery in the mute conditions compared to the audio conditions and potentially greater sensitivity to visually salient events in the periphery (such as a foot movement or bright light) in the absence of sound (although none of these differences were statistically significant; see Smith, 2015 for further critique).

There is also some evidence that the addition of film sound and especially music can influence the duration of fixations (the period of a relatively stable localization of gaze). Wallengren & Strukelj (2015) identified some evidence of a reduction in fixation duration subject to the inclusion of film music (although the effect may reverse when the soundtrack includes speech; Rassell et al. 2016), and a study by Coutrot & Guyader (2013), found that the inclusion of film sound increased the attentional synchrony of participants' eye movements, and influenced the size of the saccades (suggestive of exploratory scene viewing away from the center). This may be evidence that the sound does guide viewers' attention as predicted by Chion and others. Taken together this evidence allows us to predict that audible dialogue would be expected to capture gaze to the mouth of the speaker, music may reduce the duration of fixations, and the addition of audio generally could promote a clustered exploration of the visual scene.

In this chapter we will investigate the influence of audio on viewer gaze via two stylistically very different "found" experimental case studies, *How To Train Your Dragon* (DeBlois & Sanders 2010) and the classic Francis Ford Coppola movie *The Conversation* (1974) which was famously inspired by the work of the Oscar winning sound designer, Walter Murch. By using famous case studies of sound design we aim to demonstrate the relationship between viewer gaze and the three key elements of sound design, music, dialogue and sound effects as they would appear in Hollywood narrative movies, as well as highlight the need for future research of the audiovisual influences on overt attention using more controlled naturalistic stimuli.

How to Train Your Dragon

One of the challenges facing research into how sound design influences viewer attention is the inaccessibility of a professionally produced film's individual sound stems. Studies comparing a soundtrack's presence or absence (see above) can identify the overall influence but cannot pinpoint whether individual audio components such as sound objects or music independently influence attention. To overcome this limitation we will exploit a "found" experiment presented during a SoundWorks Collection interview with the creative team responsible for the animated film *How to Train you Dragon* (DeBlois & Sanders 2010). During this interview a short clip from the film was repeated three times to feature in isolation the separate sound stems (dialogue, music and sound effects). This exemplar of sound design provided an excellent opportunity to extract and investigate the influence of the final sound mix from each stem on eye movement behavior and affective response. The 52 second clip taken from the very beginning of the movie was viewed by forty-eight adult participants (36 Female, aged from 20 - 50 years old). Twelve participants were in each audio condition (music, dialogue, sound effects and a silent control). Each participant gave informed consent for their eye movements to be recorded (a Tobii TX300 screen-based eye tracker recording at 300Hz with video resolution of 1920x1080, 24fps), and were tasked to watch the clip with the knowledge of a later memory test (to encourage close viewing). Following the clip, they rated how the film made them feel on both a 9-point arousal and happiness scale (Bradley & Lang 1994).

How to Train Your Dragon (2010), is a highly successful DreamWorks Animation film that tells the story of a diminutive and resourceful teenage Viking (Hiccup), in a land plagued by dragons. The story follows Hiccup, who befriends and trains an intelligent dragon (Toothless), ultimately saving his village and earning the pride of his father (Stoick the Vast, the village chief). The eye tracked 52-second scene is set in Hiccup's hill-top village and the plot both introduces the different dragons that plague the people, whilst also demonstrating their destructive abilities around the village (lighting houses on fire, stealing sheep, destroying defenses). The overarching message of the clip is that there is a fight between the people who are equipped with simple weapons and the immense destructive powers of the different dragon types. The clip ends with a narrated description of the elusive and powerful Night-Fury dragon who causes the explosive demolition of a large boulder throwing catapult (containing Stoik the Vast), and is later to be revealed as the character Toothless.

In the music condition, participants watched with the associated film music (composed by John Powell), which was formed of percussive drumming and a brass refrain. Two features of the music stand out, firstly the use of a pulse like beat (marching snare sounds and low booming drums) reinforce the visual momentum of the scene both within and across cuts. Secondly, the rise and fall of the horn melody evokes awe and suspenseful emotion and the musical motif calls to mind film scores of battles scenes. The dialogue condition contained not only the speech of the characters, but also the narration (the voice of Hiccup) and all other human vocal noises (murmurs and vocal exertion sounds). With the exception of the silent condition the dialogue version was relatively limited in the amount of sound and variability. The sound effects condition contained a combination of the Foley and the sound effects for both the actions on scene for example, low rumbling explosions, impact sounds, animal noises and dragon vocalizations.

The specific sound stems each add different qualities to the film. The music adds an emotion and tempo not found in other mixes. The additive quality of music as energy and emotion would be predicted to increase enjoyment and arousal ratings for the film (when compared to the silent condition; see Gorbman, 1980). The music would be predicted to increase the dilation of pupillary response, which is modulated by arousal state changes and variance in cognitive demand (Hoeks & Levelt, 1993). The music would also be predicted to decrease fixation durations as observed in Wallengren and Strukelj (2015). The sound effects condition containing diegetic sound would be predicted to guide attention in a more tightly clustered manner than the other auditory conditions, increasing attentional synchrony through time (Coutrot & Guyader, 2013; Robinson, Stadler, & Rassell, 2015; Rassell et al. 2016). Additionally, as the representation of sound objects is believed to capture attention, when a clear audiovisual correspondence occurs this will capture gaze to that object. Finally, when characters on screen speak, gaze is predicted to cluster on the mouth more in the dialogue condition (Coutrot et al, 2012; Võ, et al., 2012; Foulsham & Sanderson, 2013; Rassell et al, 2016).

As predicted the participants in the music condition, reported a significantly higher happiness level than those in silence (revealed by a statistical t-test comparing the means between conditions; $t(22) = 3.02$, $p < .01$). There were no other significant differences in the self-report measures between the four conditions, including no difference in arousal (excitement) between those with music and silence. We did not show significantly different fixation durations between the conditions, nor any trend indicative of a shortening of fixation durations in the music condition (revealed by an Analysis of Variance; $F(3,44) = .548$, $p = .652$). Furthermore, whilst pupillary responses were highly sensitive to changes in luminance observed in Figure 1., there was no support for the prediction that any audio condition significantly altered pupil dilation.

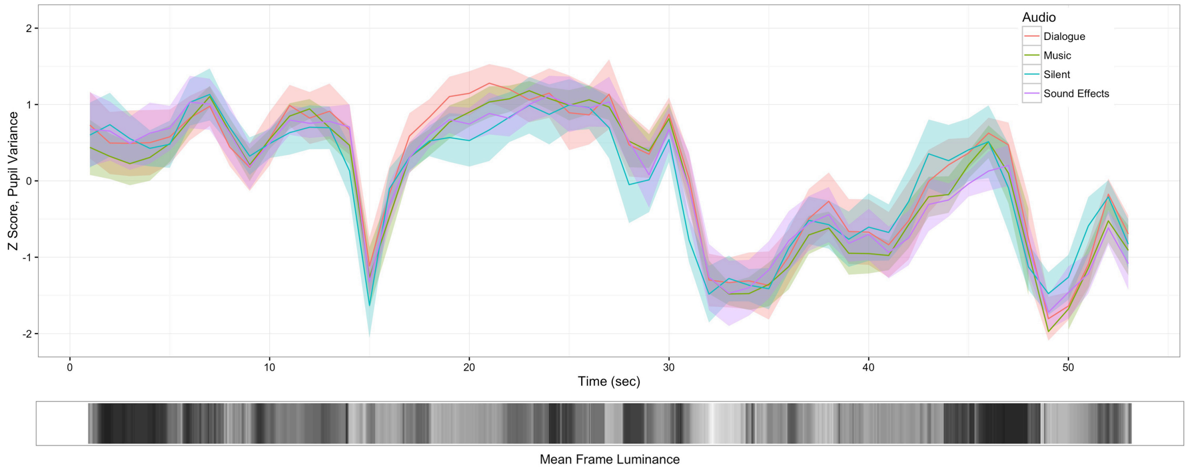

Figure 1. Normalized pupil variance across conditions (red = Dialogue, green = Music, blue = Silent, purple = sound effects) with 95% confidence intervals, and a representation of the mean luminance of each frame through time (from black = dark to light).

Analysis of the variance of gaze scan paths between the groups through time was conducted using a methodology employed in Loschky et al., (2015). This methodology takes the gaze from each frame of the movie and calculates the probability that each gaze point belongs to its own 2D spatial distribution (e.g. within the Silent condition) as well as calculating the probability between groups (Dialogue vs. Silent, Music vs. Silent, SFX vs. Silent). These probabilities are then normalized relative to the referent group's (Silent) mean and standard deviation, creating a Z-scored gaze similarity score. The Silent condition was chosen as the baseline so we could identify the additive influence of sound. Negative values indicate random or less clustering than average. Positive values indicate moments of tighter than average clustering and separation of the lines indicates that gaze in that condition is located in a different part of the screen than the Silent condition (see Loschky et al., (2015), for further details about the method). A shuffled baseline was added as a referent for what randomly distributed gaze would look like (green line in Figure 2). By shuffling the gaze data from the Silent condition and rerunning the gaze similarity analysis for this shuffled data it provides a baseline for random (i.e. asynchronous) gaze. In Figure 2, the gaze similarity means present a generally tightly clustered distribution of gaze that does not vary notably by auditory condition and are mostly more clustered than would be predicted by chance (denoted by the moments when the lines intersect with the shuffled baseline). Each of the significant moments in the plot are attributable to visual events as the groups tend to peak in unison, for example at 26 seconds the cut to a medium shot of Stoik's face produced a tight clustering of gaze to his eyes that did not differ by condition. This is further evidence for *the Tyranny of Film* (Loschky et al., 2015), i.e. that the visual editing techniques, lighting, and central framing of action produced reduced exploration of the screen space and centralized scan-path.

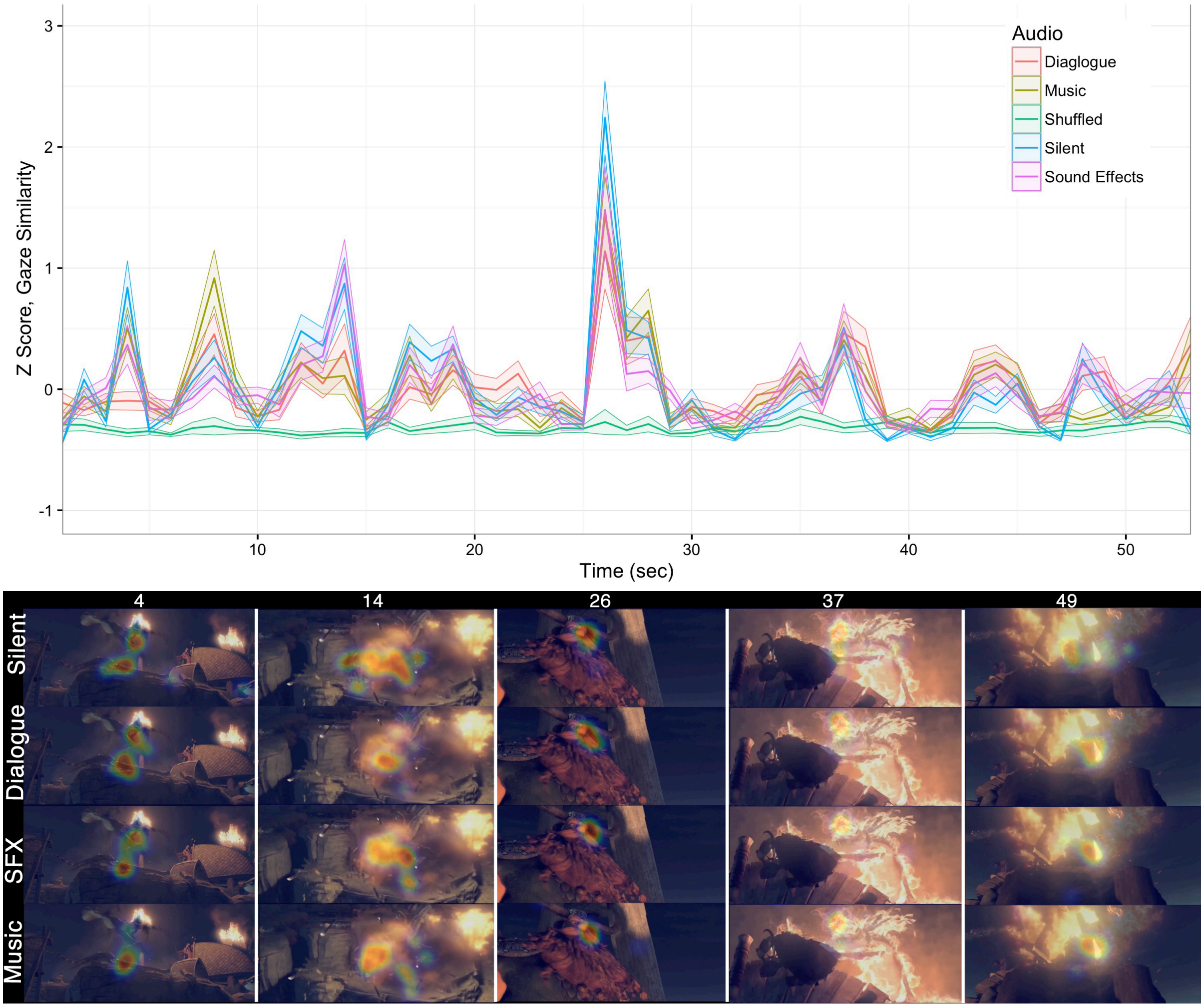

Figure 2. Gaze similarity over time from *How To Train Your Dragon* under four different

audio conditions (red=Dialogue, khaki =Music, green = silent baseline ‘Shuffled’, blue = Silent and purple=Sound Effects). Upper and lower faded color bands around each line indicate 95% confidence intervals. None of apparent differences between these bands reach statistical significance. Key frames from *How to Train Your Dragon* (DeBlois & Sanders, 2010; Copyright: DreamWorks Animation) with gaze heat-map overlaid for each audio condition are displayed at the bottom.

Sound events within the clip were isolated to test whether audiovisual representation of objects captures gaze. Regions of interest (ROI) dynamically traced the audiovisual events (for example the sheep baaing, the villager dialogue and the sound of a dragon exhuming gas) for comparison between the conditions. No significant influences of audiovisual representation on gaze to these ROIs were observed in the sound effect or dialogue conditions. What is apparent is that the editing and highly-mobile virtual camerawork was highly effective in holding attention at the screen center. This drive towards the screen center combined with highly salient character motion preceding every diegetic sound effect meant that the gaze scan-path was very conservative and not influenced by audio changes. These findings mirror prior evidence (Smith, 2013; Loschky et al. 2015; Redmond, 2015; Smith, 2015), that fast-paced, highly composed film sequences from a blockbuster narrative film do not afford the opportunities for idiosyncratic gaze exploration that would be required to observe audio influences. Although, prior studies using slower paced film clips with more scope for exploration have shown an influence of audio on spatial distribution of gaze (Coutrot et al., 2012; Võ, et al., 2012; Foulsham & Sanderson, 2013; Rassell et al., 2016). To provide greater opportunity for gaze exploration, the next case study used a classic example of innovative sound design within a single long take, long shot: the opening scene from Francis Ford Coppola’s film, *The Conversation* (1974).

The Conversation

The Conversation (Coppola 1974) is a film about Harry Caul (Gene Hackman) a renowned surveillance operative in San Francisco, who wrestles with the moral implications of the information he captures. The sound designer Walter Murch was nominated for a Best Sound Oscar award for his work on the film. Whilst the film is a fine example of the 1970s American art film which differs from *How To Train Your Dragon* on many dimensions, not least of all an active subversion of classical Hollywood formal technique and narrative style (Elsaesser, 1975), our use of the film here will focus on its famous opening sequence that serves as an antithesis of the highly dynamic and rapidly edited sequence used in our previous case study. The opening scene is unique for both its use of a single continuous shot (with a subtle use of zoom), and for the use of a solely diegetic sound track. There are no overt non-diegetic sound effects, dialogue or music. The 2 minute and 54 second scene begins with a long wide shot of Union Square, bustling with Christmas shoppers. The sequence slowly pans and zooms, initially not directly framing any particular person or interaction. The square is busy, with a band playing in the bottom right corner, a mime who is playfully mimicking passers-by, dogs barking, and generally a scattered crowd of people. It ends with a zoomed in shot of Harry Caul as he exits the square. The audio from the scene is completely diegetic, and (with hindsight) a surveillance recording that is interspersed with short periods of incoherent electronic noise as Caul tunes in to objects of interest. The general mix (aside from these moments of distortion), captures footsteps, the band playing, dancing foot-scuffs, hand-claps, dogs barking and the hubbub of a busy square. These sounds provide a unique opportunity to isolate and identify gaze differences subject to the visual correspondence with diegetic sounds. The most identifiable (and least competitive) moment is when the sound of a dog barking corresponds with the entrance of a dog from the right of the screen. The barking increases in loudness as the dog enters the screen reinforcing the audiovisual contract (Chion 1994). A second isolatable moment in the auditory mix is the 16 second period when the mix is solely the band playing (increasing in loudness, then fading out with the song end). The predictions for the study are: Firstly, an auditory representation of the dog barking will both capture attention to the screen entrance of the dog, and that those in the auditory condition will look at the dog faster than those without. Secondly, the self-reported ratings for arousal and happiness should be both happier and more excited (arousal) in the audio condition when compared to the silent condition. The third prediction is that the inclusion of audio will facilitate a more 'guided' visual attention, increasing the clustering of gaze within the group to similar screen locations in time. The fourth prediction is that during the auditory representation of the band (noticeably reduced auditory complexity), the pupil dilation reactions of the two groups to the music (in audio) and to the visuals alone (in silence) should differ indicative of differing interpretations of the scene, the isolation of the music should disambiguate the scene for those in the audio condition.

Forty-eight adults, 36 Female, 20-50 years' old, watched the first 2 minutes and 54 seconds of the opening sequence of the film. Twenty-four watched with the corresponding sound (played through headphones), and 24 in silence. Eye tracking hardware and presentation conditions were identical to the previous case study. Each participant was asked to watch the film with the knowledge that a memory test based on what they had seen will follow (although this test was not administered). After the clip, each participant rated how happy (sad - happy), and how aroused (excited - unexcited), the film made them feel on a scale from 1-9 (Bradely & Lang, 1994).

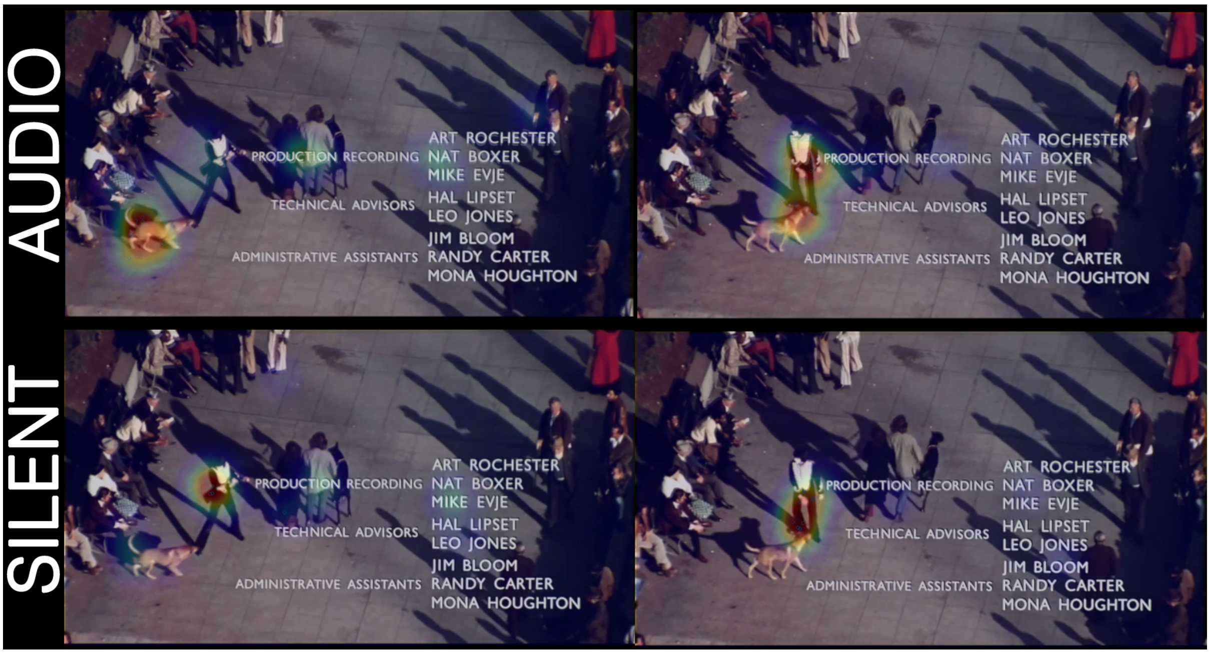
AUDIO
SILENT
PRODUCTION RECORDING ART ROCHESTER
NAT BOXER
MIKE EVJE
TECHNICAL ADVISORS HAL LIPSET
LEO JONES
ADMINISTRATIVE ASSISTANTS JIM BLOOM
RANDY CARTER
MONA HOUGHTON

Figure 3. Gaze distribution heat map for two frames (left and right column) from *The Conversation* (Coppola, 1974; Copyright: The Directors Company) that highlight the early allocation of gaze to the dog in the Audio condition (Top) compared to the Silent condition (Bottom). The red 'hot spots' indicate a clustering of multiple viewers' eye position).

As observed qualitatively in the heat-map overlay of Figure 3, the group who heard the dog barking were significantly faster to fixate the dog (mean time from the entrance of the dog to the screen = 1316.8ms) than those in silence (1527.63ms; a statistical t-test of the mean times to fixate the dog showed a significant difference, *t*(35) = -2.114, *p* = .048). Both groups had a similar proportion of participants who looked at the dog. This provides some evidence that auditory information influences visual attention to corresponding objects, although the effect is subtle mostly due to the general salience of moving objects and the need for movement to generate sound (these audiovisual objects are already visually salient). The effect of audio in this instance is a slightly earlier capture of attention, rather than the clear guidance of attention that is predicted with the inclusion of sound. The self-reported scores for happiness and arousal (Bradley & Lang, 1994), support the general prediction that audio would be more exciting and generally make people feel happier than watching in silence. Those who watched the clip with the audio (M = 3.46, SD = 1.38) reported significantly happier scores than those who watched in silence (M = 4.29, SD = 1.55), *t*(46) = 1.97, *p* = .03 (one-tailed). Also, those who watched the clip in silence (M = 6.08, SD = 2.10) reported significantly less excitement than those with the audio (M = 4.61, SD = 2.19), *t*(46) = 2.36, *p* = .012 (one-tailed).

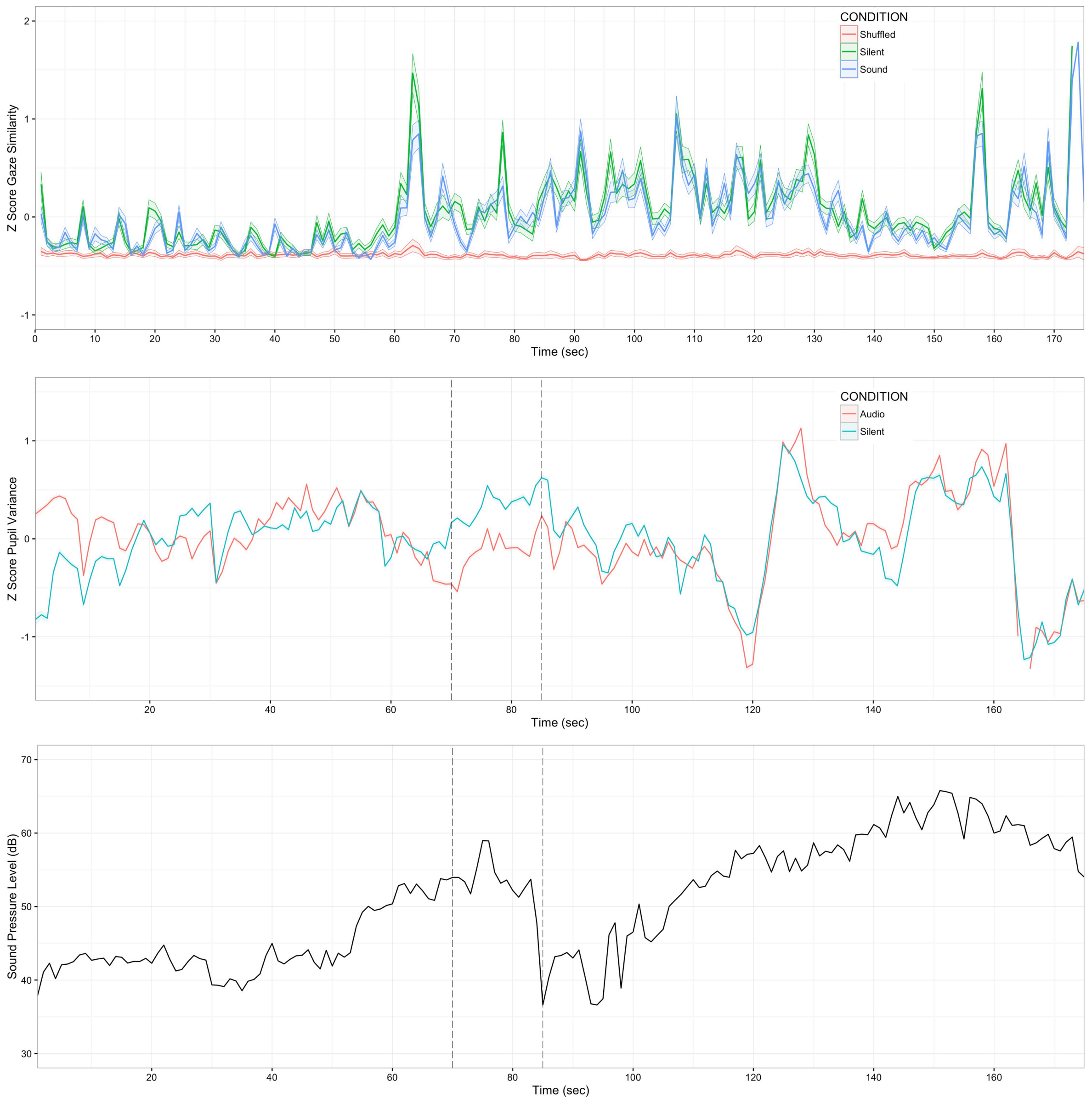

Figure 4. Gaze similarity over time from The *Conversation* with the two different audio conditions (green = Silent, blue = Sound, Red = Baseline shuffled; faded upper and lower

bounds around each line indicate the 95% confidence intervals). Normalized Pupil Variance through time (green = Silent, red = Audio). Mean sound pressure level (dB) of the audio mix through time.

As with the *How to Train Your Dragon* clip, the gaze similarity of the participants was analyzed between the two groups. This is visualized in the top panel of Figure 4. A shuffled baseline derived from the gaze data in the silent condition was again included as a referent for randomly distributed gaze. Contrary to the prediction of the study, the gaze similarity values were not significantly different between the audio and silent groups *F*(1,46) = 1.04, *p* = .3. The silent condition tended to have slightly more clustered distribution of gaze (for example the peaks at 63 and 157 seconds). There is variance over time in the clustering distributions, but the pattern of variances does not indicate an additive auditory influence, as both the silent and audio conditions peak and trough in unison through time, indicating a primary shared influence of visual events. When considering the prior analysis on the preceding effect on the dog bark, this short (below 1 second) variance is not noticeable as a peak in the gaze similarity data, as the time difference and general gaze locations are not sufficiently different in distribution. As well as the spatial distribution of gaze showing no difference between audio conditions, the timing of eye movements (measured as average duration of fixations) also failed to differ, *t*(46) = 0.384, *p* = .703, further evidence that eye movements were not generally effected by the addition of audio.

The second isolated section of the film clip utilized for analysis was from 70 to 85 seconds, highlighted in the bottom two panels of Figure 4 by two vertical dashed lines. At 70 seconds the dominant sound of surveillance equipment distortion fades, and the music of the band increases noticeably in loudness (see the third panel of sound pressure level). The band is the only identifiable auditory signal until 85 seconds when the music fades as the song ends. The divergence of the pupil change between the groups, with a reduction in dilation in the audio condition (middle panel, Figure 4), can only be attributed to the difference in sound between the groups (there was no significant difference in gaze similarity between the conditions during this period). To confirm that the pupil variance between the conditions was significantly different during the period that the band was playing the Z-score pupil variance values were tested over time between the two audio conditions (audio and silence). A 2-way ANOVA of Time by Condition had a significant main effect of Condition, $F(1,46) = 7.21$, $p = .010$, $\eta^2 = .135$, as pupil dilation was significantly greater in the silent condition than the audio. There was a significant main effect of time, $F(15,690) = 3.705$, $p < .001$, $\eta^2 = .075$ as the dilation values changed over time. There was no significant interaction between the Time and Condition, $F(15,690) = .669$, $p = .816$, $\eta^2 = .014$. Pupillary responses are sensitive to changes in mental processing demands (cognitive load) and to changes in arousal (Hoeks & Levelt 1993). With increases in the cognitive load or in arousal states the pupil dilates. Reduction in dilation is the inverse of this relationship, with reduced processing demand (or complexity) there is a reduction in pupil size (Winn 2016). The reduction in pupil dilation of the audio condition compared to the silent condition during this 16 second clip suggests that the clarity of the auditory signal during this period may have increased narrative engagement and simplified the viewing task of understanding what is being shown (comparatively the sound design of the rest of the sequence is frequently layered with multiple potential sources, footsteps, music, distortion etc.). Alternately, this could be an example of covert attention (attention in the absence of gaze to the attended object), as the clarity of the sound signal can negate the need for overt attention. This facilitation of covert attention may have simplified the cognitive demands of tracking multiple objects within the scene. This hypothesis would need to be tested using measures of covert attention, for example reaction time probes or electroencephalography studies (Nako et al. 2016)

In summary, where, when and for how long people looked at the opening of The *Conversation* was almost completely due to the visual content in the scene. The addition of audio certainly altered how people felt and processed the attended information, based on the self-report scores and the pupil variance but did not generally result in differing eye movement behavior. The notable exception is the single instance of faster allocation of gaze to the dog when the barks could be heard, a clear example of a diegetic audiovisual event capturing viewer gaze but one which seems to be rare in the film sequences analyzed here.

The two case studies presented here suggest that sound design in a fast-paced highly-edited film sequence and a slower minimally composed long shot have clear impacts on audience affective responses but virtually no impact on the timing or location of gaze. Overt attention seems to be predominantly under visual influence with the sound design accentuating and complimenting the visuals (as was previously observed in *Alexander Nevsky*; Smith 2014). But given that prior studies using dialogue film sequences have demonstrated gaze differences when audio was removed (Coutrot et al. 2012; Võ, et al. 2012; Foulsham & Sanderson 2013; Rassell et al. 2016), our current null results may either be due to the choice of scenes chosen or the audio manipulations. Given the complex dimensions that sound designers manipulate whilst crafting a film scene, the heavy-handed on/off manipulations used so far to investigate the influence of sound design in film may be missing the subtle nuance of how sound may guide attention and shape a viewer's overall experience of a film. Future studies must manipulate specific and isolatable diegetic sound effects, the moments of speech onset and musical patterns independently from the corresponding visually salient events. This controlled approach will facilitate investigation into the independent contributions of sound and visuals to the dynamics of gaze. By conducting these studies, we may start approaching a better scientific appreciation of the power of sound design.

Acknowledgements

We thank all participants for taking part in these studies. Jonathan P. Batten was funded by an ESRC PhD studentship (ES/J500021/1). Tim J. Smith was funded by a Philip Leverhulme Prize (PLP-2013-028).

References

Alexander Nevsky. 1938. [Film]. Sergei Eisenstein. Soviet Union: Mosfilm.

Bordwell, D. & Thompson, K., 2013. *Film Art: An Introduction* 10 ed., New York: McGraw-Hill Education.

Bradley, M.M. & Lang, P.J., 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental ….*

Bregman, A.S., 1990. *Auditory scene analysis: The perceptual organization of sound*, MIT Press.

Chion, M., 1994. *Audio-Vision: Sound on Screen* C. Gorbman, ed., New York: Columbia Univeristy Press.

The Conversation. 1974. [Film]. Francis Ford Coppola. USA: The Directors Company.

Cohen, A.J., 2014. Film Music From the Perspective of Cognitive Science. In *The Oxford Handbook of Film Music Studies*. The Oxford Handbook of Film Music Studies.

Coutrot, A., Guyader, N., Ionescu, G. & Caplier, A., 2012. "Influence of Soundtrack on Eye Movements During Video Exploration"*, Journal of Eye Movement Research 5*, no. 4.2: 1-10.

Coutrot, A. & Guyader, N., 2013. Exploration of dynamic natural scenes: influence of unrelated soundtracks on eye movements. *17th European Conference on Eye Movements, Lund, Sweden*.

Cutting, J.E., 2015. The Framing of Characters in Popular Movies. *Art & Perception*, 3(2), pp.191–212.

Donnelly, K.J., 2009. Saw heard: musical sound design in contemporary cinema. In, Buckland, Warren (eds.) *Film Theory and Contemporary Hollywood Movies*.

Escoffier, N., Sheng, D.Y.J. & Schirmer, A., 2010. Unattended musical beats enhance visual processing. *Acta Psychologica*, 135(1), pp.12–16.

Eisenstein, S., 1957. *The Film Sense* J. Leyda, ed., New York: Film Form.

Elsaesser, T, and Adam B., 1990. *Early Cinema: Space, Frame, Narrative*, London: British Film Institute.

Elsaesser, T. 1975. The Pathos of Failure. American Films in the 1970s. Notes on the unmotivated Hero [1975]. *Eds., Horwath, King (Hg.): The last great American Picture Show*, 279-292.

Findlay, J.M. & Gilchrist, I.D., 2003. Active Vision, Oxford University Press, USA.

Foulsham, T. & Sanderson, L. (2013). Look who's talking? Sound changes gaze behaviour in a dynamic social scene. *Visual Cognition*, *21 (7),* 922-944.

Gorbman, C., 1980. Narrative Film Music. in *Yale French Studies, No. 60. Cinema/Sound*. Yale University Press.

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

Hoeks, B. & Levelt, W.J., 1993. Pupillary dilation as a measure of attention: A quantitative system analysis. *Behavior Research Methods*, 25(1), pp.16–26.

How To Train Your Dragon. 2010. [Film]. Dean DeBlois & Chris Sanders. USA: DreamWorks Animation.

Jensen, M.S. et al., 2011. Change blindness and inattentional blindness. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(5), pp.529–546.

Le Meur, O., Le Callet, P. & Barba, D., 2007. Predicting visual fixations on video based on low-level visual features. *Vision Research*, 47(19), pp.2483–2498.

Loschky L.C, Larson A.M, Magliano JP, Smith T.J., 2015. What Would Jaws Do? The Tyranny of Film and the Relationship between Gaze and Higher-Level Narrative Film Comprehension. *PLoS ONE* 10(11): e0142474

Marr, D., 1982. Vision: A computational investigation into the human representation and processing of visual information, *Henry Holt and co. Inc.,* New York, NY, 2, 4-2

McGurk, H. & MacDonald, J., 1976. Hearing lips and seeing voices. *Nature*, 264, pp.747–748.

Monsters Inc. 2001 [Film] Pete Docter, David Silverman, & Lee Unkrich, USA: Pixar Animation.

Murch, W., 2001. *In the Blink of an Eye* 2nd ed., Silman-James Press.

Nako, R, Grubert, A, & Eimer, M., 2016. Category-based guidance of spatial attention during visual search for feature conjunctions. *Journal Of Experimental Psychology: Human Perception And Performance*, 42, 10, pp. 1571-1586

Rassell, A., Robinson, J., Verhagen, D., Pink, S., Redmond, S. & Stadler, J., 2016, Seeing, sensing sound: eye tracking soundscapes in Saving Private Ryan and Monsters Inc.. In Reinhard, C and Olson, C (ed), *Making sense of cinema: empirical studies into film spectators and spectatorship*, Bloomsbury Academic, New York, N.Y., pp.139-164.

Rayner, K., 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3), pp.372–422.

Raynauld, I., 2001. Dialogues in Early Screenplays: What Actors Really Said. In: Abel, R. & Altman, R. eds. The Sounds of Early Cinema. Bloomington: Indiana University Press, pp. 69 – 78.

Redmond, S., 2015. Eye tracking the sublime in spectacular moments of science fiction film. In Redmond, Sean & Marvell, Leon (ed), *Endangering science fiction film,* Routledge, New York, N.Y., pp.32-50

Robinson, J., Stadler, J. & Rassell, A., 2015. Sound and Sight: An Exploratory Look at Saving Private Ryan through the Eye-tracking Lens", *Refractory 6 (1).*

Saving Private Ryan. 1998 [Film] Directed by Steven Spielberg. USA: Dreamworks/Paramount.

Smith, T.J. & Levin, D. and Cutting, J.E., 2012. A window on reality: perceiving edited moving images. *Current Directions in Psychological Science 21* (2), pp. 107-113.

Smith, T.J., 2013. Watching you watch movies: using eye tracking to inform film theory. In: Shimamura, A (ed.) *Psychocinematics: Exploring Cognition at the Movies*. New York, U.S.: Oxford University Press, pp. 165-191.

Smith, T.J., 2014. Audiovisual Correspondences in Sergei Eisenstein's Alexander Nevsky: A Case Study in Viewer Attention. *Cognitive Media Theory*, pp.1–19.

Smith, T.J., 2015. Read, watch, listen: a commentary on eye tracking and moving images. *Refractory: A Journal of Entertainment Media 25* (9)

Smith, T.J. & Mital, P.K., 2013. Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. *Journal of Vision*, 13(8), pp.16–16.

Sobchack, V., 2000. What My Fingers Knew, The Cinesthetic Subject, or Vision in the Flesh. in *Carnal thoughts: embodiment and moving image culture.* University of California Press, Berkeley, USA.

Sonnenschein, D., 2001. *Sound Design: The Expressive Power of Music, Voice and Sound Effects in Cinema*, Michael Wiese Productions.

Tatler, B.W. et al., 2011. Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, 11(5), pp.5–5.

Thurlow, W.R. & Jack, C.E., 1973. Certain Determinants of the "Ventriloquism Effect." *Perceptual and motor skills*, pp.1171–1184.

Tseng, P.-H. et al., 2009. Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, 9(7), pp.4–4.

Van der Burg, E. et al., 2008. Pip and pop: Nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 34(5), pp.1053–1065.

Võ, M.L.H. et al., 2012. Do the eyes really have it? Dynamic allocation of attention when viewing moving faces. *Journal of Vision*, 12(13), pp.3–3.

Wallengren, A.K. & Strukelj, A., 2015. Film Music and Visual Attention: A Pilot Experiment using Eye-Tracking. *Music and the Moving Image*.

Winn, M., 2016. Rapid Release From Listening Effort Resulting From Semantic Context, and Effects of Spectral Degradation and Cochlear Implants. *Trends in Hearing*, 20(0), pp.1–17.