Audio-Visual Integration in Stereoscopic 3D

Lesley Deas, Laurie M. Wilcox
Dept. of Psychology
York University
4700 Keele Street
Toronto ON M3J 1P3*

Ali Kazimi
Department of Film
York University
4700 Keele Street
Toronto ON M3J 1P3†

Robert S. Allison
Dept. of Computer Science and Engineering
York University
4700 Keele Street
Toronto ON M3J 1P3‡

Abstract

The perception of synchronous, intelligible, speech is fundamental to a high-quality modern cinema experience. Surprisingly, this issue has remained relatively unexplored in stereoscopic 3D (S3D) media, despite its increasing popularity. Instead, visual parameters have been the primary focus of concern for those who create, and those who study the impact of, S3D content. In the work presented here we ask if ability to integrate audio and visual information is influenced by adding the third dimension to film. We also investigate the effects of known visual parameters (horizontal and vertical parallax), on audio-visual integration. To this end, we use an illusion of speech processing known as the McGurk effect as an objective measure of multi-modal integration. In the classic (2D) version of this phenomenon, discrepant auditory (/ba/) and visual (/ga/) information typically results in the perception of a unique 'fusion' syllable (e.g. /da/). We extended this paradigm to measure the McGurk effect in a small theatre. We varied the horizontal (IA: 0, 6, 12, 18, 24 mm) and vertical (0°, 0.5°, 0.75°, 1°) parallax from trial-to-trial and asked observers to report their percept of the phoneme. Our results show a consistently high proportion of the expected fusion responses, with no effect of horizontal or vertical offsets. These data are the first to show that the McGurk effect extends to stereoscopic stimuli and is not a phenomenon isolated to 2D media perception. Furthermore, the results show that audiences can tolerate a high level of both horizontal and vertical disparity and maintain veridical speech perception. We consider these results in terms of current stereoscopic filmmaking recommendations and practices.

CR Categories: H.5.2 [Multimedia Information Systems]: Audio input/output— [H.5.2]: Multimedia Information Systems—Video I.3.3 [Computer Graphics]: Three-Dimensional Graphics and Realism—Display Algorithms I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Human Factors;

Keywords: McGurk Effect, Stereoscopic 3D, Audio-visual Integration, Vertical Parallax, 3D Film

1 Introduction

Since the tremendous commercial success of Avatar in 2009, stereoscopic 3D (S3D) cinema as a medium has become increasingly popular. Hollywood has seen filmmakers from all genres adopt S3D as an additional element to create a truly immersive film experience. Digital technology and conversion techniques have also promoted the redistribution, and in some instances the conversion, of classic films (e.g. Jurassic Park) for viewing by new audiences. Furthermore, advances in display technology, particularly in home television and theatre options have made S3D content more widely accessible. However, the degree to which consumers will continue to embrace S3D media will largely depend on the quality of their viewing experience, which is determined by a range of factors, some content based, others display based.

Not surprisingly, research on viewer experience in response to S3D content has focused on visual factors with particular emphasis on viewer comfort. A range of factors that contribute to viewer discomfort have been well documented, and include cross-talk [Konrad 2000; Pastoor 1995; Kooi and Toet 2004], keystone/geometrical distortions [Woods et al. 1993; Stelmach et al. 2003], and conflict between accommodation and vergence [Hiruma and Fukuda 1993; Inoue and Ohzu 1997]. Of relevance to this study is the finding that excessive parallax between the left and right images are a consistent source of discomfort and fatigue [Woods et al. 1993; Wopking 1995; Yano et al. 2004; Speranza et al. 2006]. Such discomfort may result from the inability to fuse large disparities—either horizontal or vertical—into a single percept.

Although horizontal parallax is fundamental to stereoscopy, there are upper limits beyond which viewers report discomfort and degraded depth percepts. In static images, the general consensus is that disparity values up to 60-70 arcmin represent a 'comfort zone' [Wopking 1995]. This range is not absolute, as the fusional limit for binocular single vision depends on stimulus factors such as size, orientation, exposure duration, blur and other spatiotemporal properties (see [Howard and Rogers 2012]). In dynamic scenes, converging evidence suggests horizontal disparity magnitude and motion-in-depth may interact to determine comfort [Yano et al. 2004; Speranza et al. 2006; Nojiri et al. 2006]; that is, comfort levels may be determined by changing disparity. The specific correlation is unclear, however Speranza et al. [2006] suggest that controlling for the rate of change in disparity magnitude over time—rather than the magnitude of disparity—is critical to viewer comfort.

Whilst horizontal parallax between images is necessary for stereopsis, vertical parallax is an unwanted artifact that can result from either misalignment of cameras or projectors, or from the use of toed-in camera configurations [Allison 2007]. The fusion range for vertical disparities is known to be smaller than that for horizontal disparities [Ogle 1950]; however, the human eye can partially compensate for vertical disparities using vertical vergence eye movements. This mechanism can correct whole-field disparities of up to 1.5° for a central isolated target [Howard et al. 1997], and integrate disparities over a fairly large area (20° diameter) [Howard et al. 2000; Stevenson et al. 1997]. Beyond this limit, disparities exceed the fusional range and stereopsis is degraded [Stevenson et al. 1997]. As for horizontal disparities, the upper limit for fusion of vertically disparate images is variable, and depends on several factors including stimulus size and position [Howard et al. 2000], size of the display, the global pattern of vertical disparity and the presence of reference surfaces [Allison et al. 2000]. Under some

*e-mail: {ldeas, lwilcox}@yorku.ca
†e-mail: akazimi@yorku.ca
‡e-mail: allison@cse.yorku.ca

SAP 2013, August 22 – 23, 2013, Dublin, Ireland.
2013 Copyright held by the Owner/Author. Publication rights licensed to ACM.
ACM 978-1-4503-2262-1/13/08 $15.00

conditions, we are able to tolerate quite high degrees of vertical misalignment of the two eyes' images. For example, Speranza and Wilcox [2002] assessed comfort while participants viewed a 30 min IMAX film with added whole-field vertical disparity up to approximately 1° simulating a substantial projector misalignment. Reports of discomfort were found to increase only marginally with increasing vertical disparity, leading to the conclusion that whole-field vertical misalignment is not a major contributor to discomfort associated with viewing S3D film. However, they acknowledged that their results may only apply to very wide fields of view (as is the case with IMAX screens). Larger displays are known to increase the vertical vergence response and the vertical fusion range [Allison 2007], which may make vertical disparities more tolerable. Indeed, Allison et al. [2000] showed that while we are able to fuse vertically offset simple stimuli on smaller displays (i.e. a computer monitor), this can be impeded by the presence of competing peripheral stimuli.

Another potentially important, but as yet unexplored, consequence of parallax in the two eyes' images in stereoscopic 3D media is the disruption of audio-visual integration. Our ability to rapidly combine and interpret visual and auditory signals is one that we often take for granted in the natural environment where physics constrains their temporal and spatial relationship. In film this correspondence is also required, particularly in scenes with synchronized sound, where actors speak either to one another, themselves or to the audience. It is notable that the lips are small features, relative to the size of the head or body. Given that the range of disparities where binocular fusion can be obtained (both horizontally and vertically) is smaller for high-frequency, narrow stimuli than for larger, coarser features; small fine features such as the lips will be among the first to appear doubled or diplopic in a vertically misaligned S3D film. Thus, another potential consequence of vertical disparity in S3D film may be a disruption of speech perception.

In this paper, we will assess whether our ability to integrate audio-visual signals is influenced by viewing content in S3D. More specifically, we will use a well-documented phenomenon known as the McGurk Effect [Mcgurk and Macdonald 1976] to evaluate the audio-visual integration required for speech perception, while manipulating horizontal and vertical parallax.

The McGurk effect shows that when auditory and visual phonetic information are discrepant, the information can be combined into a new percept that was not originally presented to either modality [Mcgurk and Macdonald 1976]. For example, dubbing the auditory token /ba/ on to a visual articulation /ga/ typically results in the perception of a unique token, such as /da/ or /tha/. Existing research shows that this experience of a new utterance, referred to as fusion, occurs on upwards of 90% of trials [Mcgurk and Macdonald 1976; Rosenblum and Yakel 2001; Green et al. 1991; Walker et al. 1995]. The strength of this illusion and its reliance on integration of visual and auditory input makes it an ideal tool to assess the status of audio-visual integration. That is, when audio-visual integration occurs in a typical manner, we should replicate the frequency of 'fusion' responses reported in the literature. If multi-modal integration is disrupted by manipulating vertical or horizontal disparity then as these variables increase, the fusion rate (i.e. the McGurk illusion) should decline.

2 Methods

Viewers were asked to report the token they perceived in S3D trials that contained trial-to-trial variations in horizontal and whole-field vertical disparities.

2.1 Viewers

Forty undergraduate students (20 female, mean age = 22.1 years (sd: 5.3); 20 male, mean age = 21.4 (3.8)) participated in this experiment for course credit. Observers wore their prescribed optical correction during testing. All viewers had normal stereopsis, assessed by a pre-experiment test: viewers were asked to copy the three letters presented in a random dot stereogram with crossed disparity on the projection screen. Six disparities were tested: the smallest on-screen converted to 0.03°at the closest viewing distance (largest 0.26°). Participant data was only included if they were able to correctly identify the three letters with at least 0.05°of disparity.

2.2 Apparatus

The experiment was conducted in a screening room with multiple viewers at one time. The experiment was run over three sessions with 13, 13, and 14 viewers in the respective sessions. Images were projected onto a white-surface cinema screen (2.97x1.65 m) using a Christie HD6K-M system (resolution 1920x1080) via Stereoscopic Player [Wimmer 2005]. Viewers wore LC shutter glasses (Xpand 3D Cinema glasses) that alternately blocked the left and right eye view at 120 Hz in synchrony with the display of the left and right images, respectively. This provided a time-multiplexed stereoscopic display and the percept of the content in S3D. The auditory playback level was set to a moderate level. Viewers were positioned in three rows, at viewing distances of 3.4, 4.6 and 5.8 m from the screen. At the closest distance, the screen subtended 47.2°x 27.3°of visual angle and one pixel subtended 0.03°.

2.3 Stimuli

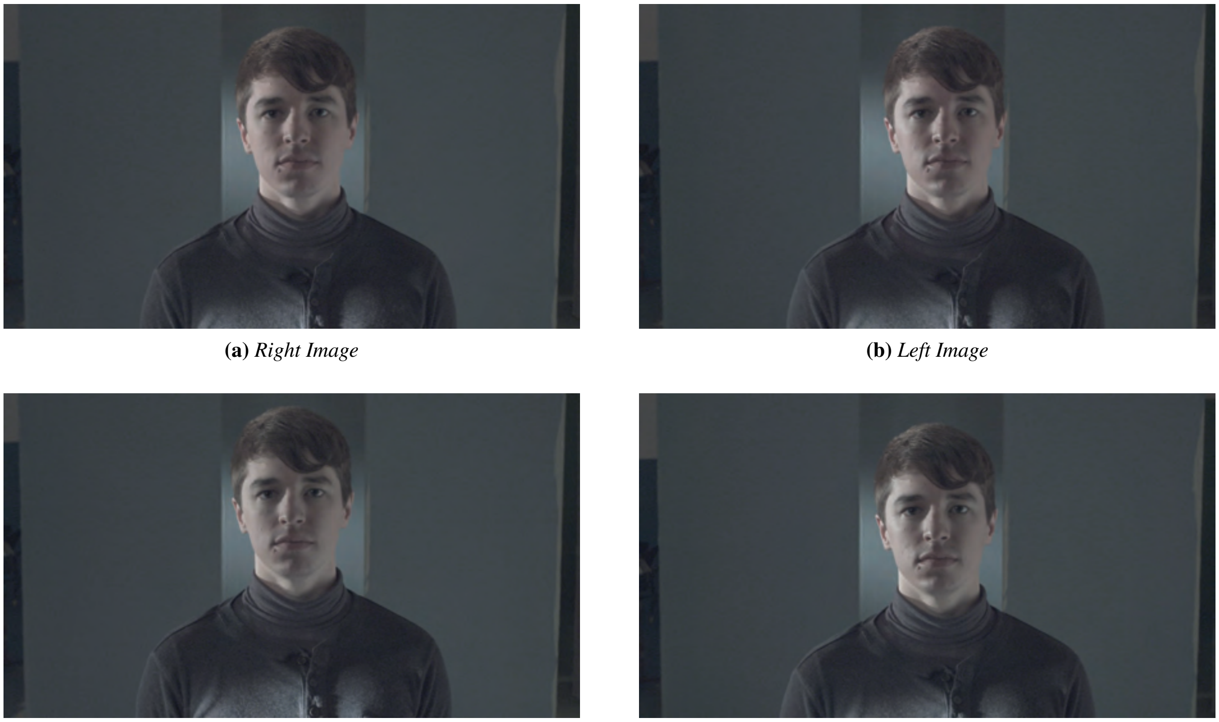
2.3.1 Stimulus materials/capture

A male native English speaker was filmed using a SI-2K digital cinema camera pair with 12mm Zeiss lenses at a distance of 5ft. Synchronized audio was recorded by a Tram TR-50 lavalier microphone, clipped to the actor's chest. The actor was lit with a 1 kW tungsten fresnel fixture, key light from screen left. Another 1 kW tungsten fresnel fixture was directed at a white-board to screen right; the soft diffused light 'bouncing' off the board was used to fill in the shadows and decrease the contrast. A classic medium shot was composed of the actor's head and chest against a background of three architectural panels with a step in depth between them. The two side panels were lit from behind, such that the front surface of the centre panel would be illuminated by the resulting bounced light to a slightly higher level with the aim of enhancing the perception of depth. See Figure 1.

Addressing the camera, the actor articulated the syllables /ba/ and /ga/ multiple times, with the clearest token selected in post-production. Each token was recorded at one of five different inter-axial distances (lateral offsets between the axes of the camera): 0, 6, 12, 18, 24 mm. Ten stimulus clips were filmed in total (2 tokens x 5 interaxial distances). Data were encoded using a Cineform codec [GoPro 2013] and transferred to a computer to be edited.

2.3.2 AV alignment

Four audiovisual combinations were generated: two congruent (/ba/ba/ and /ga/ga/) and one incongruent (Visual/ga/Auditory/ba/). Note that we only produce one incongruent pair as this combination provides the stimulus for the McGurk effect. We also generated and tested the converse condition (Visual/ba/Auditory/ga/) but do not report these data here because this combination is not prone



(a) *Right Image*

(b) *Left Image*

(c) *Right Image, Vertical Offset*

(d) *Left Image, Vertical Offset*

Figure 1: *Stereoscopic image pair of a sample stimulus from the experiment. Upper pair shows a vertically aligned pair with IA of 18 mm. Lower pair shows 12 mm IA with a vertical misalignment. Arranged for cross-eyed fusion.*

IA	Horizontal Disparity		
	(degrees)		
(mm)	Front	Middle	Back
0	0.00	0.00	0.00
6	0.41	0.31	0.25
12	0.75	0.56	0.45
18	1.13	0.84	0.67
24	1.50	1.12	0.89

Table 1: *Degrees of horizontal disparity, converted from on-screen parallax for 5 interaxial conditions. Values are calculated for three viewing distances: 3.4m (front row), 4.6m (middle) and 5.8m (back) from the screen. The screen parallax of a point on the actor's face near the mouth was used to define the horizontal disparity.*

to the classical McGurk effect so we do not have predictions for the effects of audio-visual disruption on this condition.

For the congruent stimuli, no dubbing was required and a single token segment from each native recording was selected. These tokens were edited using a video editing software program (Adobe Premiere CS5 [Adobe 2013]) to create segments that were approximately 3 seconds long, ensuring that the visual articulation started and ended with the actor's mouth closed. Each started with a 1 second fade-in from grey and ended with 1 second fade-out to grey. To create the audiovisual incongruent stimuli, a single auditory /ga/ was selected to be dubbed on the contrasting visual token /ba/. Stimuli were synchronized to align the burst onset of the dubbed auditory token with the onset of the original auditory sound. The editing software allowed this alignment to be reproduced with approximately 1 ms accuracy across trials.

To assess the effect of horizontal parallax, we constructed fifteen stimuli by factorially combining the audio-visual combinations and the camera interaxial conditions (3 combinations x 5 interaxials). Table 1 shows how the parallax in terms of on-screen pixels converts to degrees of horizontal disparity at the distances corresponding to each row: viewers were seated at three different viewing distances, therefore the same clip produced a different range of disparity values. Note that horizontal disparities vary across the image, therefore we selected a common location for all clips (a point near the mouth) to compute the horizontal or vertical pixel offsets between the left and right images.

2.3.3 Stimulus preparation

Whole-field vertical offsets of 0°, 0.5°, 0.75°, and 1° were applied to each stimulus. For every image pair, half of the offset was applied to each image, with the left image shifted upward and the right downward. Nominal vertical disparities were converted to equivalent on-screen pixels based on the assumption of standard viewing conditions that produce a 36° viewing angle (THX recommendations [THX 2013]). This distance is a nominal distance approximately equivalent to our middle row in the theatre. Observers in the near and far rows will experience more and less disparity, respectively. Table 2 shows the vertical disparities corresponding to each offset for each row.

Each audio-visual token (utterance) was duplicated and repeated three times within a trial, with a brief grey frame separating each repetition. As a result each trial had a duration of 11 s. In total, there were 60 trials (3 token combinations x 5 IAs x 4 vertical offsets) which were randomized in a playlist for each experimental session.

Nominal Offset (degrees)	Actual Vertical Disparity (degrees)		
	Front	Middle	Back
0	0	0	0
0.5	0.73	0.54	0.43
0.75	1.1	0.81	0.64
1	1.46	1.08	0.86

Table 2: *Vertical disparity in degrees, converted from on-screen parallax for 4 vertical offset conditions. Values are calculated for three viewing distances: 3.4m (front row), 4.6m (middle) and 5.8m (back) from the screen.*

2.4 Procedure

Viewers were instructed to write down the token they perceived after completion of each trial (open response). It was emphasized that they must view the full clip before responding. To monitor viewer behaviour, one experimenter was positioned to the side of the seating area and another at the back of the room. After the blocks of stereoscopic trials, an additional two control blocks of audio and visual only trials were presented. Each of the 10 native clips (2 token x 5 IA) was played as either an audio only (participants were asked to close their eyes) or visual only (presenting the right image, therefore 2D).

3 Results

3.1 Exclusion criteria

Our method is novel in McGurk research as many viewers were assessed at one time, rather than on an individual basis. As such, a limitation of this method is that we cannot guarantee viewers were attending to the screen. We specifically used the McGurk illusion as a means to study speech perception because, by definition, it requires integration of both the visual and auditory signals. With this in mind, we applied very conservative criteria to viewers responses to identify viewers who may not have been attending to the screen. To be included, viewers must have reported > 90% correct on congruent trials, and in addition, must have shown >80% responses other than audio to incongruent trials. Based on this criteria, only 8 participants were excluded (3 female, mean age = 22 (3.0); 5 male, mean age = 24.6 (8.2)). After applying our exclusion criteria, the responses of 32 viewers were analyzed (17 female, mean age = 21.7 (5.4); 15 male, mean age = 20.0 (2.9)). It is important to note that these viewers responded to the auditory token on all sixty test trials. This uniform pattern of responses is readily explained by a viewer's failure to adhere to the experiment requirements, i.e. not attending to the screen. Although the eyes do not need to fixate directly on the mouth to integrate information, the mouth must be in view and the viewer's gaze should not deviate from the actor's mouth by more than 10–20° [Paré et al. 2003]. However, we acknowledge that other factors could explain why some viewers did not perceive the illusion. For example, less susceptibility has been shown for people who frequently watch dubbed movies as they have learned, to some extent, to ignore visual articulations [Boersma 2012]. The root causes of failures to experience the McGurk effect are potentially interesting. However, the fact that the goal of the study is not to explore the McGurk phenomenon per se, but to use it as a tool, justifies the application of these exclusion criteria.

Row	Fusion	Visual /ga/	Auditory /ba/	Other
Front	95.0	0.4	4.6	0
Middle	97.1	0.7	2.1	0
Back	98.5	0.4	1.1	0
Total (n=32)	96.6	0.5	3.0	0

Table 3: *Percentage of responses to incongruent stimuli /ga/ba/, categorized as either fusion (unique response), visual (/ga/), auditory (/ba/) or other (/bga/ or /gaba/).*

3.2 Single modality — audio and 2D visual only

These trials serve as a check on the intelligibility of the individual tokens without any influence of S3D parameters. For the audio only trials, both tokens /ba/ and /ga/ were correctly identified on 100% of trials. For the visual only trials, the visual articulation /ba/ was identified with 100% accuracy. Visual /ga/ produced average accuracy of 76.2% (range: 68-88.3%) across the five trials. This visual articulation is known to be mis-identified when lip-reading as the letter /g/ is pronounced at the back of the throat and does not form on the lips [Nitchie 1919]. Examples of incorrect responses included /ah/, /ha/, and /ka/.

3.3 Both modalities — S3D

3.3.1 Congruent trials

The mean percentages (and standard errors) of correct responses for each token are as follows: /ba/ba/ total correct = 94.5% (5.0) and /ga/ga/ 99.8% (0.2). A correct response is defined as whenever a viewer responded with a token that was identical to the congruent information. Note that it is not possible with congruent trials to determine if viewers are integrating both the auditory and visual information or just attending to one modality. Thus, the incongruent trials (the McGurk effect) are crucial to the assessment of our hypothesis.

3.3.2 Incongruent trials

Responses were collated and tallied according to three categories: visual response (/ga/), auditory response (/ba/) and fusion response (unique percept). Previous research [Mcgurk and Macdonald 1976; Rosenblum and Yakel 2001; Green et al. 1991; Walker et al. 1995] suggests we should expect to see a high proportion (> 90%) of fusion responses.

The average percentages of responses for each row are shown in Table 3. Note that there were no responses that would be classified as 'other'. Since disparity scales with viewing distance, we looked for a difference in the proportion of responses across rows (i.e. viewing distance). Although suggestive of a trend, an ANOVA revealed that there was no effect of viewing distance on the proportion of fusion responses in this experimental set-up (F(2,57)=2.68, p=0.08). We therefore collapsed over rows for subsequent analyses.

To determine if our incongruent stimuli produced a McGurk effect, an ANOVA was used to test for effects of presentation condition (congruent /ba/ba/; congruent /ga/ga/; audiovisual incongruent) on number of 'incorrect' responses. Here, we define an incorrect response as one that did not match either the visual or auditory token: note that in the incongruent trials, this would constitute an expected-fusion McGurk response. This test revealed a significant effect for the presentation condition (F(1.04, 19.67) = 308.32, p<.001). Post-hoc tests (Bonferroni correction) revealed that the AV incongruent condition was significantly different from each individual congruent condition at the p < 0.001 level. Thus, these

IA	VO (deg)			
(mm)	0	0.5	0.75	1
0	100	100	96.9	93.8
6	93.8	87.5	93.8	90.6
12	93.8	100	100	90.6
18	100	100	93.8	100
24	87.5	96.9	96.9	96.9

Table 4: *Percentage of expected-fusion responses to incongruent stimuli /ga/ba/. Each cell represents a unique combination of horizontal and vertical offsets.*

results reveal a significant McGurk effect with the incongruent pairing of /ga/ and /ba/.

3.4 S3D variables

Table 4 shows the percentage of fusion responses recorded for each combination of IA and vertical offset for the McGurk fusion condition. Each trial employed a unique combination of a horizontal and vertical disparity (including a trial with no disparity). Most responses were the expected-fusion responses with all of the remaining responses indicating a single modality; no 'other' responses were recorded. The effect of each type of parallax can be assessed by considering each variable in isolation.

3.4.1 Effect of horizontal offset

Figure 2 shows the percentage of fusion responses for the McGurk combination grouped by horizontal disparity (i.e. collapsed over vertical offset). Expected-fusion responses were greater than 91% for all conditions. The highest percentage of fusion responses was recorded at IA 0 mm, and this is matched at the larger offset of 18mm. A small proportion of responses were made to a single modality, with auditory responses more common than visual.

3.4.2 Effect of vertical offset

Figure 3 shows the percentage of responses to the McGurk fusion stimulus grouped by vertical offset (i.e. collapsed over interaxial distance). Again, a high proportion, 94% or greater, of expected-fusion responses is seen for all conditions, and there is a bias toward auditory responses when reporting a single modality.

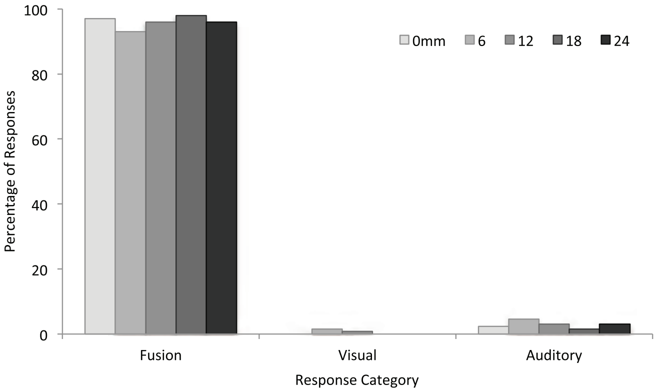


Figure 2: *Percentage of responses to incongruent stimuli /ga/ba/, grouped by interaxial distance.*

3.5 Statistics

We used a log-linear analyses to examine the relationship between our variables and response frequencies. Analysis was established with three levels (2x4x5): response type[1], vertical offset and interaxial distance. The generated model retained only a main effect of response type (likelihood ratio, $\chi^2(38) = 0, p = 1$), suggesting that there was a significant difference in the frequency of 'fusion' responses compared to 'other' responses ($\chi^2(1) = 657.23, p < 0.001$). There was no main effect of interaxial distance or vertical offset, nor an interaction with response type. In other words, the analysis seems to reveal that there is no effect of S3D variables on the proportion of fusion responses on incongruent trials. Taken together, our results show that the McGurk effect was consistently and reliably produced in stereoscopic 3D content. The proportion of expected-fusion responses was not affected by horizontal disparities produced by IAs of up to 24mm and with vertical offsets up to 1°. Furthermore, combinations of these offsets had no effect on the proportion of expected-fusion responses, when compared to trials with no offset (0 mm, 0°).

4 Discussion

In this study, we used the McGurk effect as an empirical probe to assess audio-visual integration in S3D film. Our results are the first to show that the McGurk effect is experienced in S3D content: stereoscopic visual information was integrated with audio output to generate a consistently high proportion of fusion responses. Furthermore, we show that perception of the illusion is not affected by large horizontal or vertical offsets, or a combination of such offsets. Taken together, the results of the present study show that excessive parallax does not interfere with speech processing, showing that audio-visual integration is essentially the same in S3D as in 2D content. These findings suggest that the McGurk effect reflects natural processes that occur in everyday audio-visual perception rather than a curiousity of 2D media perception.

4.1 Interaction of Offsets

The data obtained in this study are important for they reveal that on-screen horizontal and whole-field vertical disparities of up to 1.5° (at the closest viewing distance) do not interfere with the ability to process visual information and integrate it with audio information in order to perceive speech. Furthermore, we show that even when large stereoscopic effects are combined with vertical misalignment, there is similar resilience of audio-visual integration of speech signals.

4.2 Reinforcing Recommendations

Despite finding considerable tolerance of the McGurk effect to vertical misalignment, we maintain that vertical offsets should always be avoided in S3D film content to create an easily fused image. Whilst some misalignment in may occur and go unnoticed by viewers, large offsets like those used here would create a noticeably uncomfortable image and typically produce a double image (subjectively, the vertical offsets employed here straddled the fusion limit). It would be valuable to assess the relationship between diplopia points for both vertical and horizontal disparity and the McGurk effect to determine if images do need to be fused to perceive speech accurately, particularly under normal viewing conditions where S3D content is watched for much longer periods of

[1] One assumption of the analysis is that all cells have an expected frequency greater than 1. Cells for auditory and visual responses did not meet this assumption and so were collapsed into one category called 'other'.

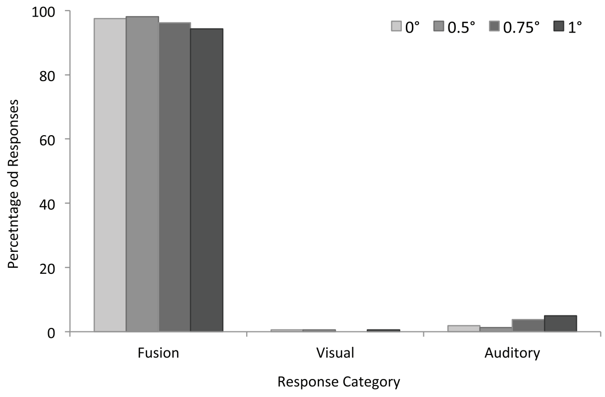


Figure 3: *Percentage of responses to incongruent stimuli /ga/ba/, grouped by vertical offset.*

time. Nevertheless, our results show no effect on the proportion of fusion responses, suggesting that audio and visual information can still be integrated at large offsets and result in accurate speech perception. However, this does not negate the desirability of minimizing vertical misalignment for reasons of comfort, fusion and good stereopsis under extended viewing conditions.

4.3 Reconsidering Recommendations

Horizontal disparity limits of up to 60-70 arcmin have become the benchmark recommendation for comfortable stereoscopic viewing [Wopking 1995; Speranza et al. 2006]. These guidelines were based on the identification of excessive horizontal parallax being a main contributor to discomfort. In line with the recommendations of Sky Television [Cassy 2013] (the first television broadcaster to issue guidelines, followed by many stereographers), we suggest that this 'limit' should be considered a 'general guideline' in filmmaking, rather than a hard constraint. Previous studies have identified causal factors for discomfort for fused images that employ disparities within the limit, for example, unnatural blur from cross-talk [Konrad 2000; Pastoor 1995; Kooi and Toet 2004] or excessive demand on the accommodation-vergence link (e.g. from fast motion in depth). Therefore, disparity should not be considered in isolation when making recommendations for comfort limits. Indeed, some studies are now addressing possible interactions between problematic factors [Yano et al. 2004; Speranza et al. 2006; Nojiri et al. 2006]. In this viewing arrangement, it is possible that larger disparities may disrupt audio-visual integration, however it would be difficult to isolate the influence that discomfort would have on perception. As such, we maintain our study shows that 'excessive horizontal parallax' does not affect an objective measurement of audio-visual perception. The largest disparity employed in this study equated to 90 arcmin (1.5° at the closest viewing distance), an offset that is well beyond the traditional limit; however, we found no effect of horizontal parallax in any condition. Therefore, the causal factor of 'excessive horizontal parallax' may be relative to the type of measurement used in assessment. It should be noted that the guidelines recommended for stereography do not reflect limits of the real world and we routinely experience disparities larger than these limits in real life. The significant disparity tolerance we found for the McGurk effect may reflect an ecological advantage of processing speech of a conversant that is not fixated, as would be common in many social encounters. In any case it reinforces the concept that the 'range' of stereoscopic processing depends on the visual task and is not a single hard number.

In sum, it is difficult to establish a fixed limit to the disparity that can applied to S3D film content. One main reason is that traditional recommendations are based on assessment of visual comfort and not objective measures of perceptual limits or abilities. Subjective assessment of comfort may be described as a passive process, identifying problematic symptoms such as eye-strain, dizziness and headaches. Such studies can provide valuable insight into potential physiological problems that might arise from extended viewing. On the other hand, objective measurements of perception require active interpretation of what is being seen on the screen, and may give more insight into what the visual system is actually processing. In our lab, we have applied objective measures to real-world stereoscopic images [Tsirlin et al. 2012], showing that the amount of depth perceived in a display is highly dependent on the quality of the stereoscopic image, which has implications for the appearance of realistic scenes. It would be interesting to use our experimental set-up to integrate both types of measurement, by assessing the influence of both subjective discomfort and our objective measure of perception.

4.4 Integrating Audio

A novel feature of this study is the consideration of the auditory modality, and its integration with stereoscopic information. We suggest that assessment of this relationship will allow a more complete understanding of viewer experience with S3D content. It is well known that auditory information can integrate with visual information and impact perception; psychophysical evidence has shown that an auditory tone played along with a dynamic object can influence the perceived path of that object (e.g. [Remijn et al. 2004; Sakurai and Grove 2009; Kang and Blake 2005]). But this is the first study to consider natural speech perception in stereoscopic film content. When presented with a synchronized dialogue scene in film, studies have shown that gaze is drawn specifically to the talker's mouth [Võ et al. 2012; Buchan et al. 2008], a reflex that is thought to help isolate and integrate the relevant information for the task at hand, i.e. to help the viewer perceive speech. Such visual localization may be necessary to perceive the McGurk illusion, as it requires integration of the visually discrepant information (although some studies suggest that gaze can deviate from the mouth by as much as 10—20°[Paré et al. 2003]). It is therefore reasonable to assume that observers attended to the speaker's mouth in the present study, and this ability was maintained when viewing stereoscopic content and not inhibited by large offsets. Indeed, we can speculate that the localized allocation of attention may have been of benefit in trials with the largest disparities (where fusion was not possible) as only the relevant information was integrated for speech perception, and the visual system was not overwhelmed. Future study will consider how the current findings extend to more dynamic scenes, where multiple actors are filmed from different viewing angles and articulate real dialogue.

5 Conclusion

Our results suggest that filmmakers can be bold in their choice of horizontal offsets in dialogue scenes, without concern that the audience's speech perception will not be degraded. In addition, while care should always be taken to avoid vertical offsets, there is a large range within which speech will not be affected by vertical misalignment. However, this range is extreme and would likely be uncomfortable to view in film for extended periods of time. We purposefully do not recommend limits for parallax in a dialogue scene as it remains to be determined how other S3D parameters interact with parallax in our stimuli.

Acknowledgements

The support of contracts from the Ontario Centres of Excellence and the Ontario Media Development Corporation is gratefully acknowledged. The authors would like to thank Karim Benzeroual, Arthur Lugtigheid, Parmis Goudarzi, Tom Hall, Renford Phillips, Joanna Popinska and Jacob Niedzwiecki.

References

ADOBE, 2013. Adobe Premiere Pro CS6, http://www.adobe.com/ca/products/premiere.html.

ALLISON, R., HOWARD, I., AND FANG, X. 2000. Depth selectivity of vertical fusional mechanisms. *Vision Research 40*, 21, 2985–2998.

ALLISON, R. 2007. Analysis of the influence of vertical disparities arising in toed-in stereoscopic cameras. *Journal of Imaging Science 51*, 4, 317–327.

BOERSMA, P. 2012. A constraint-based explanation of the mcgurk effect. In *Phonological Explorations: Empirical, Theoretical and Diachronic Issues*, B. Botma and R. Noske, Eds., vol. 548. Walter de Gruyter, 299–312.

BUCHAN, J. N., PARÉ, M., AND MUNHALL, K. G. 2008. The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception. *Brain Research 1242* (Nov.), 162–171.

CASSY, J., 2013. Sky 3D commissioning & production, http://www.sky.com/shop/tv/3d/producing3d/.

GOPRO, 2013. Create & edit. GoPro CineForm studio software, http://gopro.com/3d-cineform-studio-software-download/.

GREEN, K. P., KUHL, P. K., MELTZOFF, A. N., AND STEVENS, E. B. 1991. Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics 50*, 6 (Nov.), 524–536.

HIRUMA, N., AND FUKUDA, T. 1993. Accommodation response to binocular stereoscopic TV images and their viewing conditions. *SMPTE Motion Imaging Journal 102*, 12 (Dec.), 1137–1140.

HOWARD, I. P., AND ROGERS, B. J. 2012. *Perceiving in Depth*, vol. 2 Stereoscopic Vision. Oxford University Press, Feb.

HOWARD, I. P., ALLISON, R., AND ZACHER, J. E. 1997. The dynamics of vertical vergence. *Exp Brain Res 116*, 1, 153–9.

HOWARD, I. P., FANG, X. P., ALLISON, R., AND ZACHER, J. E. 2000. Effects of stimulus size and eccentricity on horizontal and vertical vergence. *Experimental Brain Research 130*, 2, 124–132.

INOUE, T., AND OHZU, H. 1997. Accommodative responses to stereoscopic three-dimensional display. *Applied Optics 36*, 19 (July), 4509–4515.

KANG, M.-S., AND BLAKE, R. 2005. Perceptual synergy between seeing and hearing revealed during binocular rivalry. *Psichologija 32*, 7–15.

KONRAD, J. 2000. Cancellation of image crosstalk in time-sequential displays of stereoscopic video. *IEEE Transactions on Image Processing 9*, 5, 897–908.

KOOI, F. L., AND TOET, A. 2004. Visual comfort of binocular and 3d displays. *Displays 25*, 2-3, 99–108.

MCGURK, H., AND MACDONALD, J. 1976. Hearing lips and seeing voices. *Nature 264*, 5588 (Dec.), 746–748.

NITCHIE, E. B. 1919. *Lip-reading: Principles and Practise : a Hand-book for Teachers and for Self Instruction.* Taylor & Francis.

NOJIRI, Y., YAMANOUE, H., IDE, S., YANO, S., AND OKANA, F. 2006. Parallax distribution and visual comfort on stereoscopic hdtv. In *Proc. IBC*, IBC, 373–380.

OGLE, K. N. 1950. *Researches in binocular vision*. Saunders.

PARÉ, M., RICHLER, R. C., HOVE, M. T., AND MUNHALL, K. G. 2003. Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect. *Perception & Psychophysics 65*, 4 (May), 553–567.

PASTOOR, S. 1995. Human factors of 3d imaging: Results of recent research at heinrich-hertz-institut berlin. *International Display Workshop 3*, 69–72.

REMIJN, G. B., ITO, H., AND NAKAJIMA, Y. 2004. Audiovisual integration: An investigation of the &lsquo;streaming-bouncing&rsquo; phenomenon. *Journal of Physiological Anthropology and Applied Human Science 23*, 6, 243–247.

ROSENBLUM, L. D., AND YAKEL, D. A. 2001. The McGurk effect from single and mixed speaker stimuli. *Acoustics Research Letters Online 2*, 2, 67–72.

SAKURAI, K., AND GROVE, P. M. 2009. Multisensory integration of a sound with stereo 3-d visual events. In *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2009*, IEEE, IEEE, 1–4.

SPERANZA, F., AND WILCOX, L. M. 2002. Viewing stereoscopic images comfortably: the effects of whole-field vertical disparity. *Stereoscopic Displays and Virtual Reality Systems IX,Proc. SPIE 4660* (May), 18–25.

SPERANZA, F., TAM, W. J., RENAUD, R., AND HUR, N. 2006. Effect of disparity and motion on visual comfort of stereoscopic images. *Stereoscopic Displays and Virtual Reality Systems XIII, Proc. SPIE 6055* (Feb.), 60550B–60550B.

STELMACH, L. B., TAM, W. J., SPERANZA, F., RENAUD, R., AND MARTIN, T. 2003. Improving the visual comfort of stereoscopic images. *Stereoscopic Displays and Virtual Reality Systems X, Proc. SPIE 5006* (May), 269–282.

STEVENSON, S., LOTT, L., AND YANG, J. 1997. The influence of subject instruction on horizontal and vertical vergence tracking. *Vision Research 37*, 20 (Oct.), 2891–2898.

THX, 2013. THX Certified Cinema screen placement, http://www.thx.com/professional/cinema-certification/thx-certified-cinema-screen-placement/.

TSIRLIN, I., ALLISON, R., AND WILCOX, L. 2012. Crosstalk reduces the amount of depth seen in 3D images of natural scenes. *Electronic Imaging: Stereoscopic Displays and Applications, Proc. SPIE 8288*, 82880W–82880W–9.

VÕ, M. L.-H., SMITH, T. J., MITAL, P. K., AND HENDERSON, J. M. 2012. Do the eyes really have it? dynamic allocation of attention when viewing moving faces. *Journal of Vision 12*, 13 (Dec.), 3.1–3.13.

WALKER, S., BRUCE, V., AND O'MALLEY, C. 1995. Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Perception & Psychophysics 57*, 8 (Nov.), 1124–1133.

WIMMER, P. 2005. Stereoscopic player and stereoscopic multiplexer: a computer-based system for stereoscopic video playback and recording. *Stereoscopic Displays and Virtual Reality Systems XII, Proc. of SPIE 5664A*, 400–411.

WOODS, A. J., DOCHERTY, T., AND KOCH, R. 1993. Image distortions in stereoscopic video systems. *Stereoscopic Displays and Applications IV, Proc. SPIE 1915* (Sept.), 36–48.

WOPKING, M. 1995. Viewing comfort with stereoscopic pictures: An experimental study on the subjective effects of disparity magnitude and depth of focus. *Journal of the Society for Information Display 3*, 3 (Dec.), 101–103.

YANO, S., EMOTO, M., AND MITSUHASHI, T. 2004. Two factors in visual fatigue caused by stereoscopic HDTV images. *Displays 25*, 4 (Nov.), 141–150.