Cross-Modal Influence on Binocular Rivalry

Joshua M. Lewis
Department of Cognitive Science
University of California, San Diego
San Diego, CA 92093
josh@cogsci.ucsd.edu

Adam S. Fouse
Department of Cognitive Science
University of California, San Diego
San Diego, CA 92093
afouse@cogsci.ucsd.edu

Virginia R. de Sa
Department of Cognitive Science
University of California, San Diego
San Diego, CA 92093
desa@cogsci.ucsd.edu

Abstract

Binocular rivalry occurs when two distinct stimuli, one for each eye, are presented to corresponding retinal areas. Similar to other bistable phenomena such as Necker cubes, this overlap often causes one's conscious perception to alternate between a coherent perception of one stimulus, a coherent perception of the other and sometimes a mixture of the two. Previous studies have tried to identify where rivalry occurs, and what is actually being rivaled. Some studies have provided evidence for low-level effects on rivalry, lending support to the idea that rivalry is between monocular visual streams. Other studies have provided evidence for higher-level effects on rivalry, supporting the idea that rivalry is between opposing patterns. While this debate has largely been passed on in favor of a hybrid theory of rivalry that includes effects at several levels, questions still remain about specific higher-level effects. In the present study, we look at the effect of a congruent auditory stimulus on perception of rival videos of speaking people. We find that auditory stimuli can have an effect on rivalry, indicating that cross-modal processes such as speech to lip matching or voice to face matching are among the high-level factors impacting rivalry.

Keywords: binocular rivalry; patchwork rivalry; stimulus rivalry; cross-modal; multi-modal; psychophysics.

In this paper, we investigate the role of cross modal interaction between audition and vision in determining stimulus dominance in a binocular rivalry paradigm. Binocular rivalry occurs when two distinct stimuli are presented to separate eyes, so that each eye only sees one stimulus but they overlap in one's visual field. Similar to other bistable phenomena such as Necker cubes, this overlap often causes one's conscious perception to alternate every few seconds between a coherent perception of one stimulus and a coherent perception of the other.

Historically, researchers have debated where in the visual processing stream one stimulus becomes dominant over the other and rises to conscious perception [1]. Evidence of ocular suppression has been found very early in the visual processing stream, at the Lateral Geniculate Nucleus and in V1 [2, 3, 4]. This finding supports the idea that rivalry is between monocular visual streams. On the other hand, high-level properties of the stimuli, such as visual coherence [5] and "natural" amplitude spectra [6], have been shown to affect rivalry dominance duration and strength, indicating that rivalry may be between the perceived stimulus rather than the monocular pathway. Further support for the stimulus being the object of rivalry comes from sudies that rely on interocular grouping during rivalry, in which cohesive stimuli can be perceived from parts that are divided between the eyes [7, 8]. These findings also indicate that areas of the brain further along in the visual processing stream are likely play a significant role in the phenomenon. Recent evidence from neuroimaging studies suggests that a complete answer for rivalry likely involves a hybrid of the two theories, involving both high-level and low-level visual processing systems [1].

While controversy over whether rivalry is controlled from low-level or high-level processing has largely been supplanted by an acknowledgment of the role of multiple levels of processing, questions remain about specific roles. Studies such as [4] have effectively answered the "how low?" question in the binocular rivalry literature, but the "how high?" question has remained more elusive. Attention is one potential candidate for a mechanism for bistable perception, as attention has been noted to have an effect on dominance of rival stimuli since Helmholtz [9] . Studies have shown that attention can control the rate of alternation between rival stimuli, but that selective attention showed stronger affects for ambiguous figures than for binocular rivalry [10]. However, the strength of effect on stimulus duration appears to depend on specific features of the stimuli, such as their complexity, and whether attention is focused on specific stimulus features [11, 12]. Attention seems to have the most effect on the initially dominant stimulus [13], and neurophysiological results indicate that attention can bias early processing in the visual stream [14].

Other higher-level effects on rivalry have been shown, in particular the importance of global coherence in pattern rivalry [15] and of biological motion in determining perception with both ambiguous monocular stimuli and rival binocular stimuli [16]. The biological motion result, in which upright walking figures were perceived more often than inverted figures, suggests a top-down effect where the global perception influences lower-level processing.

One interesting question is whether stimuli in another modality can influence rivalry. This question has been recently studied for bistable perception of visual and auditory objects. Hupe and colleagues looked at perception of concurrently perceived bistable (but not binocularly rivaling) visual and auditory stimuli [17], particularly the temporal proximity of auditory and visual perception shifts during perception of the parallel bistable stimuli.

In the present study, we look at the question of whether simultaneously perceived auditory input can influence perception during binocular rivalry. We hypothesize that if a subject views two rivaling videos while listening to a soundtrack appropriate to only one of the videos, the video appropriate to the soundtrack will dominate perception for a greater period of time than the other video. There is considerable evidence that normal speech recognition involves both audition and vision. For example, the McGurk effect has shown that different articulations, as seen in a video of moving lips, can affect the perception of identical-sounding syllables [18], and many studies have suggested that speech perception is inherently multimodal (see [19] for a review). Recent studies have demonstrated sensitivity in speech recognition for matching between the gender of auditory and visual sources, lending support for the idea that cross-modal integration in speech recognition involves top-down processes [20]. Cross-modal matching is robust, with the ability to match a voice to lips that are represented only by point light sources [21].

Cross-modal experience has also been shown to affect performance in a visual-auditory temporal frequency matching task [22] where subjects were better able to match auditory and visual temporal repetition rates when the match was in the context of an upright point-light walker than for scrambled and inverted point-light walkers (with the same local motions). Auditory input has also been shown to influence visual perception of the number of flashed stimuli [23, 24] and visual input (color) has been shown to influence olfactory perception [25]. All of these effects are automatic, just as one cannot ignore the visual input when looking at it in the McGurk Effect. We reason that well-associated auditory input could similarly bias visual perception in a binocular rivalry paradigm.

Methods

We performed our experiment using StereoGraphics CrystalEyes LCD shutter goggles attached to a PC running Matlab and Psychophysics Toolbox 3.0. Our CRT monitor was configured to display stimuli intended for the left and right eyes on alternating refreshes, which were coordinated with the eye alternation of the shutter goggles via an emitter attached to the GeForce QuadroFX quad-buffered graphics card in the machine. We recruited 18 subjects, all undergraduates, 7 males and 11 females, with normal or corrected normal vision and no colorblindness. One subject was removed from the study due to incorrect performance on catch trials (described below), and another was removed because they only ever pressed one of the two responses, resulting in a grand total of 16 subjects.

Our stimuli were composed of four videos. All four videos showed head shots of volunteers relating a story about a recent experience. Two videos were clips of a story told by a male actor, and the other two videos were clips of a story told by a female actor. In order to make the videos easier to distinguish, we created both red and green versions of each video by converting the videos to grayscale and using the grayscale values as brightness on the red or green color channel. On our equipment, the green versions of the videos were noticeably brighter and we therefore reduced the brightness of the green videos at presentation time to 65% of their original brightness in order to better match them with the red videos. Note that we used shutter goggles, not red/green glasses, so the colors have no impact on which eye sees which stimulus, they just serve to aid discrimination and help group the patterns. The audio tracks from each video were separated so that video and audio media could be presented independently of one another.

We used a stimulus rivalry paradigm where we presented one eye with the left half of the male video and the right half of the female video and the other eye with the right half of the male video and the left half of the female video (see Figure 1). Stimulus rivalry is believed to occur higher in the visual processing stream than ocular rivalry [1], so we use stimulus rivalry in order to give ourselves the best chance to discover a high-level cross-modal effect. Our early pilot trials with standard eye rivalry (female video to one eye and male video to the other) did not reveal a cross-modal effect.[1]

Subjects viewed 25 trials, consisting of one warmup trial (not reported) and 24 trials generated from the following counterbalanced conditions: four possible combinations of male and female videos by male in red and female in green or vice versa by male soundtrack, female soundtrack or no soundtrack. Subjects indicated which video they felt they mostly perceived by pressing keys on the keyboard for female or male. If subjects were unsure of their perception, we instructed them to press both keys or press neither key and we considered either of those responses as identical. Each trial lasted 86.6 seconds and was followed by a short (approximately 5 second) catch trial where only one of the two videos was displayed.

The experiment was run in a darkened room with the subjects seated in front of the computer described above. A fixation cross was present in the center of the video, and we instructed subjects to stay focused on the cross as much as possible. The video itself was 640 x 480 pixels in the center of a 1024 x 768 display, with a black background. The response keys were the Z and / keys on a standard qwerty keyboard. We affixed glow-in-the-dark labels to the response keys to help subjects reorient if their hands got lost in the dark.

We performed two primary analyses on our data. For the purposes of both, a congruent response is a response indi-

[1] Just before the due date for the camera ready copy of this paper we discovered a poster at Vision Sciences Society Annual Meeting presented May 10, 2010 that did find that auditory congruent stimuli could bias binocular rivalry of line drawings presented to each eye [26].

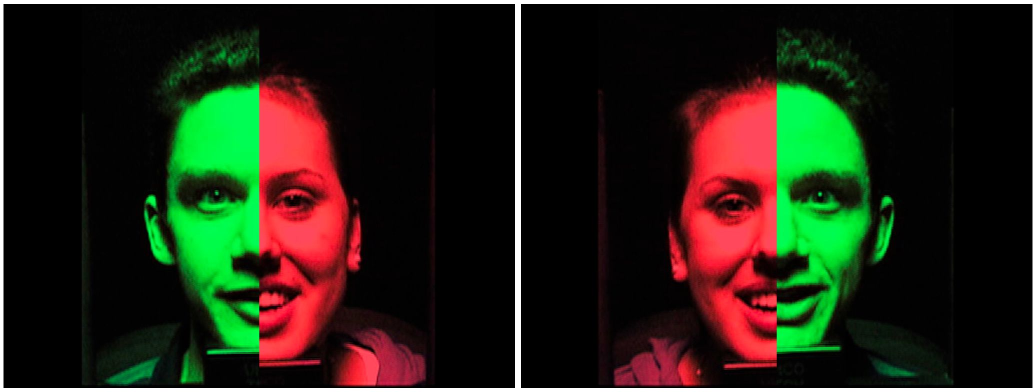

Figure 1: Sample stimulus from a single frame of the video. Left: left eye display. Right: right eye display.

cating dominance of the video associated with the currently playing audio and an incongruent response is a response indicating dominance of the video not associated with the audio. A neutral response is a response made during a trial with no audio. Our first analysis considered only the 16 trials that included sound. We subtracted the congruent dominance duration from the incongruent dominance duration and performed a positive one-tailed t-test comparison between the distribution over the subjects and a null distribution with zero mean (equal time spent on congruent and incongruent responses). Even though our trials were counterbalanced, we were concerned about two possible biases, a gender bias and a color bias. To correct for these biases we looked for a per-subject systematic bias in the no sound trials (previously unused) and subtracted the mean value of that bias from the appropriate responses in the trials with sound. We did this independently for both color and gender, resulting in three versions of this result: raw (uncorrected), corrected for color and corrected for gender. The equation for calculating this measure is as follows

$$\delta_{ci}(s) = \sum_{i \in S} \left(\sum_{j \in C_{is}} R_{ijs} - \sum_{k \in I_{is}} R_{iks} \right)$$

where $\delta_{ci}(s)$ is the difference between responses of congruent and incongruent visual percepts for subject s, S is the set of trials with sound, C_{is} is the set of congruent responses from subject s on trial i, I_{is} is the set of incongruent responses from subject s on trial i, and R_{ijs} is the duration of response j from subject s on trial i.

Second, we recorded the difference in reported dominance time of the male stimulus on trials with male sound versus trials with no sound. We did the same with female stimuli (dominance time of female stimulus on trials with female sound versus trials with no sound) and summed the results to see how much more often congruent stimuli were dominant versus their neutral counterparts in the no sound trials. We did the same comparison in the other direction to see what (dis)advantage incongruent stimuli had compared to neutral stimuli. Since these measures do not come at the expense of one another like those above (both congruent and incongruent are being compared to neutral, rather than to each other), the effect should be weaker but still an interesting basis for comparison. Also note that though there are twice as many trials with sound, we are only considering the congruent or incongruent responses from each trial. Since we consider both male and female responses from every no sound trial the comparison is even. Similar to the above, we performed a positive (for congruent, negative for incongruent) one-tailed t-test comparison between the distribution of this measure over the subjects and a null distribution with zero mean. The equation for calculating this measure (in the congruent case) is as follows

$$\delta_{csns}(s) = \sum_{i \in S} \left(\sum_{j \in C_{is}} R_{ijs} - \sum_{k \in N} \sum_{l \in FM_{ks}} R_{kls} \right)$$

where $\delta_{csns}(s)$ is the difference between congruent sound responses and corresponding no sound responses for subject s, N is the set of trials with no sound, and FM_{ks} is the set of female or male responses (as opposed to not sure) for subject s on trial k. Other terms are the same as described above and the incongruent case is a simple modification.

Results

Figure 2 shows the result (ordered by increasing effect) of our first analysis. In each of the raw, color corrected and gender corrected conditions we reject the null hypothesis that congruent stimuli are as likely to be dominant as incrongruent stimuli ($\mu = 60.25$ $\sigma = 102.37$ $p < .017, \mu = 44.42$ $\sigma = 83.59$ $p < .026, \mu = 54.33$ $\sigma = 96.44$ $p < .020$, respectively). Qualitatively the results don't change much after correction, as would be expected given the counterbalanced experimental design.

Figure 3 shows the result (in the same order as Figure 2) of our second analysis for the advantage of both congruent and incongruent stimuli compared to neutral stimuli. In the congruent case we do not obtain a significant effect, but there is a trend ($\mu = 24.26$ $\sigma = 69.21$ $p < .091$), as can be seen from

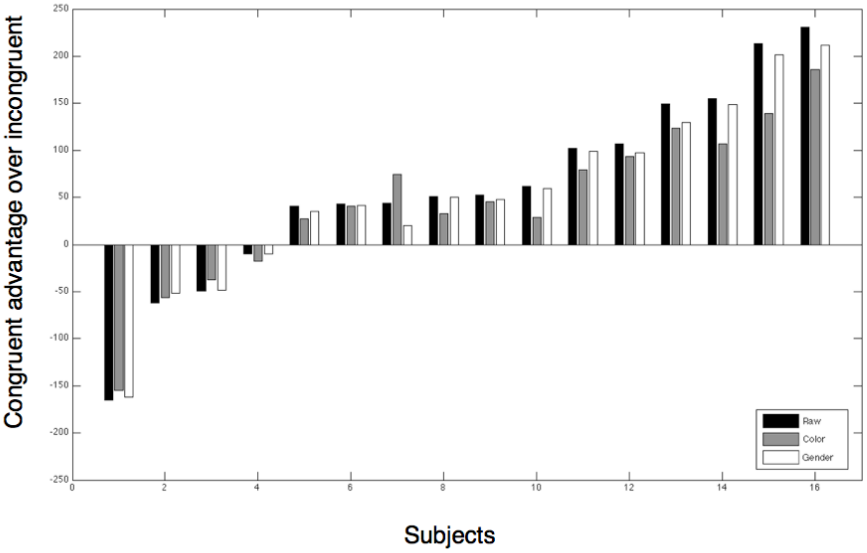

Figure 2: Difference in seconds between total congruent responses and incongruent responses, bucketed by subject in order of increasing effect. Bar color represents correction method.

the bar plot. The incongruent case does show a significant disadvantage compared to neutral ($\mu = -35.99$ $\sigma = 56.39$ $p < .011$). Notably, subject 4 changes character significantly in this analysis as compared to the previous. As a post-hoc investigation, we calculated the average absolute difference in total dominance duration between sound and no sound trials for each subject. Subject 4 had a per trial average dominance duration of 14.3 seconds less on the sound trials. Over all the other subjects the average absolute difference was 2.4 seconds with a max of 5.4. Clearly subject 4 responded much less to the trials with sound than was typical for the subject pool. If subject 4 is excluded from the hypothesis test on the second analysis there is a significant effect of congruent sound versus no sound ($\mu = 33.84$ $\sigma = 59.66$ $p < .023$) and the incongruent effect remains significant though weakened as one would expect ($\mu = -31.09$ $\sigma = 54.73$ $p < .023$). These results are very much in line with the first analysis.

Discussion

Our result represents an important step forward in mapping out the ways in which high-level processing can impact rivalry. Unlike previous high-level effects, such as global coherence and biological motion, this effect is not solely in the visual domain. Instead it is the result of matching a voice to a speaker, constituting the integration of both auditory and visual information.

It is not clear from this experiment whether the gender of the voice alone was enough to cause greater dominance of the congruent video, or whether voice to lip matching was responsible for the effect. Either of these effects would reveal an interesting cross-modal influence on binocular rivalry. Future studies that pair each speaker's video for one of their stories with the audio from the other could help illuminate the particular role of each aspect. [26] seems to show that semantically relevant sounds can bias the perception of eye rivaling static stimuli. However given that cross-modal voice to lip matching is so robust [21], we believe that the dominance effect is likely helped by the temporal coherence of voice and lips. An interesting question is whether subjects are aware of the matching even when the incongruent stimulus is dominant. This question could be addressed with an experiment that manipulates the temporal phase of the matched visual video during periods of nonperception. As soon as the subject indicates dominance of an incongruent stimulus one could switch or delay the audio track so as to put it out of sync with the congruent video. This might require less naturalistic stimuli (with pauses between words, for example) in order to execute without the audio sounding garbled. If subjects detect the lack of matching even when they're not consciously perceiving the congruent stimulus (e.g. by changing perceived dominance status), it would indicate the presence of a cross-modal blindsight for the voice to lip relationship.

Given that voice to lip matching is such a powerful effect, we were very concerned to sync the audio to the video precisely. It was not possible to do this perfectly given our experimental design, which required us to decouple the audio and video, though we came very close. We wonder whether the few subjects that showed an auditory congruence effect in the opposite direction were more temporally sensitive subjects (perhaps musically trained?) and more sensitive to slight offsets in sync and thereby biased against congruent stimuli at times when the sync is not quite right (a slightly offset audio/visual pair would be more anticorrelated than an unrelated audio/visual pair leading to a potential preference for perception of the unrelated video). An experiment where audio/video sync is manipulated across trials and subjects are

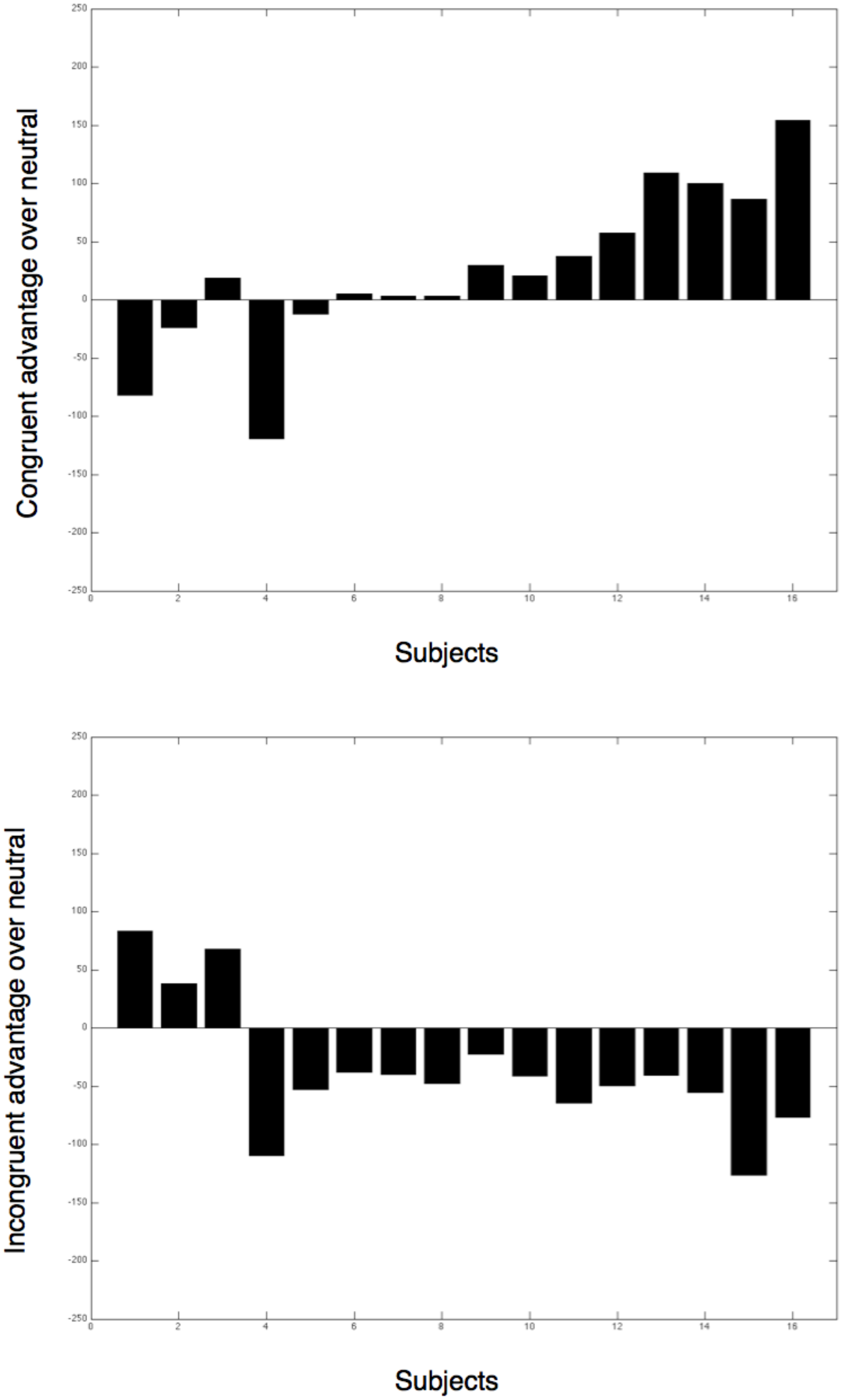

Figure 3: Difference in seconds between total congruent responses (top) and corresponding responses on no sound trials and between incongruent responses (bottom) and corresponding no sound responses, bucketed by subject in the same order as Figure 2. See the post hoc outlier analysis of subject 4 is in the results section.

screened for temporal sensitivity could help address this issue.

Another interpretation of our results might be that the auditory stimulus is causing subjects to consciously attend more to the congruent video, resulting in a greater dominance period due to attention rather than a more automatic cross-modal effect. As mentioned in the introduction, however, attention mainly seems to affect the rate of alternation [10] and the initially dominant stimulus [13], both of which have a limited impact on dominance duration. When attention does bias dominance duration it is usually when subjects are attending specific stimulus features [12]. By contrast, [26] seem to find a significant effect of commanded attention (which adds with their cross-modal interaction), but it is unclear whether they had subjects maintain fixation. Without maintaining fixation, subjects' eyes can easily wander or be specifically directed to higher contrast/complexity regions of the attended image and thus bias dominance on a low level. Since our subjects were specifically instructed to fixate on a fixation cross, and had no task related reason to remember or interpret the stories our actors told (subjects were simply told they would hear sounds during some of the trials), we do not believe that attention had a significant impact on our results. A future study carefully designed to focus on the interaction of cross-modal/attention effects (e.g. by requiring subjects to attend both to stimuli congruent and incongruent with a soundtrack) would likely help illuminate this issue further.

We believe this is an exciting result for the bistable perception field. It shows a new way in which high-level perceptual processes can interact with conscious perception and opens up new ground for researching the nature of both cross-modal interactions and bistable perception.

Acknowledgments

This work was supported by NSF IGERT Grant #DGE-0333451 to GW Cottrell/VR de Sa & NSF CAREER Grant #IIS-0133996 to VR de Sa.

References

[1] F. Tong, M. Meng, and R. Blake. Neural bases of binocular rivalry. *Trends in Cognitive Sciences*, 10(11):502 – 511, 2006.

[2] A Polonsky, R Blake, J Braun, and D Heeger. Neuronal activity in human primary visual cortex correlates with perception during binocular rivalry. *Nature neuroscience*, 3(11):1153–1159, 2000.

[3] S Lee and R Blake. V1 activity is reduced during binocular rivalry. *Journal of Vision*, 2:618–626, 2002.

[4] J. D. Haynes, R. Deichmann, and G. Rees. Eye-specific effects of binocular rivalry in the human lateral geniculate nucleus. *Nature*, 438(7067):496–499, 2005.

[5] David Alais and David Melcher. Strength and coherence of binocular rivalry depends on shared stimulus complexity. *Vision Research*, 47:269–279, 2007.

[6] D Baker and E Graf. Natural images dominate in binocular rivalry. *Proceedings of the National Academy of Sciences of the United States of America*, 106:5436–5441, 2009.

[7] I Kovács, T Papathomas, and M Yang. When the brain changes its mind: Interocular grouping during binocular rivalry. *Proceedings of the National Academy of Sciences of the United States of America*, 93:15508–15511, 1996.

[8] Derek H Arnold, Bridie James, and Warrick Roseboom. Binocular rivalry: spreading dominance through complex images. *Journal of Vision*, 9(13):4.1–9, 2009.

[9] H. von Helmholtz. *Treatise on physiological optics, Vol. III*. Dover, 1925.

[10] Ming Meng and Frank Tong. Can attention selectively bias bistable perception? Differences between binocular rivalry and ambiguous figures. *Journal of Vision*, 4(7):539–51, 2004.

[11] R Van Ee, LCJ Van Dam, and GJ Brouwer. Voluntary control and the dynamics of perceptual bi-stability. *Vision Research*, 45(1):41–55, 2005.

[12] S Chong, D Tadin, and R Blake. Endogenous attention prolongs dominance durations in binocular rivalry. *Journal of Vision*, 5:1004–1012, 2005.

[13] S Chong and R Blake. Exogenous attention and endogenous attention influence initial dominance in binocular rivalry. *Vision Research*, 2006.

[14] J Mishra and S Hillyard. Endogenous attention selection during binocular rivalry at early stages of visual processing. *Vision Research*, 2009.

[15] A Maier, N Logothetis, and D Leopold. Global competition dictates local suppression in pattern rivalry. *Journal of Vision*, 2005.

[16] T Watson, J Pearson, and C Clifford. Perceptual grouping of biological motion promotes binocular rivalry. *Current Biology*, 14:1670–1674, 2004.

[17] J Hupé, L Joffo, and D Pressnitzer. Bistability for audiovisual stimuli: Perceptual decision is modality specific. *Journal of Vision*, 8(7):1.1–15, 2008.

[18] H McGurk and J MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[19] L Rosenblum. Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science*, 17(6):405–409, 2008.

[20] A Vatakis, C Spence, and S Vecera. Crossmodal binding: evaluating the "unity assumption" using audiovisual speech stimuli. *Perception and Psychophysics*, 69(5):744–756, 2007.

[21] LD Rosenblum, NM Smith, SM Nichols, S Hale, and J Lee. Hearing a face: Cross-modal speaker matching using isolated visible speech. *Perception and Psychophysics*, 68(1):84, 2006.

[22] Ayse P. Saygin, J. Driver, and Virginia R. de Sa. In the footsteps of biological motion and multisensory perception: Judgements of audio-visual temporal relations are enhanced for upright walkers. *Psychological Science*, 19(5), 2008.

[23] Y. Kamitani L. Shams and S. Shimojo. What you see is what you hear. *Nature*, 408, 2000.

[24] Ladan Shams Shinsuke Shimojo. Sensory modalities are not separate modalities: plasticity and interactions. *Current Opinion in Neurobiology*, 11:505–509, 2001.

[25] Debra A. Zellner and Mary A. Kautz. Color affects perceived odor intensity. *Journal of Experimental Psychology: Human Perception and Performance*, 16(2):391–397, 1990.

[26] Yi-Chuan Chen, Su-Ling Yeh, and Charles Spence. Cross-modal constraints on human visual awareness: Auditory semantic context modulates binocular rivalry. Presented at the 10th annual Vision Sciences Society Meeting, 2010.