In our daily lives, we are constantly bombarded with information coming from different senses. For example, imagine you are hiking in the woods. Suddenly, you hear a snake hissing sound coming from the bush in front of you and you also see some movement behind the bush. Combining the auditory and visual information will increase the probability of noticing and correctly locating the snake, which increases the chances of running away from it. However, due to internal and external noise, the sensory signals would not be in perfect agreement, which could be manifested either as having a temporal or a spatial discrepancy. To form a coherent percept of the world, our brain integrates multisensory cues that are likely to be coming from a common source. By doing so, the precision of the estimate for the property of interest is maximized. For example, the *ventriloquism effect* happens when a ventriloquist speaks without moving her lips while the puppet’s mouth moves in synchrony with her speech and the audience perceives a “talking puppet”. A range of psychophysics experiments have demonstrated that the *ventriloquism effect* not only occurs in the real world, but also in laboratory settings using simple stimuli such as light flashes and sound bursts (Jack & Thurlow, 1973; Jackson, 1953; Pick et al., 1969). In such experiments, participants were asked to localize both auditory (e.g., a sound burst) and visual (e.g., a light flash) stimuli presented on the same trial. When there was a spatial discrepancy between the auditory and visual stimuli, participants’ localization responses shifted towards the other modality, which suggests successful integration of multisensory cues (Körding et al., 2007).

However, the brain does not always integrate information from different sensory modalities. One important factor determining integration is spatiotemporal proximity: if the cues are far apart either in space or time, they should not be integrated. The brain needs to solve the correspondence problem, that is, judging which cues are coming from the same source and which are coming from separate sources. Many studies found that humans perform causal inference in spatial localization tasks. For example, when visual and auditory stimuli that were presented on the same trial had a relatively large spatial discrepancy between them, to a certain point, participants stopped integrating the cues and made separate location estimates of the sound and light sources (Körding et al., 2007) or reported that the cues were coming from separate sources. (Wallace et al., 2004; Hairston et al., 2003).

Similarly, to integrating single sensory events from different modalities, when encountering stochastic sequences of multisensory events, the brain also needs to solve the correspondence problem. In this case, spatiotemporal correlation becomes the main factor underlying the cues to be integrated. For example, imagine seeing a person and their dog walking past you and hearing their footsteps, you would integrate the sound of the person’s footsteps with the person’s legs but not the dog’s legs. Various studies have investigated the role of temporal cross-correlation between auditory and visual signals on multisensory integration and causal inference (Locke & Landy, 2017; Parise et al., 2012, 2013; Parise & Ernst, 2016). For example, Parise and colleagues (2012) presented trains of visual, auditory, and combined audiovisual stimuli consisting of ten flashes and ten clicks over a 2 s temporal window. In bimodal trials, the temporal structure of visual and auditory sequences was varied for each trial: they were either identical (correlated trials) or random (uncorrelated trials) while their spatial locations always coincided. When asked to localize the perceived location of the stimuli., participants’ response precision in bimodal trials relative to unimodal trials was statistically optimal only when audiovisual stimuli were temporally correlated. These results suggest that humans use the temporal correlation between multisensory signals as a cue for causal inference. That is, when sensory cues from different modalities are temporally correlated, we are more likely to infer a common underlying cause and integrate them. Various other studies have found similar results using different behavioral tasks and explored the underlying mechanism with computational models (Locke & Landy, 2017; Parise et al., 2012, 2013; Parise & Ernst, 2016). However, to date, no study has investigated the role of spatial correlation in causal inference and multisensory integration. Here, spatial correlation is defined for stochastic sequences of multisensory events: a perfect spatial correlation means that the physical location of each individual event within that sequence in one modality is co-located with its concurrently presented individual event in the other modality.

In my thesis work, I aim to vary the spatial structure of stochastic auditory and visual sequences while keeping their temporal structure identical and examine how spatial correlation influences multisensory integration in the context of the ventriloquism effect. Specifically, I will introduce various spatial discrepancies between auditory and visual stimuli by varying the physical location of the centroid of each stochastic sequence. In the main experiment, I will ask participants to localize the centroid of the sequences in both modalities and ask them to judge whether the auditory and visual sequences are from one underlying source or two separate sources. If spatial correlation operates as a cue for causal inference, auditory and visual signals would be integrated more optimally when their spatial correlation is higher. I hypothesize that participants’ localization responses will shift toward the location of the centroid of the other modality, compared to their localization responses in unimodal trials and they will be more likely to report that the two cues share a common cause.

**Method**

**Apparatus and Stimuli**

The experiment will be conducted in a dark, semi sound-attenuated room. Participants will be seated comfortably with their chins rested on a chin rest.

A visual event is a brief flash of low-contrast Gaussian blob projected onto a gray background. An auditory event is a brief broadband noise burst, delivered via headphones. Each stimulus presentation is a sequence of five visual and/or five auditory events, evenly distributed over a 2 s temporal interval. The spatial locations of all visual and auditory events range from -18° from the farthest left to 18° from the farthest right on a horizontal line. Each sequence of visual or auditory events has a spatial window of 10°, that is, the locations of individual visual or auditory events are randomized within the range of 10°, centered on the centroid. Spatial structures are generated for the visual sequence and for the auditory sequence with various spatial correlations ranging from 0-1. To introduce spatial discrepancies between the visual and auditory sequences, the centroids will be placed on the left side, at the center or on the right side of the display.

To create compelling spatialized sounds, in a preliminary session the raw auditory events will be played from a loudspeaker behind an acoustically transparent screen from each of the 20 spatial positions and recorded with a pair of in-ear binaural microphones placed inside the left and right ear canals of each participant. The recording for each participant will then be cut into short clips, each containing the sound of one auditory event at a certain spatial location. By performing this procedure individually for each participant, the auditory stimuli are filtered by the individual’s head-related transfer function, thereby providing rich and ecological cues for sound localization.

To find the perceptual location of each auditory event for each participant, in a preparatory experiment, participants will localize each of the auditory events recorded from 20 spatial locations 20 times played from headphones, the order of which will be randomized. Then, the perceptual location of each auditory event can be computed by calculating the mean localization response for that location. To find the perceptual locations of 20 visual events for each participant, I will fit a linear regression to the mean auditory localization responses for each participant. The perceptual locations of auditory and visual events for each participant will be used to generate their individualized auditory and visual sequences and their respective centroids in the main experiment.

**Procedure**

This study will consist of four preparatory experiments and two main experiments. The purposes of the preparatory experiments are to find perceptual event (auditory and visual) locations for each participant based on their individual perceptual biases, to get participants familiarized with the experimental setup, the stimuli, and the task, and to obtain information about participants’ memory and motor noise.

The first main experiment is a unimodal localization task. On each trial, participants will be presented with either a visual sequence or an auditory sequence with a duration of 2 s (the order will be randomized). After each stimulus presentation, participants will localize the centroid of that sequence using the trackball on a mouse. Scrolling the trackball will move a visual cursor left and right pointing at the horizontal line where stimuli are presented. Once participants move the visual cursor to the desired location, they will left-click to confirm.

The second main experiment is a bimodal localization task. On each trial, participants will be presented with sequences of visual and auditory stimuli with a duration of 2 s, simultaneously, with various spatial discrepancies between them. After stimulus presentation, participants will localize the centroid of either the auditory sequence or the visual sequence following the prompt on the screen. Participants will be asked to localize the centroid of the auditory sequence in half of the trials while ignoring the visual sequence and localize the centroid of the visual sequence while ignoring the auditory sequence in the other half of the trials (the order will be randomized). After making a localization response, participants will be asked to judge whether the two stimuli are coming from the same source or separate sources using left or right click on the mouse following the prompt on the screen.

**Statistical analysis**

To understand how the brain detects and integrates spatially related information across continuous streams of multisensory signals and how it adapts to spatial conflicts between the sensory modalities, I plan to fit my data to a variation of the multisensory correlation detector (MCD) model, which was initially adapted from the Hassenstein-Reichardt detector for visual motion perception. The MCD model was first developed to explain how temporal correlation between stochastic sequences of multisensory signals influence integration and the model successfully replicated human perception (Parise & Ernst, 2016).

Locke, S. M., & Landy, M. S. (2017). Temporal causal inference with stochastic audiovisual sequences. *PLOS ONE*, *12*(9), e0183776. <https://doi.org/10.1371/journal.pone.0183776>

Parise, C. V., & Ernst, M. O. (2016). Correlation detection as a general mechanism for multisensory integration. *Nature Communications*, *7*(1), 11543. <https://doi.org/10.1038/ncomms11543>

Parise, C. V., Harrar, V., Ernst, M. O., & Spence, C. (2013). Cross-correlation between Auditory and Visual Signals Promotes Multisensory Integration. *Multisensory Research*, *26*(3), 307–316. <https://doi.org/10.1163/22134808-00002417>

Parise, C. V., Spence, C., & Ernst, M. O. (2012). When Correlation Implies Causation in Multisensory Integration. *Current Biology*, *22*(1), 46–49. <https://doi.org/10.1016/j.cub.2011.11.039>