Investigating the effects of audio-visual spatial congruence on multisensory integration.

Lindsey R. Kishline

A dissertation

submitted in partial fulfillment of the

Requirements for the degree of

Doctor of Philosophy

University of Washington
2020

Reading Committee:

Adrian KC Lee, Chair

Lynne Werner

Ross K Maddox

Program Authorized to Offer Degree:

Speech and Hearing Sciences

Abstract

Investigating the effects of audio-visual spatial congruence on multisensory integration.

Lindsey R. Kishline

Chair of the Supervisory Committee:

Adrian KC Lee, Sc.D.
Department of Speech and Hearing Sciences

We live in a complex world wherein we are bombarded by many different sensory inputs. Somehow our brain is able to take all of these complex and overlapping sensory inputs and make order out of our surroundings, creating our world and all the various objects in it. Combining information across modalities is behaviourally advantageous, as it has been shown to improve visual search times, provide faster reaction times, and improve accuracy. Specifically, for the purposes of this body of work, we will focus on auditory and visual integration and how each modality may influence perception, and how we can leverage the peculiarities of audio-visual integration for potentially beneficial applications to Virtual Reality. Three studies are reported in this thesis and each provides a unique contribution to our understanding of multisensory integration. Specifically these studies are focused on how the spatial relationships of both auditory and visual stimuli affect multisensory integration.

Acknowledgements:

I wish to express my deepest gratitude to my advisor, Adrian KC Lee for his help and guidance over the years. Additionally, to my dissertation committee and members of LABS^N for encouragement and support. I would like to thank the auditory perception team at Facebook Reality Labs, as well as my fellow doctoral students in the department. Finally I would like to thank my family, with a special thank you to my father Brian and brother Samuel for their invaluable and never failing support and love.

Table of Contents

Chapter 1: Introduction

We live in a complex world wherein we are bombarded by many different sensory inputs. Somehow our brain is able to take all of these complex and overlapping sensory inputs and make order out of our surroundings, creating our world and all the various objects in it. This includes grouping information both within and across sensory modalities such as audition and vision. Even though these sensory systems are coded in separate pathways, and with different levels of spatial and temporal acuity, this merging of the sensory information termed multisensory integration, often happens effortlessly and without us thinking about it.

Combining information across modalities is behaviourally advantageous. It has been shown to improve visual search times, provide faster reaction times, and improve accuracy(Frens et al.,1995; Diederich & Colonius, 2004; Gielen et al., 1983; Perrott et al., 1990; Van der Burg et al., 2008; Laurienti et al., 2006). Specifically, for the purposes of this body of work, we will focus on auditory and visual integration and how each modality may influence perception, and how we can leverage the peculiarities of audio-visual integration for potentially beneficial applications to Virtual Reality.

Three studies are reported in this thesis and each provides unique contribution to our understanding of multisensory integration, as follows:

Study 1: Generally, combining auditory and visual stimuli afford advantages to the observer, but only when these stimuli are presented approximately the same time in the same location. This observation is often referred to as the spatial-temporal window of integration. However, having the auditory and visual stimuli at the same location is not always necessary to see the benefits of multisensory integration. Study 1 asked when spatial colocation of auditory

and visual stimuli is important and explored how spatial colocation, as well as the spatial relationship of competing stimuli influence the reporting of an audiovisual illusion. Furthermore, it investigated the effect of spatial attention on audiovisual integration.

Study 2: This study explored how we can leverage the superior spatial resolution in vision using a ventriloquist paradigm. Specifically, it assessed the amount of visual spatial influence on an auditory stimulus under two different auditory spectral filters (one meant for general use and the other meant to be participant-specific). Importantly, while the ventriloquist effect has been well characterized in the horizontal plane, only a couple of studies have examined the effect in the vertical plane. Here an absolute localization task was used to look at how visual stimuli affect auditory elevation localization judgement.

Study 3: The first two studies both highlight the importance of moving towards more naturalistic and complex stimuli. However, incorporating these naturalistic stimuli in laboratory settings can be difficult. Study 3 outlines a tool designed for experiments in Virtual Reality -- an environment in which the experimenter has greater control of realistic audiovisual presentation to participants. An audio-visual speech corpus specifically was designed to leverage the unique capabilities this technology has to offer the research community, by allowing researchers to present physically impossible but ecologically valid stimuli in contrived scenarios. This corpus was created to facilitate auditory and multisensory research in immersive environments, and designed to allow experimental control in multidimensional tasks.

All three chapters were written for separately technical journal submissions.

References:

Diederich, A., & Colonius, H. (2004). Bimodal and trimodal multisensory enhancement: effects of stimulus onset and intensity on reaction time. Perception & psychophysics, 66(8), 1388-1404.

Frens, M. A., Van Opstal, A. J., and Van der Willigen, R. F. (1995). Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. Percept. Psychophys. 57, 802–816.

Gielen, S.C.A.M., Schmidt, R.A. & Van Den Heuvel, P.J.M. On the nature of intersensory facilitation of reaction time. Perception & Psychophysics 34, 161–168 (1983).

Laurienti, P. J., Burdette, J. H., Maldjian, J. A., & Wallace, M. T. (2006). Enhanced multisensory integration in older adults. Neurobiology of aging, 27(8), 1155-1163.

Perrott, D. R., Saberi, K., Brown, K., and Strybel, T. Z. (1990). Auditory psychomotor coordination and visual search performance. Percept. Psychophys. 48, 214–226.

Van der Burg, E., Olivers, C. N., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and pop: nonspatial auditory signals improve spatial visual search. Journal of Experimental Psychology: Human Perception and Performance, 34(5), 1053.

Chapter 2: Assessing the interaction of selective attention and spatial relationship of visual and competing auditory stimuli on the sound-induced flash illusion.

Authors: Lindsey R. Kishline, Adrian KC Lee, Ross K. Maddox

Abstract:

To investigate the influence of spatial colocation between audio and visual stimuli on multisensory integration, we assessed the sound induced flash illusion (SIFI) while manipulating the spatial congruence and proximity of two competing auditory stimuli and a single visual stimulus. Participants were presented with two timbrally distinct and concurrent auditory stimuli presented at different spatial locations and consisting of either one or two events. A single visual stimulus--always spatially congruent with one of the auditory stimuli--was composed of either one or two events. On every trial one of the auditory stimuli matched the number of visual events, and the other did not. Spatial attention was also manipulated by providing (or omitting) a visual symbolic cue that preceded each trial indicating the location of the visual flash. While colocation of the illusion inducing auditory stream did not have an effect on the SIFI, there was a strong effect of where the visual flash was presented on the likelihood of the SIFI being reported.

I. Introduction:

Auditory and visual stimuli are reported to be maximally integrated when they occur at the same time and in the same location. Known as the temporal and spatial 'rules' of multisensory integration, originally derived from neurophysiological data recorded in the cat superior colliculus (Meredith & Stein, 1987; Meredith & Stein, 1986), temporal and spatial

proximity of sensory stimuli are regarded as the fundamental components in promoting multisensory integration. Specifically these recordings showed that maximal responses could be elicited in multimodal neurons when concurrent auditory and visual stimuli were presented within the receptive field of the cells and presented at the same spatial location. Over the last 20 years there has been a growing question as to how well these multisensory rules translate to behavioral studies. Some have questioned how the behavior of single neurons, which were recorded in anesthetized cats, translate to the behavior and task performance of awake humans (Ma & Pouget, 2008; C. Spence, 2013).

There is substantial behavioral evidence for the temporal rule of multisensory integration. For example, adding a non-informative auditory stimulus presented at the same time a visual target appears is enough to greatly decrease visual target search times in the presence of many distractors. This phenomenon is known as the 'Pip and Pop' effect (Van der Burg et al., 2008). Temporal coincidence has also been found to lead to enhanced visual luminance detection (Frassinetti et al. 2002) and enhanced auditory detection (Lovelace et al., 2003). While the above studies presented the auditory and visual stimuli concurrently, there is a wide temporal window in which this temporal rule holds, with the caveat that the width of these temporal windows is both stimulus and task dependent (Alais et al., 2010; Bertelson & Radeau, 1981; Colonius & Diederich, 2004, 2010, 2011; Hillock, Powers, & Wallace, 2011; Hillock- Dunn & Wallace, 2012a; Keetels & Vroomen, 2005, 2007, 2008a; Keetels, Stekelenburg, & Vroomen, 2007). Stimuli presented outside of this optimal temporal window are less likely to be integrated or see behavioral benefits (Wallace & Stevenson, 2014).

For the spatial rule however, there is conflicting evidence for how important spatial congruence is for multisensory integration in behavioral studies. There have been a number of studies in support of the spatial rule (Harrington & Peck, 1998; Spence & McDonald, 2004; Bizley et al., 2012) but other equivocal observations are also as abundant (Doyle & Snowden, 2001; Colin et al., 2001; Kumpik et al., 2014; Innes-Brown & Crewther, 2009; Zampini et al., 2007).

Spence (2013) posited that the behavioral benefit of spatial congruence on multisensory integration is only evident when performing a spatial task or deploying spatial attention. While this is an elegant postulation, there are notable exceptions in which a temporal task has shown evidence for a benefit of spatial congruence (Gondan et al., 2005; Tiippana et al., 2011; Di Luca et al., 2009; Leo et al., 2011), as well as a spatial task or task that deployed spatial attention that did not show a behavioral performance based on spatial colocation (Fiebelkorn et al., 2011).

So when does spatial congruence matter? In naturalistic environments there is often stimulus competition. Specifically in acoustic scenes, multiple talkers or masking stimuli are often present and the benefits of target-masker spatial separation and the knowledge of the spatial locations of target talkers become clear (Maddox et al., 2012; Kidd & Mason, 2005; Freyman et al., 2001). However, there are only a handful of studies that have addressed audio-visual integration in competition (Van der Burg et al., 2008; Bizley et al., 2012; Cappelloni et al., 2019), and thus evidence that can shed light on the potential influences of spatial collocation on multisensory integration when multiple stimuli are present is still relatively rare (Lee et al., 2019).

A popular way to investigate multisensory integration is through audio-visual illusions, as they shed light onto the obligatory ways our brain integrates information. One particular illusion used is the sound induced flash illusion (SIFI), in which flashes and beeps are presented to participants and when two beeps are paired with a visual flash, people would often report seeing two flashes even though only one physical flash was shown (Shams et al., 2001). This original SIFI study showed that there are temporal limitations on the illusory flash, that the auditory stimuli had no effect on the visual stimulus after an asynchronous presentation of 300 ms, but was agnostic about the spatial relationship between the flashes and the beeps had an effect on the illusion percept.

In a subsequent study on the SIFI, Innes-Brown et al. (2009) asked whether the illusion was dependent upon the spatial congruence of the flashes and beeps, but they did not show any effect of the illusion based on this factor. However, another study (Bizley et al., 2012) did report an effect of spatial colocation on the SIFI. Interestingly, this study included stimulus competition by having two sets of auditory and visual stimuli separated by hemifield. While this study included stimulus competition, it also included manipulation of spatial attention and the effect of spatial congruence is in line with Spence's postulation (Spence, 2013).

The present study aims to tease apart the influence of spatial congruence, competing auditory stimuli, and spatial attention on the SIFI. Here we present two timbrally distinct and concurrent auditory stimuli, always at different spatial locations, consisting of either one or two auditory events, along with the traditional visual flash (which could consist of either one or two flash events). On every trial there was an opportunity for a sound induced flash illusion as one of the auditory stimuli had a differing number of events as the visual stream. On every trial, there is

a possibility of capturing one of two illusory percepts: fission or fusion. Fission is defined as the classical SIFI illusion of two auditory beeps and one physical flash being reported as two flashes (Shams et al., 2001). Fusion is defined as the report of one visual flash, while two physical flashes were presented along with a single auditory beep. On half the experimental blocks, spatial attention was cued to the position of the visual flash, while on the other half participants were not cued where to attend. In addition to investigating the influence of spatial colocation, we also assessed the influence of spatial proximity, i.e., whether how far the competing auditory stimulus is away from the visual stimulus could influence on the reported number of visual illusions. We anticipated more illusory reports when attention was directed to the location of the flash and the illusion inducing auditory stream is collocated (Bizley et al., 2012), as well as fewer illusory reports when the auditory stream with the matching number of events was collocated with the visual flash(es). Additionally, we expected the number of reported flash would depend on the spatial separation between the flash and the non-matching auditory events.

II. Methods

A. Participants

Twenty-two participants (14 female, ages 19-52, mean age of 29) were recruited in the study. Three participants were excluded for further analysis because two of these performed less than 90% correct responses on catch trials in which illusion was not expected (the visual stimulus was delayed sufficiently outside of the illusionary window), and one had abnormal reaction times that on average exceeded two seconds. Therefore we analyzed data from 19 participants. All participants had normal audiometric thresholds (20 dB hearing level or better at octave frequencies from 250 Hz to 8 kHz). They were compensated at an hourly rate, and gave

informed consent to participate as overseen by the University of Washington Institutional Review Board. All participants attended two experimental sessions (separated by a minimum of 24 hours and a maximum of one week). Each experimental session lasted on average 1 hour and 20 minutes.

B. Stimuli

Auditory and visual stimuli were presented using expyfun software (Larson et al., 2016).

B. 1. Acoustic

Two acoustic stimuli were generated, a noise burst and a tone complex, at a sampling rate of 24.414 kHz. The tone complex consisted of the first five harmonics of 250 Hz, all of equal amplitude and a fixed sine starting phase. The noise bursts were white noise highpass filtered at 1500 Hz using Scipy signal "firwin" filter with a 100th order Hamming window. The white noise bursts were generated afresh for each participant. Both tone complex and noise bursts were time windowed (6 ms onset/ offset ramps for the tone complex and 3 ms onset/ offset ramps for noise burst) and normalized by their root-mean-squared values to equate intensity. Both auditory stimuli had a single event duration of 30 ms, and two events within the same stream had an 80 ms inter stimulus interval giving a 50 ms silent interval between events (see **Figure 2** for timing details). There were three possible locations that the auditory streams could occur ∓ 15° off midline and at 0° on midline. Sounds were processed using non-individualized head-related transfer functions from the CIPIC database (Algazi et al., 2001). These two sounds were never co-located; however, one sound was always spatially congruent with the visual flash location. Sound locations and number of events in each stream were randomized trial to trial. Sound stimuli were delivered over insert earphones in a sound-attenuated booth through Tucker Davis

Technologies (Alachua, FL) RP2 real-time processor at 70 dB SPL with a continuous white noise masker presented at 45 dB SPL.

B.2 Visual

The visual flash was a white disk at 100% contrast subtending 1 degree of visual field presented on an LCD monitor. Flashes were shown 6 degrees below the horizon at zero degrees azimuth, and could occur at one of three locations on a given trial ∓ 15° to the right and left of midline and at 0° (center). On every trial a visual fixation dot that subtended 0.3° of visual angle was presented at eye level on the midline. Each flash had a duration of 16.7 ms and two flashes had an inter stimulus interval of 83 ms. The first flash started 20 ms after the onset of the auditory stimuli (see **Figure 2**). In catch trials, the onset of the visual flash was delayed by 300 ms putting it outside of the illusion window (Shams et al., 2002). Thus in these trials, we expect that the visual flash counts would not be influenced by the auditory events. On cued trials, the visual spatial cues were green arrows (<, >, < >) , indicating left/right/center for the flash location on the upcoming trial to indicate where attention should be directed on the following trial. Participants were instructed to maintain fixation at a fixation dot during every trial. Thus even on cued trials, they were instructed to attend the cued spatial position without moving their eyes off the fixation dot (viz., deploy covert spatial attention).

C. *Spatial configuration (Figure 1)*

Figure 1

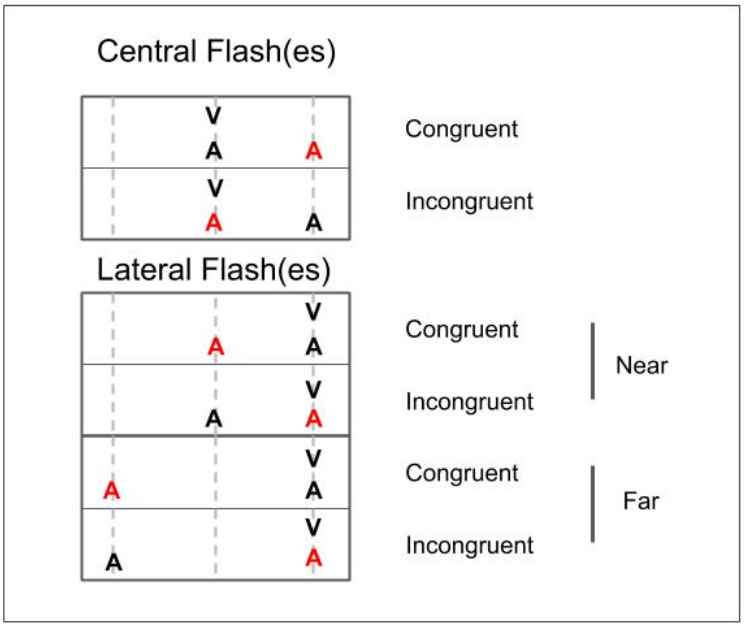


Figure 1: Six spatial conditions shown above. The visual flash location is represented by the black 'V', and the two competing auditory streams are represented by the letter 'A'. The black 'A' is the auditory matching stream, and the red 'A' represents the illusion inducing auditory stream. Two center spatial conditions, when the visual flash(es) are located at 0 degrees azimuth, one with the auditory matching stream spatially congruent (top most panel) and one with the illusion inducing stream as spatially congruent with the flash location. Four lateral spatial conditions, defined as when the visual flash(es) are located at either +/- 15 degrees azimuth. The lateral spatial conditions are split into two near conditions (the top two panels of the lateral types) and two far conditions, designated by the total spatial separation between the auditory stream that is congruent with the visual location either being one space (near - 15 degree separation) or two spaces (far - 30 degree separation). While the above figure shows six general spatial conditions, the actual experiment included all counter balanced spatial arrangements.

A total of six spatial conditions were tested: two center conditions where the visual flash(es) occurred in the central location, and four periphery conditions where the visual flash occurred at ±15° off the center. The auditory stream with the same number of events as visual flash(es) was either spatially congruent with the visual location or was one or two (either 15° or 30°) locations away. The competing auditory streams were always in different spatial locations and there was always one auditory stream that was spatially congruent with the visual. All conditions and spatial manipulations were counterbalanced for left and right sides and for type of auditory stimulus to match the events in the visual stream, and number of events, giving a total of 48 possible trial types. Together, 30 repetitions of each of the 48 trial types were tested spreading across fifteen 3-4 minute blocks. Each block had 96 experimental trials (two sets of the 48 trial types) and 12 catch trials, with presentation order randomized. For the cued experimental session, the above was the same with the addition of the spatial cue preceding the trial, and performed in a separate experimental session.

D. Trial timing (Figure 2)

Figure 2

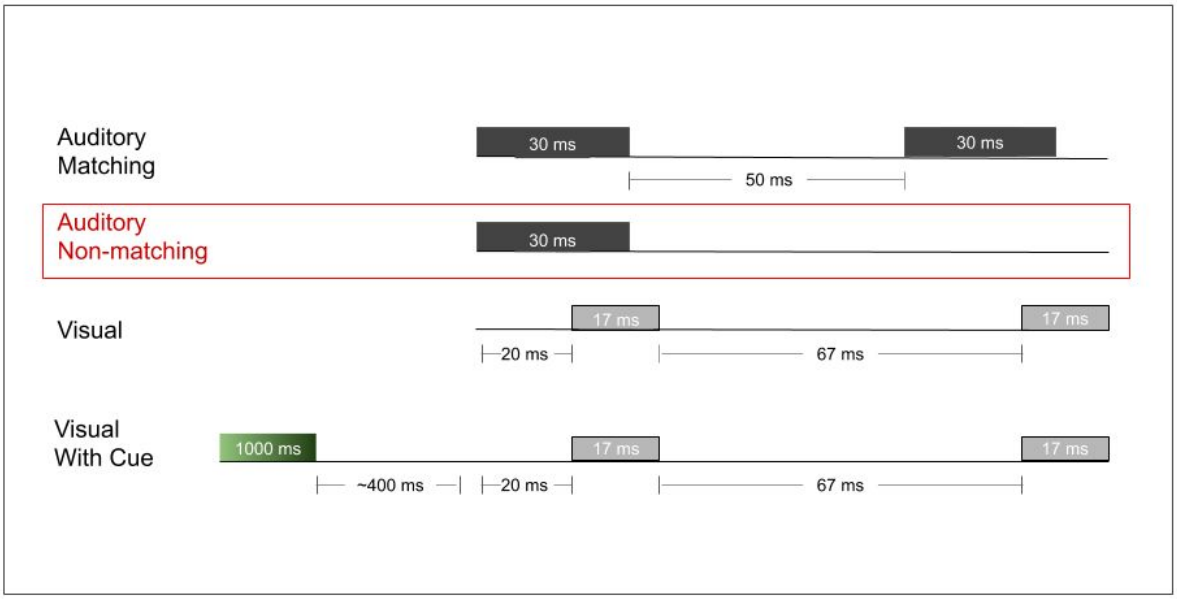


Figure 2: Trial timing is shown in the above figure. Auditory stimuli preceded visual stimuli by 20ms. Both auditory streams were presented concurrently and auditory events were separated by 50ms. Visual stimuli were 17ms long and multiple visual events were separated by 67ms. In trials that were preceded by a cue, the cue was presented for a duration of one second and occurred between 300-500ms before the start of the auditory beeps.

Both auditory streams were presented concurrently, and auditory events were separated by 50 ms while visual events were separated by 66 ms. In trials preceded by a visual cue, the cue was presented for a duration of 1 second before a jitter time of between 300-500 ms before the start of the auditory beeps. The jitter was added to ensure that the presentation of stimuli did not occur rhythmically. On non-cue trials, timing was consistent with the cued trial type, i.e.,between trials the time also consisted of a 1 second presentation of only fixation dot and the jitter time.

Every trial had the opportunity to be reported as an illusion - as the visual stimulus could either flash once or twice on any given trial and one of the auditory stimuli always had two events (beeps) and the other had one event (meaning there were never the same number of events in the tone and noise burst streams). This meant that on every trial one of the auditory stimuli had the same (matching number) of events as the number of visual flash(es), and this we refer to the Auditory Matching stream from here on out. The auditory stream that did not have the same number of events as the visual stream will be termed the Auditory Non-Matching stream, and can be thought of as the illusion inducing stream.

E. Procedure

Prior to each experiment, participants went through a two-stage training procedure. During the first stage, participants were required to score a minimum of 80% accuracy in reporting the correct number of physical flashes that occurred on each trial without auditory stimuli present, and immediate feedback was given after each trial. Nearly all participants performed this task with nearly 100% accuracy. During the second stage, participants were exposed to one minute of the experiment with sound and no feedback to acclimate to the task, and so that the experimenter could answer any questions about the task.

Participants' task was to respond as quickly and accurately as possible following the trial, and indicate the number of flashes (one or two) that they saw by pressing the "1" or "2" button on the response box. In addition to the numbered button response, reaction time for the button press was also recorded.

During the experiment, participants' eye position was tracked to check fixation on the fixation dot during each trial using an EyeLink1000 infra-red eye tracker (SR Research, Kanata,

ON). Participants were seated 50 cm from the monitor and Eyelink camera, and stabilized using a chinrest and forehead bar. Prior to each experimental block an eye calibration was performed.

F. Statistical Analysis

The probability of a participant reporting an illusion percept was modeled using generalized logistic mixed-effects regression, while the reaction times were modeled using a linear mixed-effects regression, both performed using the **`lme4`** package (Bates et al., 2014) in the R statistical computing environment (R Development Core Team, 2014).

To model the probability of reporting an illusion we fit a model to predict whether or not the participant reported having observed two flashes and pressed "2." This allowed us to fit a binomial logistic regression with random effects using the '**`glmer`**' function in the **`lme4`** package, where pressing the "2" button, represented by the term *PressedTwo,* was then dummy coded as a 1 for true and 0 for false (representing a "2" or "1" button press respectively). The model predicted whether or not participants pressed "2" based on the number of physical flashes presented on that trial (*TwoFlashes*), whether or not attention was cued (*CueAttn*), the number of auditory events in the auditory stream that was colocated with the visual flash(es) (*TwoSoundsAtTargetLoc*), as well as the spatial position of the visual flash(es) and relative distance between the auditory streams (*NonCentralFlash and NonCentralFlash_FarFoil*). Estimation of the main effects requires six coefficients (three for the spatial position predictor *ecc*, and one for *TwoFlashes*, *CueAttn*, and *TwoSoundsAtTargetLoc*). Estimation of the two-way interaction adds an additional four coefficients. Including the intercept a total of ten fixed-effect coefficients were used, plus a random intercept for the participant. Three-way interactions are not reported for this model, as the model failed to converge.

To model the reaction time of button press given all the condition and trial types a linear mixed effects regression was used, this time for the continuous outcome variable of reaction time. Reaction time was represented in milliseconds from the time of the last flash. This model predicted a participants reaction time based on the same terms mentioned above, with the addition of *PressedTwo,* which was again dummy coded with 1 for true and a 0 for false representing whether or not the participant pressed the two button to report seeing two flashes. Estimation of both the two-way and three-way interactions, along with the five main effect coefficients, plus a random intercept for the participant were used.

Model descriptions:

Illusion data model - Model in R notation:

(pressed_two ~ two_flashes*cue_attn + two_flashes*ecc + two_flashes*two_sounds_at_target_location + (1|subject))

$$Eq.\,(1) \qquad (yij) = \beta 0 + \beta 1 Tj + \beta 2 Aj + \beta 3 Cj + \beta 4 Ppj + \beta 5 Pcj + \ldots + S0j + \varepsilon ij$$

In the above equation, yj is the outcome (*PressedTwo*) for participant j, $\beta 0$ is the intercept term, and the other β terms are the coefficients estimated for the various predictors. Tj, Aj, and Cj, are binary indicators. Our 'truth' term (which represents the number of physical flashes, *TwoFlashes)* is represented as T, the attentional cue (*CueAtten*) is represented as A, and number of auditory events at the target location (*TwoSoundsAtTargetLoc)* as our colocation indicator is represented as C. Pcj indicates whether the visual was in the central or peripheral locations (*NonCentralFlash*), and Ppj indicates whether the non-collocated auditory stream was one or two spaces away from the colocated stream (*NonCentralFlash_FarFoil)*.

Pcj and Ppj are ternary spatial proximity predictors and are reverse Helmert coded shown in **Table 1.** The Pcj coefficient reflects the difference between *Vcent* (indicating a flash in the central location) and both *Vperiph.center* and *Vperiph.periph* (comparing visual center

conditions to visual peripheral conditions) (please see **Figure 1** for central versus lateral trial types). While Ppj is a coefficient that reflects the difference between *Vperiph.center* and *Vperiph.periph* (comparing the auditory non-collocated stream being in the center location, and the non-collocated stream being in the opposite peripheral location) and is an estimation of the effect of proximity (please see **Figure 1** for the lateral *near* and *far* trial types for reference). The ellipsis is an indication of the additional coefficients used for the two-way interaction of the fixed effects; S0j is the random effect for subject j, and ϵij is the error term.

Table 1:

*Helmert coding:

	NonCentralFlash	NonCentralFlash_FarFoil
Vcent	-2/3	0
Vperiph.center	1/3	-1/2
Vperiph.periph	1/3	1/2

Table 1: This table shows the Helmert coding used in the model for the contrasts between center and lateral spatial conditions, and near and far lateral spatial conditions.

Due to the way we have chosen to code the coefficients, and to the nature of regression models, the interpretation of any one coefficient is also dependent upon all the coefficients in the model and will be interpreted as such below. This means that each coefficient will represent a unique trial type, determined along each of the model coefficients; number of flashes(*TwoFlashes*), attention cue (*CueAttn*), number of sounds at visual location(*TwoSoundsAtTargetLoc*), and location of visual stimulus (*NonCentralFlash*) as well as

the non-collocated auditory stream(*NonCentralFlash_FarFoil*). The outcome variable *PressedTwo*, was chosen so that it would allow for the interpretation of both fission and fusion illusions when assessed along with the truth term (*TwoFlashes)* indicating the physical number of flash(es) presented. This allows representation of the button pressed and the number of physical flash(es) on that particular trial type, giving us both SIFI illusion types.

The effects of experimental manipulations are split into three groups of coefficients; the effects of attention, the effects of collocation, and the effects of proximity. The effects of attention are described by the main predictor *CueAttn,* indicating whether or not spatial attention was cued prior to the beginning of the trial. However, the true effect of attention will be represented by the interaction between our truth term *TwoFlashes* and *CueAttn*, which shows that when two physical flashes occurred, spatial attention was cued, and there was one sound at the target (visual flash) location, how likely subjects were to press "2" representing a fission SIFI response.

Our second experimental manipulation effect of collocation was described by the main predictor *TwoSoundsAtTargetLoc*, which indicates whether or not there was a double auditory event at the same location as the visual flash(es). This can be interpreted as how likely participants were to press "2" when attention was not cued and there was only one visual flash. In other words, when the auditory non-matching stream was colocated with the visual stream, how likely was it to influence the report of the sound induced flash illusion?

Lastly, our experimental manipulation of proximity, defined as how close the auditory non-matching stream was to the visual flashes (and the colocated auditory matching stream) represented by *NonCentralFlash* and *NonCentralFlash_FarFoil*. Previously described as the

helmert coded variables, these terms indicate if the visual flash occured in the center, and whether the illusion inducing auditory stream was either near (i.e., one spatial location away with a 15° azimuthal separation) or far (i.e., two spatial locations away with a 30° azimuthal separation). The second order and third order interaction terms will be described below in the results section.

Reaction time model - Model in R notation:

*(Reaction Time ~ TwoFlashes*CueAttn*ecc*TwoSoundsAtTargetLoc*Pressed_Two + (1|subject)*

This model was largely the same as the above Illusion model with identical predictors, with the exception that it was fit to predict reaction time in milliseconds as a continuous variable, making this model a linear mixed effects regression. The rest of the coefficients remained the same only with the addition of *Pressed_Two* as a predictor (previously the outcome variable of the above described model), with an additional 24 coefficients for the three-way interaction terms.

III. Results:

Table 2:

Modeling for (PressedTwo)				
Predictor name	**Estimate**	**Std. Error**	**Z value**	**Pr(>|z|)**
(Intercept)	-1.54	0.08	-18.35	<2e-16***
TwoFlashes	2.59	0.03	92.45	<2e-16***
CueAttn	0.03	0.02	1.44	0.15
NonCentralFlash	0.07	0.03	2.73	0.006**
NonCentralFlash_FarFoil	0.03	0.03	1.02	0.309
TwoSoundsAtTargetLoc	-0.01	0.02	-0.54	0.589
TwoFlashes:CueAttn	0.12	0.03	3.89	9.99e-05***
TwoFlashes:NonCentralFlash	-0.49	0.03	-13.96	<2e-16***
TwoFlashes:NonCentralFlash_FarFoil	-0.004	0.04	-0.11	0.911
TwoFlashes:TwoSoundsAtTargetLoc	0.004	0.03	0.12	0.908

Table 2: The table shows the output of the logistic regression model for likelihood of reporting an illusion.

A. General Illusion data

As a reminder, our first model predicts the probability of a participant reporting an illusionary percept on a trial given factors of attention, colocation, and proximity. The results presented in **Table 2** show the outcome of the generalized logistic mixed effects model and are grouped into three sections of coefficients; the effects of attention, the effects of collocation, and the effects of proximity. The numbers reported in **Table 2** are the log odds (logits) output of the logistic regression model and as the logit scale is linearized, it makes it convenient to interpret in terms of one unit increases in our predictors resulting in a one unit increase in our outcome coefficient, preserving our coding structure of the predictor variables.

The across-subject variability in illusion susceptibility is shown in **Figure 3**, where individual subject data is superimposed on the percent illusion bar graph. The number of fission versus fusion reports varied by subject, but overall a greater number of fission illusions were reported. Additionally, there was no difference in the reported number of illusion percepts based on whether the non-matching auditory stream was a tone complex or noise burst . In **Figure 3**, fission and fusion were combined under the umbrella term illusion for ease of representation. However, in terms of interpreting the output of the illusion model, fission and fusion can be examined separately. It is worth noting that even in participants with low average number of illusions reported that these are in fact accurately reported illusions as participants were required to pass a training with visual flash(es) only (no audio) and report the number of physical flashes. As reported in the above methods section, participants passed this training task at near 100% accuracy.

The intercept term is expected to be significant and is not of particular interest to this study, as it indicates that when there was only one physical flash present participants were significantly less likely to press the "2" button. This provides additional evidence that the overall rate of reporting an illusion on average was not high, and on average participants reported seeing the physical number of flash(es) presented on the screen. Additionally, *TwoFlashes* is also expected to be (and is) significant (coefficient = 2.59 logit, $p<0.001$), indicating that when there were two physical flashes presented, attention was not cued, and there was one sound at the visual target location, subjects are more likely to press "2." Intuitively this makes sense, as the instance of illusion (fusion) was a relatively rare occurrence.

Figure 3

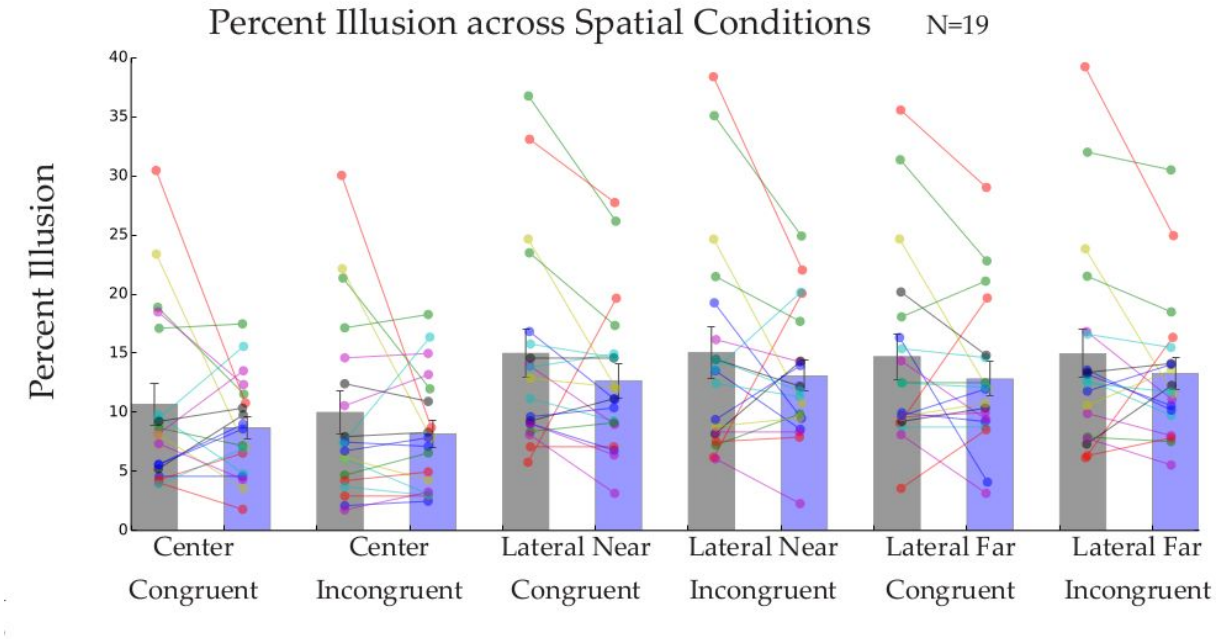


Figure 3: Total average percentage of illusions reported across all six spatial conditions for nineteen subjects. Illusion percentages for trials preceded by a cue shown in blue, and without cue shown in grey. Individual subject data imposed over the bar graph.

Figure 4

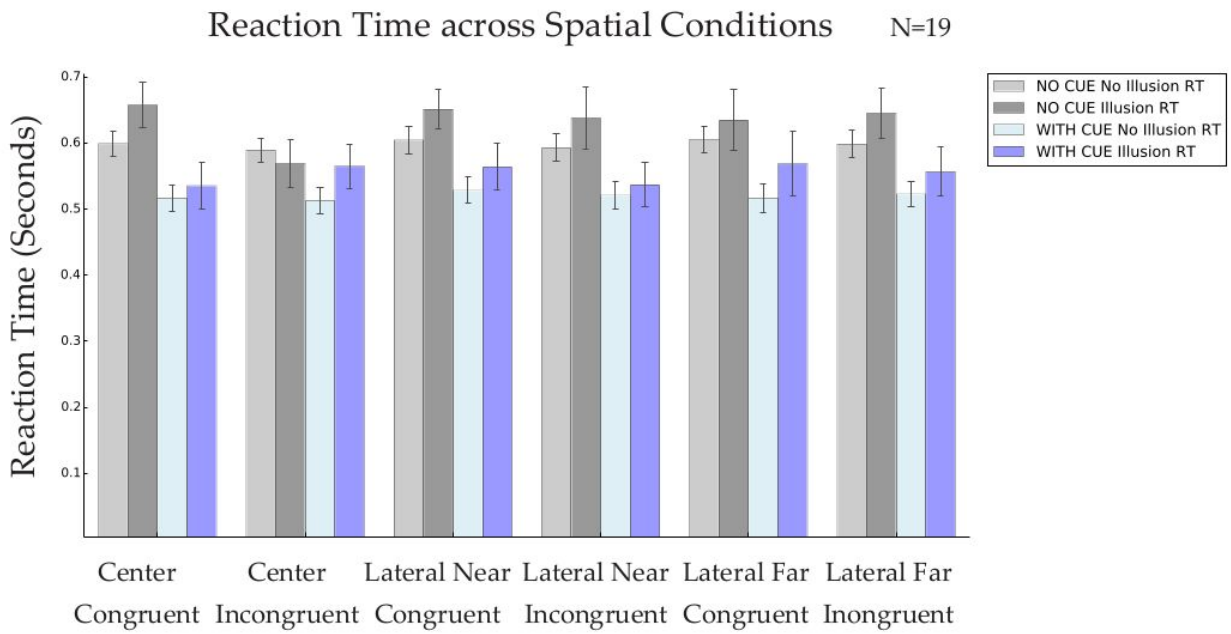


Figure 4: Averaged reaction time in milliseconds across all six spatial conditions for nineteen subjects. Grey bars representing trial types with no cue, blue with cue. Lighter bars indicate trials in which no illusion was reported.

B. Illusion model main effects

The collocation terms did not yield any significant effects; *TwoSoundsAtTargetLoc* describing trials when there were two sounds at the visual target location, one flash occurred, and attention was not cued, there was no significant effect (coefficient = -0.01 logit) on subjects pressing "2." This indicates that there was no increase in reported illusory percepts when the illusion inducing, or non-matching auditory stream, was colocated with a single physical flash.

The attention term *CueAttn* also did not show any significant effect (coefficient = 0.03 logit), confirming that attention cueing in and of itself does not influence the participants

pressing the "2" button. Interpreted against the entire model, this term indicates that when attention was cued, one flash occurred, and there was one sound at the target visual flash location, there was no increase in probability of participants pressing "2".

NonCentralFlash assesses when the flash was presented in the periphery, show that having the flash in the periphery made it slightly more likely (coefficient = 0.07 logit, $p<0.01$) for subjects to press "2" as compared to when the flash(es) occurred in the center position. In regards to the rest of the model it indicates that on trials when attention was not cued, there was one flash, and one sound at visual location, the subjects were slightly more likely to press "2". This is a small (but reliable) effect based on the effect size.

The term *NonCentralFlash_FarFoil*, representing proximity, was not significant (coefficient = 0.03 logit). Interpreted in full, this term represents the comparison between the visual in the periphery and the non-matching auditory stream was two spaces away (far) when compared with the foil being one space away (near), one physical flash occurred, attention was not cued, and one sound happened at the visual location. Surprisingly, this had no significant effect on the likelihood of participants pressing "2," and shows no significant effect of proximity of the illusion inducing auditory stream.

C. Illusion model interaction terms

The interaction of *TwoFlashes:CueAttn* term shows whether participants were significantly more likely (coefficient = 0.12 logit, $p<0.001$) to have an illusory response when attention was cued and the outcome indicates that subjects were more likely to press "2". Interpreted in full, *TwoFlashes:CueAttn* represents the trials in which two physical flashes

occurred, attention was cued, and there was one sound at the visual target location, participants were more likely to press "2.".

The interaction *TwoFlashes*:*NonCentralFlash* describes an decrease in fission illusions for trials in which the double flash occured in the periphery positions, and the results show that subjects were significantly less likely to press "2" (coefficient = -0.49 logit, $p<0.001$). The interaction term represents trials when two physical flashes occurred and they were in the periphery (as compared to the center), and attention was not cued and one sound occured in the visual target location.

The following interaction terms did not yield statistically significant results. The interaction term *TwoFlashes*:*NonCentralFlash_FarFoil*, indicates that holding all other coefficients constant, that proximity of the illusion inducing auditory stream did not affect illusory percepts. The *TwoFlashes*:*TwoSoundsAtTargLoc* term, indicates that holding all other coefficients constant, colocation of matching auditory and visual streams did not have a significant effect on which button was pressed and therefore no significant effect on the sound induced flash illusion percept.

Reaction time model data:

The reaction time model predicts the amount of time that it took the participant to make a response to the number of flash(es) that they saw by button press. Reaction time is depicted in milliseconds, and so are the numbers shown in the model output presented in **Table 3** (located at the end of the chapter for ease of reading).

Reaction time main effects:

The two illusion terms, *TwoFlashes* and *Pressed_Two* representing our fission and fusion illusions, showed that, holding all other coefficients constant, reaction times were slower than average when illusions were reported (20.56 ms and 25.54 ms slower respectively, $p<0.001$). While just the act of cueing attention (represented by the *CueAttn* term) to flash location makes subjects 32.92 ms faster at pressing the button.

In looking at the contrast terms, *NonCentralFlash* showed that participants were significantly slower than average to respond (21.56 ms slower, $p<0.001$) when the flash was in the peripheral location. Specifically, *NonCentralFlash* term shows that having the flash in the peripheral location (as compared to the center location) when one flash occurred, attention was not cued, subjects pressed "1", and one sound was at the visual target location, the participants were 21.56 ms slower to press the button. The other contrast term *NonCentralFlash_FarFoil* comparing the non-matching auditory being far or near did not have a significant effect on reaction time.

The main 'collocation' term *TwoSoundsAtTargetLoc* did not have a significant effect on reaction time. *TwoSoundsAtTargetLoc* represents trials having two events in the collocated location when one flash occurred, attention was not cued, and subjects pressed "1."

Reaction time interaction terms:

For brevity, the following are the significant interaction terms for the reaction time model, while the others being non-significant are represented in **Table 3**. Not surprisingly, cueing attention was the strongest indicator of faster reaction times with the following being significant terms.

Cueing attention produced significantly faster reaction times (17.55 ms faster than average, $p<0.001$) when the flash occurred in the peripheral locations. *CueAttn:NonCentralFlash* shows that when attention was cued to flash location and the flashes occurred in the periphery, as compared to the center location, and one flash occurred, the subjects pressed "1" then participants were 17.55 ms faster at pressing the response button.

Additionally, *TwoFlashes:CueAttn:TwoSoundsAtTargetLoc* indicates that participants were 21.61ms faster ($p<0.05$) than average at pressing the button on trials in which attention was cued, two flashes, and two beeps occurred at the same location as the flashes, but the subjects pressed "1". Shown by *TwoFlashes:CueAttn:Pressed_Two* term, participants were 22.71ms faster ($p<0.05$) at pressing the button when attention was cued to flash location, two flashes occurred, and the participants pressed "2", but one sound occurred at the visual target location. Participants were also significantly faster (by 23.65 ms, $p<0.05$) than average when attention was cued to flash location, two beeps occurred collocated with the visual, and subjects pressed "2" but there was one flash. This was indicated with the *CueAttn:TwoSoundsAtTargetLoc:Pressed_Two* interaction term.

However, there were some interaction terms with cued attention that did not show faster reaction times. *TwoFlashes:CueAttn* representing trials when two physical flashes occurred and we cued covert spatial attention, and subjects pressed "1" and one sound occurred at the visual target location, subjects were significantly (13.6 msec, $p<0.05$) slower than average in reaction time. Again this supports slower reaction times during an illusory percept, despite cueing attention. Also, strangely the term *TwoFlashes:CueAttn:TwoSoundsAtTargetLoc:Pressed_Two* indicates that participants were significantly slower (by 43.92 ms, $p<0.01$) than average when

two flashes occurred, attention was cued, two sounds occurred at the visual target location, and the subjects pressed “2”.

Interestingly, when subjects correctly reported two physical flashes, when there were two physical flashes present, even when the auditory non-matching stream was collocated with the visual flashes, represented by the *TwoFlashes:Pressed_Two* term, they were significantly faster (41.64 ms, $p<0.001$) than average in pressing the button.

IV. Discussion:

Overall the illusion model showed a significant effect on illusion reports being more likely when the visual stimulus was in the periphery, but no significant effect of the proximity of the auditory non-matching stream. While there was not a significant effect of proximity on participants' reaction times either, there did seem to be an effect of having the visual stimulus in the periphery. Specifically, participants were faster to respond when the visual stimulus was presented in the center location. When viewed in combination with the likelihood of illusory reports also changing when the flash(es) were in the periphery, there could have been an effect of visual acuity. This is in line with previous study showing that visual acuity was a large predictor of reporting the sound induced flash illusion (Kumpik et al., 2014).

Overall when an illusory percept was reported (either fission or fusion), the reaction time model revealed that reaction times were slower. The slower illusory reaction time could be seen as capturing a longer decision-making stage for pressing the button, and perhaps participants knowing that what they perceived was perhaps an illusion. Awareness of the illusory percept was not captured by our “1” or “2” button options, it is possible the participants true percept could

have been somewhere in between, as shown by a study that had a third button option representing a 'not 1 flash and not 2 flashes' (van Erp et al. 2013).

Not surprisingly, on trials where attention was cued, reaction times were faster, replicating the well-documented cueing phenomenon (Posner, 1980). The exceptions to this was seen when illusory percepts were reported *TwoFlashes:CueAttn*, showing that while knowing where the visual flash(es) would be presented provided a faster response, it was not enough to overcome the significant decrease in reaction time when participants reported seeing an illusion Additionally, the attentional cue did not have a significant effect on the likelihood of participants reporting an illusion. This was unexpected given previous evidence for spatial attention being a contributing factor to the sound induced flash illusion (Bizley et al., 2012). Nevertheless, there are a few differences between the study that found an effect of attention on SIFI and the present study. In Bizley et al (2012), there were two visual stimuli and two auditory stimuli on every trial, presented on the left/right hemisphere with no spatial arrangement ambiguity. This potentially allowed participants to completely ignore the un-attended hemisphere, and could represent the ability to ignore the left/right hemisphere and not necessarily a representation of the influence spatial collocation cues. While the current study was set up such that the non-collocated auditory stream spatial location was random and unknown to the participant, potentially making it more difficult for them to ignore as the participant would need to rely on precise spatial acuity to differentiate the auditory matching and auditory non-matching streams. Although in the current study the attentional cue indicating the location of the visual flash(es) for the trial was always valid, the participants were never told where the competing stimulus would be located on any given trial. This could represent a different effect of bottom-up (the current

study) and top-down (Bizley et al., 2012) attentional effects on spatial colocation and the probability of reporting an illusory trial.

Perhaps the most surprising finding of our study was that the colocation terms, in both the illusion and reaction time models, did not show any significant effect on participants reporting an illusory response. We had hypothesized that the spatial location of the auditory non-matching stream (illusion inducing stream) would have an effect on the illusion, and this study showed that neither the colocation or the proximity of the illusion inducing stream had a significant effect on reporting an illusion. This could be due to a couple of reasons. One possibility for not observing a spatial congruence effect is due to the auditory stimuli being very short (30 msec) in duration, and while the visual system has great spatial resolution that is perceived on very short time scales, the auditory system relies on spatial information that unfolds over time and it has been shown that localization abilities are more accurate with longer and more dynamic stimuli of at least 100 ms (Makous and Middlebrookes, 1990; Carlile et al., 1997; MacPherson & Middlebrooks, 2000). With our stimuli being only 30 ms in duration, the auditory spatial cues might not be salient enough to drive an effect of spatial congruence or proximity.

Another reason why we did not observe an expected spatial congruence effect is that we did not ask participants to attend to a particular auditory stimulus (either the noise burst or tone). While we originally hypothesized that having competing stimulus presented spatially apart would be sufficient to manipulate spatial attention, it may be that we need to direct top-down spatial attention to a particular desired target in order to see the benefits of spatial cues (Best et al., 2005; Alain & Arnott, 2000; Shinn-Cunningham, 2008).

While Bizley et al. did observe an effect of spatial arrangement on the SIFI and the current study does not find an effect of spatial congruence, there is a significant difference between the two studies that could have led to this difference in finding; Bizley et al. did not truly have spatial congruence between the auditory and visual stimuli. Instead, Bizley et al. placed the auditory stimuli at greater lateral positions (+30° and -30° relative to midline) than their visual stimuli (+10° and -10° relative to midline) in order to create a stronger hemispheric percept. However, in the current study auditory and visual stimuli could occur at the same azimuth locations and thus spatially congruent. This coupled with the ability to ignore or attend based on hemisphere and not spatial acuity, could account for the divergent findings.

Surprisingly, we did not find an effect of spatial colocation between auditory and visual stimuli on the probability of reporting an illusion. This further supports that the multisensory 'rule' on spatial congruence may have some caveats for human behavior. In alignment with Spence (2013), spatial colocation did not matter in a temporal task (report the number of flashes). However, a major difference between what Spence proposed, is that spatial attention is the determinate for the spatial 'rule' to hold in behavioral studies, and what we report here, is that even when visual spatial attention is directed we did not find an effect of spatial congruence. This suggests that there is a difference in bottom-up versus top-down attention on the spatial rule. In the current study, while attention was directed toward the visual flash(es) location, there was no prior knowledge of the spatial arrangement of the competing auditory streams. While traditionally, bottom-up type multisensory integration effects on spatial colocation are reported, perhaps as the origin of the spatial rule was described in anesthetized cats (Meredith & Stein, 1986), does not hold for conditions of stimulus competition. Further studies are needed to help

delineate the effects of bottom-up and top-down spatial attention on the spatial rule during stimulus competition.

Table 3:
Results table for Reaction Time model: *The table shows the output of the linear regression model for predicting the reaction time data.*

Fixed effects	Estimate	Std. Err	tvalue	Chisq	pval
(Intercept)	589.72843	15.49761	38.05		
TwoFlashes	20.56062	4.43056	4.64	21.53	*** <.0001
CueAttn	-32.92222	2.3066	-14.27	203.31	*** <.0001
NonCentralFlashLoc	21.56179	3.46932	6.21	38.61	***<.0001
NonCentralFlash_FarFoil	1.06189	4.03628	0.26	0.07	0.79
TwoSoundsAtTargetLoc	2.64294	2.32199	1.14	1.3	0.26
Pressed_Two	25.53743	6.20132	4.12	16.96	*** <.0001
TwoFlashes:CueAttn	13.6261	6.67503	2.04	4.17	* .04
TwoFlashes:NonCentralFlashLoc	-7.64275	10.14895	-0.75	0.57	0.45
TwoFlashes:NonCentralFlash_FarFoil	-1.22929	9.69234	-0.13	0.02	0.9
cue_attnTRUE:NonCentralFlashLoc	-17.55247	4.87748	-3.6	12.95	***< .0001
cue_attnTRUE:NonCentralFlash_FarFoil	-0.09077	5.66227	-0.02	0	0.99
TwoFlashes:TwoSoundsAtTargetLoc	9.06135	6.17244	1.47	2.16	0.14
CueAttn:TwoSoundsAtTargetLoc	-0.68124	3.25781	-0.21	0.04	0.83
NonCentralFlashLoc:TwoSoundsAtTargetLoc	-7.53589	4.90549	-1.54	2.36	0.12
NonCentralFlash_FarFoil:TwoSoundsAtTargetLoc	-1.01529	5.71076	-0.18	0.03	0.86
TwoFlashes:Pressed_Two	-41.63571	7.6348	-5.45	29.73	***<.0001
CueAttn:Pressed_Two	2.49618	8.46899	0.29	0.09	0.77
NonCentralFlashLoc:Pressed_Two	19.38944	13.20618	1.47	2.16	0.14
NonCentralFlash_FarFoil:Pressed_Two	23.38212	15.01243	1.56	2.43	0.12
TwoSoundsAtTargetLoc:Pressed_Two	11.55114	8.65	1.34	1.78	0.18
TwoFlashes:CueAttn:NonCentralFlashLoc	0.4983	15.68793	0.03	0	0.97
TwoFlashes:CueAttn:NonCentralFlash_FarFoil	16.65264	14.26835	1.17	1.36	0.24
TwoFlashes:CueAttn:TwoSoundsAtTargetLoc	-21.609	9.29663	-2.32	5.4	* .02
TwoFlashes:NonCentralFlashLoc:TwoSoundsAtTargetLoc	3.53032	14.17804	0.25	0.06	0.8

TwoFlashes:NonCentralFlash_FarFoil:TwoSoundsAtTargetLoc	5.7497	13.75514	0.42	0.17	0.68
CueAttn:NonCentralFlashLoc:TwoSoundsAtTargetLoc	5.11871	6.88903	0.74	0.55	0.46

CueAttn:NonCentralFlash_FarFoil:TwoSoundsAtTargetLoc	-7.27215	8.00529	-0.91	0.83	0.36
TwoFlashes:CueAttn:Pressed_Two	-22.70592	10.7904	-2.1	4.43	* .04
TwoFlashes:NonCentralFlashLoc:Pressed_Two	-13.50846	16.68219	-0.81	0.66	0.42
TwoFlashes:NonCentralFlash_FarFoil:Pressed_Two	-21.82871	17.94151	-1.22	1.48	0.22
CueAttn:NonCentralFlashLoc:Pressed_Two	-2.95113	18.18063	-0.16	0.03	0.87
CueAttn:NonCentralFlash_FarFoil:Pressed_Two	-1.32785	20.43445	-0.06	0	0.95
TwoFlashes:TwoSoundsAtTargetLoc:Pressed_Two	-16.45298	10.65702	-1.54	2.38	0.12
CueAttn:tTwoSoundsAtTargetLoc:Pressed_Two	-23.65267	12.02548	-1.97	3.87	*.05
NonCentralFlashLoc:TwoSoundsAtTargetLoc:Pressed_Two	12.14421	18.50684	0.66	0.43	0.51
NonCentralFlash_FarFoil:TwoSoundsAtTargetLoc:Pressed_Two	-30.12766	21.00501	-1.43	2.06	0.15
TwoFlashes:CueAttn:NonCentralFlashLoc:TwoSoundsAtTargetLoc	11.6534	21.65726	0.54	0.29	0.59
TwoFlashes:CueAttn:NonCentralFlash_FarFoil:TwoSoundsAtTargetLoc	0.70919	20.29388	0.03	0	0.97
TwoFlashes:CueAttn:NonCentralFlashLoc:Pressed_Two	30.69061	24.04317	1.28	1.63	0.2
TwoFlashes:CueAttn:NonCentralFlash_FarFoil:Pressed_Two	-27.76716	25.00698	-1.11	1.23	0.27
TwoFlashes:CueAttn:TwoSoundsAtTargetLoc:Pressed_Two	43.92347	15.23419	2.88	8.31	** <.01
TwoFlashes:NonCentralFlashLoc:TwoSoundsAtTargetLoc:Pressed_Two	-13.30758	23.35784	-0.57	0.32	0.57
TwoFlashes:NonCentralFlash_FarFoil:TwoSoundsAtTargetLoc:Pressed_Two	25.37798	25.2113	1.01	1.01	0.31
CueAttn:NonCentralFlashLoc:TwoSoundsAtTargetLoc:Pressed_Two	-1.62069	25.90758	-0.06	0	0.95
CueAttn:NonCentralFlash_FarFoil:TwoSoundsAtTargetLoc:Pressed_Two	14.40582	28.99041	0.5	0.25	0.62
TwoFlashes:CueAttn:NonCentralFlashLoc:TwoSoundsAtTargetLoc:Pressed_Two	-12.47996	33.81199	-0.37	0.14	0.71
TwoFlashes:CueAttn:NonCentralFlash_FarFoil:TwoSoundsAtTargetLoc:Pressed_Two	-1.86976	35.50741	-0.05	0	0.96

References:

Alais, D., Newell, F., & Mamassian, P. (2010). Multisensory processing in review: from physiology to behaviour. Seeing and perceiving, 23(1), 3-38.

Alain, C., & Arnott, S. R. (2000). Selectively attending to auditory objects. Front. Biosci, 5, D202-D212.

Algazi, V. R., Duda, R. O., Thompson, D. M., & Avendano, C. (2001, October). The cipic hrtf database. In Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575) (pp. 99-102). IEEE.

Best, V., Carlile, S., Jin, C., & van Schaik, A. (2005). The role of high frequencies in speech localization. The Journal of the Acoustical Society of America, 118(1), 353-363.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

Bertelson, P., & Radeau, M. (1981). Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. Perception & psychophysics, 29(6), 578-584.

Bizley, J. K., Shinn-Cunningham, B. G., & Lee, A. K. (2012). Nothing is irrelevant in a noisy world: sensory illusions reveal obligatory within-and across-modality integration. Journal of Neuroscience, 32(39), 13402-13410.

Cappelloni, M. S., Shivkumar, S., Haefner, R. M., & Maddox, R. K. (2019). Task-uninformative visual stimuli improve auditory spatial discrimination in humans but not the ideal observer. PloS one, 14(9).

Carlile, S., Leong, P., and Hyams, S. (1997). ‘‘The nature and distribution of errors in sound localization by human listeners,’’ Hear. Res. 114, 179–196.

Colonius, H., & Diederich, A. (2004). Multisensory interaction in saccadic reaction time: a time-window-of-integration model. Journal of cognitive neuroscience, 16(6), 1000-1009.

Colonius, H., & Diederich, A. (2006). The race model inequality: interpreting a geometric measure of the amount of violation. Psychological review, 113(1), 148.

Colonius, H., & Diederich, A. (2011). Computing an optimal time window of audiovisual integration in focused attention tasks: illustrated by studies on effect of age and prior knowledge. Experimental Brain Research, 212(3), 327-337.

Colin, C., M. Radeau, P. Deltenre & J. Morais. 2001. Rules of intersensory integration in spatial scene analysis and speechreading. Psychol. Belg. 41: 131–144.

Di Luca, M., T.-K. Machulla & M.O. Ernst. 2009. Recalibration of multisensory simultaneity: cross-modal transfer coincides with a change in perceptual latency. J. Vision 9: 1–16.

Doyle, M.C. & R.J. Snowden. 2001. Identification of visual stimuli is improved by accompanying auditory stimuli: the role of eye movements and sound location. Perception 30: 795–810.82.

Gondan, M., B. Niederhaus, F. R¨osler & B. R¨oder. 2005. Multisensory processing in the redundant-target effect: a behavioral and event-related potential study. Percept. Psychophys. 67: 713–726.

Harrington, L.K. & C.K. Peck. 1998. Spatial disparity af- fects visual-auditory interactions in human sensorimotor processing. Exp. Brain Res. 122: 247–252.

Hillock, A. R., Powers, A. R., & Wallace, M. T. (2011). Binding of sights and sounds: age-related changes in multisensory temporal processing. Neuropsychologia, 49(3), 461-467.

Hillock-Dunn, A., & Wallace, M. (2012). Hearing with your eyes: implications of audiovisual processing development on speech perception in noise. Audiology Today, 24(3), 40-45.

Innes-Brown, H. & D. Crewther. 2009.The impact of spatial incongruence on an auditory-visual illusion. PLoS One 4: e6450.90.

Keetels, M. & J. Vroomen. 2005. The role of spatial disparity and hemifields in audio-visual temporal order judgments. Exp. Brain Res. 167: 635–640.

Keetels, M. & J. Vroomen. 2007. No effect of auditory-visual spatial disparity on temporal recalibration. Exp. Brain Res. 182: 559–565.95.

Keetels, M., Stekelenburg, J., & Vroomen, J. (2007). Auditory grouping occurs prior to intersensory pairing: evidence from temporal ventriloquism. Experimental Brain Research, 180(3), 449.

Keetels, M. & J. Vroomen. 2008. Tactile-visual temporal ventriloquism and the effect of spatial disparity. Percept. Psychophys. 70: 765–771.

Kidd Jr, G., Arbogast, T. L., Mason, C. R., & Gallun, F. J. (2005). The advantage of knowing where to listen. The Journal of the Acoustical Society of America, 118(6), 3804-3815.

Kumpik, D. P., Roberts, H. E., King, A. J., Bizley, J. K., & G., R. (2014). Visual sensitivity is a stronger determinant of illusory processes than auditory cue parameters in the sound-induced flash illusion. Journal of Vision, 14(7), 12–12.

Frassinetti, F., N. Bolognini & E. L`adavas. 2002. Enhancement of visual perception by crossmodal visuo-auditory interaction. Exp. Brain Res. 147: 332–343.

Fiebelkorn, I.C., J.J. Foxe, J.S. Butler & S. Molholm. 2011. Auditory facilitation of visual-target detection persists re- gardless of retinal eccentricity and despite wide audiovisual misalignments. Exp. Brain Res. 213: 167–174.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2001). Spatial release from informational masking in speech recognition. The Journal of the Acoustical Society of America, 109(5), 2112–2122.

Lovelace, C. T., Stein, B. E., & Wallace, M. T. (2003). An irrelevant light enhances auditory detection in humans: A psychophysical analysis of multisensory integration in stimulus detection. Cognitive Brain Research, 17(2), 447–453.

Leo, F., V. Romei, E. Freeman, et al. 2011. Looming sounds enhance orientation sensitivity for visual stimuli on the same side as such sounds. Exp. Brain Res. 213: 193–201.

Lee, A. K., Maddox, R. K., & Bizley, J. K. (2019). An Object-Based Interpretation of Audiovisual Processing. In Multisensory Processes (pp. 59-83). Springer, Cham.

Maddox, R. K., & Shinn-Cunningham, B. G. (2012). Influence of task-relevant and task-irrelevant feature continuity on selective auditory attention. Journal of the Association for Research in Otolaryngology : JARO, 13(1), 119–129.

Maddox, R. K., Atilgan, H., Bizley, J. K., & Lee, A. K. (2015). Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. ELife, 2015(4), 1–11.

MacPherson, E. A., & Middlebrooks, J. C. (2000). Localization of brief sounds: effects of level and background noise. The Journal of the Acoustical Society of America, 108(4), 1834-1849.

Makous, J. C., and Middlebrooks, J. C. (1990). ‘‘Two-dimensional sound localization by human listeners,’’ J. Acoust. Soc. Am. 87, 2188–2200.

Meredith, M.A. & B.E. Stein. 1986. Spatial factors determine the activity of multisensory neurons in cat superior colliculus. Brain Res. 365: 350–354.12.

Meredith, M.A., J.W. Nemitz & B.E. Stein. 1987. Determinants of multisensory integration in superior colliculus neurons. I: temporal factors.J. Neurosci.7: 3215–3229. 16.

Ma, W.J. & A. Pouget. 2008. Linking neurons to behavior in multisensory perception: a computational review. Brain Res. 1242: 4–12.23.

Posner, M. I. (1980). Orienting of attention. The Quarterly Journal of Experimental Psychology, 32(1), 3–25.

Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive Brain Research*, *14*(1), 147–152.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. Trends in Cognitive Sciences, 12(5), 182–186.

Spence, C. & J. McDonald. 2004. "The crossmodal consequences of the exogenous spatial orienting of attention." In The Handbook of Multisensory Processing.G.A.Calvert,C. Spence & B.E. Stein, Eds.: 3–25. Cambridge: MIT Press.

Spence, C. (2013). Just how important is spatial coincidence to multisensory integration? Evaluating the spatial rule. Annals of the New York Academy of Sciences, 1296(1), 31–49. https://doi.org/10.1111/nyas.12121

Tiippana, K., H. Puharinen, R.Mottonen & M. Sams. 2011. Sound location can influence audiovisual speech perception when spatial attention is manipulated. SeeingPerceiving 24: 67–90.

Olivers, C. N. L., & Van der Burg, E. (2008). Bleeping you out of the blink: Sound saves vision from oblivion. Brain Research, 1242, 191–199.

Van Erp, J. B. F., Philippi, T. G., & Werkhoven, P. (2013). Observers can reliably identify illusory flashes in the illusory flash paradigm. Experimental Brain Research, 226(1), 73–79.

Wallace, M. T., & Stevenson, R. A. (2014). The construct of the multisensory temporal binding window and its dysregulation in developmental disabilities. Neuropsychologia, 64, 105–123.

Zampini, M., D. Torresan, C. Spence & M. M. Murray. 2007. Audiotactile multisensory interactions in front and rear space. Neuropsychologia 45: 1869–1877.

Chapter 3: Assessing the ventriloquist effect in elevation with personalized and generic HRTFs

Authors: Lindsey R. Kishline, Philip W. Robinson, Adrian KC Lee

I. Introduction

In order to localize sound in space, our brain takes advantage of various acoustical cues. Interaural level differences (ILDs) and interaural time differences (ITDs) are binaural cues, derived from the differences in level or time of arrival of the acoustic signals across the two ears. Spectral cues are monaural cues that result from sound interacting with the pinnae, head and torso before reaching the ear canals (Butler, 1977; Oldfield & Parker, 1984; Wenzel et al., 1993; Blauert, 1997). While localization of sound in azimuth primarily relies on ITD and ILD cues, auditory localization in the vertical plane (or elevation) relies on spectral cues (Blauert, 1997; Middlebrooks & Green, 1991; Van Opstal, 2016).

The geometry of each person's head, torso and pinna is unique, which results in a unique filter that is applied to the sound reaching their ear canals. The auditory system then utilizes these unique directionally dependent cues to calculate an auditory source's angle (Wenzel et al., 1993; Blauert, 1997). These personalized cues (or filters) are encapsulated in individualized Head-Related Transfer Functions (HRTFs). Obtaining individualized HRTFs requires a lengthy process of measurements and they can be acquired by many different HRTF measurement methods. Generally, it requires an individual sitting in the middle of an anechoic chamber whilst thousands of source directions are recorded by in-ear microphones that are specialized and usually expensive (Ben-Hur et al., 2018; Algazi et al., 2001). These methods also take

considerable time and effort to calculate. In applications such as Virtual Reality (VR) or Augmented Reality (AR) where the accurate spatialization of audio signals can play an important role, these processes can become incredibly computationally expensive to dynamically render each person's individualized HRTFs. Furthermore, the task of calculating and rendering individualized HRTFs for each headset user may also be an untenable task. Therefore, in place of individual's unique HRTFs, generic HRTFs are often used instead. Generic HRTFs are typically measured with a Knowles Electronics Mannequin for Acoustic Research or KEMAR mannequin (Gardner & Martin, 1995). These generic HRTFs can then be pre-rendered or calculated which saves significant computational power and creates an 'out-of-the-box' experience for the user of a VR head mounted device (HMD).

While generic HRTFs have been shown to create a sense of auditory space relatively accurately in azimuth locations, they do not provide reliable spatial cues for auditory sources in elevation - particularly below the horizon (Wenzel et al., 1993). Auditory localization tasks that have compared stimuli spatialized with generic and individualized HRTFs have shown increased confusion over auditory source location (Wenzel et al., 1993), and an increase in the magnitude of source localization errors (Middlebrookes, 1999). These findings provide justification for using individualized HRTFs for audio rendering applications, where elevation is important for immersive games.

In reality (as well as in VR and AR), we often experience the world through multiple senses. Is it possible that the high spatial resolution and acuity of our visual system compensate for the poorer accuracy in spatialization of auditory stimuli using generic HRTF? An observer often perceives an auditory stimulus as emanating from a visual object's location even when the

auditory stimulus is physically presented at a separate location in space. We experience this phenomenon in everyday life, e.g., when watching a puppet show (a ventriloquist puppet), someone giving a talk on a stage, or even at the movie theater (Connor, 2000; Alais & Burr, 2004). This is the classical ventriloquist effect in which a visual stimulus appears to 'capture' the auditory spatial location. Perceptually, this is the result of the near-optimal combination of auditory and visual spatial cues where in each modality spatial cues are weighted by "the inverse estimate of noisiness" (Cochran, 1937). In fact, it has been shown that if the auditory spatial cues are made more reliable than the visual spatial cues, the auditory stimulus could appear to 'capture' the visual stimulus in reverse (Alais & Burr, 2004).

While the ventriloquist effect has been well characterized in the horizontal plane (Howard & Templeton, 1966; Bertelson & Radeau, 1981; Bertelson et al., 2000; Bermant & Welch, 1976; Radeau, 1974; Warren, 1979; Weerts & Thurlow, 1971; Wallace et al., 2004; Slutsky & Recanzone, 2001), it has only been investigated in the vertical plane in a handful of studies (Thurlow & Jack, 1973; Hendrickx et al., 2015). Thurlow and Jack (1973) used a television and a speaker separated by 55 degrees, and asked participants if they heard the speech coming from the speaker or television. They reported ventriloquist effects to be greater in the vertical plane than the horizontal. Hendrickx et al. (2015), looked at vertical ventriloquism using video of a person speaking and again showed that the maximum separation between video and speaker location could be larger in elevation than in azimuth, before the ventriloquist effect breaks.

Using the ventriloquist after-effect is another way to study the influence of visual stimulus on auditory localization. Specifically, after repeated exposure to a visual and auditory

stimulus that are spatially disparate, ventriloquist after-effect studies investigate whether the perception of an auditory stimulus is biased by that same spatial mislocation (Bertelson et al., 2006; Wozny & Shams, 2011). Berger et al. (2018) showed that a repeated presentation of a spatio-temporally aligned visual stimulus paired with generic HRTFs auditory stimulus provided an auditory localization improvement. Unfortunately, they did not directly measure differences in the visual spatial capture between generic and individualized HRTFs. The auditory localization accuracy improvement they reported for generic HRTF localization in this study highlights the plastic nature of our auditory system (Shinn-cunningham et al., 1998). However, using the principle of ventriloquist after-effect to provide improvement of generic HRTF localization may not be appropriate for an 'out-of-the-box' ready-to-go VR or AR experience -- it can take an undefined amount of time for spatial re-calibration and it is not clear for how long these improvements last.

To our knowledge, there has yet to be an investigation into the ventriloquism effect in the vertical plane with various HRTFs. This is particularly interesting because auditory elevation perception relies on specific spectral information to determine sound source direction. Thus comparing spectral cues that are specific to the participant (individualized HRTF) and supposedly create the strongest auditory spatial reliability, with ones that are intended to be general (generic HRTF), should lead to differing amounts of visual spatial cue influence (according to the MLE - Alais & Burr, 2004) but has yet to be tested. We employed an absolute localization task rather than a relative localization task as is typically seen in ventriloquist studies. While the relative localization tasks provide an insight into the sensitivity of participants to a particular cue, the absolute localization tasks allow for an unrestricted assessment along

more dimensions about an external sound source and allows for more natural localization behavior (Middlebrooks & Green, 1991). Specifically, the current study investigated whether the degree of visual spatial capture (or ventriloquist effect) of an auditory stimulus in the vertical plane changes depending on whether the audio was spatialized with a generic or individualized HRTF. We compared participants' ability to localize auditory stimuli both with the auditory and visual stimuli being spatially congruent, as well as spatially incongruent with stimuli of up to 60 degrees separation in the vertical plane. While most ventriloquist studies use speech and faces or short stimuli such as a white noise burst, the current study utilizes longer dynamic stimuli, which has been shown to create a compelling perception of correspondence as in previous ventriloquist experiments (Vatakis & Spence, 2007; Welch & Warren, 1980; Jackson, 1953).

Given previous findings of localization accuracy differences based on HRTF representation (Wenzel et al., 1993; Middlebrooks, 1999), we expected that in the auditory unimodal conditions we would see a less accurate localization performance when using generic HRTFs. We also anticipated an elevated auditory perception with the generic HRTFs, as participants traditionally are poor at providing source perceptions below the horizon (Begault et al., 2001). Additionally, we hypothesized that there would be more visual spatial capture (ventriloquist effect) in the audio-visual when stimuli were spatially incongruent for those trials spatialized with generic HRTFs as the spatial cue reliability for those trials should be less than for those cues provided by individualized HRTFs.

For clarity, localization results and discussion will be presented by grouping conditions in which listeners were presented with no conflicting cues (Experiment 1) followed by conditions in which they were faced with conflicting cues (Experiment 2).

II. **Methods**

A. Participants

A total of 30 adult listeners (6 females, with a mean age of 35 and age range of 18 to 54) participated in this experiment. Of the total participants, 18 were practiced listeners and 12 were non-practiced listeners. Practiced listeners were defined as having more than two hours experience listening to their own individualized HRTFs (regardless of how that individualized HRTF was generated). Every participant had an audiogram to verify normal hearing thresholds of less than or equal to 20 dB hearing level at octave frequencies between 250 Hz and 8000 Hz. Seven of the participants were excluded from the study for the following reasons; inability to perform well on the audio only training procedure, inability to distinguish auditory stimuli in azimuth, unable to locate the visual stimulus, hand motor coordination issues for reliability placing the response circle at the desired location, which was indicated via the visual only condition and reviewed post-experiment. All participants additionally went through a scanning process to capture their individualized HRTF data (details in the following section). All participants were run under an approved protocol by the Internal Review Board at the Facebook Reality Labs, where the data were collected. Participants gave written informed consent to the experimental session and were provided compensation for their time.

B Individualized HRTF acquisition

Participants were seated inside an acoustic chamber while their individualized HRTFs were measured. Audio sweeps were played from different directions using a movable speaker. These stimuli were recorded via a pair of in-ear microphones to capture the resulting acoustic response at the entrance of the ear canal.

C Stimuli

The stimuli were presented in a Virtual Reality Unity environment using Unity 3D software (version 2018.3.8f1) that was a large dark grey colored dome, with a large yellow plane positioned approximately 2 meters from the person, in which both the visual and auditory stimuli were displayed upon (details provided below).

C.1 Auditory stimuli

The auditory stimulus, generated at a sampling rate of 24.414 kHz, consisted of pink noise generated by MATLABs R2018b DSP Toolbox 'ColoredNoise' function. The pink noise was amplitude modulated at 4 Hz, with onset and offset ramps of 10 ms. The maximum total length of the auditory stimulus was 3 seconds. The audio stimulus was spatialized using either generic HRTFs or the participant's individualized HRTFs, and combined with the known head position of the participant from the HMD (see Hardware below for details).

C.2 Visual stimuli

The visual stimulus was a white 100% contrast Gaussian blob that changed in diameter along with the amplitude modulation of the auditory stimulus. The maximum size of the blob subtended 5.5 degrees of visual field, and a minimum size of 2.7 degrees. The average size of the visual stimulus subtended roughly 4 degrees of visual field. This was done by changing the diameter of the visual blob with the amplitude envelope of the auditory stimulus. The visual stimuli were presented by the HTC Vive Headset with a refresh rate of 90 Hz.

D. Hardware

Visual stimuli were presented using an HTC Vive Head Mounted Device (HMD) with a 110 degree field of view with a refresh rate of 90 Hz. The HMD had a Tobii eye tracker installed

and was equipped with a head position tracking system that tracked real time position of both the HMD and the Vive handset controllers using LIDAR technology with sub-millimeter precision. Audio stimuli were presented over Beyerdynamic 990's open ear headphones through a RME BabyFace Pro. The eye tracking data was used to ensure that participants did not close their eyes during the trials.

E. Experimental Conditions

The experimental session was performed on a separate day from the participant's HRTF measurement session. The participants went through four phases in this testing session: i) an Auditory only training phase, ii) a Visual only phase, iii) an Auditory only experimental phase; iv) an Audio-Visual experimental phase. In the Auditory only training phase, the auditory stimulus was presented alone without the visual stimulus and participants received feedback on each trial about the location of the auditory stimulus. In the Visual only phase, only the visual stimulus was presented and the participants were asked to localize the Gausian blob that changed in diameter but without sound. In the Auditory only experimental phase, participants were asked to localize the sound source but they received no feedback. Finally, in the Audio-Visual experimental phase, the auditory and the visual stimuli were presented simultaneously and no feedback was provided.

The auditory source positions that were used in the auditory training phase were not the same positions as the experimental phases. Instead, twelve unique locations were used that spanned both hemispheres and approximately the same amount of elevation changes above and below the horizon. Additionally the training phase used only the participants individualized

HRTFs, as the majority of the participants may not be used to hearing a personalized HRTF. No training was done on the generic HRTFs as is the case for a new VR user.

There were 14 possible spatial locations that the auditory and visual stimuli could appear (see Figure 1): 0, ±10, ±20, ±30 degrees in elevation. The stimuli were also offset in azimuth by ±15 degrees. Lateral offset was used because externalization along the median plane is poor (Blauert, 1997). The auditory and visual stimuli were always located within the same hemisphere (either both positive or negative 15 degrees azimuth). All 14 locations were counterbalanced, and thus the auditory and visual stimuli could be collocated or separated up to 60 degree in elevation. Six repetitions were performed for each spatial combination, for each experimental phase, and for both generic and individualized HRTFs.

In the Audio-Visual experimental phase and the Auditory only experimental phase, the generic and individualized HRTFs were split into separate blocks of trials (counterbalanced across participants) with 92 trials in each block, and a total of two blocks for the Auditory only phase and a total of 6 blocks for the Audio-Visual phase. Participants were prompted to take rest breaks in between blocks. The entire experimental session lasted on average 50 minutes to one hour.

Figure 1

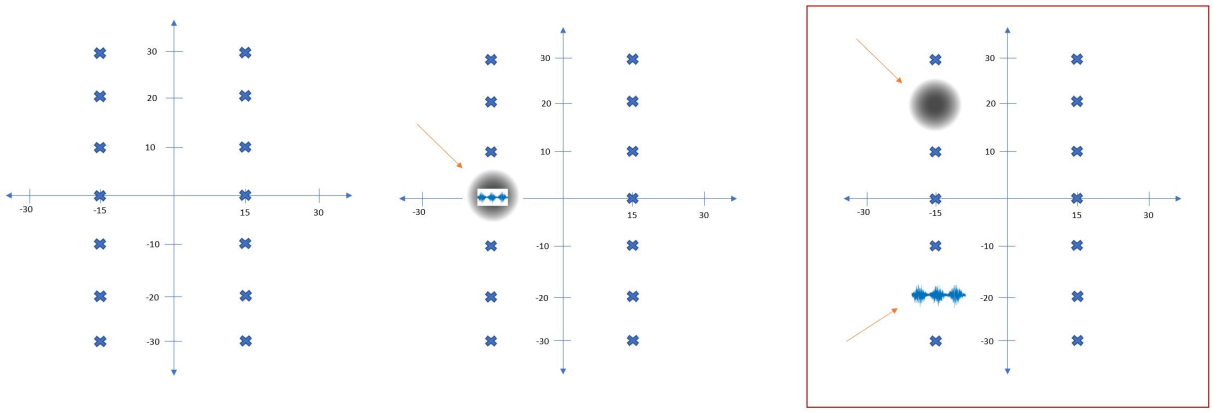


Figure 1: A. Indicates the 14 possibile locations that audio and/or visual stimuli can occur. B. Indicates an example of non-conflicting (non-ventriloquist) trial type. C. Example of a trial type with conflicting spatial cues.

F. Calibration Procedure

The participant was seated on a stool and asked to don an HTC Vive Pro with integrated Tobii eye tracking, which was adjusted for appropriate fit and comfort. The Vive had the original headphones removed and the above mentioned Beyerdynamic 990's headphones were used to deliver the auditory stimuli. The participant was asked to place their chin on a chin rest and handed a Vive controller. Then prior to the start of the experimental session an Tobii eye calibration was performed and insured successful before proceeding to the task. Head movements were restricted by the chin rest and maintained by the participant through a red laser dot that extended onto the yellow plane. Participants were asked to keep this red dot inside a small green circle on the plane directly in front of them. The trial would not start unless the red dot was within the green circle bounds. Participants were instructed to begin each trial with their eyes pointing straight ahead and to move their eyes to the visual blob in the Audio-Visual trials, or to keep their eyes open during the audio only trials. Eye tracking was used to ensure that participants kept their eyes open and looked at the blob during the Audio-Visual trials; trials were removed if participants did neither.

G. Trial structure

At the beginning of every trial, the participant was required to position ensured that their head position was such that the red head position dot was in the center of the green ring, then

they pressed the large Vive controller button pad to begin the trial. Once the trial started the stimuli would continue for up to 3 seconds before the trial ended. The participant's task on every trial was to locate the auditory source by pointing the laser pointer (coming from the end of the Vive hand controller) to intersect the yellow plane with where they wanted to lock in their response. If they answered within the 3 second time limit, a green circle would appear where they responded (indicated the location where the auditory stimuli was being played from). Once they were ready, with their head aligned, they would continue to the next trial.

H. Deriving slopes and intercepts to summarize localization results

For all the data presented below, we have collapsed across azimuth, as there were no statistical differences in localization for sounds ± 15° laterally, and taken the elevation position response coordinate for each trial type. The median elevation response was calculated for each trial type for each participant. The median elevation responses were then plotted such that the horizontal axis represents the actual stimulus location in elevation, and the vertical axis as representing the participants elevation response. In this representation, if the participant had perfect localization in elevation, all of their median responses would fall along the diagonal line.

To create a consistent metric(s) to assess localization ability across all unimodal and multimodal conditions, linear fit lines were estimated for each participant on the median responses per trial type within each unimodal conditions (audio and visual only), audio-visual collocated conditions, and linear fits for each visual spatial position in the audio-visual incongruent conditions. Linear fits were estimated for both generic and individualized HRTF audio trial types. This nets a total of 19 linear fits per participant: one for the unimodal visual

trials, two for the unimodal auditory trials, two for the audio-visual collocated trials, and 14 for the audio-visual spatially incongruent trials (discussed in Experiment 2).

Each linear fit summarizes the data collected from sound presented across the 7 elevation angles with two parameters (slope and intercept) for every participant for every unimodal and multimodal condition. The slope provides a measure of how accurate participants were in localization, as a slope of 1 would represent veridical localization responses falling along the diagonal line in each plot described above. The intercept provides an estimate of how elevated (with respect to the horizon) the localization responses were for each participant; as intercepts that do not pass through zero would indicate that the perception of the stimulus was either elevated above (positive intercept) or below (negative intercept) the horizon. Using these two parameters, we compared audio localization ability for conditions that had no conflicting spatial cues.

III. Experiment 1: NO CONFLICTING CUES

Here we examine localization ability in trials that have no conflicting spatial cues which include Visual only, Auditory only, and Audio-Visual trials in which the auditory and visual stimuli are presented at the same spatial location in both azimuth and elevation. Localization results are analyzed based on the slopes and intercepts estimated using the procedure described in Section II.H *Deriving slopes and intercepts to summarize localization results*. Localization is compared between auditory stimuli spatialized with generic and individualized HRTFs.

A. Results:

A.1.General localization errors in elevation:

Participants were highly accurate in localizing the visual stimulus and had a very small standard deviation in slope (Mean= 0.996, STD = 0.016) and intercept (Mean = 0.046, STD = 0.294). The group average of those visual only slopes and intercepts (∓1 standard deviation) are shown in Figure 2.

The average absolute elevation localization error in the Auditory only condition spatialized with generic HRTFs was mean = 15° with a STD = 7.38°, while when spatialized with individualized HRTFs mean = 15.3° and a STD = 7.35°. The average absolute elevation localization error in the Audio-Visual Collocated condition was mean = 6.78° and STD = 6.31° when spatialized with generic HRTFs, and mean = 6.43° and STD = 5.73° when spatialized with individual HRTFs.

A.2. Analysis of Variance (ANOVAs):

Table 1 [ANOVA results]

Slope ANOVA	**Df**	**Sum Sq**	**Mean Sq**	**F value**	**P value**
HRTF(Gen v Indv)	1	0.033	0.0332	0.626	0.431
AudioOnly v Audio-Visual	1	2.511	2.5107	47.354	<0.001***
Residuals	89	4.719	0.0530		

Intercept ANOVA	**Df**	**Sum Sq**	**Mean Sq**	**F value**	**P value**
HRTF(Gen v Indv)	1	5.9	5.9	0.315	0.576
AudioOnly v Audio-Visual	1	620.3	620.3	32.848	<0.001***
Residuals	89	1680.7	18.9		

Table A shows the ANOVA results of slope, Table B shows the ANOVA results of intercept.

Comparing slopes of the linear fits, a two-way ANOVA was conducted to examine the effect of HRTF (Generic, Individual) and the effect of with or without non-conflicting visual cue (Auditory only, Audio-Visual Collocated). . The main effect of HRTF was found not to be statistically significant $F(1,89) = 0.626$, $p = 0.431$. However, the main effect of the presence of a spatially congruent visual stimulus present was statistically significant $F(1,89) = 47.354$, $p < 0.001$ (see Figure 2, left).

Comparing the intercepts of the linear fits, a two-way ANOVA was also conducted to examine the effect of HRTF and non-conflicting sensory cues. Similar to the analysis for the slopes, the main effect of HRTF on the intercept of the linear fit lines was found not to be significant $F(1,89) = 0.315$, $p = 0.576$. However, the main effect of the presence of a spatially congruent visual stimulus was again found to be significant $F(1,89) = 32.848$, $p < 0.001$. S (see Figure 2, right).

Figure 2

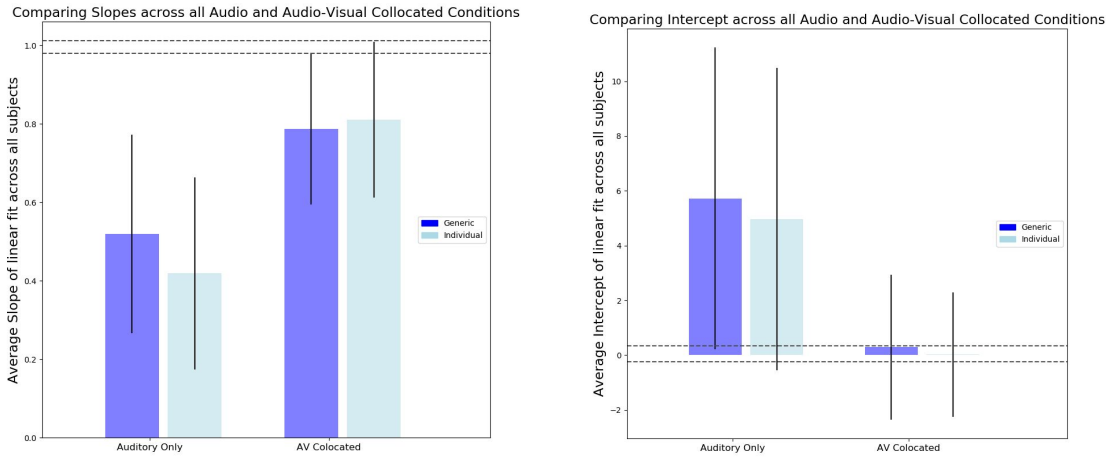


Figure 2: A. A comparison of the average linear fit slopes for all subjects in the auditory only and audio-visual colocated, non-conflicting cues conditions. Average slopes for Generic

(dark blue) and Individualized (light blue) HRTFs are shown. The dotted lines represent the range of slopes for all subjects in the visual only condition. B. A comparison of the average linear fit intercepts for all subjects in the auditory only and audio-visual colocated, non-conflicting cues conditions. Average intercepts for Generic (dark blue) and Individualized (light blue) HRTFs are shown. The dotted lines represent the range of intercepts for all subjects in the visual only condition.

B Discussion

The average auditory only condition localization errors in elevation for the spatial range that we tested (mean absolute error = 15°) were within expected ranges and did not differ from previous literature findings. While auditory elevation localization errors in a free-field setting using real speakers, range from around 8-11° for comparable elevations (Carlile, 1997; Oldfield & Parker 1984; Makous & Middlebrooks, 1990). These are often comparable to those using virtual audio, especially when using full HRTF digital filters (i.e., HRTFs not interpolated between recorded source positions; Martin et al. 2001). However when individualized HRTFs are used at locations that are interpolated between recorded source positions, there can be a substantial range of elevation errors of 13-23° (Ben-Hur et al., 2018), or even larger (Romigh et al., 2015). In comparison, free field virtual audio that uses generic HRTFs, show errors around 20° for elevation (Ben-Hur et al., 2018).

Localization in virtual environments with individualized and generic HRTFs are again dependent upon how the HRTFs are generated and typically show a large range of localization errors similar to their free field counter parts (Romigh 2015, Ben-Hur et al., 2018). Additionally, there seems to be a wide range of elevation errors that are participant dependent (Geronazzo et

al., 2018), not only in virtual environments but in free field experiments as well (Wenzel et al., 1993, Andeol et al., 2015).

Comparing localization abilities across experiments is not trivial, with the number of localization tasks (head pointing, hand pointing, virtual responses, etc) and with the wide range of localization training participants receive prior to testing (ranging from none to on the order of hours). Overall, our training being about 10 minutes in length, and absolute elevation error 15°, are expected and supported by the literature.

We did not find a significant main effect of HRTF for either the slope of the intercept of an individual's linear fit of their localization data. This was the case for auditory only and audio-visual collocated conditions. This is demonstrated in Figure 5, in which the slope for generic and individual HRTFs are plotted against each other for each participant. It is apparent in this plot that there was consistently no difference in HRTF (R^2=0.657, $p < 0.001$). While some previous literature performed in free field has reported a localization improvement in auditory localization (Middlebrooks, 1999), others have reported no differences in auditory localization performance between individualized and non-individualized HRTFs with virtual sound sources (Andeol et al., 2015). This could be due to the limited range of elevation angle tested, as it has been shown that a consistent difference between HRTFs are shown in very low elevations at around -60 to -90 degrees (Andeol et al., 2015).

There was also no significant difference between generic and individualized HRTFs when the visual stimulus was collocated with the auditory stimulus. This is perhaps less surprising since the spatially congruent visual stimulus improves the ability of participants to localize the auditory stimulus. This improvement is apparent in the significant slope increase

between the auditory only (mean slope = 0.469) and audio-visual collocated (mean slope = 0.799) conditions. This can best be explained through a maximum likelihood estimation model, which states that a location judgement with an audio-visual object would be a weighted combination of the reliability of each stimulus signal in a statistically optimal way that minimizes stimulus uncertainty (Alais & Burr, 2004). By having two estimates of stimulus location, participants are more precise on localization. And the difference in reliability of the two different auditory types (filtered with generic versus individualized HRTFs) may not be enough to affect the location estimate.

It is important to note that these audio-visual collocated trials were randomly occurring in the audio-visual experimental phase, meaning that they were interspersed throughout trials that had conflicting sensory cues where the audio and visual stimuli occurred at different elevations. Therefore, the participants were not aware of whether they were presented with conflicting or congruent sensory cues. This suggests that when the spatial information align across modalities, localization performance can be vastly improved compared to having auditory cues alone.

The intercepts were significantly different in the auditory only and audio-visual collocated conditions, albeit not significantly different based on HRTF type, with the intercepts of the auditory only conditions being significantly higher (with a mean intercept of 5.348°and std = 5.589°). In comparison, the average intercepts for the audio-visual collocated condition are essentially zero with a mean of 0.155° and with a much smaller standard deviation across participants (std = 2.50°). This suggests that the visual stimulus served as an anchor for the elevation perception of the auditory stimuli.

In summary, there is a significant improvement in auditory localization abilities when a visual stimulus is collocated with the auditory stimulus, as indicated by the higher mean slopes in the Audio-visual collocated conditions than the Auditory Only. The individual slopes estimated for each participant are also quite consistent across HRTF types suggesting a good test-retest reliability.

IV. Experiment 2: CONFLICTING CUES

Here we examine localization ability in trials that have conflicting spatial cues which include the Audio-Visual trials in which the auditory and visual stimuli are presented at different spatial locations in elevation. Localization results are analyzed based on the slopes and intercepts estimated using the procedure described in Section II: *Deriving slopes and intercepts to summarize localization results*. Localization is compared between auditory stimuli spatialized with generic and individualized HRTFs.

While we expected that the slope of the lines estimated to be close to 1 in conditions with no conflicting cues, the slope estimate across participants could be drastically different depending on how participants combined spatial information across modalities. Specifically, a consistent estimation of a unity slope regardless of which of the 7 elevation angles the visual stimulus was presented suggests that there was no influence of the visual spatial information on the participant's localization judgement. Conversely, a consistent estimation of zero slope would suggest that the visual spatial location entirely captured the auditory stimulus (i.e., the classical ventriloquist effect) and as a result, changing the elevation of the auditory stimulus had no effect on the localization judgement).

The intercept of the linear fit line again provides an estimate of how elevated (in respect of the horizon) the localization responses were for each participant; as intercepts that do not pass through zero would indicate that the perception of the stimulus was either elevated above or below the horizon. This is also a representation of visual spatial pull, as an intercept close to the visual spatial position would indicate a strong visual spatial capture. Using these two metrics, we compared audio localization ability for conditions that had conflicting (non-congruent) spatial cues between the auditory stimulus (for both Generic and Individualized renderings) and the visual stimulus.

A. Results

Table 2 [Regression results]

Slope Regression	**Estimate**	**Std. Error**	**T value**
(Intercept)	0.1438	0.0483	2.977
HRTF.Individual	-0.0254	0.0351	-0.724
Spatial.Below	-0.0129	0.0286	-0.450
Spatial.Above	0.0041	0.0286	0.141
HRTF.Individual:Spatial.Below	0.0493	0.0405	1.217
HRTF.Individual:Spatial.Above	0.0165	0.0405	0.408

Intercept Regression	**Estimate**	**Std. Error**	**T value**
(Intercept)	-0.1471	1.8081	-0.081
HRTF.Individual	0.2073	2.5052	0.083
Spatial.Below	-11.4817	2.0455	-5.613 ***
Spatial.Above	13.1597	2.0455	6.433 ***
HRTF.Individual:Spatial.Below	0.1804	2.8928	0.062
HRTF.Individual:Spatial.Above	0.4396	2.8928	0.152

Table 2: Table A shows the regression results of slope, Table B shows the regression results of intercept.

A.1 *Generalized linear mixed regression on slope*

The following analyses were performed using the **`lme4`** package (Bates et al., 2014) in the R statistical computing environment (R Development Core Team, 2014). A generalized linear mixed regression with simultaneous predictor entry was used to predict the slope of the linear fit lines on trials where the audio and visual stimuli were spatially incongruent and had conflicting spatial sensory cues. Two fixed effect predictors were used in the model. The first predictor is HRTF, which was effect coded for a one unit difference between Generic and Individualized (-0.5 = generic HRTF, 0.5 = individualized HRTF). The second predictor is Spatial, which comprised of three levels indicating the location of the visual stimulus (Above horizon, Center at horizon, and Below horizon). This was dummy coded with Center as the reference level (Center = 0, Below = 1, Above = 2). The model also included participants as the

random effect. With the final model for the linear regression being {Slope ~ HRTF*Spatial + (1|Subject)}. The model output is shown in Table 2.

Results of the main predictors show that the slopes of the linear fits with individualized HRTFs are slightly lower than slopes of generic HRTFs by -0.0254 (SE = 0.0351), holding all else constant with the visual at the horizon. However, this difference is not statistically significant. Comparing when the visual stimulus is below the horizon to when the visual stimulus is at the central location indicates that there is a slightly shallower slope (-0.0129) than the grand average. When comparing the visual stimulus being above the horizon to being at central position, there is a negligibly small increase in slope (0.004).

The interaction terms are of the most interest for our main questions. While the interaction terms were not significant, the trend indicates that when the visual stimulus is below as compared to at the horizon, individualized HRTFs had a slightly steeper slope by (0.049) than when generic HRTFs were used. Finally, there is a slightly steeper slope (0.017) when the visual stimulus is above compared to at the horizon when the auditory stimulus was spatialized with individualized HRTFs.

A.2 *Generalized linear mixed regression on intercept*

A second generalized linear mixed regression with simultaneous predictor entry was used to predict the intercepts of the linear fit lines on trials where the audio and visual stimuli were spatially incongruent. The same model and predictor coding was used as the above model for slope and also included participants as the random effect. With the final model for the linear regression being {Intercept ~ HRTF*Spatial + (1|Subject)}. The model output is shown in Table 2.

The HRTF main effect was not significant, but the trend indicates that the Intercept is slightly higher (0.207) for the individualized HRTFs than the generic HRTF when the visual stimulus is at the horizon. When the visual stimulus is below the horizon as compared to when the visual stimulus is at the central location there is a significantly lower intercept (-11.481). The second Spatial main effect indicates that there is a significantly higher intercept (13.160) when the visual stimulus is above the horizon than when it is at the central position.

The first interaction term indicates that the intercept is slightly more elevated (0.18) when the visual stimulus is below the horizon when compared to the center, and when the auditory stimulus is spatialized using individual HRTFs than when spatialized with generic HRTFs. The second interaction term indicates that the intercept is very slightly higher (0.44) when the visual stimulus is above the horizon when the audio is spatialized using individual HRTFs than when using Generic HRTFs. However, these interaction terms are also not statistically significant.

B. Discussion

The results showed no statistically significant difference in slope between trials with auditory stimulus spatialized with a generic or individualized HRTFs. This is consistent with the results in Experiment 1 results section with congruent cues that also showed no effect of HRTF type on the slope of the linear fit. Given that there was no difference seen in the auditory-only condition, it would be expected that no difference was found in the audio-visual incongruent condition due to the overwhelmingly strong influence of visual spatial cues (Alais & Burr, 2019). In order to see the influence of auditory localization cues on a visual stimulus localization performance, Alais and Burr (2004) had to blur the visual stimulus to a width of 60 degrees

(although the auditory stimulus was spatialized using only ITDs and thus was probably less reliable than with all the spatial cues provided by HRTFs).

There was also no significant difference in the slopes when the visual stimulus was above, below, or at the horizon. This suggests that the influence of the visual stimulus was the same regardless of where the visual stimulus was located in respect to the horizon.

Comparatively, the slopes in the audio-visual incongruent condition are lower (shallower) than the ones in the spatially congruent and audio only (congruent sensory cues). Slopes closer to zero imply that the participants were collapsing to the visual spatial location, and slopes that are higher (steeper) imply more influence from the auditory spatial information - with a unity slope being completely biased towards the auditory spatial location. This implies that when the auditory and visual spatial cues are incongruent, participants are more influenced by the visual location compared to when these spatial cues are congruent across modalities.

Similar to the congruent cues results, the intercept of the linear fits showed no statistical difference between auditory stimuli spatialized with generic or individualized HRTFs. This is again not surprising given that in the congruent conditions (auditory only and audio-visual congruent) there was also no difference between HRTF types.

There was a statistically significant difference of intercept across the three visual spatial conditions (Above horizon, Center at horizon, and Below horizon - See Figure 3 intercepts). This was expected and not surprising given that the slope analysis indicated most of the slopes were fairly shallow (flatter), indicating a significant effect of visual location (significant influence of the visual stimulus). Due to the way the spatial conditions were collapsed (+30°, +20°, +10° elevations collapsing to Above, and -30°, -20°, -10° elevations collapsing to Below), we could

expect that the average intercepts for complete collapse to visual stimuli to be around +20° and -20° respectively. While the mean intercepts after collapsing across HRTF type (mean Above horizon = 13.37°, mean Below horizon = -11.44°) are not that extreme - they do have a large standard deviation (stds = 8.28 and 8.79, respectively).

Figure 3:

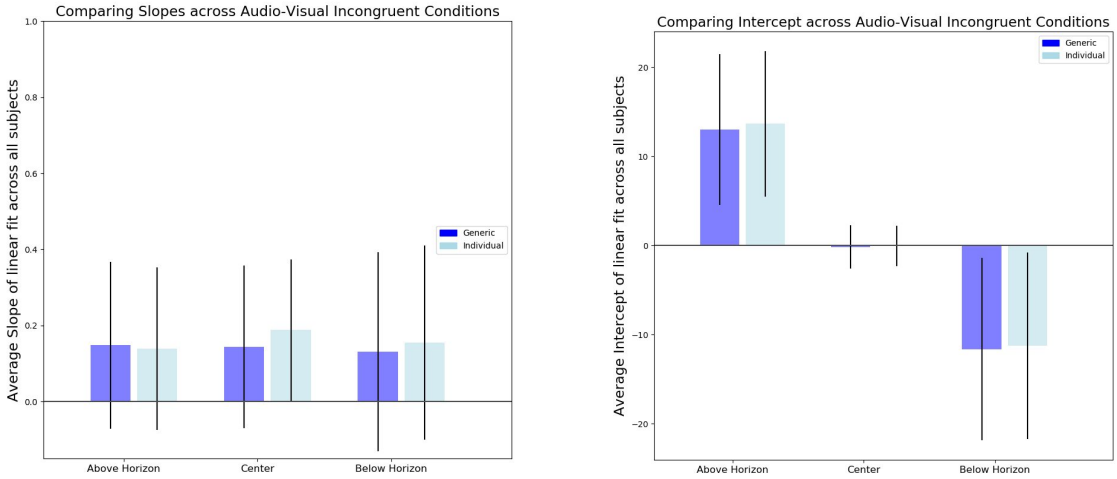


Figure 3: A. A comparison of the average linear fit slopes for all subjects in the audio-visual conflicting cues conditions. Average slopes for Generic (dark blue) and Individualized (light blue) HRTFs are shown. Slopes are averaged for visual locations above, centered, and below the horizon. B. A comparison of the average linear fit intercepts for all subjects in the audio-visual conflicting cues conditions. Average intercepts for Generic (dark blue) and Individualized (light blue) HRTFs are shown. Intercepts are averaged for visual locations above, centered, and below the horizon.

In looking at the distribution of individual participant averages for slope and intercept across spatial conditions, the participants were remarkably consistent in terms of whether they maintained steeper slopes or not. However, and perhaps more interestingly, there is a spread of

the slope estimated across participants. Together, this suggests that there is a good test-retest reliability within each participant on their localization response throughout the experiment, and each participant consistently has a different strategy on how to weight the auditory and visual spatial cues when they are incongruent.

To further explore this across-subject variability, we split the participants based on their estimated average slope across all audio-visual conditions. Participants with slopes who were one standard deviation or more from the grand mean were categorized as 'bimodally influenced' participants. (i.e., an average slope greater than 0.342). While the others were grouped into a 'unimodal visually influenced' group. This criterion conservatively pulled out 4 participants classified as 'bimodally influenced.' After splitting the participants based on slope, we replotted the bar graphs for the intercepts according to the participants groupings to further explore the data qualitatively.

Figure 4 shows the intercepts of both the participants who have been classified as 'unimodal visually influenced' (Figure 4 top) as well as those who have been classified as 'bimodally influenced' (Figure 4 bottom). Here it shows two different patterns for intercepts. The 'unimodal visually influenced' group has intercepts that follow the visual stimulus spatial location, while the intercepts of the bimodally influenced group look rather different; they are closer to zero, albeit slightly elevated on the above horizon spatial condition.

Figure 4:

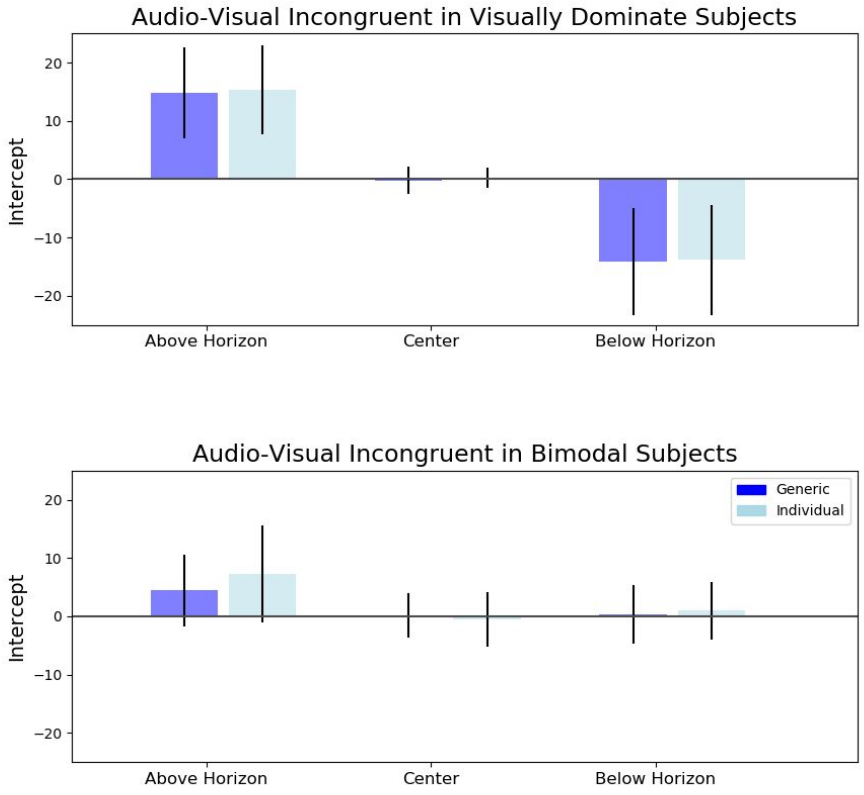
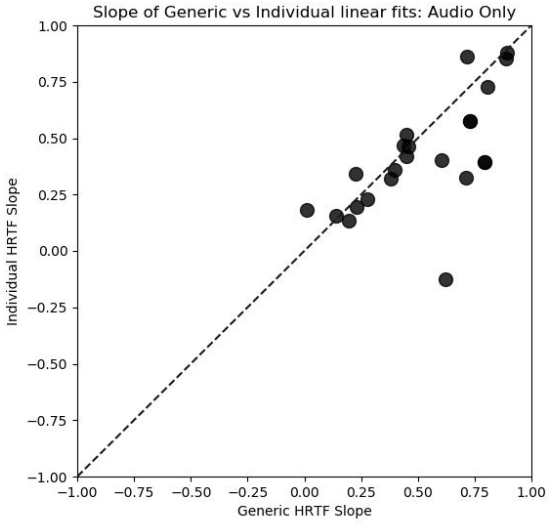


Figure 4: Top panel shows the intercepts of the participants who have been classified as 'unimodal visually influenced', here termed visually dominant participants, across visual locations above, centered, and below the horizon. Bottom panel shows those participants who have been classified as 'bimodally influenced'.

In 'unimodal visually influenced' participant group, there can be two different ways that result in the slope of the participants being close to zero. First, the participants could localize sounds based on the visual stimulus location. These participants will have a pattern of zero slope, but changing intercepts. Alternatively, the participants could be localizing close to the horizon

regardless of the visual stimulus position. These participants will have a pattern of zero slope and zero intercept, suggesting that they have a consistent percept of the auditory stimulus at the horizon and actually be minimally biased by the visual stimulus location. To look at this we assessed the distribution of intercepts by rank ordering the slope by participant for the audio only condition (see Figure 6). The participants who were classified as bimodal are colored in blue, and they are ranked as participants at the top. Those participants who were classified as 'unimodal visually influenced' and also had intercepts of zero in the audio-visual incongruent conditions, are colored in purple and are also in the bottom half of the ranking. These participants most likely also have a zero slope and a zero intercept that does not change with visual position - meaning that they fairly consistently perceive the auditory stimulus to be at the horizon and are minimally influenced by the visual stimulus. Further studies should explore individual susceptibility of the ventriloquist effect based on auditory localization ability.

Figures 5 (left) and 6 (right):



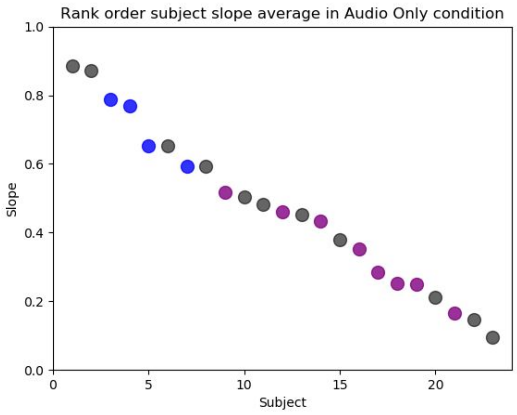


Figure 5: Comparison of all participants audio only linear fit slope for Generic versus Individualized HRTF type. The majority of participants fall along the diagonal line representing no difference in slope within participants across HRTF type.

Figure 6: A rank order of participants intercepts based on the slope of participants in the audio only condition. Participants who were classified as 'unimodal visually influenced' and had intercepts of zero in the audio-visual incongruent conditions, are represented in purple. Participants who were classified as bimodal are represented in blue.

V. Summary and Future Directions

Overall, this is one of the few studies to investigate ventriloquist effect in elevation, and to the best of our knowledge, the only study to look at how the amount of visual spatial capture might change depending on the spectral cues provided by generic versus individualized HRTFs.

We found no difference in auditory localization performance when spatialized with generic or individualized HRTFs - regardless of whether a spatially congruent or incongruent visual stimulus was present. While a somewhat surprising find in the auditory only condition, this could be due to several factors. It is quite possible that for some participants the spectral differences between their own individualized HRTFs and the generic HRTFs could be closer or further apart, affecting their perceived reliability of the auditory stimuli as well as the effect that the different HRTFs had on their localization (Schonstein & Katz, 2012). It may also be possible that the approximation of the individual HRTFs had artifacts that did not reflect the spectral filter as accurately. Although a previous study also found no difference in localization abilities (in azimuth or elevation) when using generic versus individualized HRTFs (Begault et al., 2001).

While we found no difference in auditory localization based on HRTFs, we did show that a spatially congruent visual stimulus significantly improves the accuracy of auditory localization. However when that visual stimulus is spatially incongruent, the majority of participants displayed a strong ventriloquist effect. Although not all participants were equally susceptible to the ventriloquist effect; those who we classified as 'bimodal' participants had better localization accuracy even in the face of conflicting spatial cues. This could be due to two different factors: individual differences in perceptual processing of the spectral cues, or the spectral cues themselves (Andeol et al., 2015).

In previous studies, a wide range of localization abilities across participants have been demonstrated (Makous & Middlebrooks, 1990; Wenzel et al., 1993). A source of participant variation could be attributed to the auditory stimulus in this experiment being presented anechoically, without any room acoustic information. Room acoustics or the first few early reflections have been shown to improve externalization - the perception that the audio is outside of the observer's head in the real world (Shinn-Cunningham, 2000; Begault et al., 2001). Externalization could potentially produce a better spatial correspondence of the sound and the visual stimulus displayed (emanating 2 metres in front of participants). Subsequently it is possible that some participants had a performance disadvantage if they could not associate the internalized sounds and the virtual localization plane.

Finally, in our exploratory analysis we found that there are potentially individual differences on the susceptibility of visual spatial capture. Further study is needed to delineate a way to categorize participants into those that are great auditory localizers, maximully or minimally influenced by visual spatial information, and those that may experience a significant

difference between HRTF types. A wider range of elevations should also be tested as well as including room reflections to investigate the contributions of different HRTFs on the ventriloquist effect.

References:

Alais, D., and Burr, D. (2004). " The ventriloquist effect results from near-optimal bimodal integration," Curr. Biol. 14, 257–262.

V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The cipic hrtf database. In Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the, pages 99–102. IEEE, 2001.

Andéol, G., Savel, S., & Guillaume, A. (2015). Perceptual factors contribute more than acoustical factors to sound localization abilities with virtual sources. Frontiers in neuroscience, 8, 451.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

Begault, D. R., Wenzel, E. M., and Anderson, M. R. (2001). Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. J. Audio Eng. Soc. 49, 904–916.

Blauert J (1997) 'Spatial Hearing. The Psychophysics of Human Sound Localization', 2nd edition, Cambridge MA: MIT Press

Ben-Hur, Z., Alon, D. L., Rafaely, B., & Mehra, R. (2018). Localization and externalization in binaural reproduction with sparse HRTF measurement grids. Acoustical Society of America Journal, 143, 1830-1830.

Berger, C. C., Gonzalez-Franco, M., Tajadura-Jiménez, A., Florencio, D., & Zhang, Z. (2018). Generic HRTFs may be good enough in virtual reality. Improving source localization through cross-modal plasticity. Frontiers in neuroscience, 12, 21.

Bertelson, P., and Radeau, M. (1981). "Cross-modal bias and perceptual fusion with auditory-visual spatial discordance," Percept. Psychophys. 29, 578–584.

Bertelson, P., Vroomen, J., de Gelder, B., and Driver, J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. Percept. Psychophys. 62, 321–332.

BerTelson, P., Frissen, I., Vroomen, J., & De Gelder, B. (2006). The aftereffects of ventriloquism: patterns of spatial generalization. Perception & psychophysics, 68(3), 428-436.

Bermant, R. I., and Welch, R. B. (1976). " Cross-modal bias and perceptual fusion with auditory-visual spatial discordance," Percept. Mot. Skills 43, 487–493.

Butler, R. A., and Belendiuk, K. (1977). "Spectral cues utilized in the localization of sound in the median sagittal plane," J. Acoust. Soc. Am. 61, 1264-1269.

Carlile, S., Leong, P., and Hyams, S. (1997). The nature and distribution of errors in sound localization by human listeners. Hear. Res. 114, 179–196. doi: 10.1016/S0378-5955(97)00161-5

Cochran, W.G. (1937) Problems arising in the analysis of a series of similar experiments. J. R. Stat. Soc. 4, 102–118

Connor, S. (2000). Dumbstruck: A Cultural History of Ventrilo- quism. (Oxford: Oxford University Press).

Gardner, W. G., & Martin, K. D. (1995). HRTF measurements of a KEMAR. The Journal of the Acoustical Society of America, 97(6), 3907-3908.

Geronazzo, M., Sikström, E., Kleimola, J., Avanzini, F., De Götzen, A., & Serafin, S. (2018, October). The impact of an accurate vertical localization with HRTFs on short explorations of immersive virtual reality scenarios. In 2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR) (pp. 90-97). IEEE.

Hendrickx, E., Stitt, P., Messonnier, J. C., Lyzwa, J. M., Katz, B. F., & De Boishéraud, C. (2017). Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis. The Journal of the Acoustical Society of America, 141(3), 2011-2023.

Howard, I. P., and Templeton, W. B. (1966). Human Spatial Orientation. London: Wiley.

Jackson, C. V. (1953). " Visual factors in auditory localization," Q. J. Exp. Psychol. 5, 52–65.

Makous, J. C., and Middlebrooks, J. C. (1990). " Two-dimensional sound localization by human listeners," J. Acoust. Soc. Am. 87, 2188–2200.

Martin, R. L., McAnally, K. I., and Senova, M. A. (2001). Free-field equivalent localization of virtual audio. J. Audio Eng. Soc. 49, 14–22.

Middlebrooks JC and Green DM (1991) 'Sound localization by human listeners.' Annual Review of Psychology 42:135–159

Middlebrooks, J. C. (1999a). Individual differences in external-ear transfer functions reduced by scaling in frequency. J. Acoust. Soc. Am. 106, 1480–1492.

Middlebrooks, J. C. (1999b). Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. J. Acoust. Soc. Am. 106, 1493–1510.

Oldfield, S. R., and Parker, S. P. (1984). Acuity of sound localisation: a topography of auditory space. I. Normal hearing conditions. Perception 13, 581–600.

Radeau, M. (1974). “ Adaptation au déplacement prismatique sur la base d’une discordance entre la vision et l’audition” (“Adaptation prismatic displacement based on a discrepancy between vision and hearing”), L’Année Psychologique 74, 23–24.

Romigh, G. D., Brungart, D. S., & Simpson, B. D. (2015). Free-field localization performance with a head-tracked virtual auditory display. IEEE Journal of Selected Topics in Signal Processing, 9(5), 943-954.

Schönstein, D., & Katz, B. F. (2012). Variability in perceptual evaluation of HRTFs. Journal of the Audio Engineering Society, 60(10), 783-793.

Slutsky, D. A., & Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect. Neuroreport, 12(1), 7-10.

Shinn-Cunningham, B. G., Santarelli, S., & Kopco, N. (2000). Distance perception of nearby sources in reverberant and anechoic listening conditions: Binaural vs. monaural cues. In Assoc Res Otolaryn. Meeting (Vol. 23).

Shinn-cunningham, B. G., Durlach, N. I., and Held, R. M. (1998). Adapting to supernormal auditory localization cues. Bias, I., and resolution. J. Acoust. Soc. Am. 103, 3656–3666.

Thurlow WR, Jack CE (1973) Certain determinants of the ”ventriloquism effect”. Percept Motor Skills 36(3):1171–1184

Vatakis, A., & Spence, C. (2007). Crossmodal binding: Evaluating the “unity assumption” using audiovisual speech stimuli. Perception & psychophysics, 69(5), 744-756.

Van Opstal AJ (2016) ‘The auditory system and human sound-localization behavior’, 1st Ed., Academic Press, Elsevier Publishers, Amsterdam, NL

Warren, D. H. (1979). “ Spatial localization under conflict conditions: Is there a single explanation?,” Percept. 8, 323–337.

Wallace, M., Roberson, G., Hairston, W., Stein, B., Vaughan, J., and Schirillo, J. (2004). “ Unifying multisensory signals across time and space,” Exp. Brain. Res. 158, 252–258.

Weerts, T. C., and Thurlow, W. R. (1971). “ The effect of eye position and expectation on sound localization,” Percept. Psychophys. 9, 35–39.

Wozny, D. R., & Shams, L. (2011). Recalibration of auditory space following milliseconds of cross-modal discrepancy. Journal of Neuroscience, 31(12), 4607-4612.

Welch, R. B., and Warren, D. H. (1980). “ Immediate perceptual response to intersensory discrepancy,” Psychol. Bull. 88, 638.

Wenzel EM, Arruda M, Kistler DJ, Wightman FL (1993) Localization using nonindividualized head-related transfer functions. J Acoust Soc Am 94(1):111–123

Chapter 4: A multimedia speech corpus for audio visual research in virtual reality

Authors: Lindsey R. Kishline, Scott W. Colburn, Philip W. Robinson

Intended for: JASA as a Letter to the Editor.

Abstract:

Virtual reality environments offer new possibilities in perceptual research, such as presentation of physically impossible but ecologically valid stimuli in contrived scenarios. To facilitate perceptual research in such environments, we present a publicly available database of anechoic audio speech samples paired with matching stereoscopic and 360° video. These materials and accompanying software tool, allow researchers to create simulations with up to five talkers positioned at arbitrary azimuthal locations, at multiple depth planes, in any 360° or stereoscopic environment. We describe recording conditions and techniques, contents of the corpus, and how to use the materials within a virtual reality environment.

I. Introduction

For studies investigating speech and communications in naturalistic and ecologically valid environments, the need for a research tool to allow the creation of acoustically and visually complex environments while maintaining the ability to parametrically manipulate the audio and video of multiple talkers, and their environment, is necessary (Cappelloni et al., 2019; Stecker et al., 2019). The advent of Virtual Reality (VR) technology has greatly enhanced our ability to explore questions surrounding speech intelligibility (Gonzalez-Franco et al., 2017), spatial audio

(Poirier-Quinot & Katz, 2018), auditory localization (Ahrens et al., 2019), room acoustics (Amengual Garí et al., 2019; Rungta et al., 2018, Stecker et al., 2018) and many others in complex environments (Stecker, 2019). However, until now, there has not been a speech corpus specifically designed to leverage the unique capabilities this technology has to offer the research community. This corpus was created to facilitate auditory and multisensory research in immersive environments, and designed to allow experimental control in multidimensional tasks.

II. 3D Audio-Visual Speech Corpus

This audio-visual speech corpus includes recorded sentences from five Talkers, two male and three female, all with an American English accent. The Talkers were recorded saying two different speech corpora; the coordinate response measure (CRM) (Bolia et al., 2000), and the Harvard IEEE corpus word list (Rothauser, 1969). The CRM corpus was chosen for its use in measuring speech intelligibility in multitalker environments, as well as offering the ability to "gain measures of sensitivity (d') and response bias" via a signal detection style task (Bolia et al., 2000). Additionally, because it is a widely used corpus across many fields including automatic speech recognition (ASR) technology (Cooke et al., 2006), as well as audiology (Best et al., 2012). And the IEEE corpus was chosen because of their wide use in speech intelligibility in multitalker scenarios (Hawley et al., 2004; Qin, M.K. & Oxenham, A.J., 2003; Bernstein, J. G., & Grant, K. W., 2009), and due to the construction of the sentences being phonetically balanced and low context. The CRM recordings include seven callsigns ('Laker', 'Baron', 'Charlie', 'Ringo', 'Eagle', 'Hopper', and 'Tiger'), four colors (red, green, white, and blue), and numbers one through eight. All five Talkers gave a completed all factorial combinations of the callsigns,

colors, and numbers, giving a total of 224 sentences. These sentences are recorded twice, at two planes of depth; a near plane of 81 centimeters and a far plane of 183 centimeters. The IEEE recordings include 50 list sentences per Talker, each Talker recorded unique list sentences. The Talkers recorded the same 50 list sentences at the two planes of depth. The Talker and corresponding list numbers are provided in the hosting site [].

This audio-visual speech corpus includes both a 360 recording and 180 stereoscopic video of each recorded sentence. Additionally, each 360 and 180 video also includes three versions; a black background, a black transparent background, and a greenscreen version. The greenscreen version allows the user to do cropping and placement of the Talkers into any 180 or 360 filmed background environment, and the black transparent allows the already cropped Talkers to be placed into any intended background and composited into videos with other Talkers. In total, with all five Talkers, both speech corpora, two planes of depth, both 360 and 180, with all three backgrounds (greenscreen, black, and black transparent), a total of 16,440 videos are included. The raw audio files for every recording are also included for a total of 2,740 individual audio files.

Figure 1:

Figure 1: Top panel shows all five talkers included in the corpus in order from left to right of Talker number. Bottom panel (left) shows three of the talkers at two different depth planes inserted into a 360 video background. Bottom panel (right) shows Talker 2 green screen crop.

III. Methods

All video and audio were recorded in an anechoic chamber. The anechoic chamber was a fully anechoic room manufactured by Eckel Industries with an interior clear dimension of 15 feet wide by 25 feet long by 15 feet high, and designed to be anechoic (echo free) down to 100 Hz. All talkers were filmed against a green-screen of approximately 9 feet wide by 10 feet tall with a

length of 20 feet. And lit with six lights in individual 24" x 24" Softboxes with 85W 5500K CFL bulbs to minimize shadow effects for post-processing of green-screen cropping. All talkers were chosen for an American English accent in an effort to eliminate talker accent as a potential confounding factor for those using the corpus for speech intelligibility. Five total Talkers were recorded, two males and three females.

Video:

Talkers were filmed by two simultaneously recording Vuze XR Dual VR Cameras; one recording in 3D 180x180x2 stereoscopic mode (half sphere) and the other in 360x180 mode (full sphere). All video was captured at a 5.7k 30fps video resolution and frame rate. The 360 full sphere camera was recorded at a height of 162 cm and the 3D 180 half sphere camera was recorded at a height of 165 cm. Talkers were filmed at two planes of depth from the cameras; near and far. Talkers stood approximately 81cm from the cameras for the near recordings, and approximately 183cm from the cameras for the far recordings.

Audio:

The main mic source was a Sennheiser 416 shotgun mic attached to a boom stand and located approximately 1 meter away and out of the camera view. The second source was a Shure CVL - B/C Lapel mic clipped to the talkers collar, which was attached to a Shure BLX1 H10 transmitter and placed into the talkers back pocket. This signal was received by a Shure BLX4 H10 receiver. Both sources were hard wired to a Zoom F8 field recorder operating at 48K sample frequency and 24 bit resolution.

Recording process:

Prior to recording all talkers went through a training phase for both the CRM sentences and the IEEE sentences. The training phase consisted of listening to an example sentence for timing, tone, and emphasis. Talkers were then required to repeat the training phrases with correct timing and pronunciation prior to the start of the recording sessions. During the recording sessions one experimenter listened to and maintained the timing of the sentences from the talker, and another listened to and maintained the pronunciation and quality of the sentences. Both experimenters had to accept the sentence spoken by the Talker before moving on to the next. Periodically throughout the recording sessions, the talkers would listen to the training sentences again for timing.

The recording process for the CRM sentences consisted of the Talker being given a call sign and a color, and asked to follow the CRM carrier phrase starting at the number one and ending at number eight, before being prompted by an experiment for the next color. This was done for all seven call signs and four colors. The recording for the IEEE sentences consisted of the talker reading all the sentences prior to the recording session for familiarity and pronunciation questions. During the recording of the sentences, each talker had the sentence displayed at the top of the cameras to read whilst the recording was happening. Talkers were prompted to take breaks every twenty mins.

During the recording, talkers were asked to stand on their distance mark on the green-screen paper and face the camera, maintaining eye contact between the two Vuze Cameras while speaking the sentences. The timing of recording for each sentence would consist of the following; five seconds of silence, say the sentence or phrase, then another five seconds of silence.

Before each twenty-minute recording session, the Talker was recorded holding a XRite color calibration palette to facilitate color correction in post production, their position on each camera for green-screen clearance was checked, and an experimenter would create a loud clap for post processing audio-video alignment.

Post-processing:

After completing the collection of the entire corpus, the audio and video components had several post-processing steps.

Audio:

The audio from the lavalier mic was discarded and the audio from the boom mic was used for all audio purposes. The audio was edited using Adobe Audition. The raw audio from the boom mic was volume matched using Auditions 'Match Clip Loudness' effect to a target loudness of -23 LUFS, with a tolerance of 0.1 LU (EBU128-2014) . The audio was then cleaned for noise and removal of talker generated noise manually by using Adobe Auditions 'Noise Reduction' effect at 50% and reduced by 16dB. Five seconds of silence was added at either end of the audio clips by taking a five second 'room tone' clip consisting of a static recording of the room, free of noise and aligning over the audio tracks. The final audio track edits were saved to a .wav file at 48k Hz Mono 32 bit format. The audio files were then imported to Adobe Premiere Pro 2019 for alignment with video described in the section below. The audio .wav files, without video, are also provided by this corpus.

All video:

The video from both the 360 camera and the stereoscopic 180 camera was processed in Adobe Premiere Pro 2019. The stereoscopic 180 general settings in Human Eyes 180 used a

custom editing mode with a timebase of 29.97 fps. The frame size was set to 5760h 2880v (1.0000), with a frame rate of 29.97 frames/second, and a pixel aspect ratio of Square Pixels (1.0). No fields were used and thus set to Progressive Scan. The audio settings for the videos were at a sample rate of 48000 samples/second. The VR settings for projection were Equirectangular with a monoscopic layout and captured view 360 degrees Horizontal by 180 degrees Vertical.

The 360 video general settings in Human Eyes 360 also used a custom editing mode with a timebase of 29.97 fps. The video settings for frame size, frame rate, pixel aspect ratio, and fields were the same as above, and the audio settings were again 48000 samples/second. The VR settings for projection, layout, and captured view were also the same as above.

The audio and the video clips were aligned in Audition by unlinking the original audio and video and replacing with the edited audio tracks, then aligning the recorded audio to the camera audio using the sound markers, the tracks were then imported into Premiere. The greenscreen was then cropped and transparent black background used for the black cropped videos. A three second silent sequence was ensured both before and after every talker phrase or sentence. The video format used was the Quicktime format (mov) using Apple Pro Res 4444+Alpha as the codec, applied to both sets of exported videos the 180 and 360 Black Transparent and the 180 and 360 Green screen videos.

Each video and audio file was screened for mispronounced words, extraneous noises, speed, etc. The few known issues with the video and/or audio of the corpus is listed below.

Special notes:

Here we list the known issues with Talker 5, and exceptions to the CRM sentences and limits on the IEEE sentences. Talker 5 has a removed t-shirt logo which required a matte effect to obscure said logo. The logo was covered with the following color on the close distance (161619) and far distance (20232A). The following CRM sentence combinations are not included in the corpus due to various errors such as noise, talker eye position, talker pronunciation and designated in the following format; Talker followed by the *Callsign/color/number/distance* of the sentence. Talker 1: *Laker/blue/one/close, Laker/blue/two/far, Charlie/white/one/far, Eagle/green/one/far, Eagle/green/eight/far*, Talker 2: *Laker/red/one/far, Laker/green/one/close*, Talker 4: *Baron/green/seven/close*, Talker 5: no sentences with number 8 for the far depth planes are included, *Laker/blue/three/both, Baron/white/one/close, Baron/blue/one/close, Charlie/blue/four/far, Ringo/green/seven/far, Eagle/green/seven/far, Eagle/blue/seven/far, Tiger/red/five/close*.

IV. Availability

The corpus, in its entirety, is available for free under the [] license agreement, and hosted on []. This includes the 180 and 360 Black Transparent and the 180 and 360 Green screen videos, as well as the edited Audio .wav files. In total, a complete download of the corpus would include 17,322 videos (including the XRite color bar recordings) and 3,200 audio files. Additionally, for future lighting or color adjustments, each Talker has a short video sequence holding an XRite color calibration palette, for each distance position. A JSON file is also included with a description of the files naming convention, Talker notes, and instructions for

viewing across different VR platforms, compiling the videos, and spatializing the audio for the compiled videos.

Use in VR environments:

In addition to the above, we are making available a simplistic Unity tool to allow experimenters to quickly compile multiple talkers (up to three talkers at a time), within any background, and their relevant audio files. This player will also allow experimenters to designate the spatial locations of the audio files, as well as a basic interface for behavioral data collection and data exporting. The complete details of this Unity tool will also be hosted here [] for download.

Acknowledgements:

The authors would like to thank the research assistants for their incredible work to help record and quality check this corpus, with a special thank you to Alex Gustafson, as well as to our research team for useful feedback on prototypes and early production.

References:

Ahrens A, Lund KD, Marschall M, Dau T (2019) "Sound source localization with varying amount of visual information in virtual reality". PLoS ONE 14(3): e0214603. https://doi.org/10.1371/journal.pone.0214603

Bernstein, J. G., & Grant, K. W. (2009). Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners. The Journal of the Acoustical Society of America, 125(5), 3358-3372.

Best, V., Marrone, N., Mason, C. R., & Kidd, G., Jr (2012). The influence of non-spatial factors on measures of spatial release from masking. The Journal of the Acoustical Society of America, 131(4), 3103–3110.

Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (2000). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. 107, 1065–1066.

Cappelloni, M. S., Shivkumar, S., Haefner, R. M., & Maddox, R. K. (2019). Task-uninformative visual stimuli improve auditory spatial discrimination in humans but not the ideal observer. PloS one, 14(9), e0215417. doi:10.1371/journal.pone.0215417

Cooke, M., Barker, J., Cunningham, S., Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America. 120(5), 2421-2424.

EBU - R 128-2014 Audio loudness normalisation & permitted maximum level.

Garí S. V. A, Schissler C., Mehra R., Featherly S., Robinson P. Evaluation of Real-Time Sound Propagation Engines in a Virtual Reality Framework. (2019) AES International Conference on Immersive and Interactive Audio.

Gonzalez-Franco, M., Maselli, A., Florencio, D. et al. "Concurrent talking in immersive virtual reality: on the dominance of visual speech cues". Sci Rep 7, 3817 (2017) doi:10.1038/s41598-017-04201-x

Hawley, M. L., Litovsky, R. Y., & Culling, J. F. (2004). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. The Journal of the Acoustical Society of America, 115(2), 833-843.

Poirier-Quinot D., Katz B. F. G.. “Impact of HRTF individualization on player performance in a VR shooter game II”. AES International Conference on Audio for Virtual and Augmented Reality, Aug 2018, Redmond, United States.

Qin, M. K., & Oxenham, A. J. (2003). Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. The Journal of the Acoustical Society of America, 114(1), 446-454.

Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H.R., Urbanek, G. E., and Weinstock, M. 1969. ‘‘I.E.E.E. recommended practice for speech quality measurements,’’ IEEE Trans. Audio Electroacoust. 17, 227–246

Rungta, A. , Rewkowski N., Schissler ., Robinson P., Mehra R., and Manocha D. . 2018. Effects of virtual acoustics on target word identification performance in multi-talker environments. In ACM Symposium on Applied Perception 2018, August 10–11, 2018, Vancouver, BC, Canada. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3225153.

Stecker, G. Christopher. Using Virtual Reality to Assess Auditory Performance. The Hearing Journal: June 2019. Volume 72. Issue 6. P 20,22,23 doi:10.1097/01.HJ.0000558464.75151.52

Stecker, G. Christopher; Moore, Travis M.; Folkerts, Monica; Zotkin, Dmitry; Duraiswami, Ramani. Toward Objective Measures of Auditory Co-Immersion in Virtual and Augmented Reality. 2018 AES International Conference on Audio for Virtual and Augmented Reality.

Chapter 5: Conclusions and future directions

To conclude this body of work we will revisit each chapter, discuss the overall findings, and propose future directions.

We live in a beautifully complex world; a multisensory world in which we are continuously bombarded with a multitude of sensory information. Somehow our brain manages to seamlessly take what our eyes are seeing, and what our ears are hearing, and merge these different sensory systems and cues together to create our experience of the world. How our brain does this is an incredibly complex problem, and one that is still being investigated. While from early neurophysiology experiments in animals we know that neurons in the brain show the most activity when sights and sounds occur together at the same time and at the same place in space (Meredith et al., 1987; Meredith & Stein, 1986) giving a clue to some of the fundamental ways in which our brain creates our perception of the world, it does not always reflect findings in behavioral tasks much like in Chapter 2.

Study 1: Here we explored when the spatial colocation of auditory and visual stimuli is influential to multisensory integration and investigated how spatial colocation, competing stimuli, and attention affected the sound-induced flash illusion. Surprisingly we did not find influence of spatial colocation or the proximity of a competing stimulus on the probability of reporting an illusion. These two surprising negative findings provide more evidence to question whether the spatial 'rule' of multisensory integration should be applied to behavioral studies.

Furthermore, our manipulation of spatial attention also did not find a significant effect. However, a more in-depth examination across studies suggests that there is a potential difference between bottom-up and top-down spatial attention influences. Future follow up studies should further tease apart how these two different types of spatial attention could affect multisensory integration differently.

While in the above study we set out to show an effect of spatial congruence, we also asked whether the spatial incongruence of audio and visual stimuli (as shown through the ventriloquist effect) may be used to our advantage.

Study 2: This study assessed the amount of visual spatial influence, in elevation, on an auditory stimulus under two different auditory Head Related Transfer Functions (HRTF). Surprisingly we did not find an effect of HRTF type. Nonetheless, we did find that when auditory and visual stimuli had non-conflicting spatial cues, participants showed a significant improvement of auditory localization even without the knowledge of whether they had a trial where audio-visual stimuli were spatially collocated or not. Additionally we showed that there is a very strong influence of visual spatial information on auditory localization. But importantly, the amount of visual spatial capture may vary based on individual differences. Future studies should investigate whether these individual differences can be capitalized on in Virtual Reality settings to categorize users who may not need highly accurate auditory spatial rendering.

In the above two studies highly simplistic stimuli were used to investigate how the brain uses conflicting and non-conflicting auditory and visual spatial cues. However, the world we experience in everyday life is constructed of a highly complex and overlapping bombardment of auditory and visual stimuli. So how can we bring the complexity of the real world into an

experimental setting? One way is through Virtual Reality. Chapter 4 is dedicated to assisting experimental scientists in bringing experimental paradigms into more natural and ecological settings.

Study 3: Here we described a large audio-visual speech corpus that is uniquely designed for experiments in Virtual Reality. To facilitate perceptual research in such environments, we present a publicly available database of anechoic audio speech samples paired with matching stereoscopic and 360 video. These materials and accompanying software tool, allow researchers to create simulations with up to five talkers positioned at arbitrary azimuthal locations, at multiple depth planes, in any 360 or stereoscopic environment. We anticipate that this will act as a vastly valuable tool moving forward in the multisensory and acoustic research communities.

References:

Meredith, M.A. & B.E. Stein. 1986. Spatial factors determine the activity of multisensory neurons in cat superior colliculus. Brain Res. 365: 350–354.12.

Meredith, M.A., J.W. Nemitz & B.E. Stein. 1987. Determinants of multisensory integration in superior colliculus neurons. I: temporal factors.J. Neurosci.7: 3215–3229. 16.