The ventriloquist effect on interactive moving objects in virtual reality

Palla Lorden, O.

Award date:
2021

Eindhoven, March 3rd 2020

The Ventriloquist Effect on Interactive Moving Objects in Virtual Reality

by Odin Palla Lorden

identity number 782892

in partial fulfilment of the requirements for the degree of

Master of Science
in Human-Technology Interaction

Supervisors:
Raymond Cuijpers
Armin Kohlrausch
Kynthia Chamilothori
Victoria Korshunova

Abstract

The development of virtual reality in recent decades has been valuable to perceptual research. The seamless integration of multisensory cues makes it an ideal tool to investigate sensory integration, while at the same time this research can be used to improve VR experiences. Since VR is made up of an intricate system communicating between hardware and software layers there is a large potential for latencies. Human perceptual plasticity allows for some temporal and spatial margins between multisensory cues while still being perceived as united. The current study investigated this ventriloquist effect in a bimodal versus trimodal sound delay detection experiment involving audio, video and proprioceptive/agency cues. Participants were tasked with moving an interactive audiovisual object in VR and distinguish intervals including controlled sound delays from intervals without. Results showed a significant difference between conditions where the object followed participant's pointing direction and the condition where it did not. No significant difference was found between conditions where the object was visible and the condition where it was not. These results suggest that pointing direction involves a top-down process of location anticipation that significantly decreases ventriloquism thresholds to the point that adding a third visual modality cannot decrease it further.

1. Background

1.1 Multimodality and sensory integration

1.1.1 Multisensory integration

Human perception is a complicated field that involves brain input from all our senses stemming from various organs all over our body. Each sense has a unique pathway for receiving, transmitting and interpreting stimuli. Vision, for example, works by absorbing photon energy on our retina which is sent to the visual cortex. This process compensates for color, brightness, contrast and more for an accurate representation of our environment. Similarly, with hearing, where the cochlea in the inner ear decodes vibrational energy from pressure differences in the air into neural signals. Frequency and loudness are important factors to perceive the sound field around us. These stimuli are made even more complex by

additive interactions with the environment like reflections and filtering which create artifacts that aid in interpreting said environment. The somatosensory system refers to perceiving our environment via direct skin contact. Proprioception specifically provides information about our body position and movement. We have proprioceptors in and around our limbs that interpret muscle length. The primary somatosensory cortex translates this information to relative positions in the environment where our limbs may be located at any given moment.

These three senses (vision, hearing and somatosensory) contribute to perceiving the surrounding environment, although one could argue that the benefit of multiple senses is gaining a more holistic understanding. Individual senses affect each other as can be observed by the behavior of multisensory cells. These cells are hypothesized to respond when multiple sensory input is received within a given time frame. This suggests an inherent preference for multisensory stimulation in human perception. Moreover, mutual cancellation of multiple inputs had been observed to decrease the multisensory response significantly as well (Calvert, Hansen, Iversen & Brammer 2001). Multisensory integration has been experimentally investigated in response time studies with multimodal stimuli. A decrease in response times was found when presenting participants with bimodal over unimodal stimuli, and a further decrease was observed between bimodal and trimodal stimuli. Todd (1912) suggested an “energy summation” mechanism to explain this reduction in reaction times. Recently, alternative explanations have been proposed as summarized by Diederich & Colonius (2004). A study on multisensory cells suggested that their firing rate increases when perceiving stimuli at multiple inputs, hypothesizing that presenting a multimodal stimulus produces sensory activations in differing neural pathways. This builds a response initiated by the first signal to reach the cell (Raab 1962). Based on that neural race condition the model predicts the decrease in response times in bimodal over unimodal stimuli. This is referred to as the separate activation model. An alternative model is coactivation based on the hypothesis that unimodal signals are combined before generating a united response (Grice, Canham & Boroughs 1984). However, it is still unclear at what processing level coactivation occurs.

An experimental study similar to the current study supported findings of multisensory integration as well. An audio-visual light localization task was performed better when both modalities were involved as opposed to only spatial audio (Bolia, D'Angelo & McKinley 1999). This shows that the addition of the audio stimulus contributed positively to the localization task and supports some kind of multimodal integration during sensory processing. It can be taken a step further by adding proprioceptive localization cues for a total of three sensory modalities. On a neural level an interaction of this kind would depend on the presence of trimodal multisensory neurons that respond to auditory, visual and proprioceptive stimuli. A definitive categorization of multisensory neurons is still an open problem, but since their effect is observable the benefit of integrating multisensory localization cues is open for investigation. Virtual reality (VR) is a technology where all these cues occur and is by nature a technology that allows for real-time interaction. This makes it the perfect candidate to investigate multisensory localization on moving objects. Previous literature has investigated localization of moving objects in VR (Wiker 2018), but the movements did not come from participant interaction which is the goal for the current study.

1.1.2 The Ventriloquist effect as tool for sensory integration research

To study basic mechanisms of multisensory integration it is ideal to have some robust perceptual phenomenon that can be relied upon to give reproducible results. One such phenomenon is the ventriloquist effect where one stimulus captures location and/or timing of another. The perceived directional information for an audio-visual stimulus will be dominated by the visual direction. Classic ventriloquism is observable if the audio and visual component appear from separate directions and refers to the illusion that cues from both modalities originate from the visual source regardless of their actual location. This shows that visual location cues can take preference over simultaneous auditory location cues in some situations. The ventriloquist effect can be exploited in VR situations where visual and audio cues are meant to come from the same location but due to system limitations or artifacts it is not possible in a most accurate manner. Of course, there are limits to how far the incongruency in location can be stretched for this illusion to still occur and these perceptual thresholds will be investigated in this study.

The ventriloquist effect has been shown to appear between stimulus pairings of many modalities. The classic example is visual capture of audition (Bertelson & Radeau 1981), but it also applies to audio-tactile (Bruns & Röder 2010) and visuo-tactile (Samad & Shams 2016) stimulus pairings. Consequently, the ventriloquist effect has shown to be a popular candidate to investigate multisensory integration. One review paper summarized various manifestations of the capturing effect between multiple stimulus pairs, specifically differentiating between immediate and aftereffects (Chen & Vroomen 2013), although the current study will focus on the immediate effects. They also discussed the mediation of attention, congruency and cognitive factors. This shows the depth of the ventriloquist effect and its sensitivity to both bottom-up and top-down processes which are considered in the current study.

1.2 Virtual reality

1.2.1 VR as a tool for sensory integration research

New experimental methods benefit from unique stimulus presentations to explore new perceptual grounds. This is why the development of VR in recent decades has been valuable to research of this kind. The seamless integration of multisensory cues makes it an ideal tool to investigate sensory integration. The strongest addition of VR is the motion tracking that allows adjusting the field of view to the viewer's head movements in real time. Tracking the hands and subsequently displaying them in the virtual environment also adds the modality of proprioception and agency to the experience. Combined with the 3D visual effect provided by the stereoscopic presentation of the head-mounted display (HMD) and the 3D audio effect from binaural sound, it allows an experimental setting that has high ecological validity compared to regular screens and input methods with less degrees-of-freedom (mouse, keyboard). All of this contributes to a more immersive experience which is an aspect that consumer VR seeks to maximize. Additional to the number of sensory modalities is display quality, in multimedia this is sometimes referred to as fidelity.

1.2.2 Fidelity, immersion and presence

Levels of fidelity are an important factor in consumer VR experiences and major contributor to VR equipment value. VR fidelity can be broken down into display fidelity and interaction fidelity. The

former refers to the quality of perceptual aspects like audio, graphics and tracking. And the latter refers to how convincing the interactions with the virtual environment are. These two aspects of fidelity dictate the realism to a large degree and in turn the immersion of the VR experience of the user. A review study discussed how objective measures of immersion contribute to subjective presence as measured by questionnaires. They looked into what levels of quantitative and qualitative fidelity is beneficial to reach a certain degree of presence for many different use cases including military training, phobia therapy, medical training and VR gaming (Bowman & Mcmahan 2007). The relevant takeaway is that the effect of immersion on subjective presence was mediated by the type of VR application. Nevertheless, the effect was always positive so maximizing levels of fidelity to increase immersion is a good starting point for making VR experiences better. This is again also reflected in the Oculus best practices guide which explicitly states the importance of "ensuring less than 20 ms delay between head movements and corresponding visual updates, in addition to maximizing screen refresh rate, to avoid negative impacts on user comfort and presence" (Oculus 2016).

1.2.3 Embodiment and agency

Where presence represents the more technical capacity of VR by optimizing perceptual fidelity, the more cognitive components to self-representation in a virtual environment correspond to embodiment and agency. A classic experiment regarding embodiment is the rubber hand illusion where an artificial hand is placed next to a participant's own (out of view) hand. When both hands receive tactile stimulation at the same time, this can induce an experience of embodiment such that the participant feels that the artificial hand is part of their actual body (Botvinick & Cohen 1998). In experiments like this embodiment is often expressed in subjective measures of body ownership, agency and proprioception. Agency refers to participants' sensory predictions regarding the outcome of their intended actions. Subjective agency can be manipulated by introducing a delay between an action and the stimulus resulting from it. This can affect the feeling of ownership towards an action regardless of the causality awareness. An example of this is a study on tickle responses where a mechanical device was used to introduce a delay between tickling actions and the tactile stimulus. The larger the delay the higher the tickle rating due to the action

being experienced as external (Blakemore, Wolpert & Frith 2000). Since in VR there exists inherent input delays the experienced agency has potential for variability. Support for this was found in a study where delays were introduced between an action modality (button press and voice command) and display modality (HMD). Higher delays resulted in a lower sense of agency (Winkler, Stiens, Rauh, Franke & Krems 2019). Another study showed that “motor performance and simultaneity perception are affected by latencies above 75 ms. Although sense of agency and body ownership only decline at a latency higher than 125 ms, and deteriorate for a latency greater than 300 ms, they do not break down completely even at the highest tested delay” (Waltemate, Senna, Hülsmann, Rohde, Kopp, Ernst & Botsch 2016). This suggests that sense of agency is robust in VR even at delays far above system latencies of current consumer VR equipment.

1.3 Spatial audio and sound localization in virtual reality

1.3.1 Tools for spatial audio in VR

Several methods exist to present virtual audio sources to a listener. In an exploratory VR audiovisual localization study performed by myself two methods were investigated: Ambisonics and vector-based amplitude panning (VBAP). Both these methods require multi-speaker setups to reproduce the localized sound cues. What makes Ambisonics unique is that it’s an encoded signal independent from loudspeaker amount or positions, it gets decoded at playback (Gerzon 1985). VBAP is based on the phantom source illusion where playing the same sound cue simultaneously from three speakers laid out in a triangle shape can virtualize the source anywhere on a plane intersecting the position of these speakers based on the loudness balance between them (Pulkki 1997). No significant difference between these auralization methods was found regarding the temporal synchrony task during the study (Palla Lorden 2019). A downside to multi-speaker auralization is that it’s uncommon for consumers to be able to produce them at home due to the equipment requirement. On top of that, when used in combination with VR, the HMD has a significant effect on localization performance. The HMD is a physical object attached to the head so naturally it interacts with the sound waves before they reach the ears, altering the

embedded temporal information (Ahrens, Lund, Marschall & Dau 2019). Thus, for practical and reproducibility reasons the current study will use stereo headphone playback.

To present virtual audio sources in VR with headphones the sound stimulus needs to be processed the way it occurs in the real world. In a similar way to how stereoscopy allows for 3D presentation of a visual environment, binaural audio utilizes stereo presentation to synthesize 3D sound. The way sound localization in human perception works involves several sources of information including temporal, magnitude and spectral cues (Blauert 1984). Specifically the former two require comparing information between the two ears, namely interaural time differences (ITD) and interaural level differences (ILD). These two cues are at the core of the century old duplex theory (Raleigh 1907). The spectral cues are obtained from sound interactions with body parts, mainly the head and pinna (outer ear). Sound stimuli reflecting and diffracting off these body parts are uniquely filtered depending on direction (Blauert 1984). Binaural audio in VR can be synthesized by applying a head-related transfer function (HRTF) to the sound stimulus based on these three localization cues. As the shape of a human body is unique to a person, the HRTF is as well. Calculating an accurate HRTF for every single person is a long process that is not feasible in the consumer VR market, so the industry has resorted to utilizing standardized HRTFs. A recent study has investigated the accuracy of these standardized HRTFs. They found that when presenting participants with congruent audiovisual stimuli for 60 seconds, it was enough for auditory localization performance to become on par with personalized HRTFs (Berger, Gonzalez-Franco, Tajadura-Jiménez, Florencio & Zhang 2018). These results are in line with previous research that argues that our cognitive representation of acoustic space is highly plastic (Carlile 2014).

1.3.2 Effect of motion on sound localization

Apart from the information that makes up an HRTF filter there are also less explicit sources of information that aid in sound source localization. One of these is the movements of the listener's head. An experimental study on sound localization tested virtual sources that were either stationary with respect to the outside world or stationery relative to the listener's head. Their findings showed that virtual sources whose position was fixed relative to the world are more likely to be externalized than those fixed relative

to the head, independent from the fidelity of the individual impulse responses (Brimijoin, Boyd & Akeroyd 2013). This confirmed that head rotations indeed play an important role in the localization of sound, specifically to disambiguate front-back reversals. An experimental study on the ventriloquist effect in augmented reality also found that the magnitude of head rotations had a significant effect on the thresholds for the ventriloquist effect to hold (Kytö, Kusumoto & Oittinen 2015).

Information from the motion of the sound source itself can also be used for localization. The way that this situation differs from the situation where the listener moves their head relative to a stationary sound source is that during the former our auditory system is not informed by the vestibular system. Head rotations are sensed by the semicircular canals, but when the head is stationary and the sound source moves, the direction in which ITD and ILD change may not disambiguate front-back confusions (Macpherson 2013). But there is a way to reduce the front-back ambiguity by putting the patterns of movement of the sound source in control of the participant. In a sound localization experiment where the head position was kept fixed, the condition where the participants themselves could anticipate the movements of the sound source by controlling it with cursor keys on a keyboard had fewer front-back reversals than the condition where the experimenter controlled the movements (Wightman & Kistler 1999). Whether freehand interaction also mediates the ventriloquist effect is unknown and the current study seeks to investigate that.

1.4 Temporal and spatial ventriloquism

1.4.1 Temporal ventriloquism

Since stimuli from different observable modalities of a single event have specific relations with regards to timing and localization in the real world, our perceptual system has certain expectations regarding these (Gibson 1979). For example, the audio and visual cues from a balloon pop are perceived as a single event when both stimuli reach our senses with some degree of simultaneity. Integrating both audio and visual timing information is an important perceptual process since they are usually not independent. That also means that human perception is aware of the physical properties (speed of sound and light) of both modalities and allow some margin between their timings. This margin is well studied in

many temporal synchrony experiments, a meta-analysis found that two experimental procedures generally support a greater tolerance to audio delay (~80 ms) as opposed to audio lead (~30 ms) depending on conditions (Kohlrausch & van de Par 2005). This can be seen in in figure 1 where the white symbols represent a two-category synchrony procedure ('synchronous' or 'asynchronous' response), black symbols represent a three-category procedure ('audio first', 'synchronous' or 'video first') and the grey symbols represent a two-category temporal order judgement (TOJ) procedure ('audio first' or 'video first'). The TOJ procedure produced unreliable results between participants so it seems like this kind of response structure is better avoided in tasks pertaining to synchrony detection.

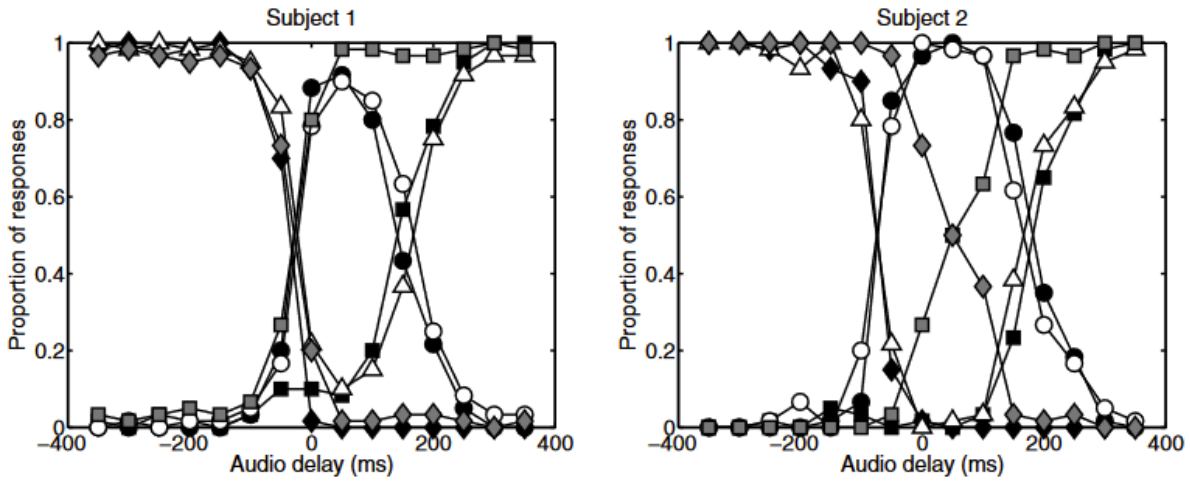

Figure 1

Audio-visual synchrony experiments for two participants performing three different experimental procedures. White symbols represent a two-category synchrony procedure ('synchronous' or 'asynchronous' response), black symbols represent a three-category procedure ('audio first', 'synchronous' or 'video first') and the grey symbols represent a two-category temporal order procedure ('audio first' or 'video first') (source: Kohlrausch & van de Par 2005).

1.4.2 Spatial ventriloquism and multimodality

Other than timing synchrony, multisensory interactions also deal with localization integration. To study this effect one can create stimuli containing conflicting information regarding event location and

letting the ventriloquist effect perceptually unite them. A previously mentioned study by Kytö et al. (2015) investigated the distance threshold between modalities of an audio-visual stimulus in an augmented reality environment. Participants were tasked with matching the location of a visual speech bubble with the location of a sound stimulus: a voice repeating a sentence. The stimulus locations were static and occurring within a 90° field of view in front of the participant. When the disparity between audio and visual POIs was below 15°, the visual and audio cues were perceived as united. Above 35° both modalities were perceived as separate (Kytö, Kusumoto & Oittinen 2015). The (in this case) 15° upper bound for perceived unity and 35° lower bound for perceived disparity will be referred to as the ventriloquism thresholds.

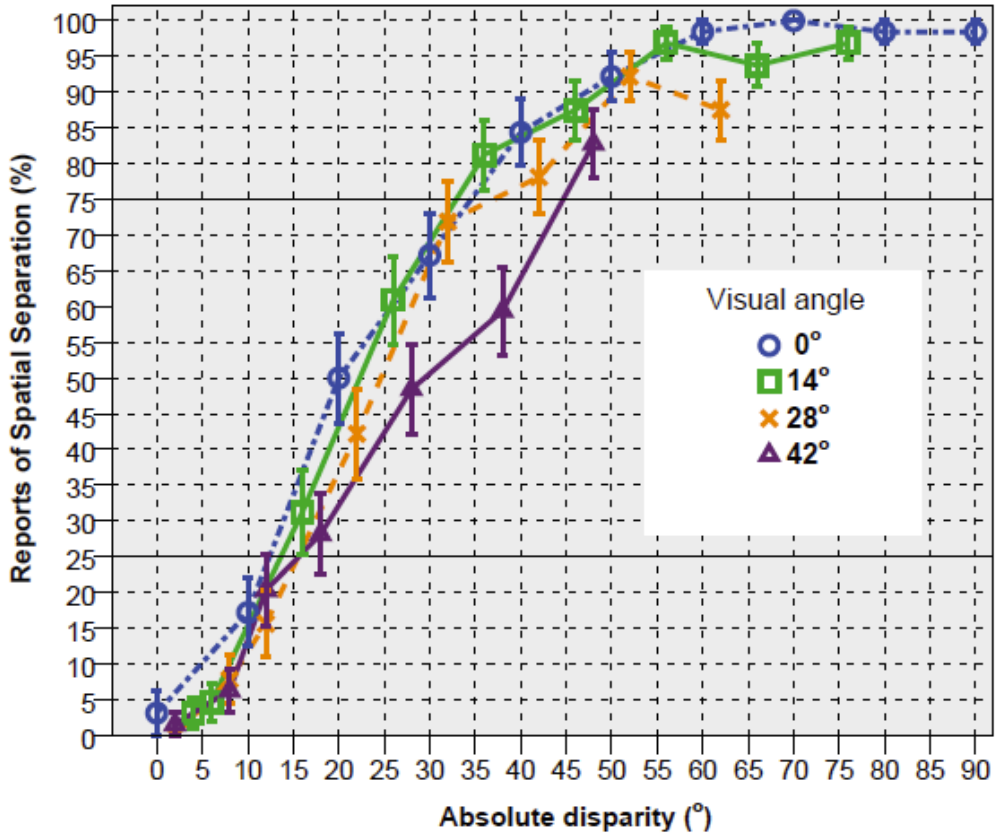

Figure 2

The rate of reports of spatial separation as a function of absolute disparity after 2-AFC procedure (source: Kytö, Kusumoto & Oittinen 2015).

As mentioned before, the ventriloquist effect occurs between more stimulus pairings than just the audio and visual modalities. In an experimental study, participants were presented with spatially displaced tactile stimuli during an auditory localization judgement task. The experimental setup resembles the one shown in figure 3. The main finding of the study was that concurrent tactile stimulation modulates the perceived location of sounds. The authors argue that the most likely explanation is an interaction at a perceptual level between tactile and auditory spatial information (Caclin, Soto-Faraco, Kingstone & Spence 2002). In a later study that took electroencephalogram (EEG) recordings during a similar experiment supports similar results by exploiting retinotopy. The EEG recordings showed that "the occurrence of the audio-tactile ventriloquist illusion is associated with a biasing of cortical activity toward the location of the discrepant tactile stimulus" (Bruns & Röder 2010). They argue that audio-visual and audio-tactile ventriloquism are mediated by a common neural mechanism.

1.4.3 Bayesian approach

Up until now the ventriloquist effect has been referred to as one modality capturing another, in practice it is less binary than that. It is generally understood that the modality having the best acuity will take preference and thus dominate the capturing effect. In audiovisual spatial ventriloquism, the visual modality has better spatial acuity which causes the auditory source location to be biased towards the visual source location (Bertelson & Radeau 1981). Alternatively, in temporal ventriloquism the auditory stimulus has higher acuity with regards to temporal resolution which causes the perceived onset to be biased towards the audio stimulus (Hartcher-O'Brien & Alais 2011). The Bayesian approach seeks to bring more nuance to this theory by introducing the two factors of likelihood and the prior. The idea is that "auditory and visual cues are combined in an optimal way by weighting each cue relative to an estimate of its noisiness, rather than one modality capturing the other" (Chen & Vroomen 2013). Likelihood represents sensory noise for a given modality. In the case of spatial ventriloquism, when visual acuity is high it captures audio source location. When visual stimulus becomes blurry or otherwise less accurate, audio captures visual location as demonstrated by Alais & Burr (2004).

The prior represents the probability of an event occurrence. In a study investigating causal inference it is argued that a low-level cognitive process takes into account the probability whether two sensory cues originate from the same location (Körding, Beierholm, Ma, Quartz, Tenenbaum & Shams 2007). This is then taken into account together with likelihood when judging location unity.

1.4.4 Motion synchrony

When considering bimodal disparities on objects in motion the distinction between temporal and spatial delay becomes fuzzy since velocity is a function of both time and distance. Audiovisual synchrony perception during motion has previously been shown to rely on congruent direction (Soto-Faraco, Lyons, Gazzaniga, Spence & Kingstone 2002). When the movement direction of a visual cue is incongruent with the direction of a synchronous audio cue, the visual movement will capture audio direction (or, depending on interstimulus intervals, cause ambiguity). In a similar experiment the effect of apparent visual motion on the performance of a TOJ task was investigated (Kwon, Ogawa & Miyake 2014). Participants were shown two successive flashes of light spaced apart 5.6° and on the second flash an auditory cue was played with a delay (positive or negative). The TOJ task was to identify this delay between the second flash and the auditory cue. Their results showed that the just-noticeable difference (JND) of this delay was smaller against the control condition where only the second flash was presented (no apparent motion). Furthermore, the point of subjective simultaneity (PSS) also shifted towards a negative delay (audio lead) of -5 ms where the control condition without apparent motion showed a PSS of 12 ms. This shows that motion has an effect on audiovisual synchrony for discrete stimuli.

In a study where two speakers, two LEDs and two tactile actuators were spatially aligned as seen in figure 3, participants responded to simultaneously presented stimuli for the auditory stream and the target stream. In the first experiment the target stream was visual motion, in the second experiment it was tactile motion. For both experiments the authors found support for audition capturing motion direction of the target stimulus. “However, the underlying mechanisms for the two types of capture effects may be different. In experiment 1, typical spatial ventriloquism can account for the findings, while in experiment 2, temporal ventriloquism is at work” (Chen & Zhou 2011). This indicates that motion capture can be

invoked by both temporal or spatial ventriloquism depending on the internal acuity weighting of stimuli. Given that for continuous motion the onset timing of stimuli is undefined, it is expected to cause spatial ventriloquism to dominate over temporal.

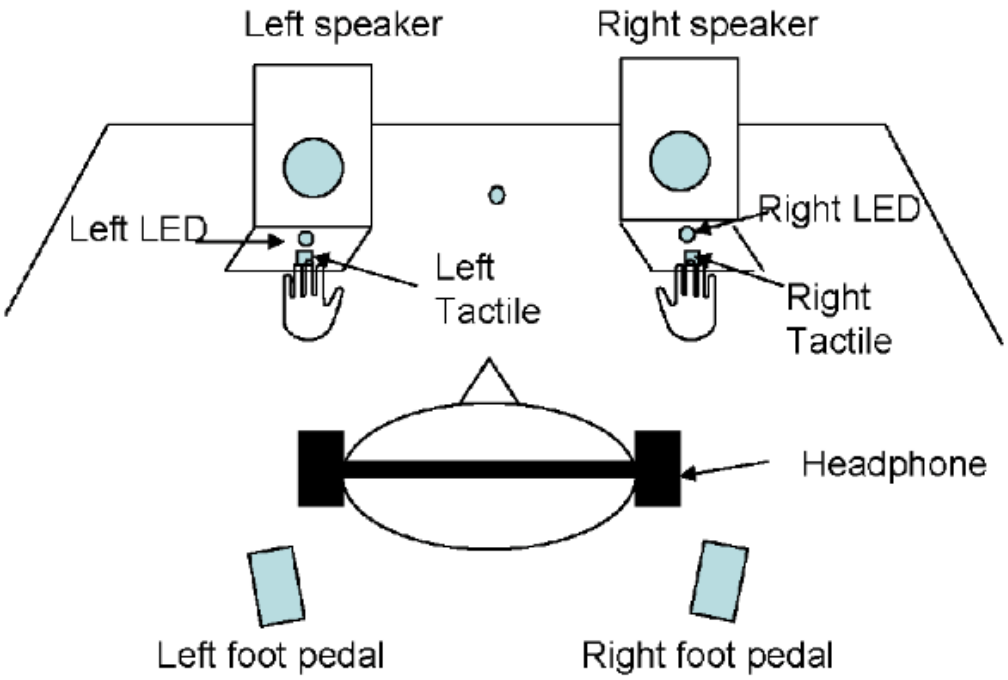

Figure 3

Common setup for multimodal motion capture experiments. The participant places two fingers on the tactile actuators, which are placed just in front of the two speakers. Two LEDs are co-located with the two actuators, respectively. Participants make their responses by lifting the left foot pedal for leftwards target motion or the right foot for rightward motion (source: Chen & Zhou 2011).

Following the previously mentioned study on the ventriloquist effect in motion from Soto-Faraco et al. (2002), they performed a replication using continuous motion rather than discrete motion. The main finding from the experiment in that study is that “the cross-modal dynamic capture effect observed in previous experiments using apparent motion displays generalizes to the case of continuous motion”. It only held for visual motion capturing audio, not vice versa (experiment #6, Soto-Faraco, Spence &

Kingstone 2004). This supports the hypothesis that spatial ventriloquism dominates over temporal ventriloquism in the case of continuous motion due to the absence of stimulus onset timing information.

To investigate the effect of motion on a VR audiovisual synchrony task the previously mentioned exploratory study was performed which will be considered a pilot for the current study (Palla Lorden 2019). An experimental setup was created where in the middle of a spherical 64 speaker array a participant was seated wearing an HTC Vive HMD. The speaker array allowed for accurate recreation of sound fields with the use of VBAP or Ambisonics. Participants performed the audiovisual synchrony task on a freehand moving object in VR. Freehand movement meant that the location of the object was determined by the pointing direction of the controller. The aim of this study was to investigate multidirectional continuous motion and its effect on audiovisual localization unity. Not enough participants were collected to conduct a proper statistical analysis. Inconclusive results indicated an asynchrony threshold value between 100 and 150 ms of perceptual audio delay. The current study seeks to replicate said experiment and expand upon it by investigating the interaction effect of the freehand movement.

2 Introduction

Previous literature has provided an understanding on the temporal and spatial ventriloquist effect between multiple sensory modality pairs (Bertelson & Radeau 1981; Bruns & Röder 2010; Samad & Shams 2016). An ideal-observer model has been formulated to describe how factors of likelihood and the prior, inspired by Bayesian statistics, predict what sensory modality will take dominance and capture the other(s) (Alais & Burr 2004; Körding et al. 2007; Hartcher-O'Brien & Alais 2011; Chen & Vroomen 2013). During experimental studies on audiovisual synchrony the capturing of motion direction was shown to depend on temporal ventriloquism for discrete audio stimuli and on spatial ventriloquism for continuous audio stimuli (Soto-Faraco et al. 2002; Soto-Faraco et al. 2004). So, in the case of a continuously moving audiovisual object the spatial ventriloquism illusion is expected to cause location capturing where a disparate sound source location becomes biased towards visual location and thus perceived as spatially united. The current study will investigate if the ventriloquism thresholds as

measured in VR will coincide with thresholds found by studies that investigated non-moving audiovisual spatial ventriloquism. This is the 3rd condition (AV) in table 1.

Similar to the findings of Kytö et al. (2015) the freedom of head movement is expected to increase sound source localization acuity (Brimijoin et al. 2013). Furthermore, the movement of the target source itself also contributes to this acuity, but only when movement direction can be anticipated by the participant (Wightman & Kistler 1999; Macpherson 2013). So, the current study will allow for a third modality: to dictate the real-time position of the target object source with pointing direction. This modality provides potentially multiple cues which may not be directly distinguishable before experimental investigation: agency and/or proprioception. For now, it will be referred to as pointing direction, with it the acuity of sound information is increased which lowers the perceptual noisiness estimation. When two cues have low noise the Bayesian likelihood factor may not disambiguate between them. Neither sense dominates and the perceived object location follows mean location due to the ventriloquist effect (Alais & Burr 2004). The increased sound source localization acuity from motion anticipation (agency) is expected to decrease asynchrony thresholds. This is the 1st condition (APV) in table 1.

When introducing pointing direction, one may expect some audio-motor or visuo-motor ventriloquism to occur. This is different from experimental audio-tactile ventriloquism research since there is no explicit tactile stimulus. To shed light on the effect of proprioception in the APV condition, it will be repeated without visual cues during the second condition (AP) in table 1. In this case location disparity only occurs between auditory source and pointing direction. The audio-motor ventriloquism threshold in this condition is expected to be higher than the results from APV and AV condition, because pointing direction is based on VR input. With controller input the inherent system delays need to be taken into account, system delays have been shown to be noticeable for visuo-haptic stimuli over 50 ms (Di Luca & Mahnan 2019). Given that both modalities may have elevated perceptual noise due to this, audio-motor ventriloquism is expected to have a high threshold for perceived synchrony. But if the thresholds

for the AP condition are actually lower than the AV condition, it means that motion anticipation has a stronger contribution towards perceived location synchrony than audio-visual capturing.

Table 1

Overview of experimental conditions and corresponding modalities present during task.

Condition	Present location cues		
	Auditory target source	*Pointing direction*	*Visual source*
1 (APV)	yes	yes	yes
2 (AP)	yes	yes	no
3 (AV)	yes	no	yes

Participants in the current experiment will each perform three tasks corresponding to the three conditions in table 1 in random order, this way the delay detection performance for each condition is measured and can be compared. The main research question is: is the ventriloquist effect in VR influenced by the amount and modality of sensory cues?

Hypotheses:

There is a difference in audio delay detection thresholds between the three conditions. (H1)

Visual location source will have bigger effect on delay detection thresholds than pointing direction. (H2)

The inclusion of a third modality results in lower audio delay detection thresholds compared to two. (H3)

3 Method

3.1 Design

To investigate sound delay detection threshold differences between the conditions a within-participant repeated measures experiment design was utilized. All participants performed three tasks, one for each condition (APV, AP, AV). Every task consisted of a dynamic number of trials. Each trial was a Two-Alternative Forced Choice (2-AFC) between two stimuli where the delayed stimulus was the target. The sound delay started at 278 ms and was adjusted after each trial following the 1-up 2-down method in steps of 1 frame time (11.11ms). For every incorrect judgement the delay was adjusted upwards (less

difficult), for every two consecutive correct judgements the delay was adjusted downwards (more difficult). After 7 reversals on the correct/incorrect streaks the task was stopped and the values for the last 4 reversals were averaged for the resulting threshold. The adaptive staircase equilibrium for this 1-up 2-down method is at P_c (probability correct) = 0.707. On the psychometric curve for a 2-AFC procedure this corresponds to d' = 0.77.

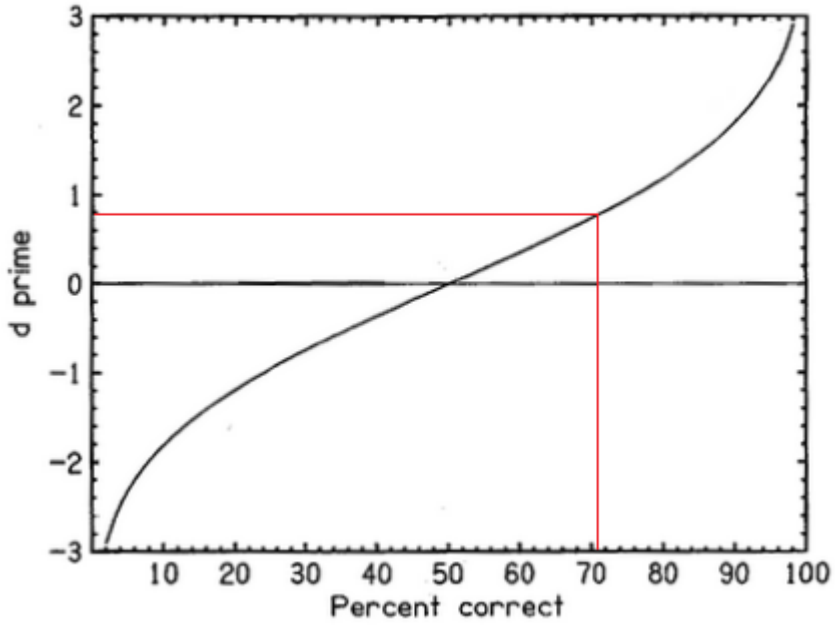

Figure 4

Psychometric curve for 2-AFC procedure with corresponding d' = 0.77 and P_c = 0.707 for a 1-up 2-down method.

3.2 Participants

The experiment for this study was performed during the COVID-19 pandemic of 2020 so special precautions had to be taken to prevent physical human interaction. For this reason, the experiment was developed as a standalone application that could be distributed to participants for at-home execution. Participants were recruited online with the requirement that they had access to an HTC Vive on a Windows PC. The briefing was done by email containing a short description of the procedure and instructions on how to run the application. Each participant digitally signed a consent form informing them of their (anonymized) data being used for academic research. A total of 17 participants conducted

the experiment, of which 11 are male and 6 female. Mean age was 33.1 years, ranging from 22 to 59 years. All participants reported normal vision and normal hearing.

3.3 Materials & stimuli

Since the experiment was performed at participants' homes the physical setting was variable. The experiment was designed to work seated or standing still to ensure any play space was accommodated. Any Windows PC with an HTC Vive was supported. The HTC Vive has a resolution of 1080x1200 pixels per eye, a 110° field of view and a refresh rate of 90 Hz. The experiment was designed to be lightweight to ensure a constant framerate of 90 fps among a large range of possible hardware. Frametimes were monitored to confirm this. For the pointing direction and response confirmations the HTC Vive controllers were utilized. The touchpad click button was mapped to cycling between intervals and the trigger click button was mapped to interval confirmation. Only one controller was required to perform the experiment. Participants were free to use any pair of stereo headphones.

The experiment application was built in Unity 3D. The temporal resolution of the virtual environment was tied to the Unity 3D Engine physics simulation which occurs at the same interval as the framerate (90 Hz). This meant that the smallest measurable delay threshold was 11.11 ms.

The audio stimulus was a continuous broadband white noise rendered using the Steam Audio plugin from Valve, it was the only free audio rendering engine in Unity 3D that allowed for custom HRTF datasets. The binaural rendering was based on a generic HRTF from the KEMAR dataset (Gardner & Martin 1995). The directivity of the sound source was omnidirectional by setting dipole weight to 0. HRTF interpolation was set to bilinear to ensure smooth audio during head movements.

The spatial blend setting in Steam Audio was set to 1 for a fully spatialized audio source, binaural rendering was enabled and any other processing like air absorption, occlusion, reflections, reverb and distance attenuation were disabled. To match this free field sound replication the visual environment was left empty with only a floor to stand on. A horizon was visible and a blue sky with a neutral light source as per the default skybox in Unity 3D (figure 5). A minimal amount of visual rendering also ensured

stable frame times by reducing graphic processing load. In front of the participant a digital timer was displayed to indicate the remaining time to confirm their response.

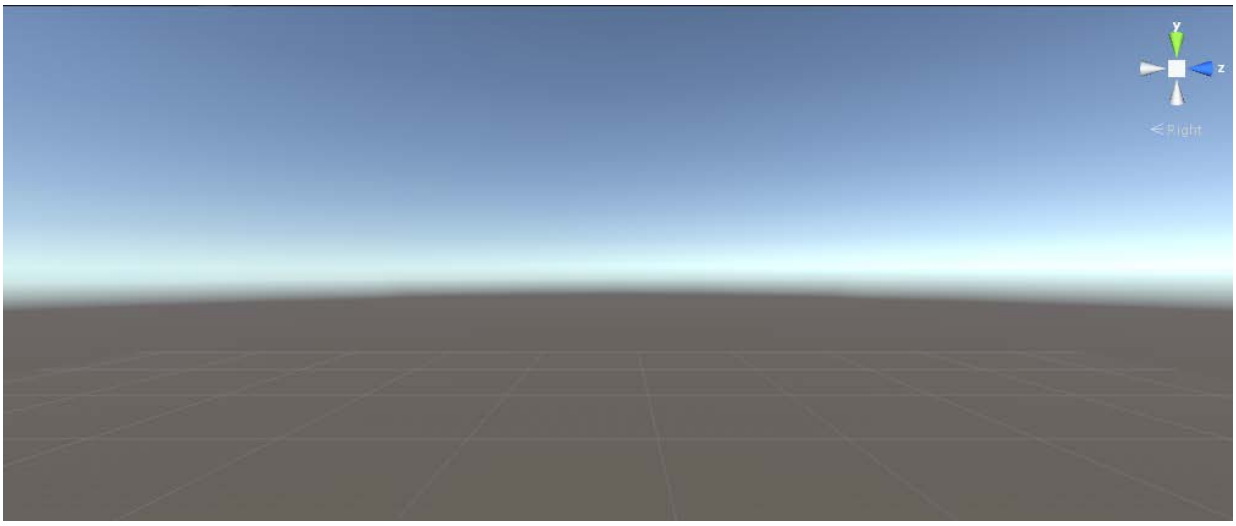

Figure 5

Unity 3D default skybox.

3.4 Procedure

Participants downloaded and started the custom-built VR application on their home computers while wearing their VR head-set. The application itself starts in a virtual environment with a menu where participants are informed of their progress towards the experiment and a reminder of the controller button mappings. When the participant indicates they are ready, the first of 4 tasks start. The first task was a tutorial phase where they were familiarized with the procedure and response mechanisms. During this phase the participants were performing the APV condition. They were standing inside a semi-transparent sphere with 2.4 m radius, wherever their pointing direction intersected with this sphere determined the location of the source-object in real time. This object emitted a continuous localized sound stimulus and was visually represented by a small sphere 30 cm in diameter (7° visual angle). Two intervals existed for each trial: one with a sound delay, one without. Participants could cycle between intervals as often as they liked and were free to use any movement for the pointing direction. Color coding (red and blue) was used as a visual aid to distinguish between the two intervals. Once the participant had decided on the delayed

interval, they confirmed their choice to move on to the next trial. During the first tutorial task 10 consecutively correct answers were required to move on to task two. The next three tasks tested the three conditions shown in table 1. They differed from the tutorial by having a time limit per trial (20 s) and a scaling difficulty following the 1-up 2-down procedure. After each task they would return to the menu environment where they were informed of their progress and encouraged to take a short break. The dynamic nature of the difficulty scaling made the experiment length variable, average length to completion was 35 minutes. After all 4 tasks were completed the application generated a file with the experiment results on their PC which they sent over with the consent form. Each participant was compensated for their time an amount of 9,50 euro (some participants refused compensation).

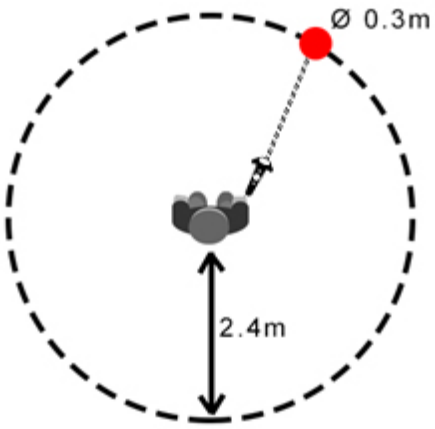

Figure 6

Illustration showing the virtual environment. Dotted circle is a 2.4 m radius sphere surrounding the participant and on the surface of this sphere the 0.4 m diameter target (illustrated in red) moves.

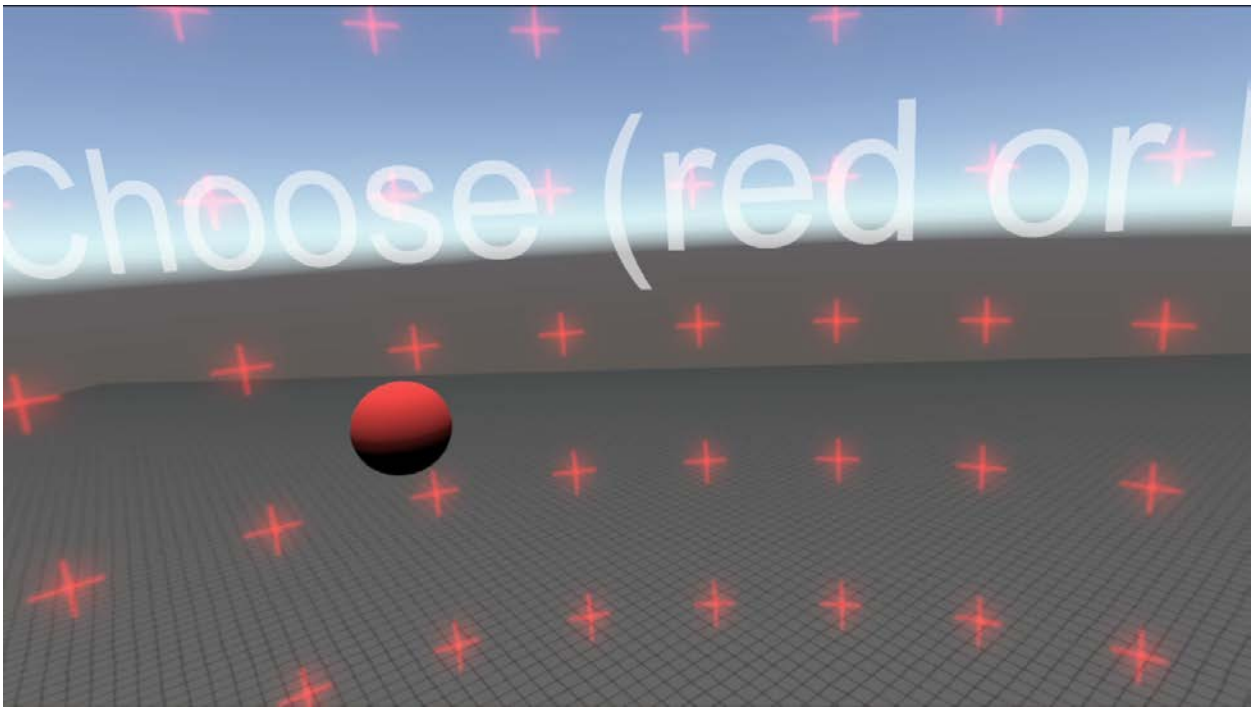

Figure 7

First person view (screen capture) from participant's perspective. The red target is following the participant's pointing direction on the surface of the large sphere. The color of the target (opaque) and the outer sphere surface (transparent cross pattern) both indicate the currently selected interval (red or blue).

4 Results

Every statistical analysis was conducted on a 0.05 significance level (two-tailed). Every effect size was interpreted according to the conventions by Cohen (1992). For all statistical tests the assumption criteria were examined. One outlier was identified based on the participant reporting a misunderstanding of the task, this brings the total to 16 participants.

The audio delay detection threshold is a relative threshold measured as a JND. An absolute delay was not measured during this study but an extensive study towards HTC Vive Pro app-to-display latencies has been completed (Le Chénéchal & Chatel-Goldman 2018). In this paper they tested a mean app-to-display delay of 31 ms using the same game engine as the current study (Unity 3D). The mean

app-to-headphone delay they measured was 66 ms. This means that the absolute delays were at least (66 - 31 =) 35 ms higher than measured here.

The APV condition resulted in a threshold of M = 99 ms, SD = 76 ms with a range of [0 ms, 256 ms] (A delay of 0 ms occurred when the participant's threshold was lower than the experiment's temporal resolution of 11 ms, this happened once during 1 trial). A M = 92 ms, SD = 56 ms threshold in a [14 ms, 214 ms] range was measured for the AP condition and a M = 162 ms, SD = 84 ms threshold with [42 ms, 303 ms] range for the AV condition as shown in figure 8. All analyses performed with N = 16.

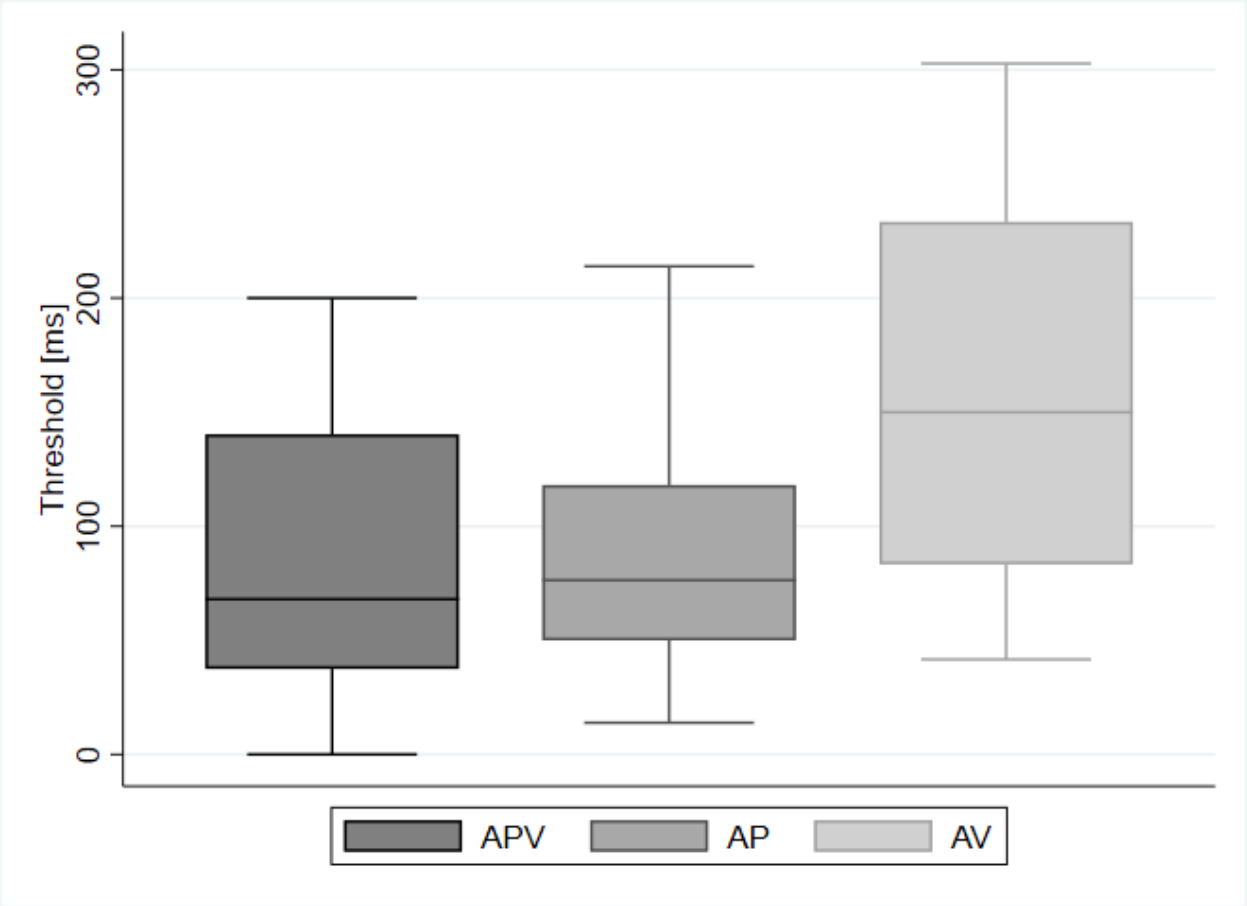

Figure 8

Boxplot indicating audio delay detection thresholds in ms for APV (Auditory source, pointing direction, visual source), AP (auditory source, pointing direction) and AV (auditory source, visual source) conditions. The box indicates median value and 75th and 25th percentile range, the lines indicate upper and lower values.

The main hypothesis was a delay detection threshold difference between the three conditions. To investigate it a one-way repeated measures ANOVA between the three conditions was performed. Sphericity was violated so the Greenhouse-Geisser correction available in Stata was applied. A significant difference was found with $F(1.85, 27.8) = 18.07$, $p < .0005$ with a large effect size $R^2 = 0.85$. The effect size for the three conditions was $h_p{}^2 = 0.54$. With these results we reject the null hypothesis that there existed no difference between conditions.

To investigate the three conditions separately Tukey's post-hoc analysis was performed. This indicated that the contrast between the APV and AP condition was extremely small (0.87, $p = .997$) and not significant compared to the contrast between APV and AV (67.0, $p < .0005$), AP and AV (66.0, $p < .0005$) which was large and significant. This means that taking away participant's ability to use pointing direction added the only statistically significant change to their thresholds during this experiment.

Spatial ventriloquism thresholds are often expressed in terms of visual angle on participant's field of view. The temporal thresholds from the current experiment were converted to visual angles to allow comparisons to previous literature. Only temporal delay was controlled which during the task manifests as spatial separation due to movement. Freehand pointing direction movements from the participants made speed variable. To determine mean target speed, total distance travelled was divided by time elapsed as recorded by Unity 3D. This was done for each condition per participant. By taking this mean movement speed data and multiplying it with their temporal threshold the mean distance threshold was determined. To express the distance threshold as visual angle the following formula was utilized: $V = 2 \cdot tan^{-1}(\frac{S}{2 \cdot D}) \cdot (\frac{180}{\pi})$ where V is the visual angle [deg], S is the distance threshold [m] and D is the static distance [m] between eyes and target (2.4 m). This formula holds under the assumption that participants are following the target source with their gaze, meaning that the distance threshold is perpendicular to the line of sight. The resulting visual angle threshold can be seen in figure 9, the AP condition is not included because it did not involve a visual target source.

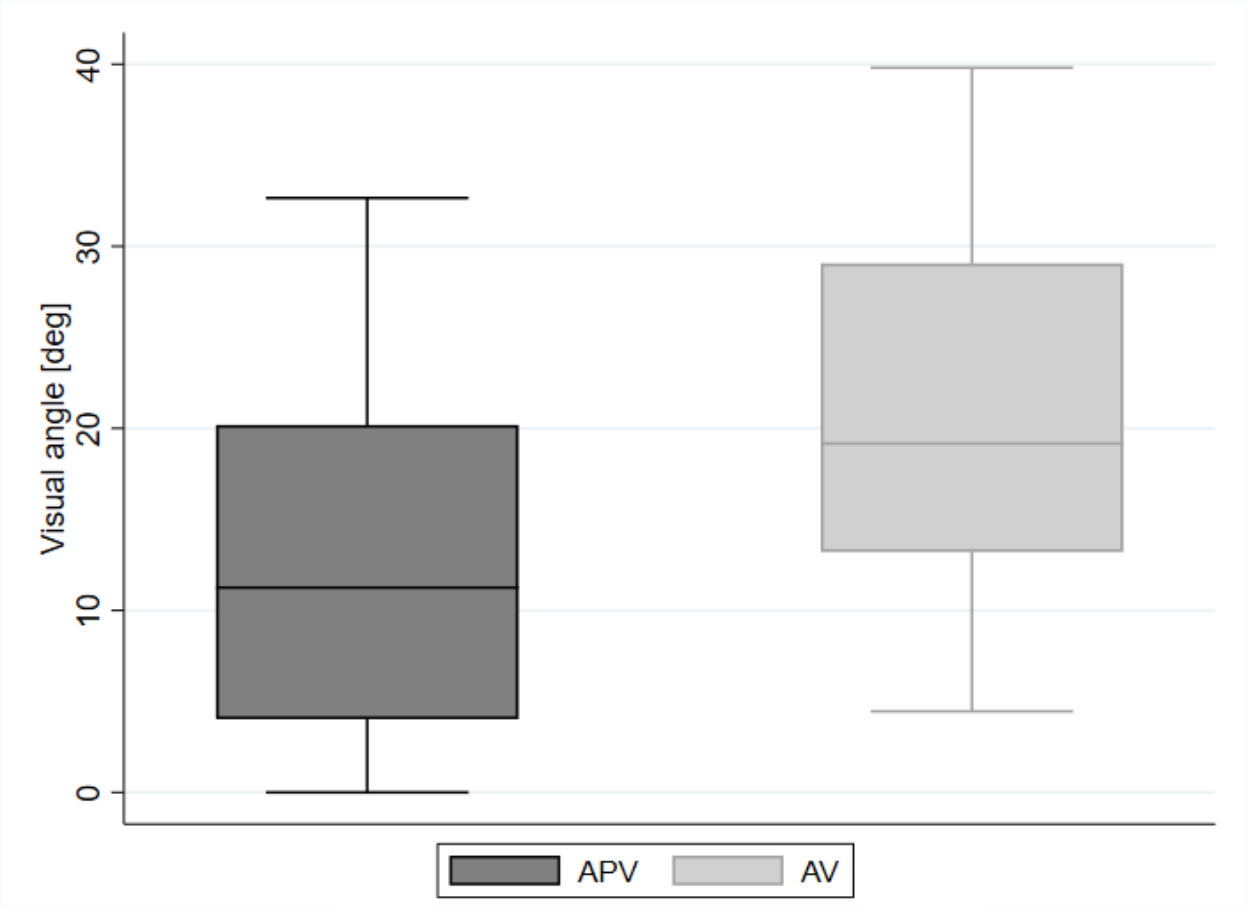

Figure 9

Boxplot indicating audio visual angle detection thresholds in degrees for APV (Auditory target source, pointing direction, visual source) and AV (auditory target source, visual source) conditions. The box indicates median value and 75th and 25th percentile range, the lines indicate upper and lower values.

Since temporal delay was kept constant and freehand pointing direction movements could vary in speed it may be interesting to look at the effect of speed on detection threshold. The mean movement speed of the target source object in the virtual environment was 6.3 m/s, SD = 2.6 m/s with a range of [2.1 m/s, 11.7 m/s]. A linear regression between movement speed and detection threshold revealed no significant relationship $F(1, 30) = 1.03$, $p > 0.31$ with $R^2 = 0.03$. This means that the speed at which participants were moving the target source had no effect on their delay detection task performance.

5 Discussion

5.1 Findings

5.1.1 Main hypothesis

The results from this experiment found support for differences in sound latency sensitivity when presented with different cues. The fact that the two conditions APV and AP did not produce significantly different results may be interpreted as the visual cue not contributing enough compared to the pointing direction. It is unlikely that this is caused by cue saliency since the temporal and spatial resolution limits are determined by the software for the target cues. So despite the visual cues' resolution being software limited (pointing direction is not), when comparing it to the target sound cue this advantage in saliency disappears. It does raise the question why pointing direction cues seem to provide a bigger advantage than visual location cues.

To investigate this, it is helpful to consider the perceptual and cognitive differences between pointing direction and visual cues when judging sound location. As mentioned before, between the audio and visual location cues the ventriloquist effect may occur when information is incongruent. Between audio location and pointing direction it is more complicated as the source of the mental representation from pointing direction is debatable. Two sources will be considered which may potentially compete or synergize: on one hand there is the bottom-up process of proprioception of the participant's hand in VR. On the other hand, there is the autonomous part of pointing. As opposed to the former, this is a top-down process as it involves the participant's intent when directing the position of the target cue. As discussed by Chen & Vroomen (2013) there exists at least a modulating influence of attention (top-down processing) on the ventriloquist effect which is required for intentional pointing movements. Wightman & Kistler (1999) found when a listener knows the direction of a sound source movement by giving them control, it decreases front-back confusions indicating a sensory acuity improvement. In the experiment from the current study the action of pointing the controller with intent is an example of agency in VR since it results in sensory predictions towards the upcoming location.

Both the bottom-up proprioceptive location source and the top-down agency determined source are subject to the input lag of the system. Meaning that the mental representation of the location source and the software representation of it occur between (4), (5) and (1) in figure 10. But the proprioceptive cue may eventually have less saliency due to a limitation from the experiment: the target source does not appear at the exact location of the participant's hand, rather at an extension of it on the outer sphere in the virtual environment. The consequences of such an extension have been shown to be detrimental to proprioceptive accuracy (Mine, Brooks Jr & Sequin 1997). Given this fact it may be valid to assume that the top-down processing resulting from agency takes preference over the proprioceptive cue.

5.1.2 Audio-visual delay detection threshold

The AV condition from this experiment is essentially a replication of earlier audio-visual ventriloquism threshold studies. The limits for the audio delay detection threshold during visual temporal order judgement tasks has been quantified to be around 80 ms (Kohlrausch & van de Par 2005). And a measurable visual capturing as a result from temporal ventriloquism has been shown to occur at 100 ms (Bresciani & Ernst 2007) and up to 300 ms (Slutsky & Recanzone 2001). These are in line with the audio delays applied in the current experiment: 155 ms with a standard deviation of 82 ms. The limits for spatial ventriloquism have been studied in real environments and has been quantified to be between 20° and 38° visual angle depending on the experimental conditions (Hairston, Wallace, Vaughan, Stein, Norris & Schirillo 2003; Jack & Thurlow 1973; Witkin, Wapner & Leventhal 1952). When investigated in an augmented reality environment the disparity should be at least 30° azimuth in order to perceive the audio and visual sources as separate (Kytö, Kusumoto & Oittinen 2015). After converting audio delay to visual angle using mean speed data the threshold from the current experiment was 15° with a standard deviation of 10°.

The reason that both spatial and temporal thresholds from the current results are on the lower end of previous literature's findings could be the presence of both these phenomena concurrently. Since movement involves both distance and time cues it may reduce the limits where the illusion of unity can occur. Additionally, in the previous experiment by Kytö et al (2015) separation was limited to the

horizontal plane whereas the current experiment allowed for any direction. This explanation relies on the assumption that multiple cues (time, location and direction incongruencies) may integrate into a more accurate perception of the differences between sensory modalities (Ernst 2006), consequently decreasing the threshold for the illusion of unity due to reduction of variance in the perceptual estimate.

5.1.3 Multisensory integration

Since this experiment involved multi-sensory stimuli it is interesting to consider the integration of these signals during the audiovisual synchrony task. The results from Diederich & Colonius (2004) "favored an explanation of the multisensory enhancement effects in terms of a coactivation mechanism that combines activations from the different modalities to jointly trigger a response". In their study they utilized audio, visual and tactile stimuli to show a decrease in reaction times when additional modalities were presented. Two other studies involving those same sensory modalities found a diminishing effect or even no advantage at all. In the first study this was applied to perceiving sequences of events (Bresciani, Dammeier, Ernst 2008) and in the most recent study to rhythmic synchronization (Johnson, Hsu, Ostrand, Gazzaley & Zanto 2020). The results from the current study are comparable to the latter two studies since no benefit of trimodal stimuli over bimodal was observed. It suggests that the ventriloquist effect as performed in this experiment suffers at least diminishing returns from the addition of a third modality. This does not hold for the AV condition. As mentioned before and discussed by Diederich & Colonius (2004): "a possible neural basis for such an [trimodal] interaction would be provided by the existence of trimodal multisensory neurons that are sensitive to visual, auditory, and somatosensory stimulation simultaneously". But comparing the delay thresholds from the current study between the APV and AP condition suggests that visual cues do not contribute significantly to the multisensory integration that is hypothesized to be present. A more likely explanation would be that the potential top-down pathway of the pointing direction retrieved location bypasses the processing channels responsible for multisensory integration.

5.2 Limitations

The limitations of this experiment start with the accuracy of the threshold measuring method. This experiment used a 2-AFC adaptive staircase method with a 1-up 2-down rule. Adaptive techniques like this are widely used in psychoacoustics to determine a specific point on a psychometric curve. However, there are slight variations of it that differ in efficiency and effectiveness when finding the threshold (Kollmeier, Gilkey & Sieben 1988). The method used in this experiment was mainly chosen for practical reasons. A 3-AFC method would complicate an already arduous task since intervals could only be displayed consecutively. A 1-up 3-down rule could have increased efficiency in turn decreasing experiment duration, but would make accidental mis-clicks more impactful. Following the suggestion by Kollmeier et al., experimental design criteria like length, complexity and reproducibility were weighed more heavily than the efficiency of the threshold estimation.

The largest compromise that had to be made for this study was the experiment setting. Instead of a lab the experiment was performed by participants at their home VR setup. This made the setting less controlled than ideal and prevented in situ intervention when needed. To reduce variability in the experiment it was designed to only work for one type of VR equipment. It ensured that all participants experienced the same visual stimuli without variation in refresh rate, field of view etc. The HTC Vive unfortunately does not include a standard earpiece so sound playback devices could not be controlled for. The impact of a headphones on listener HRTF experience is minor but present. Every participant in this experiment confirmed the use of either in-ear or circumaural headphones. Both these types of headphones perform well for HRTF localization tasks as opposed to tube-insert and bone-conduction headphones (Schonstein, Ferré & Katz 2008). Individual differences between headphones may still have minor influence on HRTF performance (Boren & Roginska 2011), but the within-subject design of this experiment makes these variations most likely less prominent.

To present the auditory stimuli to the participants an HRTF spatialized broadband noise was generated. In general, two options for HRTF spatialization are available: personalized and generic. For this study generic HRTFs were utilized to fit the method of the experiment better so it could be deployed

on a portable basis without preconfiguration. Support for the effectiveness of these generic HRTFs in a similar experimental design has been shown before. In this study a significant localization performance improvement was found when presenting participants with spatio-temporally aligned visual and auditory stimuli when compared to auditory stimuli in isolation (Berger, Gonzalez-Franco, Tajadura-Jiménez, Florencio & Zhang 2018). This implies a strong audiovisual capturing potential when utilizing generic HRTFs. The current study utilized the same HRTFs as Berger et al (2018), namely the KEMAR (MIT) dataset (Gardner & Martin 1995). Several other options are available e.g., SADIE (Armstrong, Thresh, Murphy & Kearney 2018) and CIPIC (Algazi, Thompson, Duda & Avendano 2001). In a convention paper all three databases were tested in a localization performance experiment. They noticed a decline in azimuth accuracy when exceeding 60° elevation difference from the transverse plane (Flanagan & Calle Benitez 2018). Unfortunately, they did not investigate differences between the databases, but the results from that study align with the qualitative investigation of the current experiment. After manual review of the current experiment's movement data a preference for mostly horizontal movements was noticeable with low elevation variance from the transverse plane.

The target auditory stimulus in the experiment was a continuous broadband noise with similar energy over the entire frequency spectrum (0 - 24 kHz). The decision for a noise stimulus was made following the suggestion of Berger et at (2018) where they suggest that broadband noise should be used for a spatial recalibration of acoustic space in VR. Additionally, earlier research has shown that a broader bandwidth noise stimulus increases binaural localization performance (Butler 1986). They found a noise bandwidth of 6 kHz centered around 8 kHz is required to prevent front-back confusions. Moreover, the inclusion of low frequencies (when using broadband noise) increased lateral localization acuity. Considering this, broadband noise stimuli seem ideal for binaural localization studies despite low ecological validity. The experiment utilized a continuous signal to eliminate signal onset which is known to be a primary localization cue (Perrott 1969).

The experimental setup involved several hardware and software layers processing information at different speeds. At the input side this means that tracking information from the lighthouses may arrive at

a slightly different time than button press information from the controller. At the output side this means that there may be a difference between audio and video latency at the headphones and HMD respectively. Previous research has measured all latencies involved in an interactive system similar to the one from this experiment; Unity 3D as software and HTC Vive as hardware (Le Chénéchal & Chatel-Goldman 2018). Figure 10 visualizes where different latencies occur in the system. (1) is the software latency, (2) is hardware latency (simulation-to-display), (3) is stimulus propagation time, (4) cognitive processing time and (5) is hardware latency (action-to-simulation). The timings posing a limitation towards this experiment are 1, 2 and 5. The 2018 paper measured an app-to-display latency of 31ms and an app-to-headphone latency of 66ms, both these latencies occur between 1 and 2 in figure 10. The controller-to-app latency measured 14ms and occurs between 5 and 1 in figure 10. Since this experiment investigated the JND between audio and visual stimuli the differences in hardware latency are important to consider since these measurements suggest a constant 35ms delay between both modalities. Another important consideration is that participants were using their own computers to run the experiments, each computer may potentially measure differently for these latencies. To minimize the impact of slow computers the app was monitoring for framerate and frametime fluctuations to report on abnormalities. Each participants' results reported consistent frametimes under 1ms deviation for 99.9% of frames.

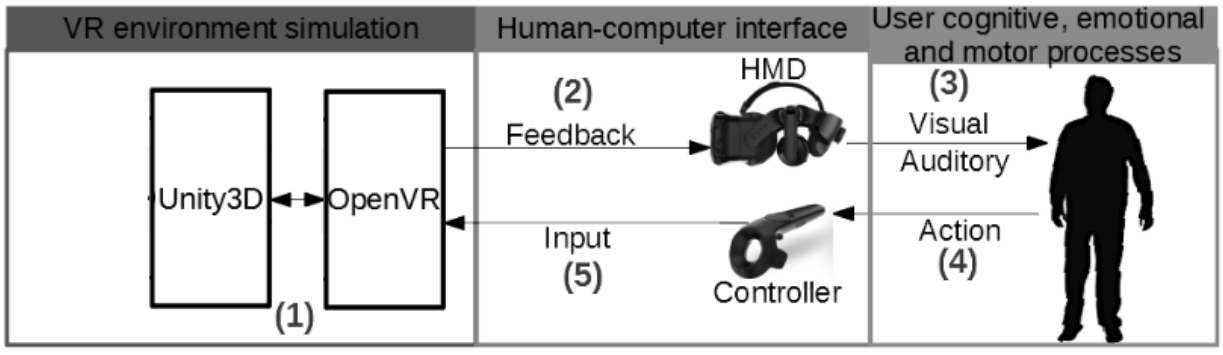

Figure 10

Illustration of the interactive system loop showing the different steps involved in latency introduction (Source: Le Chénéchal & Chatel-Goldman 2018).

5.3 Future research and conclusion

The experiment from the current study has left some results open to debate. To resolve the ambiguity of the pointing direction modality the stimulus could be altered. As mentioned before the proprioceptive cue saliency can be increased by only presenting the source at the location of participants' limbs (Mine, Brooks Jr & Sequin 1997). However, when an auditory stimulus is involved using this method there has to be adjustment for distance cues. To eliminate the predictive effect that freehand movement gains from agency it should be performed statically. This way the participant will not be able to utilize movement direction to aid in forming a mental source location. When applying both these suggestions to a replication of the current study, more convincing conclusions can be made regarding the involvement and effect of the proprioceptive modality without an interaction from top-down processes. However, the current experimental method reflects the current application of VR environments in popular use (media, simulations, etc) more. This study then contributes to understanding of these applications.

For latency-sensitive experiments in VR more robust system latency measurements are preferred. The most obvious benefit comes from the fact that absolute delay threshold measurements can be made rather than JNDs solely. Additionally, it helps with adjusting for system latency variance. Unfortunately, there is still room for improvement with regards to consumer VR equipment to facilitate scientific research better. Integrated measurement points at the DAC (digital-audio converter) and VDC (video display controller) on the equipment as done by Le Chénéchal & Chatel-Goldman (2018) would make system latency measurements trivial. The software used in this experiment could also be improved to facilitate experimental research better, mainly with regards to documentation. Currently the Steam Audio plugin documentation leaves to be desired with regards to information about signal processing e.g., no HTRF curves or other specific details about post-processing are published.

The main takeaway from the current study for the VR multimedia industry is to minimize delays to deliver an optimal experience. Although this is already a heavily researched topic (Bowman and Mcmahan 2007; Oculus 2016; Waltemate, Senna, Hülsmann, Rohde, Kopp, Ernst & Botsch 2016; Le Chénéchal & Chatel-Goldman 2018; Winkler, Stiens, Rauh, Franke & Krems 2019), the current study

sheds light on human perceptual delay sensitivity within the conditions of the experiment. It is clear that when freehand interaction is involved the delay sensitivity increases and the threshold for the ventriloquist effect breaking down decreases. This is supported by Waltemate et al. (2016) where they suggest a robust sense of agency in VR even at higher delays, but perceptually the participants inferred the presence of delays mainly from errors in the motor task and their performance. Knowing this, creators of VR multimedia should pay attention to the involved sensory modalities and degrees of freedom within their product when considering delay margins.

6 References

Ahrens, A., Lund, K. D., Marschall, M., & Dau, T. (2019). Sound source localization with varying amount of visual information in virtual reality. PloS one, 14(3).

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. Current biology, 14(3), p257-262.

Algazi, V. R., Thompson, D. M., Duda, R. O., & Avendano, C. (2001, October). The CIPIC HRTF database. In WASSAP ‘01 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics. p99-102.

Armstrong, C., Thresh, L., Murphy, D., & Kearney, G. (2018). A Perceptual Evaluation of Individual and Non-Individual HRTFs: A Case Study of the SADIE II Database. Appl. Sci. 2018, 8, 2029.

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, p21.

Bertelson, P., & Radeau, M. (1981). Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. Percept. Psychophys. 29, p578-584.

Blakemore, S. J., Wolpert, D., & Frith, C. (2000). Why can't you tickle yourself?. Neuroreport, 11(11), p R11-R16.

Blauert, J. (1984). Spatial Hearing: The Psychophysics of Human Sound. MIT Press.

Bolia, R. S., D'Angelo, W. R., & McKinley, R. L. (1999). Aurally aided visual search in three-dimensional space. Human factors,41(4), p664-669.

Boren, B., & Roginska, A. (2011, October). The effects of headphones on listener HRTF preference. In Audio Engineering Society Convention 131. Audio Engineering Society. Paper 8537

Botvinick, M., & Cohen, J. (1998). Rubber hands ‘feel’ touch that eyes see. Nature, 391(6669), p756-756.

Bowman, D. A., & McMahan, R. P. (2007). Virtual reality: how much immersion is enough?. Computer, 40(7), p36-43.

Bresciani, J. P., & Ernst, M. O. (2007). Signal reliability modulates auditory-tactile integration for event counting. Neuroreport, 18, p1157-1161.

Bresciani, J. P., Dammeier, F., & Ernst, M. O. (2008). Tri-modal integration of visual, tactile and auditory signals for the perception of sequences of events. Brain research bulletin, 75(6), p753-760.

Brimijoin, W. O., Boyd, A. W., & Akeroyd, M. A. (2013). The contribution of head movement to the externalization and internalization of sounds. PloS one, 8(12), e83068.

Bruns, P. (2019). The ventriloquist illusion as a tool to study multisensory processing: an update. Frontiers in integrative neuroscience, 13, p51.

Bruns, P., and Röder, B. (2010). Tactile capture of auditory localization: an event-related potential study. Eur. J. Neurosci. 31, p1844-1857.

Butler, R. A. (1986). The bandwidth effect on monaural and binaural localization. Hearing research, 21(1), p67-73.

Caclin, A., Soto-Faraco, S., Kingstone, A., & Spence, C. (2002). Tactile “capture” of audition. Perception & psychophysics, 64(4), p616-630.

Calvert, G. A., Hansen, P. C., Iversen, S. D., & Brammer, M. J. (2001). Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. Neuroimage,14(2), p427-438.

Carlile, S. (2014). The plastic ear and perceptual relearning in auditory spatial perception. Frontiers in neuroscience, 8, p237.

Chen, L., & Zhou, X. (2011). Capture of intermodal visual/tactile apparent motion by moving and static sounds. Seeing and Perceiving, 24, p369-389.

Chen, L., & Vroomen, J. (2013). Intersensory binding across space and time: a tutorial review. Atten. Percept. Psychophys. 75, p790-811.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), p155-159.

Cullen, J. K., Collins, M. J., Dobie, T. G., & Rappold, P. W. (1992). The effects of perceived motion on sound-source lateralization. Aviation, Space, and Environmental Medicine, 63(6), p498-504.

Di Luca, M., & Mahnan, A. (2019, July). Perceptual limits of visual-haptic simultaneity in virtual reality interactions. In 2019 IEEE World Haptics Conference (WHC), p67-72.

Diederich, A., & Colonius, H. (2004). Bimodal and trimodal multisensory enhancement: effects of stimulus onset and intensity on reaction time. Perception & psychophysics, 66(8), p1388-1404.

Ernst, M. O. (2006). A Bayesian view on multimodal cue integration. In G. Knoblich, I. M. Thornton, M. Grosjean, & M. Shiffrar, (Eds.), Human body perception from the inside out, p105- 131. Oxford: Oxford University Press.

Flanagan, P., & Calle Benitez, J. S. (2018). Localization of Elevated Virtual Sources Using Four HRTF Datasets. In Audio Engineering Society Convention 145. Paper 10115.

Gardner, W. G., & Martin, K. D. (1995). HRTF measurements of a KEMAR. The Journal of the Acoustical Society of America, 97(6), p3907-3908.

Gerzon, M. A. (1985). Ambisonics in multichannel broadcasting and video. Journal of the Audio Engineering Society, 33(11), p859-871.

Gibson, J. J. (1979). The Ecological Approach to Visual Perception. Psychology Press.

Grice, G. R., Canham, L., & Boroughs, J. M. (1984). Combination rule for redundant information in reaction time tasks with divided attention. Perception & Psychophysics, 35(5), p451-463.

Hairston, W. D., Wallace, M. T., Vaughan, J. W., Stein, B. E., Norris, J. L., & Schirillo, J. a. (2003). Visual localization ability influences cross-modal bias. Journal of cognitive neuroscience, 15(1), p20-29.

Hartcher-O'Brien, J., & Alais, D. (2011). Temporal ventriloquism in a purely temporal context. Journal of Experimental Psychology: Human Perception and Performance, 37(5), p1383.

Jack, C. E., & Thurlow, W. R. (1973). Effects of Degree of Visual Association and Angle of Displacement on the ”Ventriloquism” effect. Perceptual and Motor Skills, 37, p967-979.

Johnson, V., Hsu, W.-Y., Ostrand, A. E., Gazzaley, A., & Zanto, T. P. (2020). Multimodal sensory integration: Diminishing returns in rhythmic synchronization. Journal of Experimental Psychology: Human Perception and Performance, 46(10), p1077-1087.

Kohlrausch, A., & van de Par, S. (2005). Audio-visual interaction in the context of multi-media applications. In Communication acoustics, p109-138. Germany: Springer, Berlin, Heidelberg.

Kollmeier, B., Gilkey, R. H., & Sieben, U. K. (1988). Adaptive staircase techniques in psychoacoustics: A comparison of human data and a mathematical model. The Journal of the Acoustical Society of America, 83(5), p1852-1862.

Kwon, J., Ogawa, K. I., & Miyake, Y. (2014). The effect of visual apparent motion on audiovisual simultaneity. PloS one, 9(10), e110224.

Kytö, M., Kusumoto, K., & Oittinen, P. (2015). The ventriloquist effect in augmented reality. In 2015 IEEE International Symposium on Mixed and Augmented Reality, p49-53.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. PLoS one, 2(9), e943.

Le Chénéchal, M., & Chatel-Goldman, J. (2018, November). HTC Vive Pro time performance benchmark for scientific research. In ICAT-EGVE 2018, hal-01934741.

Macpherson, E. A. (2013). Cue weighting and vestibular mediation of temporal dynamics in sound localization via head rotation. In Proceedings of Meetings on Acoustics ICA2013 (Vol. 19, No. 1, p. 050131). Acoustical Society of America.

Mine, M. R., Brooks Jr, F. P., & Sequin, C. H. (1997). Moving objects in space: exploiting proprioception in virtual-environment interaction. In Proceedings of the 24th annual conference on Computer graphics and interactive techniques, p19-26.

Oculus, V. R. (2016). Oculus Best Practices. Available online at: http://static.oculus.com/documentation/pdfs/intro-vr/latest/bp.pdf

Palla Lorden, O. (2019, July). Continuity of moving auditory and audio-visual objects.

Perrott, D. R. (1969). Role of signal onset in sound localization. The Journal of the Acoustical Society of America, 45(2), p436-445.

Pulkki, V. (1997). Virtual sound source positioning using vector base amplitude panning. Journal of the audio engineering society , 45 (6), p456-466.

Raab, D. (1962). Statistical facilitation of simple reaction time. Transactions of the New York Academy of Sciences, 24, p574-590.

Samad, M., & Shams, L. (2016). Visual-somatotopic interactions in spatial perception. Neuroreport 27, p180-185.

Schonstein, D., Ferré, L., & Katz, B. F. (2008). Comparison of headphones and equalization for virtual auditory source localization. Journal of the Acoustical Society of America, 123(5), p3724.

Slutsky, D. A., & Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquist effect. Neuroreport, 12(1), p7-10.

Soto-Faraco, S., Lyons, J., Gazzaniga, M., Spence, C., & Kingstone, A. (2002). The ventriloquist in motion: Illusory capture of dynamic information across sensory modalities. Cognitive brain research, 14(1), p139-146.

Soto-Faraco, S., Spence, C., & Kingstone, A. (2004). Cross-modal dynamic capture: congruency effects in the perception of motion across sensory modalities. Journal of Experimental Psychology: Human Perception and Performance, 30(2), p330.

Van Eijk, R. L., Kohlrausch, A., Juola, J. F., & van de Par, S. (2008). Audiovisual synchrony and temporal order judgments: effects of experimental method and stimulus type. Perception & psychophysics, 70(6), p955-968.

Waltemate, T., Senna, I., Hülsmann, F., Rohde, M., Kopp, S., Ernst, M., & Botsch, M. (2016). The impact of latency on perceptual judgments and motor performance in closed-loop interaction in virtual reality. In Proceedings of the 22nd ACM conference on virtual reality software and technology (pp. 27-35).

Wightman, F. L., & Kistler, D. J. (1999). Resolution of front-back ambiguity in spatial hearing by listener and source movement. The Journal of the Acoustical Society of America, 105(5), p2841-2853.

Wiker, E. D. (2018). Human Perception of Aural and Visual Disparity in Virtual Environments. Doctoral dissertation, Virginia Tech. http://hdl.handle.net/10919/85015

Winkler, P., Stiens, P., Rauh, N., Franke, T., & Krems, J. (2020). How latency, action modality and display modality influence the sense of agency: a virtual reality study. Virtual Reality, 24(3), p411-422.

Witkin, H. A., Wapner, S., & Leventhal, T. (1952). Sound localization with conflicting visual and auditory cues. Journal of experimental psychology, 43(1), p58.