Compensating for Distance Compression in Audiovisual Virtual Environments Using Incongruence

Daniel J. Finnegan
Centre for Digital Entertainment
University of Bath
D.J.Finnegan@bath.ac.uk

Eamonn O'Neill
Department of Computer Science
University of Bath
E.ONeill@bath.ac.uk

Michael J. Proulx
Department of Psychology
University of Bath
M.J.Proulx@bath.ac.uk

ABSTRACT

A key requirement for a sense of presence in Virtual Environments (VEs) is for a user to perceive space as naturally as possible. One critical aspect is distance perception. When judging distances, compression is a phenomenon where humans tend to underestimate the distance between themselves and target objects (termed egocentric or absolute compression), and between other objects (exocentric or relative compression). Results of studies in virtual worlds rendered through head mounted displays are striking, demonstrating significant distance compression error. Distance compression is a multisensory phenomenon, where both audio and visual stimuli are often compressed with respect to their distances from the observer. In this paper, we propose and test a method for reducing crossmodal distance compression in VEs. We report an empirical evaluation of our method via a study of 3D spatial perception within a virtual reality (VR) head mounted display. Applying our method resulted in more accurate distance perception in a VE at longer range, and suggests a modification that could adaptively compensate for distance compression at both shorter and longer ranges. Our results have a significant and intriguing implication for designers of VEs: an incongruent audiovisual display, i.e. where the audio and visual information is intentionally misaligned, may lead to better spatial perception of a virtual scene.

Author Keywords

Distance perception; spatial audio; head mounted display; virtual environment; binaural audio; incongruent display.

ACM Classification Keywords

H.5.1. Multimedia Information Systems: Artificial, augmented, and virtual realities.

INTRODUCTION

Distance perception is a fundamental element of spatial perception. As virtual reality technology matures, its application domains are expected to broaden to such diverse areas as medical (i.e. remote surgery), military (e.g. remote manned drones) and emergency services training simulations. Many of these applications would benefit from our perception of space in the virtual world being as close as possible to our perception of space in the corresponding physical world. Distance perception is particularly important when VR is used to simulate real world scenarios in which an action must be done quickly and accurately, e.g. reaching for an object; jumping an obstacle; moving to a target. All require an understanding of the virtual space and distances to and from points [26, 41].

Previous research in spatial perception has shown that humans often underestimate or compress distances. Research into distance perception in VR has shown that the compression is significantly amplified compared to the real world [13, 22]. This poses a challenge for VR applications: how can we effectively simulate environments which require spatial perception similar to that of the real world, when the same environments reconstructed digitally in VR are more perceptually compressed?

To further complicate the issue, most research in VR has focused on spatial perception with visual displays [32, 19, 38]. However, in studies that utilized spatial auditory displays, similar compression of distance has been shown to occur [42]. Rébillat et al. investigated distance perception in audiovisual environments, and found that distance was also compressed [29]. Thus, distance compression in audiovisual VR is a *multisensory* problem, involving both visual and auditory perception and the interaction between the two. For an example of an interaction between modalities other than audition and vision, see [35].

Spatial perception is adaptive; as people move from an extended period within a VE to the real world, perceptual artifacts from the virtual world carry over to the physical world [38]. By reducing distance compression in the virtual world, we can reduce the differences between the virtual and real worlds, enabling more seamless transitions between the two.

In this paper we make 2 main contributions. First, we propose a design for virtual distance perception based on *incongruent* presentation of objects in a virtual environment. Throughout this paper, we define incongruent presentation as the intentional misalignment of audio and visual information in a scene, from the perspective of an observer. Secondly, based on this design, we propose and test a method for the systematic incongruent presentation of audiovisual stimuli to sup-

port distance perception in a VE that is closer to real world distance perception.

We begin by presenting some background work on the problem of distance compression. Next, we present the theory behind our solution in the context of previous research. We then report an empirical evaluation of our proposed solution through an experiment designed to investigate the effects of this solution on the perception of distance in VR. Finally, we discuss the implications of our results for designers of virtual environments.

BACKGROUND & RELATED WORK

Compression in Visual Environments

Previous work has highlighted a perceived compression of distance in VEs [38, 30, 21]. In these studies, for a given task there is typically a discrepancy in participants' responses within the VE compared to the real world. When asked to make a distance estimate, people typically provide varying estimates, under the same conditions, in virtual and real environments. While research shows that individual differences do exist for distance compression, it remains a general phenomenon across the population [18].

Distance compression has been attributed to various factors, such as the response measurement method used, the interaction task involved, as well as the cues available to the observer [30]. Identifying a finite set of factors influencing distance compression within virtual environments has proven difficult, and has made a concrete answer to the question 'Why do humans generally compress distance inside a virtual environment?' rather elusive. See [30] for an extensive review of the factors believed to be related to distance compression in the visual system.

The technology used to render the VE has also been considered as a factor in distance compression. In head mounted display (HMD) studies, the weight of the HMD itself has been suggested as contributing to distance compression [40]. Other researchers have linked this compression to the measurement method used [13]. Piryankova et al. compared various display technologies to investigate distance compression under different technological conditions; 2 out of the 3 VE technologies used in this study showed no significant influence on the error rate compared to the real world, for a number of varying egocentric distances to the target [28]. However, the third technology, a semi-spherical Large Screen Immersive Display (LSID) showed a significant difference in the per cent error (correct distance judgments vs. incorrect judgments) compared to an analogous real world setting. The results across the conditions suggest that distance compression is not simply caused by hardware technology. In discussing their results, Piryankova et al. speculated that a wider field of view (FOV) and resolution provided by the LSID may in combination reduce distance compression, rather than specific hardware. Jones et al. found that a large FOV (150° x 88°), or simply stimulating the visual periphery via bright light, reduced distance compression. Their results were comparable to real world spatial perception [14].

Though an exhaustive list of the the main factors contributing to distance compression is not yet known, distance compression remains a common phenomenon. Variation in results across different research means that no single causal factor can be identified. It is not enough simply to blame the technology; more work is necessary to understand the different contexts in which distance is compressed to varying degrees and to develop designs accordingly.

Compression in Acoustic Environments

Compression is not only a problem of visual perception. Zahorik et. al [45] and Kolarik et al. [17] give detailed summaries of previous research in auditory distance compression. Intensity, direct-to-reverberant ratio, and frequency spectrum are the best known cues to influence distance perception. Of the three, intensity is the most reliable cue for relative distance. Familiarity has also been shown to influence perception of distance. For example, we are familiar with the intense roar of an airplane and we typically don't expect the sound it makes, however loud it may be, to be in our near vicinity. In this instance, we may even overestimate the distance between ourselves and the airplane. The influence of familiarity is evidence of top-down processing of distance compression, involving a cognitive bias in perception. (See [30, 32, 17] for examples of familiarity as a distance cue.)

Similar to the design of visual based VEs, there are many technological factors to be considered in the production of virtual audio even before considering perceptual factors. In virtual acoustics, techniques such as binaural capture, where the acoustics of a room are captured with paired microphones tucked in the inner ear of a mannequin's head, enable virtual reproduction via headphones of audio signals within a particular acoustic environment. In headphones-based spatialisation, headphone response, binaural impulse capturing and processing, and the performance of the software have all been considered to impact acoustic spatial perception in virtual auditory displays [31].

Kearney et al. demonstrate evidence that higher order ambisonics technology, a form of 3D audio that implements multiple channels by decomposing the soundfield at a specific point into spherical harmonic functions (i.e. functions defined in terms of spherical coordinates), results in similar compression to that of the real world [15]. Spatial audio has a wide variety of applications, such as an interactive display using speaker arrays to implement a spatial music mixing room [25], and has been shown to provide an immersive experience. Ambisonics decoding over speaker arrays requires a 'sweetspot', meaning that the listener's head is required to remain fixed at an acoustically optimal position in space [4]. In order to provide for more flexible head movement (since such head movement is typically desirable in HMD-based VR applications), we chose to use binaural spatial audio operating over headphones. Through digital signal processing techniques and geometric manipulations, visual and auditory distance cues can be modified to alter the impression of the virtual space.

Distance Cue Manipulation

Kuhl et al. attempted to reduce visual distance compression in virtual environments [19]. Their technique, which they termed *minification*, involved manipulation of the geometric field-of-view in order to render objects artificially further from the observer. Their results demonstrated that participants who experienced the minified spaces underestimated distances less than a control group who received no such geometric manipulation. Later research investigated the effects on distance judgments of calibrating the pitch of the HMD [20], however no statistically significant effects were found for pitch on reducing distance compression. The authors suggested that further research could investigate possibly negative effects of HMD calibration on other aspects of spatial perception, eg. cues such as relative size of objects etc.

Zahorik describes two important results with regard to auditory distance estimation from his experiments in source position and stimulus type [44]. First, he showed that distance compression was independent of source position and stimulus type. When presented with a noise burst and a speech signal, distance estimates were shown to follow a power function fit, compressing the distance between the observer and the stimuli. This effect was observed as the angular position of the target stimuli differed from the observer's front facing direction. In a second experiment investigating the weighting of direct-to-reverberant (D-R) ratio (i.e. the ratio between the energy in the direct signal from the source to the observer and the reflection of that signal within the environment) and intensity in making distance estimates, the weights of the two cues were 'found to change substantially as a function of source signal type, source direction, and to a lesser extent, source distance' [44]. The conclusion was that D-R ratio is most likely used by the human auditory system to indicate changes in absolute distance. Discrimination between multiple closely positioned stimuli seems to rely heavily on intensity differences [45].

Given that we have control over the distance cues we present in our VEs, we can begin to consider ways in which we may manipulate the spatial environment in order to influence the observer's perception. In binaural environments, digital signal processing provides abilities to alter the intensity, frequency and reverberation present in the audio signal. Füg et al. modified binaural distance cues to study the effect upon distance perception in a virtual reconstruction of the environment's acoustics [10]. After capturing the binaural room impulse response (BRIR), an acoustic 'signature' of the room, 2 algorithms were applied to two distinct distance cues; the initial time delay gap (ITDG) and the energy decay curve (EDC). The ITDG is the time difference between the first direct sound and the initial reflection. The EDC is closely related to the reverberation time (RT_{60}), the time taken for the source signal to fall by 60 dB within a given environment. The algorithms applied involved direct manipulation of the ITDG and the energy remaining in the room after a set time.

Analysis of their results demonstrated no interaction effect between the stimuli, but an interaction effect across the modified and unmodified BRIRs was observed [10]. Manipulation of the binaural distance cues affected distance perception, supporting the hypothesis that distance perception may be controlled by direct manipulation of the intensity and reverberant distance cues; a controlled, *algorithmic* manipulation of distance perception in auditory environments. However, since this was a perceptual listening test consisting of auditory stimuli alone, it remains unclear how this manipulation will affect perception in a multisensory environment such as when using a VR HMD with audiovisual displays. See [16] for more attempts at manipulating distance perception in audio.

Incongruent multisensory environments

Through manipulation of distance cues across different modalities (in this paper, we study the audiovisual modalities), it is possible to render VEs that are *not* spatially coherent. When audio cues and visual cues are rendered intentionally misaligned to one another, we call this an *incongruent* environment. Incongruent environments can shrink and/or expand dimensions across modalities. For example, a distance of 5 meters may be represented as 5 meters visually, yet the same distance may be rendered in audio through a slight drop in intensity, intentionally ignoring physical laws regarding sound propagation in space. Conversely, an acoustic field may be mapped to a virtual visual environment that is larger or smaller than the original physical environment which it represents.

Zhou et al. incorporated 3D sound into their investigations of distance perception in incongruent augmented reality (AR) environments [46]. They focused on the intensity of a binaural source as their primary distance cue for manipulation, scaling the intensity in order to exaggerate the observer's perceived distance from the source. Their results showed that 3D audio had a significant effect on participants' ability to distinguish the relative depth of two competing audiovisual stimuli, reporting an improvement of correct distance judgments of around 250% compared to a visual only condition. They coupled their perceptual results with a questionnaire to elicit qualitative data from the participants. The audio objectively helped the participants to discriminate more accurately yet, qualitatively, more than half the participants surveyed were unclear whether the audio aided their judgment. From a psychological viewpoint, the integration of the audio stimuli with the visual stimuli results in a better estimate, even though it seems participants were not consciously aware of the benefit of the audio stimuli.

In incongruent perception studies, Gorzel et al. presented participants with incongruent, collinear audiovisual stimuli [12]. Binocular images were taken of a range of loudspeaker positions directly in front of a reference viewing point, in order to emulate photorealism in their study. A pink noise burst was presented virtually over headphones using captured BRIRs. An experimental task asked participants to state whether the sound came from in front of, behind, or the same location as a photoreal visual representation of a loudspeaker. Their results show that for a visual distance range of 2, 4 and 8 meters, misaligned audio was still perceived as consistent with the visual object, despite being rendered at an incongruent position. Perceptual binding (i.e. the audio and visual com-

ponents of the target being perceived together as a whole object) was maintained despite the incongruence between the visual and auditory stimuli. The authors concluded that there is evidence to suggest an incongruence margin between auditory and visual stimuli exists. Within this margin, stimuli are perceived as a single target entity. Outside this margin, however, the binding of the stimuli breaks down and two distinct targets are perceived.

Incongruities have been investigated by other researchers in the perception of distance. In particular, Sun et. al investigated the effect of visual and proprioceptive (in this case, the strength of effort required to move a bicycle) incongruence in a distance estimation task [35]. They demonstrated an improvement in visually specified distance estimates when the proprioceptive information was inconsistent with visual feedback provided through optic flow.

In a study of depth perception with stereoscopic TV displays, Turner et al. investigated the effect of incongruent audiovisual stimuli on distance estimation [37]. They found a significant effect of incongruent presentation of audiovisual stimuli. Participants judged a stereoscopic visual image as closer to them when a temporally coherent sound was played at a closer position over speakers which were placed physically closer to the observer. This provides evidence to suggest that incongruence between stimuli can be used to add depth to a scene, with a significant margin of incongruence where the stimuli are still integrated (or 'binded' to use the appropriate psychological terminology) as a single, multimodal stimulus.

Contrast these results with those of Chan et. al who demonstrated a negative impact of incongruent audiovisual stimuli in a target localization task [8]. In their study, participants were tasked with locating a target across two distinct spatial regions (peripersonal and extrapersonal spaces). They found that a spatially incongruent auditory stimulus affected the ability of participants to localize a visual stimulus but only in the periphery, where auditory perception is known to be more accurate than vision [6]. However, it is important to note that this study was carried out in a physical environment (lights and loudspeakers as in [37]), and that the task was not to make a distance judgment. Indeed, this is noted by the authors themselves in their discussion. Thus it is interesting that in addition to the factors noted earlier, the task at hand or the context of the judgments being made may also influence distance estimation, and this may be applicable only in physical, rather than virtual, environments.

Audio can be used to add depth to a scene but more research is needed to investigate the interactions between the manipulation of individual visual and auditory distance cues in an audiovisual environment. Manipulation of these cues will lead to variance in the estimates provided by the human visual and human auditory systems (HVS and HAS) respectively. To ask participants to make a single, multimodal distance estimate in such environments is equivalent to asking them to provide a combined estimate provided by the HVS and HAS. The next section describes a theory that addresses how humans create a combined estimate under multimodal conditions, known as multisensory integration.

MAXIMUM-LIKELIHOOD INTEGRATION THEORY

In researching how humans integrate information through the haptic and visual senses, Ernst & Banks suggested that humans integrate information from both senses in a statistically optimal fashion [9]. This optimal integration is termed Maximum-Likelihood Integration theory (ML). Using the example of ML from [9], when asked to estimate the dimensions of a virtual cuboid, a person will make an estimate of how wide and tall the object *feels* (haptic estimate) and an estimate based on how wide and tall the object *looks* (visual estimate). They then integrate the haptic and visual information together to produce a combined estimate of the width and height of the cuboid.

However, the haptic and visual information available are inherently noisy signals. When a discrepancy is observed (e.g. the haptic feedback indicates the cuboid is 20 cm wide yet visually the cuboid looks much wider), ML theory states that more 'weight' will be applied to the less noisy information. As the human visual system is much more accurate than haptic feedback through touch, a heavier weighting will be applied to the visual estimate, biasing the global estimate of 'How wide is the cuboid?' towards the visual estimate.

However, the individual estimates provided by the visual and haptic systems are not available to us as researchers. All we can directly measure is the combined global estimate provided by the human observer. In order indirectly to measure the individual estimates, we can manipulate the weighting system by artificially adding noise to one or more of the individual sensory signals. As noise is added to one sensory signal, the level of uncertainty in the corresponding estimate rises, and thus the weight applied to it is reduced. As we manipulate the level of noise in the signal, we can compare the global estimates made when different levels of noise are present in each signal. This provides a way to determine the individual estimates made through each sensory modality.

Ernst & Banks showed that this theory holds for visual and haptic modalities, but there is evidence to suggest that it also holds in audiovisual environments [5]. We use ML in this paper to determine the individual estimates made by the human visual system and the human auditory system when people are asked to estimate distances in virtual environments. By determining the individual estimates, we can determine the role that the audio and visual components of a virtual object each play in people's distance judgments.

Creating an incongruent environment, where the object's audio and visual distance cues are positioned at different depths to one another, then allows us to manipulate the contributions of the visual and auditory signals to the observer's combined estimate of distance.

DESIGN OF INCONGRUENT ENVIRONMENTS

In order systematically to position the components of a target object or stimulus, i.e. its audio and visual components, we need a method for computing *how far* the components should be positioned apart from each other in order to reduce perceived distance compression. By anchoring to the visual component of a stimulus, we can position the auditory

component by offsetting it based on the visual component's position. Next, we discuss how a systematic offset may be computed given the visual position of a target and the desired distance we want the observer to perceive.

Incongruent Positioning

Anderson et al. investigated distance compression in virtual auditory environments. In their work they provide the following exponential function for describing the degree to which humans compress distance:

$$\hat{y} = k\phi^{\alpha} \tag{1}$$

where $\hat{y}$ is the perceived target position, ϕ is the actual target position, and α and k are the slope and intercept respectively [3]. If the ϕ, α, and k parameters are a good representation of distance compression in VEs, they describe mathematically an equation between the actual distance between the observer and the target, and the perceived distance between the observer and the target. Given any 3 parameters to the equation, we can solve for the fourth. If we know the perceived position of the target $\hat{y}$, the slope of the function α, and the intercept coefficient k, we can solve for the actual position of the target.

We can move the variables over the equality sign in order to compute a value for ϕ based on a given value for a perceived position $\hat{y}$. This changes the semantics of the variables a little: rather than $\hat{y}$ acting as a perceived distance or position, it now represents the *desired distance* we want the observer to perceive. α, $\hat{y}$, and k maintain their semantics from the original equation. In this study, values for α and k were taken to have the values 2.22 and 0.61 respectively, based on the work by [3].

Once this positional offset has been computed, we can pass it into the binaural system's auditory distance rendering (ADR) algorithm. Combined with the visual rendering system, we can produce an audiovisual environment that is incongruent. This is the method we propose for the systematic positioning of incongruent stimuli in order to design an audiovisual VE that takes account of humans' compression of distance.

In order to derive the positioning function, we begin with the function given by [3] and expressed above in Equation 1. Dividing through by k and taking the inverse of the function gives us:

$$\phi = \left(\frac{\hat{y}}{k}\right)^{\frac{1}{\alpha}} \tag{2}$$

Using Equation 2, we can systematically position the audio component of a virtual object incongruently to its visual component.

EXPERIMENTAL INVESTIGATION

In our experiment, we assessed whether incongruence of collinear audiovisual stimuli affected distance perception in a virtual environment. The experiment was composed of a series of conditions involving unisensory and multisensory stimuli, with the virtual environment presented using state-of-the-art HMD hardware.

Previous studies have applied techniques involving absolute distance judgments, however, experiments involving verbal estimates of absolute distance judgments have shown a cognitive bias in participants' concepts of different metrics [23]. Hence in our experiment we used a discrimination task, a common approach in psychophysics research. A discrimination task enables us to determine the variance attributed to the weighting of sensory stimuli on the task, essential for applying the maximum likelihood theory described earlier. The experimental procedure was designed to measure the estimates for both the auditory stimuli and the visual stimuli individually.

Hypotheses

Our experimental task was designed to capture distance estimates provided by participants within the VE. We know that (at least at a range of more than a few meters), the visual component of an object will tend to produce distance compression. If we place the audio component at a greater distance from the observer, but within the incongruence margin suggested by [12], the auditory sensory signal should be integrated with the visual sensory signal to produce a combined distance estimate that is closer to the intended distance.

If we can determine the respective weightings of the 2 individual signals, we will be able to specify the positions at which we should place the visual and audio components of an object to give the desired distance perception. Drawing on ML, we would expect the auditory modality to be weighted more as the visual signal becomes less reliable. Thus, by artificially adding visual noise to the display, we can observe how the weights applied to the auditory and visual modalities change, and thereby measure the individual sensory estimates before their integration to a combined estimate.

Hence we had two distinct hypotheses, namely:

H1: Rendering the audio at an incongruent position further from the observer than the visual stimulus (IV), will result in more accurate distance perception (DV) compared to conditions where both stimuli are at the same position (congruent conditions).

H2: In incongruent conditions, an increase in visual noise within the display (IV) will lead to a shift in the sensory signal weights towards the audio modality (DV).

Participants, Apparatus and Design

Data were collected from 18 participants (7 of whom were female), with a mean age of 28. Participants were a mixture of postgraduate students and full time employees of a small company. None of the participants declared any hearing impairments and 4 had corrected vision (i.e. they wore glasses or contact lenses). All participants took part in this experiment on the basis of written, informed consent approved by the University of Bath's Psychology Research Ethics Committee, Reference 13-204, and they were free to opt out of the study at any time and without delay. The participants were not reimbursed with monetary payment for their time, nor did they receive course credit for their participation.

We used an Oculus Rift Development Kit 2 HMD[1] and audio was rendered using a custom plugin that we built for the Unity Game Engine[2]. A pair of Sennheiser HD201 Lightweight Over-Ear Binaural Headphones was used as the audio display device. Our plugin integrates the SoundScape Renderer (SSR)[3], a GPL licensed software implementation for binaural audio [1], with Unity for spatial audio rendering over headphones. Each participant was seated, with their chin resting on a chin rest to prevent head movement during the trials. The machine used to simulate the VE was a Macbook Pro (13-inch, Mid 2012 model) with a 2.9GHz Intel i7 processor, 16GB RAM and an Intel HD Graphics 4000 card, running OS X Yosemite 10.10.3.

The experiment used a repeated measures design, manipulating 4 independent variables (IVs): Modality, Visual Noise, Congruence, and Target Range. The modality factor was manipulated across 3 levels: visual-only, audio-only, and audiovisual. Visual noise was implemented at 3 levels via a gaussian blur, applied in real time to the camera view texture through a custom fragment shader written in the OpenGL Shading Language (GLSL), and applied to the camera's render callback function in Unity's rendering pipeline. Blur was implemented by a gaussian spread over the rendered scene in each frame. This approximates a gaussian blur by sampling the texture at each pixel and taking the average of the neighboring pixels. This neighboring spread was kept constant at 4 pixels to make a 9x9 grid. The blur was implemented iteratively, with the number of iterations determining the blur level. The AV1 conditions used 2 iterations, AV2 conditions 3 iterations, and AV3 conditions 5 iterations. An example screenshot of what the participants saw inside the headset is shown in Figure 1. The virtual environment consisted of the stimuli, a white plane acting as the floor, and a blue ceiling.

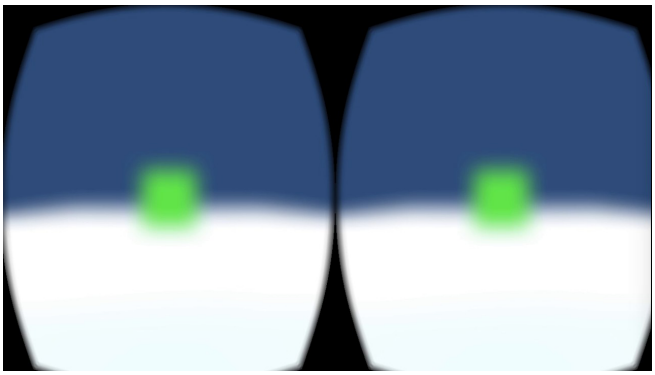

Figure 1. An example screenshot of an audiovisual condition in our experiment, with the visual noise at the highest level (AV3).

The Congruence IV determined whether auditory and visual elements of a target object were presented at the same position (congruently) or not (incongruently). In 4 conditions, the auditory stimulus was positioned the same distance from the observer as the visual stimulus. In another 4 conditions, the auditory stimulus was offset from the visual component by applying the positioning function derived above (see Equation 2). The experiment had 9 conditions in total: 6 audiovisual (3 visual noise levels x 2 congruent/incongruent conditions), 1 visual-only, and 2 audio-only conditions. A noise free audiovisual condition was not included as it does not allow for computing the relative weights of the auditory and visual signals using ML in order to determine their respective distance estimates. Conditions were presented in randomized order across participants in order to minimize order and training effects.

The range we were interested in was approximately the 10 meters in front of the observer as this is similar to the range used in related work, e.g. [3]. Each trial presented the target to the participant twice, with a brief (500 ms) disappearance between presentations. One of these presentations was fixed at a reference distance from the observer. The other presentation was positioned based on a staircase algorithm (see Procedure) which computed a distance offset from the reference distance. The initialising parameter of 2.5 meters, stepping down to 0.5 meters, for the staircase algorithm gave a total distance range of 0.5 to 9.5 meters in front of the observer, and split this total range into near (0.5 to 5.5 meters) and far subranges (4.5 to 9.5 meters). This partition into near and far subranges gave us an IV which we called the Target Range, with 2 levels. The midpoints of the near and far subranges were at 3 meters and 7 meters respectively and provided the reference distances. In the congruent conditions, both the auditory and visual stimuli were presented at these reference distances. In the incongruent conditions, the visual stimulus was at the reference distance with the audio stimulus offset by the incongruent positioning function.

The stimuli presented to each participant were the same, and consisted of a visual cube, an auditory pink noise burst, or both concurrently. A pink noise burst was chosen as it distributes the same power across each octave. This avoids conflating pitch in higher octaves with magnitude [24], as frequency spectrum is known to be a distance cue [45, 17]. The distance cues available to the participant were relative size (for the visual stimulus) and intensity (for the audible stimulus).

Procedure

Upon entering the laboratory, participants were invited to sit down opposite the experimenter, where they were handed the HMD and asked to position it until they could see a cube clearly through the HMD viewport. The experimenter then carefully adjusted the position and tightness of the strap until the participant was comfortable, ensuring the participant's pinnae were not occluded. Next, the experimenter carefully placed the headphones over the participant's ears, and helped the participant to engage the chin rest before beginning the experimental conditions. Before commencing, all participants were subjected to an interpupillary distance (IPD) measurement phase. This phase calibrated the HMD for the viewer's individual IPD, and was measured using a utility packaged with the Oculus Rift SDK.

[1] https://www.oculus.com/en-us/dk2/

[2] http://unity3d.com/

[3] http://www.spatialaudio.net

The experimental task involved, for each trial, presentation of the target (audio, visual, or audiovisual depending on the condition) at a particular distance for 500 ms. The target then disappeared and reappeared at a different distance 300 ms later. The participants' task was to indicate, using a button press on a standard computer gamepad, whether they perceived the first appearance or second appearance as closer to them. In order to choose the next distance for the target, trials followed a 3-up-1-down staircase method. 3 correct answers resulted in reducing the relative distance between each target presentation and a single incorrect answer increased the relative distance. Guidelines from [11] were followed as closely as possible in designing the staircase algorithm implemented here.

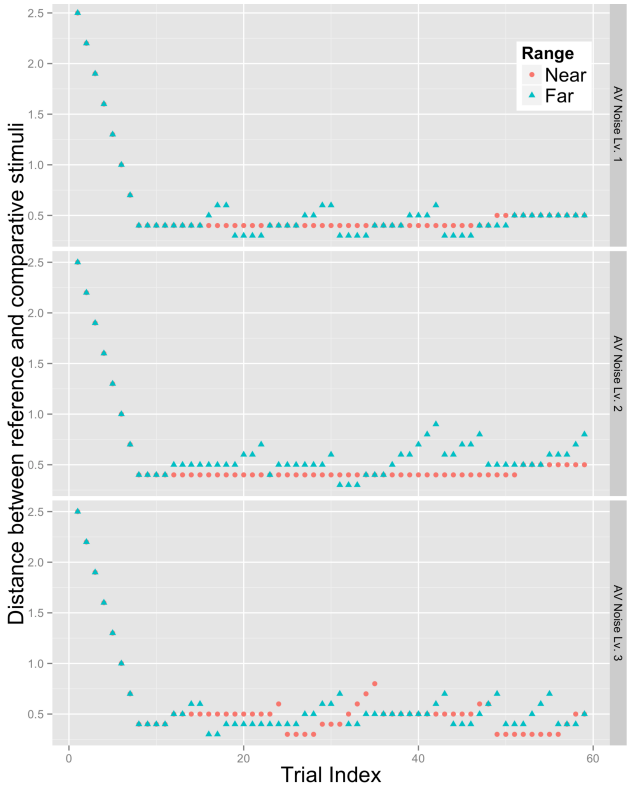

Figure 2. Staircase results for a single participant in an experimental session. Data are shown for all congruent conditions, in both the near and far ranges. Trials are graphed against the distance between the stimuli on the left y-axis. The level of noise added to the visual scene is displayed on the right y-axis

Two staircases were implemented based on the Target Range, one for each subrange. Figure 2 shows representative staircases, giving the results for a single participant across the AV1, AV2 and AV3 congruent conditions. Each condition consisted of 59 trials in each of the near and far subranges.

At the end of each condition, participants had a rest period (signaled by a red cube appearing in the center of the display until dismissed with a double tap of the gamepad's shoulder buttons) in which they were free to remove the headset and take a break before continuing to the next condition. When they were ready to continue, they were instructed to position their head so that the environment appeared with the white plane acting as a horizon in the vertical center of the viewport, and the red cube stimulus was directly in front of them (0 ° azimuth). Participants were asked to keep their eyes open during the audio-only conditions even though there were no visual stimuli in these conditions. The entire experiment took 60 ($\pm$ 15) minutes to complete.

RESULTS

Data from 18 participants were evaluated, resulting in 19,116 data points across all 9 conditions of the experiment, with 118 trials for each condition, and each participant having 1062 trials. Conditions where the audio and visual stimuli were **congruent** are termed *CON*. Conditions where the audio and visual stimuli were **incongruent**, i.e. offset with the positioning function of Equation 2, are termed *INCON*. All data were processed and all plots were produced using statistical packages (notably ggplot) from the R Language and Environment for Statistical Processing [36, 39].

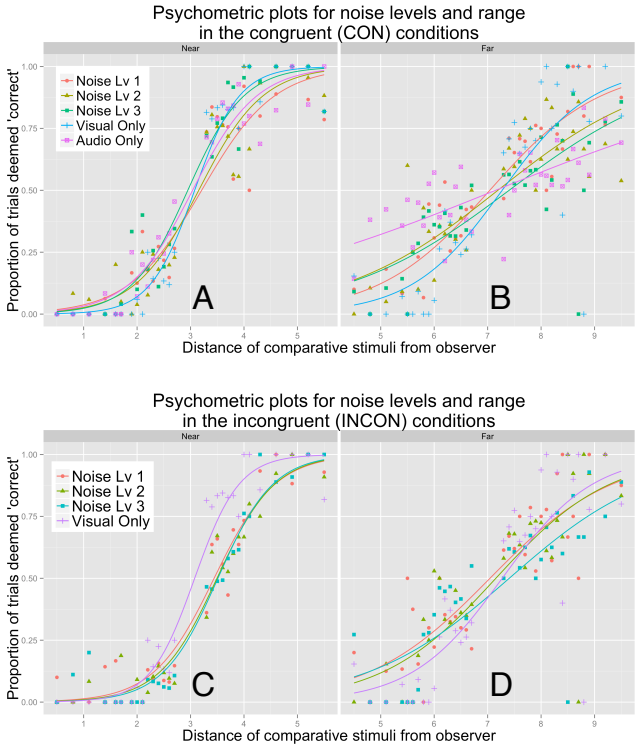

Figure 3. Psychometric functions in both near and far ranges, averaged over all 18 participants. Panel A shows results for the near congruent trials, panel B shows results for the far congruent trials. Panels C & D show results for the near and far incongruent trials respectively. The audio-only condition is excluded from the incongruent condition as no visual anchor was present and thus the audio stimulus cannot be 'incongruent' to a visual stimulus.

Figure 3 displays psychometric function data, taking the average from 18 participants. These functions are plotted in terms of *CON* & *INCON* conditions across the three levels of the visual noise factor. All participants' results were aggregated on visual noise level, range, and distance of target. The Y-axis represents the proportion of trials where the reference interval was perceived as *closer* than the comparative interval. Also,

functions are plotted with respect to the position of the visual stimulus, which acted as the anchor for the audio stimulus. In order to plot the data, an individual trial was considered 'correct' if the participant identified the reference trial interval as being closer to the participant than the standard trial interval. ML integration weights were taken from the thresholds (at 82% correctness) of general linear model (binomial) fits to the data. The functions are as predicted from **H2**; note that the slope of the functions increase as the noise in the visual modality is increased in the *INCON* conditions. This implies that the weights have shifted to the audio stimulus, and the incongruence between the audio position and the visual position is affecting the participants' distance estimates (**H1**).

Noise Level	Threshold	Slope	χ^2	Audio Weight
Near Congruent				
Lv 1	4.32	0.84	$\chi^2(27) = 14.037$, $p < 0.01$	N/A
Lv 2	4.13	0.84	$\chi^2(27) = 16.053$, $p < 0.01$	N/A
Lv 3	3.83	1.06	$\chi^2(27) = 18.371$, $p < 0.01$	N/A
Far Congruent				
Lv 1	8.62	0.58	$\chi^2(34) = 11.021$, $p < 0.01$	N/A
Lv 2	9.41	0.41	$\chi^2(34) = 6.311$, $p < 0.05$	N/A
Lv 3	9.71	0.40	$\chi^2(34) = 5.735$, $p < 0.05$	N/A
Near Incongruent				
Lv 1	4.39	0.95	$\chi^2(27) = 16.011$, $p < 0.01$	0.41
Lv 2	4.38	1.02	$\chi^2(27) = 16.820$, $p < 0.01$	0.39
Lv 3	4.38	1.02	$\chi^2(27) = 17.094$, $p < 0.01$	0.39
Far Incongruent				
Lv 1	8.78	0.53	$\chi^2(34) = 9.459$, $p < 0.01$	0.81
Lv 2	8.77	0.57	$\chi^2(34) = 10.580$, $p < 0.01$	0.83
Lv 3	9.40	0.46	$\chi^2(34) = 7.378$, $p < 0.01$	0.91

Table 1. Table of χ^2 results for the *CON* and *INCON* conditions (goodness of fit) shown in Figure 3. Weights were computed for the *INCON* conditions only. Threshold and slope are shown for each individual noise level in the visual display.

Table 1 shows χ^2 results for various binomial models constructed based on the distance of the target stimuli from the observer, and the effect of range on the psychometric functions. The χ^2 values indicate that the psychometric functions presented are good fits to the data. The threshold values in the table represent the distance from the observer when trials were answered correctly 82% of the time. **H1** predicts these thresholds to be reduced in the *INCON* conditions compared to the *CON* conditions.

Mean slope values for individual psychometric functions of all 18 participants were tested for the effect of incongruence. A significant effect of incongruence on the slopes of the far ranges, for all 18 participants across 6 (*CON* & *INCON*) audiovisual conditions, was observed, $t(102) = -1.84, p < 0.05, r = 0.18$. A non-significant result was obtained for the near range, $t(94) = 0.50, p = 0.69$. Thus **H1** is supported by the results of our analysis in the far range but not in the near range. Incongruence resulted in more accurate distance estimates when audiovisual targets were presented in the far range.

Audio modality weights in Table 1 were computed for the *INCON* conditions using the following equation from [9] (adapted for our experimental modalities):

$$w_A = (PSE - S_V)/(S_A - S_V)$$

where w_A is the weight with respect to the auditory modality, PSE is the point of subjective equality, or the point at which people are uncertain (chance level), and S_V and S_A are the visual and auditory estimates respectively. All conditions in the far range show a shift in weight to the audio modality ($> 80\%$ for audio), supporting **H2**. This shows that participants relied more heavily on the audio than on the noisy visual information. The opposite was observed for the near range; the weight dropped from noise level 1 to noise level 2 and then remained constant. With weights under 0.5 in the near range *INCON* conditions, it is assumed that participants still relied on the visual information even though the display was heavily degraded, but this calls for further research and investigation.

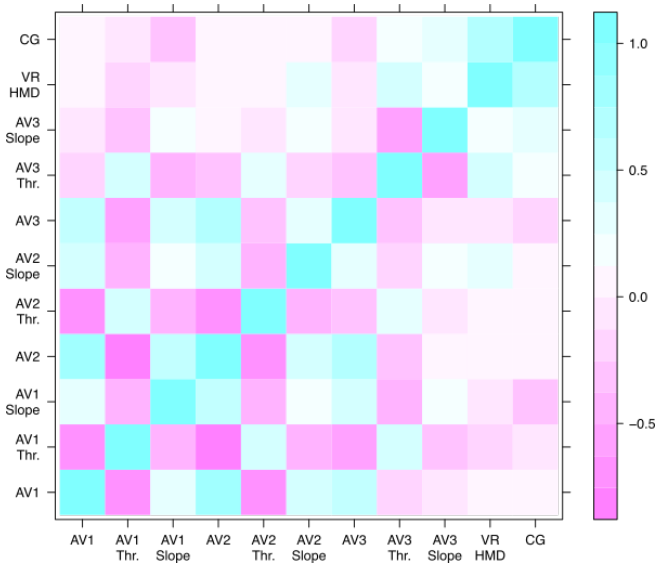

Figure 4. Pearson correlation matrix between mean accuracy in the experimental task, the slope and threshold of AV conditions, prior experience with a VR HMD, and prior experience playing computer games. The low correlations between accuracy and HMD usage, and between accuracy and game play experience, are indicators that our method is unrelated to either factor.

Figure 4 shows the results of correlations between prior experience playing computer games, prior experience with virtual reality head mounted displays, slope and threshold of psychometric functions, and mean accuracy across all the AV *INCON* conditions. The AV1, AV2 and AV3 rows represent the participants' accuracy in the audiovisual conditions, with CG (computer games experience) and VR HMD (experience with virtual reality headset displays). The correlation is low ($r < 0.5$), implying that the results of our experiment generalize across the population rather than being skewed to a subset who frequently play computer games or who are familiar with VR head mounted displays. There is no evidence to suggest that our method relies on mastery of computer games, and it is independent of prior experience with head mounted displays.

DISCUSSION

We have applied a psychophysical analysis to explore how humans compress distance in virtual environments using HMD technology. Psychometric functions vary in two main characteristics: their slope and their 50% correctness threshold (chance level). As can be seen from Figure 3 and Table 1, column 3, the slopes of the psychometric function have fallen comparing the far range to the near range *CON* conditions (Panels A & B), and similarly in the *INCON* conditions (Panels C & D). A lower slope indicates a less restrictive dynamic range; the data vary more between the threshold value and the point of subjective equality (PSE) [7]. Participants were less accurate in their estimates (lower slope) for far range than for near range *CON* conditions, which is expected given that distance estimates are less reliable further away due to compression. However, comparing the ranges across congruence and incongruence, the data show higher slopes (Table 1, column 3) in the *INCON* conditions compared to the *CON* conditions. This increase in slope means participants were *more* accurate in their estimates when presented with incongruent stimuli.

In the *CON* Noise Level 1 condition, the slope of the function is 0.84, while in the *INCON* Noise Level 1 condition it is 0.95 (Table 1, column 3). An increased slope was observed excluding the near Noise Level 3 conditions, and the far Noise Level 1 condition. Participants were more accurate in the *INCON* conditions compared to the *CON* conditions. The increase in psychometric function slope observed between identical visual noise conditions across *CON* and *INCON* conditions means participants were more accurate in the *INCON* conditions compared to the *CON* conditions, thus supporting **H1**.

The results in Table 1 indicate that the weights for the audio modality were much higher in the far range than in the near range (Table 1, column 5). The near range audio weights are all under 0.5, meaning that the audio modality estimate accounted for less than 50% of the total distance estimate. In the far range, the audio modality estimate rose as the visual noise increased, to above 90%, meaning that participants biased their estimates to the audio modality much more heavily than the visual modality. Hence, **H2** is supported only in the far range.

Threshold values decreased in the *INCON* noise levels 2 & 3 conditions compared to their corresponding *CON* conditions (Table 1, column 2). Participants could better discriminate between positions in trials where the audio and visual stimuli were incongruent, however, this effect is observed only for the far range. This finding implies that incongruent presentation did not reduce distance estimate errors when the targets were close to the observer (3 ± 2.5 meters).

Distance overestimation, where objects are perceived as further away than they actually are, occurs in both the visual and auditory domains when targets are presented close to the observer. The crossover point, that is the point at which the observer typically moves from underestimating to overestimating and vice versa, is influenced by the constant parameters k and α fitted to the positioning function [3]. This crossover point is closely related to the specific distance tendency (SDT), the point where targets are perceived when the observer is given minimal distance cues. Anderson & Zahorik found the crossover point to be 3.23 meters in an audio-only condition, and our function is based on their parameters [3]. They do not report a combined audiovisual crossover point. Our findings here suggest that the audiovisual crossover point is at a similar distance.

Our method was designed to compensate for distance compression by presenting audiovisual targets incongruently, specifically by using the visual component of a target as an anchor and systematically positioning the audio component of the target further away from the observer. Our method was applied regardless of the egocentric distance between the observer and the target. In the experiment, near range trials had a reference point of 3 meters. If the crossover point is indeed $\approx$ 3.23 meters, then we may infer that in the near range condition our function had the effect of worsening distance overestimation. As the thresholds are shifted to the right in the near range (and far range Noise Level 1 & 2 conditions), it is plausible that our function has adversely affected distance estimates under these conditions. If the crossover point can be reliably determined, our function could be modified and, instead of simply reducing distance compression by pushing the audio back (as we have shown to be effective in the far range), it could adapt to the range and reduce distance expansion in the near range by pulling the audio in front of the target's visual component.

More research is needed to investigate potential negative effects that might be introduced by using incongruence as a design tool in VEs. For example, if there is any interaction between egocentric and exocentric distance perception, it might be affected by incongruence in one domain or the other. If incongruence leads to reduced egocentric distance compression, it is unknown what effect (if any), this may have on our ability to internalize spatial maps of a scene. There is also evidence to suggest that visual experience affects internal spatial representation [27]. Further research is needed to investigate how manipulation of the audiovisual signals (in this case, intentional incongruence) may affect this internalization mechanism. Further research could also investigate the effects, if

any, of incongruence on commonly reported issues with VEs such as motion sickness and sense of presence.

In our experiment, all stimuli were constrained to the frontal view. There is evidence to suggest that localization accuracy varies as the angle between the observer's direction and the source of the sound shifts away from the 0° azimuth. Through post-hoc analysis of a real world experiment, Chan et al. provide evidence of higher accuracy in localizing a multimodal stimulus in both visual-only and audiovisual incongruent conditions [8]. Zahorik has demonstrated for auditory distance perception that the weights applied to various distance cues changed substantially for various positions from the frontal plane [43]. Further research is needed to investigate whether our results hold in audiovisual virtual environments where targets appear at various positions around the observer.

Basing our method on ML to compute the weights of the auditory and visual signals meant that it was not possible to include a noise free audiovisual condition in our analysis. Future research might apply other methods to investigate different audiovisual conditions, however, there cannot be a noise free audiovisual condition in an absolute sense. The quality of the visual signal is relative and will vary depending on, for example, which HMD is used.

This study immersed participants in a sparse, minimally populated VE within a laboratory setting. It remains to be shown whether such results can be replicated in more realistic settings which could include a variety of visual cues, auditory cues and audiovisual targets.

IMPLICATIONS FOR VE DESIGN

We hope that our results will prompt discussion on incongruent design in VE development. As the technology matures, and designers become more familiar with techniques and tools, and begin to experiment, incongruence becomes an interesting prospective tool for addressing the problem of distance compression. While more research is needed to develop more sophisticated incongruent methods, our work paves the way for others to experiment with incongruent environments. Future improvements to our method will cover incongruent presentation across different modality combinations (e.g. visual-haptic targets) and more interactive and cluttered environments.

A key feature of a VE is to be flexible, permitting users to engage in a range of behaviors. As an example of where our method could be applied, consider a VE designer's task to produce a VR movie from a first person perspective. While the environment and the narrative can be designed to follow a linear trail, at any point during the experience the observer can move freely about the scene. Hence, during a particular eventful scene, the observer may be further from the event than the designer had anticipated. If the observer is further than the crossover point from the event, she will start to compress the virtual space, and perceive the event differently from intended. Using tools that automate incongruent positioning calculation, the designer could design the scene as she sees fit, with the scene being generated automatically at runtime to accommodate for the observer's distance compression.

As VR moves away from audiovisual environments and begins to incorporate movement and interaction with physical, tangible props (recently termed substitutional reality [33]), interaction across the virtual and real worlds will become a more common phenomenon. Mixed reality will enable VR to move from an entertainment platform to a business platform, facilitating remote face-to-face meetings, virtual offices and creative spaces. These applications can benefit from psychophysical research into multisensory integration and human perception. For example, for virtual meetings, speech could be rendered incongruently to the respective 3D avatars of the participants, adapted to their relative positions, to compensate for the ventriloquist effect [2, 34]. Our findings suggest exciting future research to investigate how environments using different sensory modalities may benefit from different forms or degrees of incongruence.

CONCLUSION

Our findings suggest that intentionally rendering the auditory and visual components of objects incongruently to one another could improve distance perception in a virtual environment. We have derived and tested a method for positioning the auditory and visual elements of an audiovisual target incongruently to each other. Our results show that our method was successful when the target was at longer range; participants were more accurate in a distance discrimination task when the auditory and visual components of the targets were incongruent.

At closer range, where distance *expansion* may have affected the observers' perception, the positioning function actually seems to have made the distance estimates worse, which corroborates previous work suggesting a crossover point at $\approx$ 3.23 meters. If this crossover point can be confirmed for audiovisual targets in VEs, we can refine our method such that it is range adaptive, i.e. adapting for distance expansion up to the crossover point and for distance compression beyond the crossover point.

Our findings have implications for applications and the design of VEs. VR is an exciting but immature field, and we are still learning the techniques required to implement successful VEs. Designers should create VEs that are carefully tailored to human spatial perception. Investigating incongruence to understand its potential effectiveness in compensating for distance compression can inform engineers in developing tools for VE designers to enhance the mapping between the designer's intentions and the user's perceptions.

ACKNOWLEDGMENTS

Daniel J. Finnegan's research is funded by the UK's EPSRC Centre for Doctoral Training in Digital Entertainment (CDE), EP/L016540/1, and by Somethin' Else. Eamonn O'Neill's research is partly funded by CAMERA, the EPSRC Centre for the Analysis of Motion, Entertainment Research and Applications, EP/M023281/1. Michael Proulx's research is partly funded by EPSRC grant EP/J017205/1, Design Patterns for Inclusive Collaboration. We thank Guy McCusker and Rob McHardy for their insightful comments.

REFERENCES

1. Jens Ahrens, Matthias Geier, and Sascha Spors. 2008. The SoundScape Renderer: A unified spatial audio reproduction framework for arbitrary rendering methods. In *Audio Engineering Society Convention*. Audio Engineering Society. http://www.aes.org/e-lib/browse.cfm?elib=14460
2. David Alais and David Burr. 2004. Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Current Biology* 14, 3 (2004), 257–262. DOI: http://dx.doi.org/10.1016/S0960-9822(04)00043-0
3. Paul W. Anderson and Pavel Zahorik. 2014. Auditory/visual distance estimation: accuracy and variability. *Frontiers in Psychology* 5 (2014), 1–11. DOI:http://dx.doi.org/10.3389/fpsyg.2014.01097
4. Daniel Artaega. 2013. An ambisonics decoder for irregular 3D loudspeaker arrays. In *Journal of the Audio Engineering Society*. Audio Engineering Society. https://www.researchgate.net/publication/256802720
5. Peter W. Battaglia, Robert A. Jacobs, and Richard N. Aslin. 2003. Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America. A, Optics, image science, and vision* 20, 7 (2003), 1391–1397. DOI: http://dx.doi.org/10.1364/JOSAA.20.001391
6. Daphne Bavelier, Matthew W.G. Dye, and Peter C. Hauser. 2006. Do deaf individuals see better? *Trends in Cognitive Sciences* 10, 11 (2006), 512–518. DOI: http://dx.doi.org/10.1016/j.tics.2006.09.006
7. Leslie E. Cameron, Joanna C. Tai, and Marisa Carrasco. 2002. Covert attention affects the psychometric function of contrast sensitivity. *Vision Research* 42, 8 (2002), 949–967. DOI: http://dx.doi.org/10.1016/S0042-6989(02)00039-1
8. Jason S. Chan, Corrina Maguinness, Danuta Lisiecka, Annalisa Setti, and Fiona N. Newell. 2012. Evidence for crossmodal interactions across depth on target localisation performance in a spatial array. *Perception* 41, 7 (2012), 757–773. DOI: http://dx.doi.org/10.1068/p7230
9. Marc O. Ernst and Martin S. Banks. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415, 6870 (2002), 429–433. DOI:http://dx.doi.org/10.1038/415429a
10. Simone Füg, Stephan Werner, and Karlheinz Brandenburg. 2012. Controlled Auditory Distance Perception using Binaural Headphone Reproduction Algorithms and Evaluation. *Tonmeistertagung - VDT International Conference* November (2012). https://www.researchgate.net/publication/235707006
11. Miguel A. Garca-Pérez. 1998. Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. *Vision Research* 38, 12 (1998), 1861–1881. DOI: http://dx.doi.org/10.1016/S0042-6989(97)00340-4
12. Marcin Gorzel, David Corrigan, Gavin Kearney, John Squires, and Frank Boland. 2012. Distance Perception in Virtual Audio-Visual Environments. *25th UK Conference of the Audio Engineering Society: Spatial Audio In Today's 3D World* (2012), 1–8. http://www.researchgate.net/publication/230751682
13. Timofey Y. Grechkin, Tien Dat Nguyen, Jodie M. Plumert, James F. Cremer, and Joseph K. Kearney. 2010. How does presentation method and measurement protocol affect distance estimation in real and virtual environments? *ACM Transactions on Applied Perception* 7, 4 (2010), 1–18. DOI: http://dx.doi.org/10.1145/1823738.1823744
14. Adam J. Jones, Edward J. Swan II, and Mark Bolas. 2013. Peripheral Stimulation and its Effect on Perceived Spatial Scale in Virtual Environments. *IEEE Transactions on Visualization and Computer Graphics* 19, 4 (2013), 701–710. DOI: http://dx.doi.org/10.1109/TVCG.2013.37
15. Gavin Kearney, Marcin Gorzel, Frank Boland, and Henry Rice. 2010. Depth Perception in Interactive Virtual Acoustic Environments Using Higher Order Ambisonic Soundfields. *Proc. of the 2nd International Symposium on Ambisonics and Spherical Acoustics, 2010, May 6-7, Paris, France* (2010). https://www.researchgate.net/publication/228448070
16. Hae Young Kim, Yôiti Suzuki, Shouichi Takane, and Toshio Sone. 2001. Control of auditory distance perception based on the auditory parallax model. *Applied Acoustics* 62, 3 (2001), 245–270. DOI: http://dx.doi.org/10.1016/S0003-682X(00)00023-2
17. Andrew J. Kolarik, Brian C. J. Moore, Pavel Zahorik, Silvia Cirstea, and Shahina Pardhan. 2015. Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss. *Attention, Perception, & Psychophysics* (2015). DOI: http://dx.doi.org/10.3758/s13414-015-1015-1
18. Scott. A. Kuhl, Sarah H. Creem-Regehr, and William B. Thompson. 2006a. Individual differences in accuracy of blind walking to targets on the floor. *Journal of Vision* 6, 6 (2006), 726–726. DOI: http://dx.doi.org/10.1167/6.6.726
19. Scott A. Kuhl, William B. Thompson, and Sarah H. Creem-Regehr. 2006b. Minification influences spatial judgments in virtual environments. *Proceedings of the 3rd Symposium on Applied Perception in Graphics and Visualization - APGV '06* 1, 212 (2006), 15. DOI: http://dx.doi.org/10.1145/1140491.1140494
20. Scott A. Kuhl, William B. Thompson, and Sarah H. Creem-Regehr. 2009. HMD calibration and its effects on distance judgments. *ACM Transactions on Applied Perception* 6, 3 (2009), 1–20. DOI: http://dx.doi.org/10.1145/1577755.1577762

21. Chiuhsiang Joe Lin, Bereket Haile Woldegiorgis, Dino Caesaron, and Lai-Yu Cheng. 2015. Distance estimation with mixed real and virtual targets in stereoscopic displays. *Displays* 36 (2015), 41–48. DOI: http://dx.doi.org/10.1016/j.displa.2014.11.006

22. Jack M. Loomis, José A. Da Silva, Naofumi Fujita, and Sergio S. Fukusima. 1992. Visual space perception and visually directed action. *Journal of Experimental Psychology: Human Perception and Performance* 18, 4 (1992), 906–921. DOI: http://dx.doi.org/10.1037/0096-1523.18.4.906

23. Jack M. Loomis and John W. Philbeck. 2008. *Measuring spatial perception with spatial updating and action*. Taylor and Francis, Chapter 1, 1–43. http://www.tandf.net/books/details/9780805862881/

24. Catarina Mendonça, Jorge A. Santos, Guilherme Campos, Paulo Dias, and João P. Ferreira. 2010. On the Impact of Training HRTF-Based Auralisation. *Interacção 2010 4ł Conferência Interacção Pessoa-Máquina* (2010), 1–5. http://webs.psi.uminho.pt/lvp/site/Publications

25. Jörg Müller, Matthias Geier, Christina Dicke, and Sascha Spors. 2014. The boomRoom. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*. ACM Press, New York, New York, USA, 247–256. DOI: http://dx.doi.org/10.1145/2556288.2557000

26. Gaëtan Parseihian, Christophe Jouffrais, and Brian F. G. Katz. 2014. Reaching nearby sources: comparison between real and virtual sound and visual targets. *Frontiers in Neuroscience* 8, September (2014), 1–13. DOI:http://dx.doi.org/10.3389/fnins.2014.00269

27. Achille Pasqualotto and Michael J. Proulx. 2012. The role of visual experience for the neural basis of spatial cognition. *Neuroscience & Biobehavioral Reviews* 36, 4 (2012), 1179–1187. DOI:http://dx.doi.org/10.1016/j.neubiorev.2012.01.008

28. Ivelina V. Piryankova, Stephan De La Rosa, Uwe Kloos, Heinrich H. Bülthoff, and Betty J. Mohler. 2013. Egocentric distance perception in large screen immersive displays. *Displays* 34, 2 (2013), 153–164. DOI: http://dx.doi.org/10.1016/j.displa.2013.01.001

29. Marc Rébillat, Xavier Boutillon, Étienne Corteel, and Brian F. G. Katz. 2012. Audio, Visual, and Audio-visual Egocentric Distance Perception by Moving Subjects in Virtual Environments. *ACM Trans. Appl. Percept.* 9, 4 (2012), 19:1–19:17. DOI: http://dx.doi.org/10.1145/2355598.2355602

30. Rebekka S. Renner, Boris M. Velichkovsky, and Jens R. Helmert. 2013. The perception of egocentric distances in virtual environments - A review. *Comput. Surveys* 46, 2 (2013), 1–40. DOI: http://dx.doi.org/10.1145/2543581.2543590

31. Monika Rychtáriková, Tim van den Bogaert, Gerrit Vermier, and Jan Wouters. 2009. Binaural sound source localization in real and virtual rooms. *Journal of the Audio Engineering Society* 57, 4 (2009), 205–220. http://www.aes.org/e-lib/browse.cfm?elib=14814

32. Cynthia S. Sahm, Sarah H. Creem-Regehr, William B. Thompson, and Peter Willemsen. 2005. Throwing versus walking as indicators of distance perception in similar real and virtual environments. *ACM Transactions on Applied Perception* 2, 1 (2005), 35–45. DOI: http://dx.doi.org/10.1145/1048687.1048690

33. Adalberto L Simeone, Eduardo Velloso, and Hans Gellersen. 2015. Substitutional Reality. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. ACM Press, New York, New York, USA, 3307–3316. DOI: http://dx.doi.org/10.1145/2702123.2702389

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

35. Hong-Jin Sun, Jennifer L. Campos, and George S. W. Chan. 2004. Multisensory integration in the estimation of relative path length. *Experimental Brain Research* 154, 2 (2004), 246–254. DOI: http://dx.doi.org/10.1007/s00221-003-1652-9

36. R Core Team. 2015. *R: A language and environment for statistical computing*. https://www.r-project.org/

37. Amy Turner, Jonathan Berry, and Nick Holliman. 2011. Can the perception of depth in stereoscopic images be influenced by 3D sound? *Displays* 7863, February (2011). DOI:http://dx.doi.org/10.1117/12.871960

38. David Waller and Adam R. Richardson. 2008. Correcting distance estimates by interacting with immersive virtual environments: effects of task and available sensory information. *Journal of Experimental Psychology: Applied* 14, 1 (2008), 61–72. DOI: http://dx.doi.org/10.1037/1076-898X.14.1.61

39. Hadley Wickham. 2009. *ggplot2 : Elegant graphics for data analysis*. http://had.co.nz/ggplot2/book

40. Peter Willemsen, Mark B. Colton, Sarah H. Creem-Regehr, and William B. Thompson. 2004. The effects of head-mounted display mechanical properties and field of view on distance judgments in virtual environments. *ACM Transactions on Applied Perception* 6, 2 (2004), 1–14. DOI: http://dx.doi.org/10.1145/1498700.1498702

41. Haojie Wu, Daniel H. Ashmead, and Bobby Bodenheimer. 2009. Using immersive virtual reality to evaluate pedestrian street crossing decisions at a roundabout. *Proceedings of the 6th Symposium on Applied Perception in Graphics and Visualization - APGV '09* 1, 212 (2009), 35. DOI: http://dx.doi.org/10.1145/1620993.1621001

42. Pavel Zahorik. 2000. Distance localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America* 108, 5 (2000), 2597. DOI: **http://dx.doi.org/10.1121/1.4743664**

43. Pavel Zahorik. 2002a. Assessing auditory distance perception using virtual acoustics. *The Journal of the Acoustical Society of America* 111, 4 (2002), 1832–1846. DOI: **http://dx.doi.org/10.1121/1.1458027**

44. Pavel Zahorik. 2002b. Direct-to-reverberant energy ratio sensitivity. *The Journal of the Acoustical Society of America* 112, 5 (2002), 2110–2117. DOI: **http://dx.doi.org/10.1121/1.1506692**

45. Pavel Zahorik, Douglas S. Brungart, and Adelbert W. Bronkhorst. 2005. Auditory Distance Perception in Humans: A Summary of Past and Present Research. *Acta Acustica united with Acustica* 91, 3 (2005), 409–420. **http://www.ingentaconnect.com/content/dav/aaua/2005/00000091/00000003/art00003**

46. Zhiying Zhou, Adrian David Cheok, Xubo Yang, and Yan Qiu. 2004. An experimental study on the role of software synthesized 3D sound in augmented reality environments. *Interacting with Computers* 16, 5 (2004), 989–1016. DOI: **http://dx.doi.org/10.1016/j.intcom.2004.06.014**