Pitch Modulation Via Sonified Referencing to Support Reception of Virtual Spatial Communication

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

ABSTRACT

Spatially-complex communication is important to people living in a physical world. Virtual environments are suitable for visually representing spatial environments, such as the position, size, shape and orientation of objects to individual viewers. However, when team workers are in need of sharing such information, challenges emerge around how people perform spatially referential communication to contextualize spatial relationships, especially when visual cues available are partial or cognitively demanding to process. We propose diversifying perceptual assistance with tailored auditory cues to spatiotemporally support real-time reception and perception of spatial references. We developed an experimental prototype AuralTrace - an audiovisual pointer using audio spatiality and pitch modulation to support temporal representation of spatial reference trajectories. Our experiment investigated effects of varying audiovisual affordances of a pointer to communicate on addressee's remote learning of spatial information. Findings show AuralTrace has positive impact on learning, and pitch modulation supports spatial communication in cognitively constructive manners.

Author Keywords

Virtually-mediated collaboration; spatial communication; pitch change; multimodal interface; spatial audio.

CSS Concepts

• **Human-centered computing~Empirical studies in HCI**; *Auditory feedback*; Collaborative and social computing devices;

INTRODUCTION

Rising expansion of many operations worldwide have led to more people operating remotely. Spatial communication is no exception, and often necessary for most physical activities involving object manipulation. Fields like interior design, strategic team-based sports, and emergency response have often used virtual environments (VE) for remote operation, as they can spatially represent the task setting [5, 20, 31, 39]. During spatial communication for teaching or collaboration, the spatial information conveyed in some of these contexts can become more complex than just isolated statements about individual objects and may involve knowledge that encompasses different relationships between spatial entities. For example, in more heuristic-based discussion and planning of an interior space, higher-level relational information tends to be the focus of discussion (e.g. where the bathroom should be located in relation to the locations of bedroom and kitchen) instead of more strictly defined information (e.g. whether the TV should be 2.5 meters in front of the east wall). This concept can be extended to activities like determining which motion trajectories to take under uncontrolled spatial circumstances during sports games [31], or learning time-dependent anatomical relations in surgery training [10]. In this paper, we will operationalize 'complex' spatial information as such regarding the relations between spatial entities (objects or parts of the environment) — in contrast to piecemeal, fixed statements of spatial information (such as the '2.5 meters' statement above).

VEs are a common medium, and suitable for visually representing spatial aspects of the environment. However, when considering VE use beyond the individual, VEs include some constraints which may affect collaborative teamwork using such spatial information. For instance, the dearth of available perceivable stimuli in a VE is shown to affect individuals' sense of presence and ease of operation [23, 36]. In addition, spatial perception accuracy is inevitably diminished in VEs due to the lack of available real-world level 3D visual information [37]. Even in virtual reality (VR), compressed depth perception may affect accuracy of overall spatial knowledge [17].

Historically, shared informational context has been shown in many forms to be important for grounding in communication relating to physical tasks, which are usually spatial in nature [12, 14, 15, 18]. Particularly, gestures and pointing references are shown to facilitate deictic referencing and support communication of difficult-to-describe features for spatial tasks [14]. But apart from more physically-oriented forms of reference such as gesturing, current ways to generate and resolve references in VEs are still limited. In this paper, we wish to find different aspects in which referencing-heavy communication can be supported, which is achievable via exploring different modalities of shared sensory information in virtually-based communication for

CHI 2020, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04...$15.00.
DOI: https://doi.org/10.1145/3313831.XXXXXXX
*update the above block and DOI per your rightsreview confirmation (provided after acceptance)

physical tasks. More specifically, underused nonvisual channels like audio can be made better use of through studying more non-traditional mediums of auditory support, such as abstract sound [24].

Auditory Channels for Referential Communication

During a session of remote, synchronous communication for physical or spatial tasks, a large amount of information may be exchanged, for the purpose of discussion or learning. In these cases, people must use cognitive resources in a rather complex manner, such as resolving references, sensemaking, remembering, thinking, and taking actions. Such a process may be iterated multiple times, and the entire session history also has to be retained in memory for people to ultimately make use of the exchanged information. Past work on multimodal interfaces has discussed the possibility of designing to minimize cognitive load, which can free up a user's mental resources to perform better with more situational awareness [25, 27, 28].

We wish to provide perceptual-level support for spatial task interaction without the overhead of additional learning or operation, given the outlined complexity. We observe that the interaction context of spatial communication is inherently spatiotemporal, rendering the auditory channel as an ideal means of support [40]. The spatial aspect is more evident from the nature of spoken information and physical references made. The temporal aspect comes from the verbal limitation of multiple spatial features having to be described in sequence over time. For example, in the statement *"The bathroom should be adjacent to the bedroom, but the kitchen must be across from them"* the meanings of these spatial relations are not temporal, but the expression of them in a conversational context is inherently so.

Hence, the importance of listeners' ability to sustain awareness and retention of these dynamically-communicated relations opens up a design opportunity for manipulating the temporal presentation of spatial referencing information in ways feasible for listeners to perceive and consume such inter-related pieces of spatial information sequentially. Past work has shown that auditory perception could be more temporally-precise than visual [40], and also facilitates perception of multiple spatial features (orientation and position) in a wide range. Existing approaches with spatial audio in VEs have targeted blind users [1, 30], and there have been applications of audio pitch which show its potential as an intuitive spatial guiding signal [2, 32]. But so far, no work has integrated spatiality and pitch in multimodal auditory support for spatially-relevant remote communication.

To support reception of virtually-communicated complex spatial information, we propose AuralTrace, a referencing pointer which accommodates the perceptual spatio-temporality of spatial communication by integrating properties of sound with existing visual information. Visually resembling a red laser pointer, AuralTrace is a virtual VR-based pointer that dynamically manifests pitch and spatiality of sound in correspondence to the movement and location of an individual's pointer references. A constant tone is played whenever the laser pointer is active, and the pitch of the sound tone modulates (between high and low) directly based on the change in direction and speed of pointer movement, actively drawing user attention particularly to the shape, changing direction, and trajectory of pointer referencing (see Supplementary A, C1 for details of pitch modulation). The spatial properties of AuralTrace's sound directly correspond to the direction and position of the referenced location (sound source) in relation to the listener.

In this paper, we investigate how perceivers' reception of virtually-communicated complex spatial information is affected by AuralTrace, compared to pointers with less types of audio/visual cues, as well as possibilities of how these perceivers might use or be affected by the content-relevant nonverbal auditory information conveyed by the speaker through pointer references.

We conducted a within-subjects experiment evaluating performance of an interior spatial arrangement task which was done by using spatial information learned by perceiving remotely live-communicated spatial guidelines from a speaker. During communication, the speaker used AuralTrace or a pointer with other audio/visual cue variations (visual-only, visual with spatial audio but no pitch modulation) to support making spatial references during communication. The results of our study showed significant improvement under the AuralTrace condition in performance on the arrangement task. Because participants directly used communicated guidelines to perform the arrangement, this improvement implies that AuralTrace improved communication reception and shows that pitch modulation may be a key factor in aiding attention to and reception of live, in-the-moment virtual spatial referencing.

BACKGROUND AND RELATED WORK

This work aims to bridge perceptual-level insights about audiovisual perception with higher-level applicational goals such as communication understanding by designing a perceptual-level tool (multimodal pointer) to be used in a higher-order informational context (i.e. virtual task-based communication in our study). Past HCI work demonstrates how auditory properties can remotely support spatially-oriented tasks, in both individual and social contexts. Evidence of its neural correlates also upholds our motivation to integrate auditory pitch and spatiality with visual cues to support reception of remote spatial communication. We will present the relevant literature in this order.

Spatial audio can support spatial awareness in both individual and group contexts. It (compared to non-spatial audio) improved memory and comprehension of group desktop conferences when paired with visual representation of the conference [4]. Recent works have also used spatial audio as feasible directional prompting for augmented reality headsets, and guidance of directional attention in 360° VR videos [3, 9]. These works acknowledge the key role sound cues play in guiding spatial attention and awareness.

The ability of pitch to intuitively modulate human response has been used in some applications. Individuals' detection of and reactions corresponding to changes in pitch occur naturally in both social and individual situations. Pitch modulation was found to reduce habituation to the repetitive sound over a short sequence of time [21], and subtle differences in pitch contour of interlocutor voice in a conversational context have been shown to affect subjective perception of trust, dominance, and persuasiveness [26, 38].

Pitch's utility for guiding spatial referencing has also been shown in some recent works. Freehand movements made to a sequence of auditory tones demonstrated crossmodal correspondence of higher pitch to higher locations in mental space [34], while other works show similar association to different types of semantic structures and shapes [13, 22, 35]. Existing literature on crossmodal spatiotemporal processing also supports the high flexibility of perceptual associations that can be formed with pitch modulation [16]. As task-relevant cues are also better able to improve grounding and awareness [15], spatiotemporally-relevant audiovisual cues may similarly benefit information reception.

Audiovisual Perception in Spatio-Temporal Processing

Auditory and visual information interact with and influence each other's cue identification at the perceptual level. Study of auditory neurons found visual input to the auditory cortex in the context of multisensory stimuli enhances visual-auditory spatial processing, showing that auditory spatial perception adapts to visual influences [8]. But, an auditory illusion with changing pitch can also alter visual motion perception [22]. These alternating roles demonstrate that spatial processing of the two senses is integrative.

Past work has also demonstrated benefits of adding auditory cues to VE visuals for cognitive map construction. Cognitive map is defined as one's mental representation of a spatial environment, including features like landmark size, distance, and relative orientations. Individuals' cognitive mapping of VE spatial features had higher accuracy when they explored the space with object-tied spatial auditory cues, compared to lack of spatial cues [41]. Auditory spatial input in the presence of audiovisual cues also supports holistic spatial processing, together with the inverse [8].

In spatiotemporal processing, auditory perception is more temporally-precise than visual [40]. Multisensory cues are preferentially weighted according to a modality's unique advantage in the spatial or temporal domain. Optimal cue-integration (OCI) theory shows that dominant sensory pathways shift based on which is more precise under the given circumstances: "for temporal processing, vision is far less precise than audition… Therefore, OCI predicts that audition should dominate over vision in the time domain. [Audition] also alters visual perception in a variety of temporal discrimination tasks." When learners follow spatiotemporal cues as references in task-based communication, this audiovisual cue combination can compensate for visual disadvantage in the time domain.

Auditory Processing of Spatial Sequential Information

As mentioned previously, complex spatial communication involves conveying sequences of inter-related pieces of spatial information. Remotely, this involves much deictic referencing due to the context-dependence of such information, and both awareness and reception of the ongoing references made are key for receiver understanding and memory of object relationships. Since shared visual context can strengthen these aspects between remote communicators, we generalize this important information type into "perceptual context" and integrate meaningful auditory cues relevant to such referencing. Work on auditory perception of sequential information and pitch modulation informs potential to further enhance auditory channels, beyond just spatiality.

Humans perceive auditory cues as conceptual sequences over time. Perceived cues are organized into representations known as "auditory objects", which are seen by the brain as single units [7]. Based on an object's distinctness from others across time, auditory objects can also form salient auditory "streams to facilitate conceptual processing of sequential information, which can be important for receiver processing of spatially-relevant utterances and auditory references.

Auditory stream perception strongly facilitates learning of sequential input, affording better recall and learning of temporal information patterns than visual and tactile perception [11]. This suggests that auditory streams can be explicitly assigned learnable meaning. Explorations in abstract sound streams, which lack inherent meaning, have shown that humans flexibly associate abstract sounds with semantic representations independently of language [24].

Representation of Changing Pitch and Flexibility of Perceptual Association

The modulation of changing direction in audio pitch over time is often perceived as a curving path moving through space, allowing it to potentially sonify the movement shape of a visual pointer path. Pitch change direction has been shown to influence direction of visual search for moving targets. Contrasting concurrent and sequential presentation of pitch change and visual search showed that the above result only occurred with concurrency, implying crossmodal interaction to be the cause, opposed to priming [29].

Though previous works on pitch change are confined to a 2D domain, humans possess plasticity in perceptual association learning. Hidaka et al.'s review of crossmodal interactions notes, "Recent findings clearly suggest that new perceptual associations could be established between arbitrary inputs through crossmodal spatiotemporal processing." [16] There are also flexible interpretations of how high/low pitch reflect as spatial structures; some are associated with motion direction while others to distance or size [13, 22, 35]. This diversity and flexibility of spatial associations make pitch change-driven perception of 3D visual motion a possibility.

This flexible and innate ability to process sequences of pitch change implies the learnability of associating changing pitch with dynamic visuospatial pointer references. Segmentation of "tone streams" is the same mechanism for identifying tone differences in human language [33]. Freehand movement made to sound with changing pitch was also found to bias towards spatial locations matching pitch direction [34].

Some emerging applications use changing pitch to sonify the shape of visual information, but mostly for enhancing manual tasks, such as targeting [2, 32]. In the current study, we create an abstract sound "stream" which assigns the property of changing pitch with the meaning of pointer movement. Our study explores its utility for receivers of spatial communication in an interactive context.

METHOD

Experiment Task

Our study uses a new task design to investigate the effect of different pointer modality combinations on reception of communicated relational spatial information. Each participant was paired in a learner role with a confederate expert in a spatial arrangement task which consisted of two phases. First, a learner perceived the expert expressing a series of rules based on spatial relations, which were about how to organize the furniture in an interior space. Then, they would apply the communicated knowledge on an arrangement task of the interior space.

Interior spatial arrangement was chosen as the task type for this study. This activity encompasses different combinations of relative object placement and orientations which can characterize features of the complex spatial information.

The spatial arrangement task consisted of two main phases: (1) in VR, the learner perceives the expert communicate a series of spatial guidelines on how to arrange the space with a referencing pointer and (2) the learner arranges furniture in a floor plan of the space by dragging and rotating furniture objects on a desktop interface. Phase 1 acquires knowledge on how to arrange the space in an interpersonal setting. Phase 2 applies the communicated knowledge via arrangement.

Each task defined five **'areas'** learners had to arrange according to seven **'rules.'** An **'area'** is defined as a class of furniture items that must be placed together, such as the "home entertainment area" including floor lamp, TV, and sofa (see Fig. 1 center). **'Rules'** would inform how areas or furniture items must be placed in relation to one another or to a landmark in the space. Each task environment would also contain 2-3 landmarks fixed w.r.t the space (e.g. doorways) which would be used in addition to furniture objects for relational referencing in the rules. An example of a rule is "Dressing and Sleeping areas should be closer to the Bathroom [a landmark], and to each other, than any other area." Rules and shape of each environment were designed so that positions, relative distances and orientations of areas and items could be clearly broken down for task scoring. Prior to Phase 1, participants were given a floorplan of the space and informed the names and identities of the five areas (including furniture items) and landmarks.

In Phase 1, participants were present in the same VR environment as and were facing the confederate expert. Life-size furniture models corresponding to the ones in the task were placed in the Phase 1 VR environment to assist participants to reference their likeness and size in the space. The furniture model locations lacked any correspondence to the content of the seven rules. Each task varied in its application scenario (apartment, conference, office), floorplan shape, landmarks, furniture, areas, and rules. Detailed task examples can be found under Supplementary D. Here, a confederate expert communicates task-relevant complex spatial information to the learner while both are present in the same VR environment, before delegating task implementation (reliant on communicated information) to the learner.

In Phase 2, learners completed an arrangement of the interior space using the same floor plan shown to them prior to Phase 1 (see Fig. 1 center; Fig. D in supplemental) on a desktop computer. They manipulated furniture objects by clicking and dragging with a mouse to position the objects on a bird's eye view map of the interior space, which would always be located on the left side of the display.

We designed the arrangement task to accommodate measurement and scoring of a performance variable tied to fulfillment of the spatial relations embodied by the expert's rules. Hence, three variations of the task were designed to counterbalance our three experiment conditions. They shared pre-defined, quantifiable characteristics to enable scoring of spatial relations fulfillment, but had different interior floorplans, areas, and rules, requiring them to be learned separately. Our measurement targeted implementation of spatial relations as a way to observe external effects of the quality of received communication. This measurement and task style were inspired in part by Kraut et. al's 2002 study, which employed a helper and worker to complete a tile-based visual puzzle based on positioning of the different-colored tiles [19]. However, in our experiment the positioning of objects is scored based on nature and extent of their spatial relationships instead of an absolute 'ground truth' configuration. Given some level of ambiguity present when evaluating degree of adherence to a spatial relation (e.g. what counts as "being in a row formation"), we calculated inter-coder reliability for scoring in this task (see Measurement).

Experiment Conditions

Experimental Prototype - AuralTrace

In order to investigate the combined effect of visual modality, audio spatiality, and pitch modulation on the reception of communicated complex spatial information, we developed an experimental prototype (named AuralTrace) which is a VR pointer, similar in nature to a laser pointer, that combines all three components: the red line, spatial audio source, and pitch modulation. Our experiment compares and contrasts the levels of inclusion of these three components in order to observe how the addition of pitch modulation to visual pointer referencing, compared to just spatial audio or the lack thereof might affect reception of relational (complex) spatial communication.

In consideration of the pitch modulation needed for AuralTrace, we use sine wave as the default auditory output due to its easily-recognizable changes in pitch. The base sound emitted (across conditions) by all audio-inclusive pointers in the study is a 250Hz sine wave, which sounds like a droning, unchanging tone. This sound was chosen after piloting user feedback with a range of sine waves within normal hearing range determined that it was the most suitable default pitch, given the range within which pitch modulation occurs (shifts higher or lower) based on pointer movement (see Supplementary A for a conceptual diagram). The limits of this range were determined from prototype testing, where testers expressed the higher and lower limits of audible pitch, under the context of verbal interaction being present. All 3D audio spatialization, binaural rendering, and visual effects were achieved with existing SDKs (Steam Audio and Unity). Rendering of dynamic pitch modulation for AuralTrace was manually implemented, and will be described below. Video/audio samples of each experiment condition (including AuralTrace) can be found in Supplementary C1-C3.

Pitch Modulation. Pitch is dynamically modulated based on the movement of user referencing behavior, namely the movement trajectory of the pointer. Orientation and velocity data are read from HTC Vive controllers during each successively rendered frame, and audible jitter in pitch modulation (which makes the pitch change sound 'grainy') is smoothed by tuning frame-wise pitch values to a moving average of the pitch modulation and controller values from the last three frames. Moving averages of angular velocity and orientation are used. Since direction of pitch modulation correlates to changes in direction of pointer movement, micro-fluctuations in controller tracking can cause the jitter.

Pitch modulation is mainly determined by two factors: direction of change (rising/falling pitch) and rate of change (speed at which pitch rises/falls). Direction of change is toggled based on frame-wise controller orientation; every pointer trajectory starts off with rising pitch, which reverses direction whenever orientation differs by >45° (a hand-tuned threshold) from the last-obtained orientation moving average. See Supplementary B for the formula and method used in calculating pitch modulation.

Three conditions

We applied our spatial arrangement task to examine the effects of pointer modalities by having the confederate use the laser pointer to support their communication of spatial references. Experiment conditions manipulated the amount and types of audiovisual modalities included in the VR pointer used by the expert. Conditions consisted of three types of audiovisual affordances emitted by the pointer in three incremental conditions, which built up to AuralTrace.

Visual-Only. The most basic pointer variant was an opaque red line, which would be cast from the controller in its pointing direction (see Fig. 1 left). The pointer would only be on while the controller button was long-pressed. A red line was chosen to represent the pointing visual based on the red color of traditional laser pointers, and that many default pointing devices in VR were also a single solid line projected from the cursor (e.g. SteamVR Home, Google Daydream).

Spatial Audiovisual (Spatial AV). The second variant included two modality components; the visual red line and a spatial audio source at the location pointed to by the controller. As mentioned above, the spatial audio source here was a 250Hz sine wave unchanging in tone.

AuralTrace. The AuralTrace variant included all three modality components; the red line, spatial audio source, and pitch modulation. Pitch modulation was done directly on the spatial audio source by changing the frequency of the 250Hz sine wave based on pointer movement speed and direction.

Despite the impression that these conditions give of a clear benefit to the user by stacking affordances, we believed that observable differences in performance between these combinations would depend on how suitable the modality is for the task; it is not clear from past work whether pitch modulation or even spatial audio would be an effective support in this task context. Performance was determined from the adherence of the participant-made arrangement to the expert's 'spatial relation rules.'

Experiment Design

A within-subjects design, where learners perform the three arrangement tasks with a different pointer condition each time, allows us to account for individual differences in memory or spatial ability when exploring the effect on learners of AuralTrace use in complex spatial communication. Participants' performances are compared with their own across the three conditions. Three different interior spatial arrangement tasks were designed to counterbalance the conditions in a Latin Square design.

We wish to observe effects on reception of information before extending to more complex scenarios, such as bidirectional interaction between expert and learner where both are talking and using the pointer tool. Hence, an expert-learner setting was chosen instead of a more bidirectional interaction modality due to the explorational nature of this tool. To allow focus of experiment observation on learner reception of communicated information, only the expert used the pointer during the communication phase, and learners did not verbally interact with the expert.

A confederate was also chosen for the expert role to control individual differences between interlocutors; ours performed

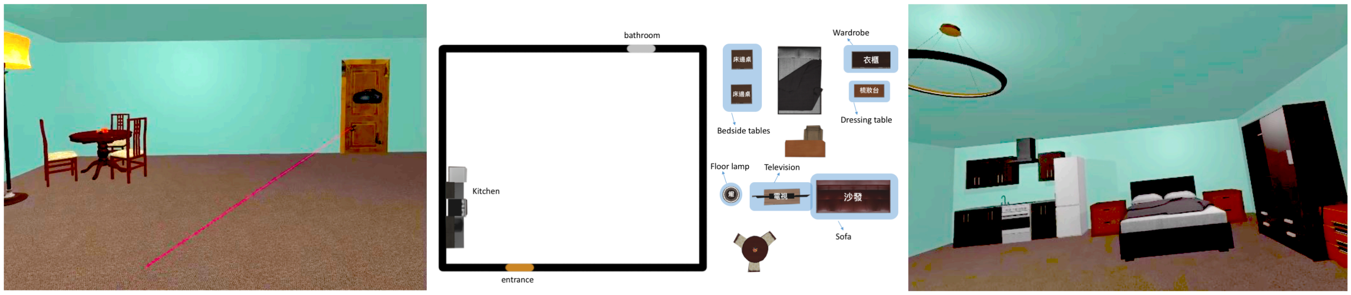

Figure 1. From left to right. Appearance of the visual laser pointer for each condition; A floor map (left) and relevant furniture for Phase 2 of one of the 3 designed tasks; A snapshot of the VR environment corresponding to the floor map to the left of it. The kitchen landmark on the floor map can be seen on the far left of this snapshot. (Note: this is the apartment task scenario.)

verbatim from a pre-written script of the spatial rules, which also included when and what gestures or references to make with the pointer. Mapping of these actions to locations in speech were explicitly specified, and the confederate was blinded to pointer audio from any of the three conditions to eliminate influence from knowing the pointer condition.

VR is used as the remote communication environment to enable binaural projection of 3D spatial sound, and to obtain results on the currently most perceptually-affordable form of VE. VR affordance of the freedom to take any audiovisual perspective of the environment is necessary for evaluating full effect of spatial sensory cues on learners. All VR material, including AuralTrace and other pointers, were created with HTC Vive, SteamVR, and Unity.

Participants

27 participants (14 females, aged 20-27) were recruited from the vicinity of the author's institute. Individuals with 0.8 (20/25) or above corrected visual acuity and lack of diagnosed hearing conditions were selected for better assessment of pointer's audiovisual perceptual information.

Procedure

Participants were first informed of the experiment procedure and provided written consent. Then, they underwent a brief, 2-3 minute familiarization session of AuralTrace and other pointer modality conditions by tracing some sound trajectories which were demonstrated by the experimenter. The purpose of this brief exposure was to facilitate understanding and association of the pitch-trajectory relationship, given that AuralTrace is an abstract sound. The exposure included the conditions of modality subsets to maintain fairness. Afterwards, participants would experience the three arrangement tasks in succession, with a different pointer condition for each task. Order of both task type and conditions were counterbalanced. In addition to default compensation given for participating in our experiment (approx. $5 USD), additional compensation was given directly based on task score (up to $9 USD per task) in order to motivate participants to perform equally well on all three tasks. A post-experimental semi-structured interview was conducted at the end to ask participants their perception of the impact and feedback of each pointer type. This entire procedure spanned approximately 1.5 hours.

Measurement

The following measurements were collected to help answer our research questions regarding how the effects of AuralTrace on learner reception of communicated information compare to other perceptual variants, and how learner perceivers might process the auditory information conveyed through pointer references.

2D Maps for Task Arrangement

Instead of directly evaluating learners in Phase 2 by implementing furniture arrangement in the same 3D VR interface used in Phase 1, we chose to implement learner-arranged maps on a 2D floor map to ascertain that learners really use conceptual spatial knowledge instead of visual or perceptual memory of the VR environment to complete the arrangement task. Previous work demonstrated sketch maps to successfully measure cognitive maps or spatial acquisition from VEs [6], and we choose to have participants arrange 2D furniture objects on the floor map in an analogy to sketch maps. The conversion of VR to floor map modality between Phase 1 and 2 ensures that participants use spatial concepts and not perceptual memory in task implementation.

Task Performance Scoring

Grading rubrics were created for each task by breaking down the requirements in the 7 rules into 'spatial components.' Learner-arranged maps were scored by their adherence to the rubric criteria. We define a spatial component (SC) as a spatial requirement w.r.t object or area which was conveyed in the guidelines. For example, in rule "Office and Meeting areas must be closer to the Restroom than any other area" there are 3 SCs; (1) Office area closest to Restroom, (2) Meeting area closest to Restroom, and (3) no other area is as close to Restroom as these two.

Scores incremented based on how many SCs were fulfilled by the learner-arranged map, and each task totaled 12 points. Each SC had a fixed point allocation, which was determined based on 2 factors: complexity and proportion of utterance.

Complexity was determined from pilot testing our created tasks, where we found that some SCs were inherently more complex or abstract than others, making them more difficult to understand. For example, one rule instructs participants to maximize *freedom of mobility* when arranging some furniture items, which was explained by the expert as the

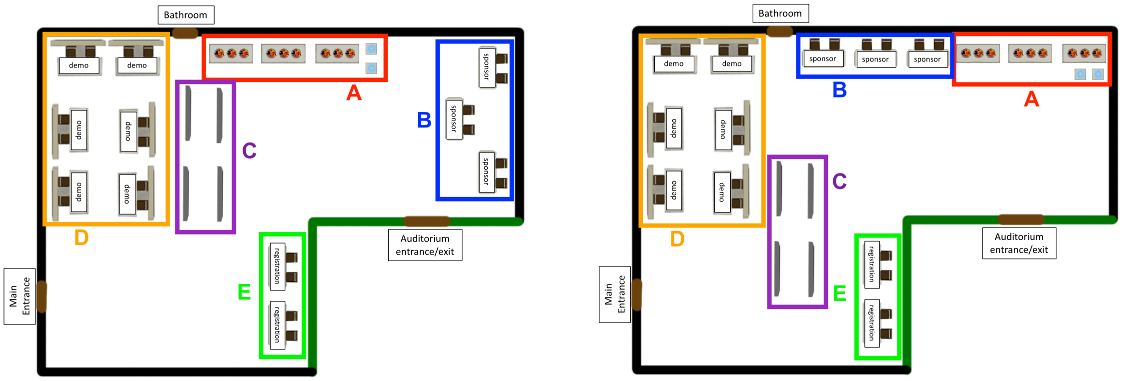

Figure 2. Examples of how SCs are consulted for point deduction for different arrangements (from left to right). See the left column below for a description of how point deduction occurs for these two examples.

amount of paths a person in the space could choose to take from a location. This SC had higher point allocation than simpler SCs, such as ‘A must be next to B’.

Proportion of utterance (PoU) was the ratio of words an SC took up in an utterance. We adjusted point allocation for SCs with similar complexity but notable difference in PoU. Smaller PoU resulted in smaller point allocation, and vice versa for larger PoU. The motivation for this is that in spite of the number of concepts in an utterance, processing of the concepts is still likely grouped by utterance as a unit. Below is an example in Table 1:

SC	Points	SC	Points
Dressing area close	0.5	Demo area close	0.3
Sleeping area close	0.5	Poster area close	0.3
Both are close	1	Break area close	0.3
		All are close	1.1

Table 1. Left: *“Dressing and sleeping areas should be closest to the bathroom.”* Right: *“Demo, break and poster areas should be closest to the auditorium.”*

The SC concept in both utterances are similar, but the number of SCs per utterance differs — so the SC point allocation here is adjusted proportionally to PoU.

SCs which could be given partial points, such as proximity-based SCs, were graded in 0.25pt increments. Binary SCs such as “place in a row” were given full or no points.

Figure 2 illustrates examples of how some sample arrangements would be scored, based on their adherence to specific spatial rules. In the left example, a SC specifies that tables in area B must be in a row (0.75pts), but since the tables are not aligned, the full 0.75pts are deducted. In the right example, a SC specifies that area B must be closer to area A and the auditorium entrance than any other area (2pts). Since B is closest to A but still closer to D and C than it is to the auditorium entrance, half of the points are deducted (1pt). Another SC specifies that the space between area E and the door must be clear (1pt), but C is blocking the way so full point ratio (1pt) is deducted.

Given the numerical task score, we calculated inter-rater reliability using Spearman’s rank correlation coefficient. 8% of the task data was randomly sampled and graded by an independent coder, and when compared to main coder-graded data the inter-rater reliability was $\rho(6) = 0.986$ ($p < 0.001$), demonstrating satisfactory reliability.

Time taken to complete arrangements was also measured to observe potential interactions between task score and time taken, and check our results. If a condition resulting in higher score also resulted in longer map times, the benefit of condition on score would be less distinguishable.

Post-Experimental Interview

A semi-structured interview was conducted after completing all three tasks to obtain more insight on how subjects experienced and perceived the pointer referencing. Map scoring measured external performance, but we asked participants to specify and compare the impact they felt from different pointer types as they tried to use the communicated information. They were asked to identify any perceived differences regarding the effect of the pointers, such as how they felt when they saw/heard combinations of audiovisual feedback, and whether their impression of it was positive or negative. We also asked how these pointers might have affected their understanding of more complex spatial information, such as expert-specified spatial relationships.

RESULTS

We first present quantitative results to answer “whether” using AuralTrace to communicate helps the reception of task-relevant spatial information. Then we show qualitative interview insights to give notion of “how” participants might have processed the information from AuralTrace.

Task Performance

A one-way repeated measures ANOVA was conducted to compare the effect of pointer type on task performance in AuralTrace, spatial-only (Spatial AV), and visual-only pointer conditions. We observed a significant effect of pointer media type, $F[2, 52] = 5.99$, $p < 0.05$. Three paired samples t-tests were conducted to make post-hoc comparisons on task score, between conditions, and Cohen’s d was also calculated. Table 2 shows comparison of scores and significance (Asterisks (*) indicate significance (p<0.01)). The t-test showed our participants having better task performance score in AuralTrace than Spatial AV, $t(26) = 3.15$, $p \leq 0.004$, d=0.6. Participants' mean task performance score was also better in AuralTrace than Visual-Only, $t(26) = 2.88$, $p \leq 0.008$, d=0.62. Considering $d = 0.5$ is regarded as a 'medium' effect, there exists noticeable effect size of $d \geq 0.6$ for the difference between AuralTrace task scores and the others. The significance sustains with Bonferroni-corrected $\alpha_{altered} = \frac{0.05}{3} = 0.0167$. There was no different task performance score between Spatial AV and Visual-Only, $t(26) = -0.205$, $p > 0.1$, $d = 0.024$.

Pointer	M	SD	Condition Pair	*t*	*p*	*d*
AuralTrace	**9.92**	1.19	AT/Spatial AV	3.15	0.004*	**0.6**
Spatial AV	8.96	1.30	AT/Vis-Only	2.88	0.008*	**0.62**
Visual-Only	9.03	1.46	Spatial AV/Vis-Only	-.205	0.839	.024

Table 2. (Left) Task scores. (Right) Paired-samples t-test values and effect size (Cohen's d). (AT = AuralTrace)

This effect size, and the significantly higher mapping score in AuralTrace condition compared to the two others, suggest that the audiovisual perceptual affordances of pointer referencing do have an effect on learner task performance. Specifically, including pitch-based referencing with spatial audio and visual cues leads to much better performance when implementing communicated spatial information, compared to just spatial audio with visual cues or just visual cues alone.

Repeated measures ANOVA and paired samples t-tests comparing pointer conditions to task mapping time in seconds (AuralTrace: M=316.13s, Spatial AV: M=325.64s, Visual-Only: M=359.12s) found no significant effects, all p's > 0.1. The lack of any mapping time effect implies that higher task score corresponding with AuralTrace condition is more likely a direct effect of the pointer itself, rather than an intermediate effect of learners spending longer to contemplate arrangement during the task.

Subjective Perception of Pointer Effects

Interview analysis exposed common trends between participant responses, highlighting remarks which shed light on how AuralTrace may have influenced learners to perform better, creating arrangements that fulfilled more of the spatial relational criteria. We contrasted participant feedback with their task performance to more comprehensively depict possible effects of AuralTrace.

Contradictions in Subjective Feedback and Performance

Initial review of the interview content revealed some unexpected patterns. This included the observation that not all participants noticed the different types of audio pointers. Given four main types of responses regarding pointer audio type when we asked participants to compare how they felt about the different pointer audio, we looked back at the task scores for participants who asserted the presence of sound did not help, and for those who did not notice differences between AuralTrace and the Spatial AV conditions.

Response Type	Individuals
Didn't notice the two audiovisual conditions were different	7
Noticed a difference in the two audiovisual conditions and felt positive impact	8
Noticed a difference in the two audiovisual conditions but felt it made no impact	8
Noticed a difference in the two audiovisual conditions and felt it impacted them negatively	4

Table 3. Participant Consciousness of Task Effect.

Of the seven individuals who reported not noticing a difference between conditions, five performed better under AuralTrace (71%). Two of three individuals who reported that using pointers with sound was disruptive to them still performed best in the AuralTrace condition. An example of what was said: *"The (pointer with sound was very disruptive… when I'm trying to memorize what the expert is saying I also hear the sound and am easily drawn in by it"*

A similarly high ratio (7/8, 87%) of individuals who noticed the difference between conditions but specified it made no difference for them also performed best under AuralTrace. As some put it: *"I just think there is no difference. It felt like a meaningless sound effect in the background."* - Subj. 7

Descriptive Feedback on Audiovisual Referencing

Feedback specific to AuralTrace denoted specific advantages not present in other remarks pertaining only to spatial audio. Participants reported finding the auditory trajectory of the pointer to be more "meaningful" when AuralTrace pitch modulation was included. They also recognized more specialized and meaningful types of pointer information, all directly related to the pitch modulation. It helped participants better differentiate the spatial changes in drawn trajectories, such as rounding corners or changing surfaces of contact. Participants also more clearly distinguished different meanings of pointer usage, such as circling objects for highlighting and outlining spatial areas or perimeters. Below are some examples:

"Pitch change made me pay particular attention to the meaning it represents… Sound with pitch resembled more than just a notification of if something was being pointed at. I could follow along better with the referencing; it helped me know more clearly what [the expert] was talking about and understand area relationships more quickly." - Subj. 23

"I think pitch is more convenient for knowing what is being drawn in space, and you can draw more quickly on purpose to evoke someone's attention. I also perceived the shape and size of outlined areas more easily with pitch." - Subj. 3

"Pitch change can support my imagining of the trajectory speed - for example, when the pointer moves slowly it's more likely that small things are being pointed at and the pitch is low, but when the pointer moves quickly it lets me know when larger targets are being referenced because the pitch rises

more quickly. *Information feels more straightforward and easy to understand*." - Subj. 19

These descriptions of how pitch change affected participant learning are in line with our motivations for supporting spatial communication with AuralTrace. Subjects 23, 19 and 3 indicated pitch led them to assign more meaning to the referencing behavior, which should facilitate better attention to and understanding of conceptual content. Subject 23 also specified that pitch made the sound resemble more than just a notification of referencing, which is how some other participants described spatial sound without pitch.

This is in line with the literature showing that the dynamics of changing pitch lead people to attend to and interpret its meaning. The use of language such as 'drawing' and 'shape' of areas adheres to previous work on human perception of pitch as drawn curves. Subject 19 describes in detail why pitch change enabled them to more saliently and effortlessly imagine the trajectory of the pointer reference and more easily understand content. AuralTrace was shown to be perceived as a particularly meaningful and informational signal, facilitating easier and faster understanding of information relevant to those references.

DISCUSSION

In this study, we proposed AuralTrace as a perceptually-targeted strategy to support communication of complex spatial information, and in an initial experiment explored its impact on the reception and usage of communication aided with the tool. Study results analysis supported our idea that AuralTrace has promising effect on spatial information reception in a social referencing context, boosting learners' implementation scores compared to referencing pointers using more common modalities. Our interview uncovered more insight on how perceivers used and were affected by AuralTrace's information in the task context. Participants' diverse perspectives on AuralTrace excavate new questions that we are excited to address in continuing work.

Reception and Implementation of Communicated Relational Spatial Concepts

Through analysis of both the performance measure and interview feedback, we essentially saw that the participant insights on how pitch modulation helped them are very likely relevant to the cognitive processes such as understanding and memory which would affect their performance on the arrangement task. To successfully use the communicated spatial rules for the task, participants would have to simultaneously encode in memory and keep track of what the interlocutor was talking about, in addition to understanding all of the content to be able to encode it at all.

Participants seemed to report that multiple aspects of this thinking and retaining process were influenced by the pitch modulation. Their account that they paid more attention to the meaning of the references being made and understood conversation content (e.g. spatial relationships) more easily (Subj 23, 19) if true, would clearly facilitate the process of mentally managing and establishing memory of this information. Ability to follow dynamic referencing more effectively (Subj 23) and perceiving spatial references with more clarity (Subj 19) both would also help with better understanding and retaining of information; struggle in just following and processing communicated information in its most primitive form (moving pointer locations and spoken words) before being able to assign meaning (turning locations to places and words to content) would disrupt the processing of conversational content at a very primary phase.

One possible explanation for the lack of improvement from *Visual-Only* by the inclusion of spatial audio in *Spatial-AV* could be that the relational nature of the spatial task was so key that aid provided by pitch modulation surpassed the benefits of localization cues afforded by spatial audio. Another is that the granularity of spatial relational components used for arrangement task scoring was not fine enough to measure differences that otherwise could have been observed between inclusion of spatial audio in *Spatial AV* and lack thereof in *Visual-Only*.

Effects of Dynamic Pitch Perception on Understanding Higher-level Spatial Information

Significant enhancement in performance of learners who perceived interactively communicated spatial information with AuralTrace as a referencing aid is not only consistent with literature on perception of pitch change, but also bridges insights from existing works and extends current knowledge. Previous works on flexibility of cross-modal perceptual association demonstrated open-endedness of pitch change interpretation [16]. The current study demonstrates that three-dimensional spatio-temporal perception of pitch change information (through dynamic pointer referencing), when it is directly relevant to communicated conceptual spatial information, can aid learning of this high-level information in a social and virtual communication context.

Namely, it has shown that pitch, when paired with supporting stimuli (spatial sound and pointer visuals) can be perceived as three-dimensional trajectories in a discrete spatial context (our VR interior space). This is a new kind of perceptual mapping resulting from the potential of perceptual associative learning to form 'arbitrary' mappings, which was demonstrated in past works showing that people could map pitch to different abstract or 2D structures [13, 22, 35]. However, it has not been previously shown that individuals can successfully map changing pitch to 3D trajectories in space. In our study, subjects were able to benefit from effects of associating pitch change with three-dimensional trajectories in a virtual environment (VE) after just a brief introduction at the start of our experiment. We were able to see in interview results that participants mapped the pitch to the "shape and size" of "drawn" trajectories, exhibiting the concept of a discrete spatial structure. Because the pointer references spanned multiple locations in the VR setting, we also know that this mapping occurred in 3D space. In this work, we used changing pitch to represent an auditory stream, the meaning of whose sequential input was then learned by the listener to represent the moving location pointed to by AuralTrace. This new association is made possible by the flexibility of perceptual associations with pitch change [16].

Our findings extend how the property of pitch can be used to support applications in higher-level thinking. Past applications have focused on more manually-oriented uses of pitch sonification [2, 32], but the current study demonstrates a performance advantage that it brings to interactive remote spatial communication. Additionally, this performance advantage can occur quite effortlessly, sometimes even without the knowledge of the learner — as shown from the interview feedback. These results can be applied in relevant communicative contexts to support acquisition and implementation of higher-level knowledge.

Perceptual Context for Communication & Collaboration

Previous work on shared visual context demonstrated its importance for spatial referencing [14, 19]. Here, we propose shared 'perceptual context' can further bolster core processes of interpersonal communication by enhancing sensory channels whose features are particularly relevant to the task at hand, such as audiovisual channels for highly-referential and spatial tasks. That pitch-based abstract sound can be perceptually harnessed to learn communicated spatial concepts in VR provides a strong case that supporting grounded perception of spatial information (pointer referencing) relevant to higher-order semantic information (expert rules) can lead to more effective learning of it.

Participant feedback reflected both heightened awareness and grounding of conveyed information during communication with AuralTrace. They recalled that pitch change helped them follow along with expert communicators, and be more aware of the references being drawn in the space. They also reported that the salient, meaningful context of trajectory movement imparted by pitch change actively aided them to attend to the content conveyed by the expert, effectively grounding expert-conveyed information to the learner. Enhancement of shared perceptual resources relevant to spatial communication, contextualized by sequential following of interlocutor-made deictic references improved learner application of information, demonstrating that perceptual information can encode and ground intentions in multiple ways that go beyond mere presence of these channels (e.g., basic audiovisual or visual cues).

Design Implications for Application Domains

AuralTrace was designed to tailor the shared perceptual context of a remote communicative space to the needs of spatial relational communication, but there also exist many other social contexts that may benefit from targeted augmentation of perceptual context. A key advantage is the likely-low cognitive overhead required of participants to benefit, due to automatic integration of perceptual-level information. When we look at how it can be used in aforementioned domains such as team sports, the directing quality of pitch modulation in a pointer like AuralTrace can assist information following and reception when tracing an explanation of different dynamic trajectories that multiple players must take in a new strategic play. Or for interior design, designers could trace different paths with AuralTrace to more clearly delineate to each other how a spatial setup might affect the flow of a one's navigation through the space.

Abstract sound can be applied to many other interpersonal contexts which generally involve dynamic structures in space, such as dance choreography. Sonification is a growing field, but there are endless possibilities for its use in contexts apart from clear-cut data visualization. With the growth and complexity in types of computer-mediated collaboration, more activities can consider the role meaningful pitch change and other sensory artifacts could play in remote interactions.

Limitations and Future Work

The present study is but a preliminary probe into the possibility of using pitch-based abstract sound for communicative support. Our current design is very limited in that it only explores effects on half (the learner) of what could be an interactive dyad, and limits the possibilities of interaction between communicating individuals. Having observed that pitch modulation does have an effect on receiving end, we plan to study next the interaction of a dyad communicating with this kind of feature and observe its effect on thought and action of both individuals.

The details of our experiment design also contained some limitations, such as the form in which we chose to implement the visual and audio for the pointers. Though we referenced more common VR software for designing the red laser visual cue, there are likely newer or more experimental methods of visual pointing in VR that afford better visibility. The 250Hz sine wave implemented as the basic audio cue was also found to be annoying for some users. We did not create more highly refined designs for this exploratory study, but this work would benefit from more rigorous comparison with state-of-the-art visual affordances and more careful selection of sound types based on pleasantness of experience.

CONCLUSION

We explored how multisensory information in the form of pitch-based spatial audiovisual pointer (AuralTrace) used by interlocutors to communicate spatial information in a virtual environment affected learner reception and implementation of the content. Results of a within-subjects experiment comparing AuralTrace with other pointer types when used in communication regarding interior spatial arrangement showed that it was linked to significantly better performance, and could also facilitate attribution of meaning to physical references. These insights shed light on the feasibility of lower-level mechanisms like pitch modulation to support acquiring higher-level spatial knowledge, connecting features of existing literature and forming implications for the design of interfaces supporting higher-level thinking.

ACKNOWLEDGMENTS

Left blank for review.

REFERENCES

[1] Amandine Afonso, Alan Blum, Christian Jacquemin, Michel Denis, and Brian F. G. Katz. 2005. A Study of Spatial Cognition in an Immersive Virtual Audio Environment: Comparing Blind and Blindfolded Individuals. Retrieved December 31, 2017 from https://smartech.gatech.edu/handle/1853/58424

[2] Miguel A. Alonso-Arevalo, Simon Shelley, Dik Hermes, Jacqueline Hollowood, Michael Pettitt, Sarah Sharples, and Armin Kohlrausch. 2012. Curve shape and curvature perception through interactive sonification. *ACM Transactions on Applied Perception* 9, 4: 1–19. https://doi.org/10.1145/2355598.2355600

[3] Paulo Bala, Raul Masu, Valentina Nisi, and Nuno Nunes. 2019. “When the Elephant Trumps.” https://doi.org/10.1145/3290605.3300925

[4] Jessica J. Baldis and Jessica J. 2001. Effects of spatial audio on memory, comprehension, and preference during desktop conferences. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI ’01*, 166–173. https://doi.org/10.1145/365024.365092

[5] Giuliano Benelli, Maurizio Caporali, Antonio Rizzo, and Elisa Rubegni. 2001. Design concepts for learning spatial relationships. In *Proceedings of the 19th annual international conference on Computer documentation - SIGDOC ’01*, 22. https://doi.org/10.1145/501516.501522

[6] M. Billinghurst and S. Weghorst. The use of sketch maps to measure cognitive maps of virtual environments. In *Proceedings Virtual Reality Annual International Symposium ’95*, 40–47. https://doi.org/10.1109/VRAIS.1995.512478

[7] Jennifer K Bizley and Yale E Cohen. 2013. The what, where and how of auditory-object perception. *Nature reviews. Neuroscience* 14, 10: 693–707. https://doi.org/10.1038/nrn3565

[8] Jennifer K. Bizley and Andrew J. King. 2008. Visual–auditory spatial processing in auditory cortical neurons. *Brain Research* 1242: 24–36. https://doi.org/10.1016/J.BRAINRES.2008.02.087

[9] Brendan Cassidy, Janet C. Read, and I. Scott MacKenzie. 2019. An Evaluation of Radar Metaphors for Providing Directional Stimuli Using Non-Verbal Sound. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI ’19*. https://doi.org/10.1145/3290605.3300289

[10] Alan R. Cohen, Subash Lohani, Sunil Manjila, Suriya Natsupakpong, Nathan Brown, and M. Cenk Cavusoglu. 2013. Virtual reality simulation: basic concepts and use in endoscopic neurosurgery training. *Child’s Nervous System* 29, 8: 1235–1244. https://doi.org/10.1007/s00381-013-2139-z

[11] Christopher M. Conway and Morten H. Christiansen. 2005. Modality-Constrained Statistical Learning of Tactile, Visual, and Auditory Sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31, 1: 24–39. https://doi.org/10.1037/0278-7393.31.1.24

[12] Sarah D’Angelo and Darren Gergle. 2016. Gazed and Confused. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI ’16*, 2492–2496. https://doi.org/10.1145/2858036.2858499

[13] Zohar Eitan, Asi Schupak, Alex Gotler, and Lawrence E. Marks. 2014. Lower Pitch Is Larger, Yet Falling Pitches Shrink. *Experimental Psychology* 61, 4: 273–284. https://doi.org/10.1027/1618-3169/a000246

[14] Susan R. Fussell, Robert E. Kraut, and Jane Siegel. 2000. Coordination of communication. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work - CSCW ’00*, 21–30. https://doi.org/10.1145/358916.358947

[15] Darren Gergle, Robert E Kraut, and Susan R Fussell. 2013. Using Visual Information for Grounding and Awareness in Collaborative Tasks. January 2013: 37–41. https://doi.org/10.1080/07370024.2012.678246

[16] Souta Hidaka, Wataru Teramoto, and Yoichi Sugita. 2015. Spatiotemporal Processing in Crossmodal Interactions for Perception of the External World: A Review. *Frontiers in integrative neuroscience* 9: 62. https://doi.org/10.3389/fnint.2015.00062

[17] Robert Konrad. 2015. What is the vergence-accommodation conflict and how do we fix it? *XRDS: Crossroads, The ACM Magazine for Students* 22, 1: 52–55. https://doi.org/10.1145/2810048

[18] Robert E. Kraut, Susan R. Fussell, and Jane Siegel. 2003. Visual Information as a Conversational Resource in Collaborative Physical Tasks. *Human–Computer Interaction* 18, 1–2: 13–49. https://doi.org/10.1207/S15327051HCI1812_2

[19] Robert E. Kraut, Darren Gergle, and Susan R. Fussell. 2002. The use of visual information in shared visual spaces. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work - CSCW ’02*, 31. https://doi.org/10.1145/587078.587084

[20] Mei-Po Kwan and Jiyeong Lee. 2005. Emergency response after 9/11: the potential of real-time 3D GIS for quick emergency response in micro-spatial environments. *Computers, Environment and Urban Systems* 29, 2: 93–113. https://doi.org/10.1016/J.COMPENVURBSYS.2003.08.002

[21] Yi-Chen Lee, Fu-Yin Cherng, Jung-Tai King, and Wen-Chieh Lin. 2019. To Repeat or Not to Repeat? In *Proceedings of the 2019 CHI Conference on Human*

Factors in Computing Systems - CHI '19. https://doi.org/10.1145/3290605.3300743

[22] Fumiko Maeda, Ryota Kanai, and Shinsuke Shimojo. 2004. Changing pitch induced visual motion illusion. *Current biology : CB* 14, 23: R990-1. https://doi.org/10.1016/j.cub.2004.11.018

[23] Katerina Mania and Alan Chalmers. 2002. The Effects of Levels of Immersion on Memory and Presence in Virtual Environments: A Reality Centered Approach. *CyberPsychology & Behavior*. https://doi.org/10.1089/109493101300117938

[24] Adrien Merer, Sølvi Ystad, Richard Kronland-Martinet, and Mitsuko Aramaki. 2011. Abstract Sounds and Their Applications in Audio and Perception Research. . Springer, Berlin, Heidelberg, 176–187. https://doi.org/10.1007/978-3-642-23126-1_12

[25] Cosmin Munteanu, Pourang Irani, Sharon Oviatt, Matthew Aylett, Gerald Penn, Shimei Pan, Nikhil Sharma, Frank Rudzicz, Randy Gomez, Benjamin Cowan, and Keisuke Nakamura. 2017. Designing speech, acoustic and multimodal interactions. In *Conference on Human Factors in Computing Systems - Proceedings*, 601–608. https://doi.org/10.1145/3027063.3027086

[26] Laya Muralidharan, Ewart J. de Visser, and Raja Parasuraman. 2014. The effects of pitch contour and flanging on trust in speaking cognitive agents. https://doi.org/10.1145/2559206.2581231

[27] Sharon Oviatt. 2006. Human-centered design meets cognitive load theory: Designing interfaces that help people think. In *Proceedings of the 14th Annual ACM International Conference on Multimedia, MM 2006*, 871–880. https://doi.org/10.1145/1180639.1180831

[28] Sharon Oviatt, Rachel Coulston, and Rebecca Lunsford. 2004. When do we interact multimodally? 129. https://doi.org/10.1145/1027933.1027957

[29] Stacey Parrott, Emmanuel Guzman-Martinez, Laura Orte, Marcia Grabowecky, Mark D Huntington, and Satoru Suzuki. 2015. Direction of Auditory Pitch-Change Influences Visual Search for Slope From Graphs. *Perception* 44, 7: 764–78. https://doi.org/10.1177/0301006615596904

[30] Lorenzo Picinali, Amandine Afonso, Michel Denis, and Brian F.G. Katz. 2014. Exploration of architectural spaces by blind people using auditory virtual reality for the construction of spatial knowledge. *International Journal of Human-Computer Studies* 72, 4: 393–407. https://doi.org/10.1016/J.IJHCS.2013.12.008

[31] G. Pingali, A. Opalach, Y. Jean, and I. Carlbom. 2001. Visualization of sports using motion trajectories: providing insights into performance, style, and strategy. In *Proceedings Visualization, 2001. VIS '01.*, 75–544. https://doi.org/10.1109/VISUAL.2001.964496

[32] Hessam Roodaki, Navid Navab, Abouzar Eslami, Christopher Stapleton, and Nassir Navab. 2017. SonifEye: Sonification of Visual Information Using Physical Modeling Sound Synthesis. *IEEE Transactions on Visualization and Computer Graphics* 23, 11: 2366–2371. https://doi.org/10.1109/TVCG.2017.2734327

[33] J R Saffran, E K Johnson, R N Aslin, and E L Newport. 1999. Statistical learning of tone sequences by human infants and adults. *Cognition* 70, 1: 27–52. Retrieved January 14, 2019 from http://www.ncbi.nlm.nih.gov/pubmed/10193055

[34] Alejandro Salgado-Montejo, Fernando Marmolejo-Ramos, Jorge A. Alvarado, Juan Camilo Arboleda, Daniel R. Suarez, and Charles Spence. 2016. Drawing sounds: representing tones and chords spatially. *Experimental Brain Research* 234, 12: 3509–3522. https://doi.org/10.1007/s00221-016-4747-9

[35] R N Shepard. 1982. Geometrical approximations to the structure of musical pitch. *Psychological review* 89, 4: 305–33. Retrieved January 14, 2019 from http://www.ncbi.nlm.nih.gov/pubmed/7134331

[36] Kay M. Stanney, Ronald R. Mourant, and Robert S. Kennedy. 1998. Human factors issues in virtual environments: A review of the literature. *Presence: Teleoperators and Virtual Environments*. https://doi.org/10.1162/105474698565767

[37] 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: 61–72. https://doi.org/10.1037/1076-898X.14.1.61

[38] Tzu-Yang Wang, Ikkaku Kawaguchi, Hideaki Kuzuoka, and Mai Otsuki. 2018. Effect of Manipulated Amplitude and Frequency of Human Voice on Dominance and Persuasiveness in Audio Conferences. *Proceedings of the ACM on Human-Computer Interaction*. https://doi.org/10.1145/3274446

[39] Xiangyu Wang and Jerry Jen-Hung. Tsai. 2011. *Collaborative design in virtual environments*. Springer.

[40] Ilana B Witten and Eric I Knudsen. 2005. Why seeing is believing: merging auditory and visual worlds. *Neuron* 48, 3: 489–96. https://doi.org/10.1016/j.neuron.2005.10.020

[41] Seraphina Yong and Hao-Chuan Wang. 2018. Using Spatialized Audio to Improve Human Spatial Knowledge Acquisition in Virtual Reality. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion - IUI 18*, 1–2. https://doi.org/10.1145/3180308.3180360