Tap the ShapeTones: Exploring the Effects of Crossmodal Congruence in an Audio-Visual Interface

Oussama Metatla
Queen Mary University of London
o.metatla@qmul.ac.uk

Nuno N. Correia
Goldsmiths, University of London
n.correia@gold.ac.uk

Fiore Martin
Queen Mary University of London
f.martin@qmul.ac.uk

Nick Bryan-Kinns
Queen Mary University of London
n.bryan-kinns@qmul.ac.uk

Tony Stockman
Queen Mary University of London
t.stockman@qmul.ac.uk

ABSTRACT

There is growing interest in the application of crossmodal perception to interface design. However, most research has focused on task performance measures and often ignored user experience and engagement. We present an examination of crossmodal congruence in terms of performance and engagement in the context of a memory task of audio, visual, and audio-visual stimuli. Participants in a first study showed improved performance when using a visual congruent mapping that was cancelled by the addition of audio to the baseline conditions, and a subjective preference for the audio-visual stimulus that was not reflected in the objective data. Based on these findings, we designed an audio-visual memory game to examine the effects of crossmodal congruence on user experience and engagement. Results showed higher engagement levels with congruent displays with some reported preference for potential challenge and enjoyment that an incongruent display may support, particularly for increased task complexity.

Author Keywords

Crossmodal congruence, spatial mappings, user engagement, user experience, games, audio-visual display.

ACM Classification Keywords

H.1.2 User/Machine systems: Human Factors; H.5.2 User Interfaces: Auditory (non-speech) feedback, Screen design (e.g., text, graphics, color), Input devices and strategies (e.g., mouse, touchscreen), Evaluation/methodology; K.8.0 General: Games

INTRODUCTION

Multisensory perception is an activity that we do everyday when we combine signals from various sensory channels to make sense of our environment and to act in it. One of the mechanisms that we use to fuse input from multiple sensory channels is referred to as *crossmodal interaction* [35]. A key feature of a crossmodal display is that it relays the same information through two or more senses, for example, when we find it easier to recognise speech when we can see the speaker's lip movements. Research aiming to apply findings from crossmodal perception to interface design has focused on designing support for interaction in complex environments, for example in the design of monitoring systems and warning signals [29, 35], on designing sensory substitution devices for people with sensory disabilities, such as the vOICe system, which uses sonification to convert images into sound [15], and on supporting collaboration between people with different sensory abilities [42, 16]. However, whilst it is increasingly feasible to support crossmodal interaction in a range of general purpose devices, e.g. tablet computers and smartphones provide touch, visual, and speech interaction, little work has considered the implications of crossmodal displays on user experience and engagement. Therefore, we propose that research into the design of effective crossmodal interfaces should consider a wider range of user experiences. In particular, evaluations of crossmodal displays should emphasise elements of both user performance and engagement to provide deeper insights into the application of crossmodal mappings to interactive experiences. This paper contributes to bridging the gap between studies of crossmodal user performance and engagement by examining the effects of crossmodal congruence on performance and engagement in the context of a memory task supported by combinations of audio and visual displays on touch-screen devices. A first study examines the effects of different levels of crossmodal congruence on how audio-visual cues support the mapping of spatial ordering. A second study examines the application of these crossmodal mappings in the design of an audio-visual memory game, focusing on evaluating user experience and engagement with the crossmodal gameplay.

BACKGROUND

Crossmodal interaction underlies the phenomenon by which signals from one sensory modality can affect the process-

CHI'16, May 07 - 12, 2016, San Jose, CA, USA

ACM 978-1-4503-3362-7/16/05...$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858456

ing of information perceived through another modality. One famous example of this phenomenon is the "McGurk" effect [14] where the auditory phoneme "ba" is perceived as "da" when paired with the visual stimuli of lips movements pronouncing "ga". The ideas behind crossmodal interaction stem from advances in cognitive neuroscience, specifically new understandings of brain plasticity and sensory substitution, which refer to the capacity of the brain to replace the functions of a given sense by another sensory modality [1]. Recently, there has been increasing interest in the study of these types of crossmodal interactions between sensory information, and their implications for user interface design. For instance, Ju-Hwan and Spence [12] demonstrated that the presentation of sounds can modulate the number of vibrotactile targets that a person will perceive, particularly when they perform secondary attention-demanding tasks. Shams *et al.* [31] also demonstrated how people's perception of flashing lights can be manipulated by sounds, with people seeing a single flash of light as consisting of two flashes when these are presented simultaneously with multiple auditory beeps. Sensory modalities are therefore far from working as independent modules and findings from these and similar studies challenge the notion that their interaction follows a hierarchy in which vision dominates the sensory experience. Massaro [13] suggests that while all modalities contribute to perceptual experience, it is most influenced by the sensory channel that mediates the least ambiguous information. In the context of this paper, this suggests that different visual and auditory mappings can influence the perception of spatial information and that different combinations may result in more efficient and engaging interactions.

Congruency and crossmodal correspondences

Research examining multi-sensory experience often use the term *congruence* or *crossmodal correspondences* to refer to non-arbitrary associations that exist between different modalities and the consequences that these have on human information processing. For instance, studies found crossmodal correspondences between high-pitched sounds and bright, small objects positioned at higher locations in space, and between low-pitched sounds and darker, bigger rounder objects at lower locations [2, 26]. Other studies found congruent mappings between pitch and vertical location, size and spatial frequency [5]. Spence highlights a further distinction between *semantic* and *synaesthetic congruency* to differentiate between sensory stimuli that vary in terms of their identity and/or meaning, and those that refer to *"correspondences between putatively nonredundant stimulus attributes or dimensions that happen to be shared by many people"* [34]. A number of researchers have demonstrated the benefits of exploiting crossmodal congruency for better user interface design. Hoggan and Brewster [9], for instance, examined the relationships between individual visual button features such as size and height with audio/tactile properties. They showed that perceived quality of touchscreen buttons was correlated to congruence between visual and audio/tactile feedback used to represent them. Fewer researchers have looked at user experience - Huang *et al* developed the MelodicBrush system in which they explored how crossmodal mappings between the shapes of Chinese calligraphy and musical tones can enhance user experience during artistic creation [10], but their system did not ground mapping choices in empirical data.

Attention, memory & motor learning

To explore crossmodal interaction we are interested in how semantically congruent audio and visual stimuli can convey spatial information and guide users' attention when locating items on interactive touch-screen devices as there is extensive evidence supporting the existence of crossmodal links in spatial attention (for reviews, see [33, 36]). In particular, a number of lab-based studies have demonstrated how the presentation of crossmodal as opposed to unimodal cues can significantly facilitate the capture of a person's spatial attention [37]. Stefanucci and Proffitt [38] examined the impact of crossmodal cues on memory tasks involving visual and auditory stimuli with a focus on whether congruency effects between learning and retrieval phases improves retention. Their findings indicated that the presence of sounds provided a strong cue for binding visual display to the information learnt and hence improve retention. Studies have also shown that crossmodal concurrent feedback can enhance motor learning, and positive effects are often explained by a reduction of workload [7]. For example, visual feedback could facilitate learning of spatial aspects of the movement, while auditory feedback could support learning of temporal aspects [32]. Curiously, concurrent crossmodal feedback has been found to enhance performance in the acquisition phase, but the performance gains are lost in retention tests. This finding is explained by the guidance hypothesis which states that permanent feedback during acquisition leads to a dependency on the feedback [30]. The guidance forces learners to ignore their intrinsic feedback which is based on proprioception [32] - our sense of bodily movement and position in space.

SCOPE

The work presented in this paper extends this line of research by examining more complex crossmodal stimuli that support a spatial ordering memory task and by exploring crossmodal congruences from two perspectives: i) task performance, and ii) user engagement. Study 1 builds on previous work on crossmodal perception that demonstrated congruence effects between the auditory feature of pitch and the visual features of size and vertical location [2, 26, 5], as well as the impact of crossmodal feedback on motor learning and the retention of spatial information post-acquisition [7, 30]. The aim is to evaluate users' ability to determine spatial orderings of a sequence of items on the basis of audio, visual, and audio-visual stimuli, in the context of a memory task. Study 2 explores user experience and engagement with a crossmodal memory game building on the results of the first study, and recent work on the evaluation of user engagement in game applications [23, 24, 41]. To do this, we added a number of gamification elements to the apparatus used in the first study that were inspired by current design practices in mobile games. These included the introduction of a game progression logic based on increasing levels of difficulty and scores with corresponding visual and auditory indicators [28]. This is described in more details in later sections of the paper.

STUDY 1: CROSSMODAL MAPPINGS

Apparatus

To examine the impact of crossmodal congruence on mappings of vertical location, motor learning and retention, we designed an interface that consists of a visual and an auditory display component for output and a touch-based component for input. The experimental apparatus was developed as an application that runs on an Apple iPad. It divides the screen horizontally into different sections (top of Figure 1), with each section corresponding to a unique shape and a unique tone (bottom of Figure 1) we refer to as *ShapeTones*.

Visual mappings & congruence levels

We designed three types of visuals to map screen sections; we refer to these as *arbitrary*, *size*, and *spikes* (Figure 1). In the spikes mapping, we used a basic circular shape and increased the amount of spikes attached to it to correspond to a given section; e.g. the shape for section three has three spikes. There is therefore an immediately perceivable relationship between the shapes and the physical layout of the screen, which constitutes a *congruent* mapping. In the size mapping, we used a single shape and we varied its size to correspond to each screen section. The gradual change in size therefore corresponded to the progression of sections, with lower sections corresponding to larger objects [2, 26]. However, compared to the spikes mapping, the exact mapping from a given size to a section has to be inferred. This mapping is therefore *semi-congruent* with the physical layout. In the arbitrary shapes mapping, different shapes are assigned arbitrarily to correspond to each screen section. We designed these shapes so that they bear no obvious relationship to the sections and are therefore *incongruent* with the physical layout of the screen.

Auditory mapping

Tones were mapped vertically to screen sections: lower pitches to lower sections, and higher pitches to higher sections (Figure 1). This mapping is based on crossmodal correspondences between vertical location and pitch [2, 26, 5]. We used musical notes and the sine wave timbre as they are common tones. After trying different scales in terms of tone discernibility with iPad speakers, we chose a mid-range octave: the G4 major scale.

Touch-based input

Users interact with this application by tapping on corresponding sections on the screen to reproduce the spatial order of a sequence of items conveyed to them through ShapeTones.

Experimental design

We manipulated level of congruency as an independent variable in a between-subjects experimental design. Participants were divided into three groups with each group performing the experimental task using one of the three visual mappings; participants used the *spikes mapping* in the *congruent condition*; the *size mapping* in the *semi-congruent condition*; and the *arbitrary mapping* in the *incongruent condition*.

We also manipulated display type in a within-subjects experimental design. Participants in each group performed the experimental task under three within-subjects conditions; an *audio-visual condition*; *a visual-only condition*; and an *audio-only condition*. The audio-only and visual-only conditions were used as controls to provide baselines to compare crossmodal and unimodal displays, i.e. to examine the effects of the visual and auditory mappings when used independently as a means for judging locations on the touch-screen. A between-subjects design thus ensured that each participant is only exposed to one visual mapping/congruence level, while a within-subjects design ensured that each participant's performance with a given crossmodal congruence level is compared against their own performance on unimodal displays. The combination of between/within-subjects designs also avoids confounding learning effects and fatigue.

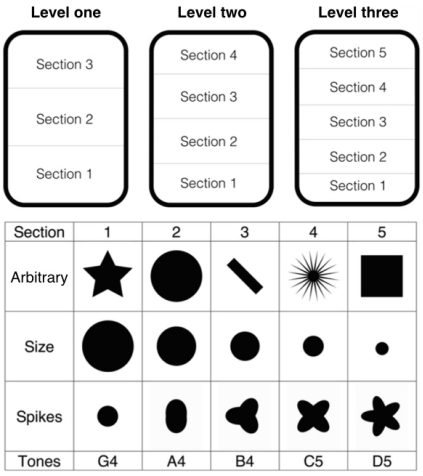

Figure 1. Crossmodal mappings.

Experimental task details

The experimental task was a memory task in which participants were presented with a sequence of three ShapeTones and were asked to reproduce the order of that sequence by tapping the corresponding sections on the touch-screen. This task builds on previous work in the area of point estimation [17] and provides a potential for broader use, e.g. in games. ShapeTones were presented one at a time at the centre of the touch-screen at a speed of 0.3 seconds per item chosen on the basis of previous studies on rapid identification of auditory and graphical stimuli [18, 27]. Depending on the experimental condition, participants were asked to watch and/or listen to a sequence of three shapes and tones and to reproduce the order in which these occurred by tapping on corresponding sections on the touch-screen.

Figure 2 exemplifies the structure of the experimental task. Participants tapped on a "play" button to start the sequence, watched and/or listened to a sequence, then tapped on the touch-screen to reproduce its order. No feedback was presented while tapping the order of the sequence, but the participants' input was played back to them at the end of the tapping (in the form of ShapeTones in the audio-visual condition, shapes only in the visual-only condition, and tones only in the audio-only condition). This was then followed by an in-

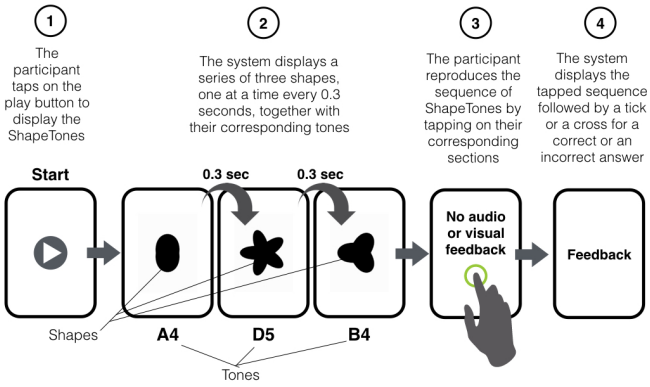

Figure 2. Experimental task

dication of whether their sequence was correct or not (a tick for a correct sequence, and a cross for an incorrect sequence). To avoid ceiling effects in each condition, participants performed sets of experimental tasks using three different complexity levels as shown in Figure 1. We used, three, four and five sections in each level of complexity respectively. The stimuli consisted of three ShapeTones in all complexity levels. Each participant performed 10 trials in each set, totalling 30 trials per condition; thus giving 90 trials per participant and a total of 3240 trials for the whole study.

Experimental setup & procedure

Participants were briefed about the study, signed consent forms and completed an initial questionnaire about demographic details, their musical training (in terms of years of practice), and experience with touch-screen devices. They were then randomly assigned to one of the three groups. Care was taken to ensure different musical abilities were broadly distributed between groups. Before the trials began, participants were trained on the particular display they were going to use. Unlike the tests, training was such that participants could tap around the touch-screen and receive audio, visual or audio-visual feedback that corresponded to the location of where they tapped. They were instructed to take as much time as they needed to memorise the tones and the shapes used for the particular condition they were about to do. Training typically lasted up to 5 minutes. Once familiar with the display, participants performed three trials similar to the actual the testing phase (Figure 2). These were not included in the analysis and were intended to help participants develop their proprioceptive skills. Participants then performed ten trials in each of the three conditions (audio-only, visual-only, and audio-visual) in a given level of complexity before moving on to the next level. They were allowed to familiarise themselves again with the shapes and tones before the start of each new set of trials. We administered short questionnaires and conducted informal interviews at the end of each level to collect feedback. Conditions were counterbalanced. An entire session lasted between 45 minutes and an hour.

Participants

36 participants took part in this study (19 female, 17 male, *mean age* = 27.9, SD = 4.4). They were a mixture of university staff (academic and non-academic), undergraduate and postgraduate students and members of the public. Participants received a cash incentive for participating. Seven participants rated their musical experience as expert, eight as intermediate, fourteen as beginner, and six had no musical training. All had experience with using touch-screen devices.

Dependent variables & measurements

The dependent variables were the scores and completion times. Scores were calculated based on the number of correct sequences reproduced by the participants. Completion times were measured as the duration from the time the participants pressed the "play" button to the instant they tapped the third and final point in a given sequence.

Hypotheses

S1H1: Level of congruence will have an effect on participants' performance: in particular, based on existing literature on crossmodal mappings [2, 5, 9, 26, 29], we expected a congruent display using the spikes mapping to lead to better performances than a semi-congruent display using the size mapping and an incongruent display using the shapes mapping. We also expected the semi-congruent display to yield better performances than the incongruent display.

S1H2: Type of display will have an effect on participants' performance: in particular, based on existing literature on the advantages of audio-visual over unimodal displays [4, 38], we expected that the effects of the level of congruence will be more apparent in the audio-visual conditions.

Results

We used single-factor ANOVAs with level of congruence as a factor (three levels: congruent, semi-congruent, and incongruent) to analyse differences in times and scores across groups, and repeated-measures ANOVAs with display type as a factor (three level: audio-visual, audio-only, and visual only) to analyse differences within each group. In both cases, we used Fisher's LSD for post-hoc comparisons of main effects. We used a confidence level of $\alpha = 0.05$ for all tests.

Scores across groups

Level one (three sections)

There was no significant main effect of level of congruence on participants' scores in the audio-visual ($F(2, 34) = 0.104$, $p = 0.902$) visual-only ($F(2, 34) = 1.578$, $p = 0.222$) and audio-only conditions ($F(2, 34) = 0.54$, $p = 0.588$).

Level two (four sections)

There was no significant main effect of level of congruence on participants' scores across groups in the audio-visual condition ($F(2, 34) = 2.834$, $p = 0.074$). In the visual-only condition, there was a significant main effect of level of congruence on scores ($F(2, 34) = 4.276p = 0.023$, $\eta^2 = 0.21$). Post-hoc tests showed that participants in the congruent condition (spikes: $mean = 7.16, sd = 1.99$) scored significantly higher than participants in the semi-congruent condition (size: $mean = 5.33, sd = 1.37$) ($p = 0.015$). Participants in the incongruent condition (random: $mean = 7.18, sd =$

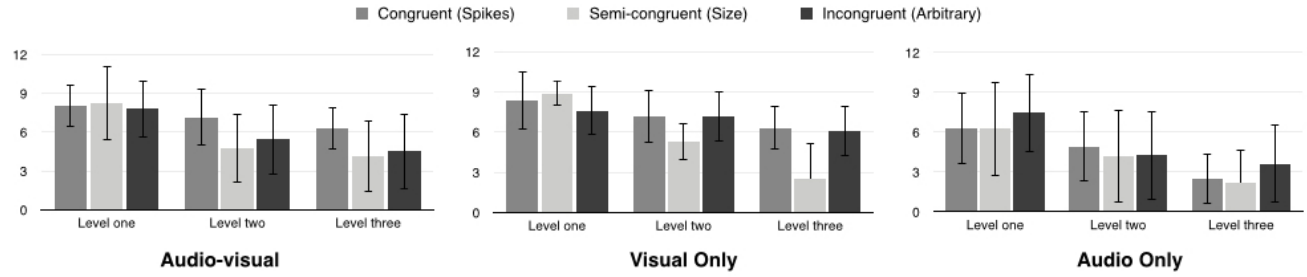

Figure 3. Scores across groups and complexity levels in the audio-visual, visual-only and audio-only conditions

1.88) also scored significantly higher than those in the semi-congruent condition ($p = 0.013$). There was no significant difference between the congruent and the incongruent conditions ($p = 0.984$). There was also no significant main effect of levels of congruence on participants' scores in the audio-only condition ($F(2, 34) = 0.192, p = 0.826$).

Level three (five sections)

There was no significant main effect of level of congruence on participants' scores across groups in the audio-visual condition ($F(2, 34) = 2.565, p = 0.093$). In the visual-only condition, there was a significant main effect of level of congruence on participants' scores across groups ($F(2, 34) = 12.097, p < 0.001, \eta^2 = 0.43$) and post-hoc tests showed that participants in the congruent condition (spike: $m = 6.33, sd = 1.66$) scored significantly higher than participants in the semi-congruent condition (size: $mean = 2.5, sd = 2.67$) ($p < 0.001$), and that participants in the incongruent condition (random: $mean = 6.09, sd = 1.86$) scored significantly higher than participants in the semi-congruent condition ($p = 0.001$). There was no significant difference between participants scores in the congruent and incongruent conditions ($p = 0.746$). There was no significant main effect of level of congruence on participants's scores across group in the audio-only condition ($F(2, 32) = 1.082, p = 0.351$).

Figure 3 summarises these results, which show that: as complexity increased, participants who used the congruent display (spikes mapping) performed significantly better than those who used the semi-congruent and incongruent displays in the visual-only conditions; Participants who used the incongruent display performed significantly better than those who used the semi-congruent display in the visual-only conditions; Augmenting the baseline visual mappings with audio output in the audio-visual conditions seems to have eliminated the observed effects of congruency levels.

Task completion times across groups

Level one (three sections)

There was no significant main effect of level of congruence on task completion times in the audio-visual ($F(2, 34) = 0.456, p = 0.638$), visual-only ($F(2, 34) = 0.679, p = 0.514$), and audio-only conditions ($F(2, 34) = 0.496, p = 0.614$).

Level two (four sections)

There was no significant main effect of level of congruence on task completion times across groups in the audio-visual condition ($F(2, 34) = 0.436p = 0.65$). In the visual-only condition, there was a significant main effect of level of congruence on task completion times across groups ($F(2, 34) = 3.72, p = 0.035, \eta^2 = 0.18$) and post-hoc tests showed that participants in the incongruent condition (random: $mean = 4649.1ms, sd = 971.6ms$) spend significantly longer time to complete the task than those in the semi-congruent condition (size: $mean = 3655.5ms, sd = 1179.2ms$) ($p = 0.04$) and those in the congruent condition (spikes: $mean = 3458.6ms, sd = 1162.8ms$) ($p = 0.015$). There was no significant difference between task completion times in the congruent and semi-congruent conditions ($p = 0.667$). There was also no significant main effect of level of congruence on task completion times across group in the audio-only condition ($F(2, 34) = 0.303, p = 0.74$).

Level three (five sections)

There was no significant main effect of level of congruence on task completion times across groups in any of the audio-visual ($F(2, 34) = 2.977, p = 0.065$), visual-only ($F(2, 34) = 2.792, p = 0.076$) and audio-only conditions ($F(2, 32) = 1.805, p = 0.181$). The above results showed that participants who used the congruent and semi-congruent display were significantly faster than those who used the incongruent display, but this was the case only in the visual-only condition in complexity level two (where the screen was divided into four sections). Again, the introduction of audio output in the audio-visual condition seems to have eliminated these significant effects.

Results within each group

We combined data from all levels of complexity to analyse scores and task completion times within each group.

Congruent group

There was a significant main effect of display type on participants' scores in the congruent group ($F(2, 22) = 13.307, p = 0.001, \eta^2 = 0.547$). Post-hoc tests showed that participants in this group scored significantly higher in the visual-only condition ($mean = 21.41, sd = 4.99$) compared to the audio-only condition ($mean = 13.75, sd = 6.48$) ($p = 0.004$). Their scores in the audio-visual condition ($mean = 21.58, sd = 4.88$) were also significantly higher than in the audio-only condition ($p = 0.002$). Differences between their scores in the audio-visual and visual-only conditions were not statistically significant ($p = 0.861$).

There was also a significant main effect of display type on participants' task completion times ($F(2, 22) = 16.584$, $p = 0.001$, $\eta^2 = 0.601$). Post-hoc tests showed that participants spent significantly longer times to complete the task in the audio-only condition ($mean = 5574.5$, $sd = 2100.2$) compared to the visual-only condition ($mean = 3537.5$, $sd = 1250.9$) ($p = 0.001$), and to the audio-visual condition ($mean = 4325.7$, $sd = 1898.3$) ($p < 0.001$). There was no statistically significant difference between task completion times in the visual-only and audio-visual conditions ($p = 0.071$). The above results show that participants performance was best when using a visual-only display and that the combination of audio-visual output in a congruent display increased performance times without significantly improving scores.

Semi-congruent group

The effect of display type on participants in the semi-congruent group was not significant for scores ($F(2, 22) = 2.216$, $p = 0.133$) but it was significant for task completion times ($F(2, 22) = 3.369$, $p = 0.043$, $\eta^2 = 0.249$). For the latter, post-hoc tests showed that participants spent significantly longer times to complete the task in the audio-only condition ($mean = 4579.9$, $sd = 1608.6$) compared to the visual-only condition ($mean = 3670$, $sd = 1002.1$) ($p = 0.042$). Difference between the visual-only condition and the audio-visual condition ($mean = 4456.8$, $sd = 1507.6$) were also statistically significant with participants spending longer time in the audio-visual condition ($p = 0.009$). There was no statistically significant difference in task completion times between the audio-only and the audio-visual conditions ($p = 0.779$). These results show that combining auditory and visual output in a semi-congruent display increased performance times and levelled the scores across the three types of displays.

Incongruent group

There was a significant main effect of display type on participants' scores in the incongruent group ($F(2, 22) = 6.212$, $p = 0.013$, $\eta^2 = 0.383$). Post-hoc tests showed that participants scored significantly higher in the visual-only condition ($mean = 17.81$, $sd = 5.61$) compared to the audio-only condition ($mean = 20.9$, $sd = 4.15$) ($p = 0.011$) and to the audio-visual condition ($mean = 15.36$, $sd = 8.15$) ($p = 0.024$). The differences in scores between the audio-only and the audio-visual conditions were not statistically significant ($p = 0.18$). There was no significant main effect of display type on task completion times in this group ($F(2, 22) = 1.082$, $p = 0.358$). These results show that combining auditory and visual output in an incongruent display did not have a significant impact on scores and performance times.

Discussion

Our hypothesis that the congruent spikes mapping leads to better performances was only partially confirmed. Participants in the congruent group scored significantly higher than participants in the other groups but only when using a visual-only display. One of the most interesting findings in Study 1, which goes against our initial hypothesis is that the effects of the level of congruence seem to have been cancelled by the introduction of audio to the baseline visual conditions. Differences between participants' performances across groups in the audio-visual conditions were not statistically significant, which suggests participants relied on the audio output to compliment or compensate for the discrepancies in congruence levels used in the size and arbitrary mappings.

We note that a number of participants reported that they sometimes chose to ignore the shapes in the audio-visual conditions. This in turn suggests that those participants relied on the audio output as a primary source for determining spatial orderings of sequences of items, which should mean that they would perform well in the audio-only conditions. The objective data contradicts this analysis, however, showing performances in the audio-only conditions to be overall worse across all complexity levels. The shape mappings used in the audio-visual conditions supported better performances albeit at the expense of more effort.

But our hypothesis that participants would perform significantly better when using audio-visual as opposed to unimodal displays was also not fully supported since participants' scores across the three groups were consistently and often significantly higher in the visual-only conditions. These findings contrast those reported in the literature which often report advantages of crossmodal over unimodal cues in recognition and retention tasks [4, 38]. Interestingly, subjective feedback from the majority of participants did not reflect the analysis obtained from the objective data. For example, many participants across the three groups described how the speed of presentation of the shapes made the task more difficult to complete in the visual-only conditions and that the addition of tones improved this experience. In a recent study, Guastellow et al [8] found that auditory and visual stimuli presented at intervals of about 300ms often produce miss errors in one or the other channel, which could explain the lower scores we obtained in the audio-visual conditions. A possible explanation for these seemingly contradictory accounts is that participants' answers in the interviews and questionnaires reflected perceived as opposed to actual difficulty. The addition of the tones to the crossmodal display may therefore have improved their confidence without necessarily impacting their scores.

Our expectation regarding the semi-congruent mapping was also not confirmed. We expected the semi-congruent size mapping to provide better support for remembering spatial locations than an incongruent arbitrary mapping, but our results showed this the opposite to be the case. The size mapping we used exploits previously reported crossmodal correspondences between vertical location, pitch and object size [2, 26, 5], but the type of task we used in our study could be a possible explanation for why these correspondences did not yield better performance. Whereas crossmodal correspondences have been studied almost exclusively in laboratory settings with simple cues where participants often deal with single or dual items [6, 33], our results show that retention of a sequence of multiple items appears to be more challenging and thus requires more careful design of crossmodal support. Indeed, as complexity increased, participants in the semi-congruent group highlighted that whilst they were able to identify that an extreme location had occurred in a sequence (i.e. small and large shapes), they found it increas-

ingly challenging to accurately reproduce full sequences, particularly those including the middle ranges of the screen (sections two, three and four). So, we suggest that whilst requiring significantly more time to complete, the distinctive visual characteristics of the shapes used in the arbitrary mapping provided a better mapping in this case.

Interestingly, a number of participants from the incongruent group highlighted that whilst they found it challenging to focus on both the shapes and the tones in the audio-visual condition, they also felt that this challenge made the task more enjoyable and engaging. None of the participants in the other groups expressed this opinion when asked about their experiences and preferences. Thus, subjective feedback indicates that, although the incongruent display did not offer complimentary information, the challenge of combining incongruent information across auditory and visual modalities increased enjoyability and engagement with the task.

From the interviews we found that there were two distinct types of responses to the addition of tones to the crossmodal displays. The first was that tones were treated as a dominant output mode, with the shapes ignored or used as a secondary source of spatial information. This was often reported to be the case in the incongruent arbitrary mapping group. The second was that participants preferred to use the shapes as the dominant source of spatial information with tones used as a secondary channel. This was often the case in the congruent and semi-congruent displays. We also observed that participants tended to switch to this "complimentary strategy", where reliance on the secondary modality increased, as the task increased in complexity. These observations are inline with claims that crossmodal perception is most influenced by the sensory channel that mediates the least ambiguous information [13] and that the positive effects of crossmodal concurrent feedback can be explained by a reduction of workload [7]. Our results confirm these findings and highlight that levels of congruency can be a factor in determining complementarity of information display.

STUDY 2: USER ENGAGEMENT

Given the subjective feedback reported in Study 1, we ran a second study focusing on engagement and user experience. This complements the focus on performance-based measures in Study 1, and follows a trend within HCI studies to take experiential issues into account, emphasizing *"the experience of using the technology, rather than the focus on the task that is characteristic of many other approaches HCI"* [11] and aiming to understand *"how the user makes sense of the artefact and his/her interactions with it at emotional, sensual, and intellectual levels"* [43]. This trend has often been ignored in the study of crossmodal displays. In order to facilitate an experience that would be more conducive to engagement and enjoyment, and in response to the reported appeal of 'challenge' identified in Study 1 (particularly with arbitrary shapes), we adapted our test application into a game. With a few exceptions, digital games use the potential of visual display for aesthetic appeal and for elements of game design more than auditory display [20]. Among the exceptions in the field of mobile games are the *Papa Sangre* series

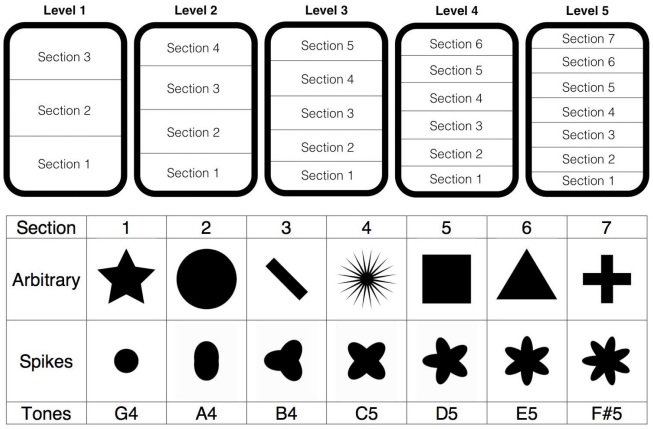

Figure 4. Levels and sections (top), crossmodal mappings (bottom)

[40] and *Dark Echo* [19], which are audio-focused games that cannot be played without sound. Nacke *et al* [20] have reported on the importance of auditory display for gameplay experience across different experiential dimensions (immersion, tension, competence, flow, negative and positive affect, and challenge). We were therefore also interested in examining the role of crossmodal display in gameplay.

Apparatus

We added a number of game design elements based on current design practices in games, particularly mobile games. Levels are a common concept – as the player succeeds in a task, she/he moves to a higher level, often with a higher degree of difficulty. We added a level identity and introduced a progression logic where the environment and challenge remains unchanged unless the player passes the challenge – in which case, the difficulty level will increase. The levels, sub-levels and the progression logic we added were as follows:

1. A player moves to a next sub-level upon successful completion of a trial. A new sequence of ShapeTones is then generated.
2. After ten sub-levels, a new level starts, with an additional ShapeTone – one vertical section is added to the initial three, and so forth, up to seven (Figure 4).
3. A training area is presented to the player at the beginning of each level for testing the new ShapeTones.
4. If the player fails a trial, the sub-level does not progress and the same ShapeTone sequence is played again.

Another common element in games is the score. The usage of scores and visual metaphors such as stars, to indicate degree of success or progress, are common gamification instruments [28]. We added a two-tier score – a star score (1-3 stars, dependent on performance on that level), and a numerical score. The scores and progression feedback we added were:

1. A trial score of one to three stars for passing a sub-level based on speed of playing back the correct sequence (one for slower, three for faster).

Figure 5. Game images, from left to right: arbitrary shapes version; spikes version; feedback after a trial; followed by global score.

2. Win and lose graphics and sounds for the end of each trial. (Figure 5).

3. The sub-level score accumulates in a global score, presented to the player at the end of each trial (Figure 5).

All other functionalities and application design remained unchanged from Study 1.

Study Design

We aimed to examine user engagement with two versions of the resulting game. We used the spikes and arbitrary shape mappings for each version because these emerged as the most successful visuals mapping in terms of performance in the first study. Several instruments have been developed to measure engagement in games, such as the Game Engagement and Game Experience Questionnaires [21]. Some of these instruments give particular attention to the concept of flow, such as the GameFlow model [39]. However, a range of diverse features contribute to engagement in games [3], we therefore adopted for an instrument that takes into account this diversity, the User Engagement Scale (UES) [24], more specifically the UESz version [41], which unlike other game engagement instruments, has been empirically validated. UES is a self-report measure consisting of a 31-item measured as a 5-point Likert scale that takes into account multiple dimensions of engagement: aesthetic appeal, perceived usability, felt involvement, novelty, focused attention and endurability. [24]. The UES has also been used in comparative studies [22]. Wiebe *et al.* [41] revised the UES for use in games (renamed as UESz) by organising the measures into four subscales: Focused Attention (FAz), Perceived Usability (PUz), Aesthetics (AEz) and Satisfaction (SAz). Later studies on the UES agree with the UESz revised set of subscales [22].

We used a within-subject design and invited participants to play the two versions of the game for 10 minutes each (10 minutes was the minimum duration of gameplay in similar studies deploying the UESz [41]). Before each 10 minute session, participants could play with the game for a short while (typically around 3 minutes), to get acquainted with it. The sequence of versions of the game was randomised and counterbalanced. We then asked participants to fill in UESz questionnaires for each version of the game and logged their scores for later analysis. An additional reason to use UESz as the sole questionnaire was to avoid respondent fatigue, as respondents already had to answer the UESz twice – one for each version of the game. Finally, we conducted a short interview focusing on crossmodal issues, usability and overall satisfaction with the game. We asked questions about the appeal of the tones and shapes, and how well they were connected. We also asked what was more important to play the game: tones, shapes, or both; and if it could be played with only audio or visuals. We also asked participants if they found something frustrating, and if they would play the game again.

Subscale	Arbitrary	Spikes	Significance
FAz	3.4 (0.77)	3.8 (0.85)	W = -2.4, p = 0.016
PUz	2.8 (0.81)	3.3 (0.78)	W = -2.764, p = 0.006
AEz	3.1 (0.62)	3.6 (0.63)	W = -2.371, p = 0.018
SAz	3.8 (0.75)	4.0 (0.84)	W = -1.963, p = 0.05

Table 1. *Mean* (*S D*) of UESz subscales for arbitrary shapes and spikes.

Hypothesis

S2H1: The congruent spikes mapping will be more engaging than the incongruent arbitrary shapes mapping.

Participants

Twelve participants took part in this study different from those who took part in the first study (10 male and two female, *mean age* = 36.6, *sd* = 6.7). Participants were a mixture of university staff and students. All participants received a cash incentive for participating. When asked about previous experience in games, on a scale of 1 (not at all experienced) to 5 (very experienced), only one participant declared to be very experienced, with two additional ones answering 4 (*mean* = 2.8). Most of the participants (seven) considered themselves to be very experienced as musicians, with two additional ones answering 4 (*mean* = 3.9). Only one participant considered himself to be very experienced as a visual artist, with three additional ones answering 4 (*mean* = 3.3).

Results

We used a Wilcoxon signed-rank test to analyse data from the UESz questionnaires and Student t-test to compare logged scores. Results from the questionnaire revealed a statistically significant preference for the spikes mapping in all four UESz subscales (Table 1). This preference was higher for Focused Attention (FAz), Perceived Usability (PUz) and Aesthetics (AEz) with differences of 0.4 to 0.5 between means, and less marked with Satisfaction (SAz), with a difference of 0.2 between means. Participants also performed better using the spikes mapping, reaching on average a higher levels (*max level mean* = 39.5, *sd* = 4.6) than with the arbitrary shapes mapping (*max level mean* = 33.1, *sd* = 8.1). This difference was statistically significant ($t = -2.579$, $p = 0.026$).

Based on interview responses, 10 participants preferred the spikes mapping and stated that the relationship between tones and shapes was more effective with spikes. Two participants did not find the two visual mappings to be very different, with one expressing a preference for arbitrary shapes because they were more distinguishable, and another stating that he did not identify any relationship between shapes and tones. One participant who preferred the spikes mapping stated that arbitrary shapes *"are more noticeable, but harder to concentrate*

[on]". Another participant considered that the spikes mapping became more difficult to distinguish in higher levels.

When asked what was more important to play the game, tones or shapes, six of the participants answered that they mostly relied on tones; four stated that they played it mostly as a visual game; one participant mentioned that he alternated between focusing on tones and shapes depending on the visualisation type; and another mentioned that he almost did not notice the visuals. Independently of the main modality, eight of the 12 participants stated that they would use the secondary modality as a backup when the difficulty level was higher. Sample statements, from visual-focused participants: *"when I got lost I relied on sound"*, *"I used sound as a backup"*, *"rely on image then rely on sound as a backup"*, *"sound was used as a check, as a support"*; and from sound-focused participants: *"I would get a visual as something to refer back"*, *"visual element gave me a confirmation"*, *"when you got the first one wrong and do it again, visuals become more important"*. Two of the participants who played the game mostly as a sound game highlighted the importance of visual feedback for seeing the screen and where the fingers were placed. Ten of the participants consider that they could play the game with any single modality (audio or visuals only), with two answering that they would not be able to play without sound.

Regarding usability, participants were asked if anything frustrated them in the game. Nine of the 12 participants mentioned that the strictness of where to tap on the screen to reproduce a given ShapeTones, and the fact that there were no visual aids for this frustrated them. This frustration would increase in higher levels of the game. As one of the participants put it: *"I got the relationship [between the ShapeTones] right but the position wrong – there was a mismatch between my head and the screen"*. Another participant stated that this frustration *"is part of the fun"*. The same frustration was also conveyed when participants were asked to suggest further improvements – six of the participants suggested showing the ShapeTones sections (permanently or only temporarily). We observed that participants used different strategies for solving this issue and achieving a higher precision, by a strict positioning of the hand or by moving the device. Some remarks in the interviews confirm this, e.g. *"I tried to hold it in a different way, shifted and treated it as a piano"*. When asked if they would play the game again, eight of the 12 participants answered affirmatively, with two answering *"maybe"* and two negatively, one of which stating *"I'm not much of a player"* and the other mentioning lack of *"entertainment value"*. Four of the participants mentioned that they would recommend it as a pedagogical game for musical training.

Discussion

The study confirmed our hypothesis S2H1 that the congruent spikes mapping is more engaging than the incongruent arbitrary shapes mapping. The results from the UESz questionnaire point in this direction in all four subscales, although the results from the interviews reveal some slight variations. Mostly, the interviews confirmed the results from the questionnaires, manifesting preference in terms of aesthetics and crossmodal correspondence with spikes and tones. This is illustrated by statements as *"Spikes is more gratifying, easier to play"*. However, one of the participants showed a preference for arbitrary shapes in general, as shapes with a spikes mapping *"were more similar"*. Another participant stated a preference for arbitrary shapes in higher levels of the game (with higher number of ShapeTones) – he argued that in higher sections spikes were harder to disambiguate (it was harder to distinguish shapes with 6 or 7 spikes, for example), while the distinctiveness of arbitrary shapes became more useful. This might point to a problem with recalling the spikes mapping beyond a certain number of spikes. One of the participants who reported preference for the spikes mapping mentioned that the challenge posed by arbitrary shapes could make it more interesting for repeated play. In relation to the perceived importance of audio-visual display in the game, most of the participants (11 of 12) reported relying on both modalities to play the game. Independently of the main modality (audio or visual), most of them (eight) would rely more on the secondary modality as the difficulty increased, as a backup or additional check. This is in line with literature on the importance of audio for user experience in games [20].

Some participants reported frustrations during the study. A common element of frustration was the inability to see the sections, which caused more missed tones as the levels increased. However, one of the participants mentioned that this frustration was *"part of the fun"*. Although the game has a vertical orientation, two of the participants tilted the device, diagonally or horizontally, to better align their hands and fingers with the tablet. When asked about these strategies, one participant mentioned that he was trying to keep a constant hand alignment to the tablet. He observed that accidentally moving the tablet would misalign his hand, leading to a need to *"recalibrate"* his hand. Another participant mentioned that he was trying to align the tablet horizontally as a piano, a musical metaphor which he was familiar with. Two of the participants would hum back in tone a sequence after it was played, and before tapping. When asked why they did this, they replied that it would help memorisation and repetition. It represents a kind of auditory sketching before committing to a sequence. These elements – importance of keeping or removing frustrating elements, spatial strategies outside the frame of the tablet, auditory sketching before playing – could point towards future research directions.

GENERAL DISCUSSION & CONCLUSIONS

In this paper we examined crossmodal congruence in the context of a memory task in which we evaluated users' ability to determine spatial orderings of a sequence of items on the basis of audio, visual and audio-visual stimuli. Two studies were reported which explored task performance and user experience of crossmodal interaction with congruent, incongruent, and semi-congruent displays. In this section we summarise and compare the insights gained from these studies.

Congruent mappings are preferred, but the addition of audio cancelled its advantages: Findings from Study 1 showed that while a congruent spikes mapping led to better results in terms of task performance, its advantages were cancelled out by the addition of audio output. Findings from Study 2,

on the other hand, showed that the combination of audio output with a spikes mapping led to more user engagements as measured by UESz. Both studies also revealed problems with the spikes mapping when the complexity of the task increased (levels three, four, and five) and some preferences for the incongruent arbitrary shapes mapping with respect to the challenge and engagement of crossmodal gameplay. Therefore, there could be a threshold at which the clarity and effectiveness of the congruent mapping is saturated. Whilst requiring significantly more time to complete, the distinctive visual characteristics of the shapes used in the arbitrary shapes could provide a better mapping in those cases. Interestingly, the use of the size mapping as a semi-congruent display yielded poor results, even though it was based on crossmodal correspondences between vertical location, pitch and object size [2, 26, 5]. The type of task, in this case recalling the order of a sequence of items, as opposed to identifying a single item, challenged the effectiveness of these particular crossmodal correspondences and therefore calls for more careful design when using this mapping in crossmodal interfaces.

Preference for crossmodal display expressed, but not always confirmed by data: Most of the participants from Study 1 expressed a preference for audio-visual display. In Study 2, the majority of the participants also preferred using both modalities for playing the crossmodal game. However, scores were higher in Study 1 in the visual-only conditions, which contradicted the subjective feedback and observed interaction strategies in both studies. Studies of crossmodal support for spatial attention and motor learning often point out the positive effects of concurrent feedback. However, in general, little work has examined retention tests without audiovisual feedback [32] as we report. The presented studies therefore contribute a systematic evaluation of crossmodal feedback in the context of multimodal information processing. Indeed, our results point toward a subjective preference for crossmodal as opposed to unimodal interaction when task complexity increases. This is evidenced by the diminished effects of levels of congruency observed when auditory output was introduced in the audio-visual displays in Study 1. These findings are inline with accounts of self-management of working memory resources that is associated with multimodal interaction when there is an increase in cognitive demands [25].

Emergence of complimentary strategies using a primary and secondary modality: The above insight is related to a further observation that was also common to the two studies. In both studies, we have seen some users who prefer visuals and others who prefer audio as the primary mode, though both make more use of the secondary mode as task complexity increases. Further research should examine correlations between preferred primary mode and users background and demographics, e.g. musical training or preferred learning style.

Incongruent crossmodal mappings can sometimes be appealing: In both studies, incongruent crossmodal mappings were sometimes associated with positive effects, namely by presenting a challenge and a level of difficulty that rendered the interaction more interesting for some participants. It was *"part of the fun"*. These observations point towards an alternative dimension of crossmodal interfaces when seen from the perspective user experience and engagement, and not merely task performance. Further studies of user engagement through crossmodal interaction should therefore consider addressing this dimension in design.

Contributions, limitations & further research

The presented studies confirmed findings of previous research on the positive performance effect of congruent display for a new task - the memory task. We also found that the addition of auditory display impacts the effects of the levels of congruency and that participants increasingly relied on multiple modalities as task complexity increased. We showed how task and user experience and engagement measures could be used to inform the design of crossmodal interaction which had not been attempted previously. We also demonstrated the deployment of the UESz in a new domain (crossmodal games) where we found it to be an effective measure of engagement.

There are limitations to these findings, however. First, the relatively small number of participants and the specific type of task used in both studies make it unclear how these findings would generalise to other types of interactions. Second, while participants showed superior performances when using the congruent spikes mapping, it is difficult to predict how successful this particular mapping would be for higher levels of complexity, for example when spikes discernibility and hence the ability to count them becomes more challenging as they represent more complex levels (e.g. beyond 10 spikes). Third, while the addition of audio output was perceived as useful, we only used one type of auditory display and did not vary its congruency mappings. It therefore remains unclear how different levels of congruency of the audio output will change the obtained results, for example by using different timbres, or multiple tones that could also be counted to correspond to different levels on the screen. Fourth, we have displayed the ShapeTones such that they are shown in a neutral position on the screen. It would be interesting to examine how displaying ShapeTones in their corresponding sections on the screen would impact participants performances on retention tasks. Finally, in relation to measuring engagement, we have used only one type of questionnaire. Using additional types of measurements could therefore lead to more insights into users engagement with crossmodal displays.

Nonetheless, our findings raise several questions which we would like to explore further. Firstly, further investigation is needed into the relationships between congruity of display, preferred modality, task complexity, and performance. Secondly, explorations of how 'challenging' aspects of crossmodal mappings can be used to enhance playful user experiences are needed. Thirdly, exploring how the role of crossmodal elements outside the device, such as proprioceptive mappings, could inform the design of engaging crossmodal interaction. Finally, our long term aim is to explore how crossmodality could be used to inform the design of engaging experiences for people with a variety of sensory capabilities.

ACKNOWLEDGMENTS

This work was supported by EPSRC grant EP/J017205/1 and EU Marie Curie fellowship FP7 REA grant 627922.

REFERENCES

1. P Bach-y Rita. 1988. Brain plasticity. *Rehabilitation Medicine. Mosby* (1988), 113–8.
2. Elisheva Ben-Artzi and Lawrence E Marks. 1995. Visual-auditory interaction in speeded classification: Role of stimulus difference. *Perception & Psychophysics* 57, 8 (1995), 1151–1162.
3. Elizabeth A. Boyle, Thomas M. Connolly, Thomas Hainey, and James M. Boyle. 2012. Engagement in digital entertainment games: A systematic review. *Computers in Human Behavior* 28, 3 (May 2012), 771–780. DOI: http://dx.doi.org/10.1016/j.chb.2011.11.020
4. Alfred O Effenberg. 2005. Movement sonification: Effects on perception and action. *IEEE multimedia* 2 (2005), 53–59.
5. Karla K Evans and Anne Treisman. 2010. Natural cross-modal mappings between visual and auditory features. *Journal of vision* 10, 1 (2010), 6.
6. Thomas Ferris, Robert Penfold, Shameem Hameed, and Nadine Sarter. 2006. The implications of crossmodal links in attention for the design of multimodal interfaces: A driving simulation study. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 50. Sage Publications, 406–409.
7. Mark A Guadagnoli and Timothy D Lee. 2004. Challenge point: a framework for conceptualizing the effects of various practice conditions in motor learning. *Journal of motor behavior* 36, 2 (2004), 212–224.
8. Stephen J Guastello, Katherine Reiter, Matthew Malon, and Anton Shircel. 2015. When auditory and visual signal processing conflict: cross-modal interference in extended work periods. *Theoretical Issues in Ergonomics Science* 16, 3 (2015), 232–254.
9. Eve Hoggan, Topi Kaaresoja, Pauli Laitinen, and Stephen Brewster. 2008. Crossmodal congruence: the look, feel and sound of touchscreen widgets. In *Proceedings of the 10th international conference on Multimodal interfaces*. ACM, 157–164.
10. Michael Xuelin Huang, Will WW Tang, Kenneth WK Lo, CK Lau, Grace Ngai, and Stephen Chan. 2012. MelodicBrush: a novel system for cross-modal digital art creation linking calligraphy and music. In *Proceedings of the Designing Interactive Systems Conference*. ACM, 418–427.
11. Joseph 'Jofish' Kaye, Kirsten Boehner, Jarmo Laaksolahti, and Anna Stahl. 2007. Evaluating experience-focused HCI. In *CHI '07 extended abstracts on Human factors in computing systems (CHI EA '07)*. ACM, New York, NY, USA, 2117–2120. DOI: http://dx.doi.org/10.1145/1240866.1240962
12. Ju-Hwan Lee and Charles Spence. 2008. Feeling what you hear: Task-irrelevant sounds modulate tactile perception delivered via a touch screen. *Journal on Multimodal User Interfaces* 2, 3-4 (2008), 145–156.
13. Dominic W Massaro. 1998. Illusions and issues in bimodal speech perception. In *AVSP'98 International Conference on Auditory-Visual Speech Processing*.
14. Harry Mcgurk and John Macdonald. 1976. Hearing lips and seeing voices. *Nature* 264, 5588 (Dec. 1976), 746–748. DOI: http://dx.doi.org/10.1038/264746a0
15. Peter BL Meijer. 1992. An experimental system for auditory image representations. *Biomedical Engineering, IEEE Transactions on* 39, 2 (1992), 112–121.
16. Oussama Metatla, Nick Bryan-Kinns, Tony Stockman, and Fiore Martin. 2012. Supporting cross-modal collaboration in the workplace. In *Proceedings of the 26th Annual BCS Interaction Specialist Group Conference on People and Computers*. British Computer Society, 109–118.
17. Oussama Metatla, Nick Bryan-Kinns, Tony Stockman, and Fiore Martin. 2015. Sonification of reference markers for auditory graphs: Effects on non-visual point estimation tasks. In *PeerJ PrePrints*.
18. Micah M Murray, Sophie Molholm, Christoph M Michel, Dirk J Heslenfeld, Walter Ritter, Daniel C Javitt, Charles E Schroeder, and John J Foxe. 2005. Grabbing your ear: rapid auditory–somatosensory multisensory interactions in low-level sensory cortices are not constrained by stimulus alignment. *Cerebral Cortex* 15, 7 (2005), 963–974.
19. Shaun Musgrave. 2015. 'Dark Echo' Review - Silence Is Golden, And So Is This Game. (Feb. 2015). http://toucharcade.com/2015/02/11/dark-echo-review/
20. Lennart E. Nacke, Mark N. Grimshaw, and Craig A. Lindley. 2010. More than a feeling: Measurement of sonic user experience and psychophysiology in a first-person shooter game. *Interacting with Computers* 22, 5 (Sept. 2010), 336–343. DOI: http://dx.doi.org/10.1016/j.intcom.2010.04.005
21. Kent L. Norman. 2013. GEQ (Game Engagement/Experience Questionnaire): A Review of Two Papers. *Interacting with Computers* 25, 4 (July 2013), 278–283. DOI: http://dx.doi.org/10.1093/iwc/iwt009
22. Heather O'Brien and Paul Cairns. 2015. An empirical evaluation of the User Engagement Scale (UES) in online news environments. *Information Processing & Management* 51, 4 (July 2015), 413–427. DOI: http://dx.doi.org/10.1016/j.ipm.2015.03.003
23. Heather L. O'Brien and Elaine G. Toms. 2008. What is User Engagement? A Conceptual Framework for Defining User Engagement with Technology. *J. Am. Soc. Inf. Sci. Technol.* 59, 6 (April 2008), 938–955. DOI:http://dx.doi.org/10.1002/asi.v59:6

24. Heather L. O'Brien and Elaine G. Toms. 2010. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology* 61, 1 (Jan. 2010), 50–69. DOI: http://dx.doi.org/10.1002/asi.21229
25. Sharon Oviatt, Rachel Coulston, and Rebecca Lunsford. 2004. When do we interact multimodally?: cognitive load and multimodal communication patterns. In *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 129–136.
26. Geoffrey R Patching and Philip T Quinlan. 2002. Garner and congruence effects in the speeded classification of bimodal signals. *Journal of Experimental Psychology: Human Perception and Performance* 28, 4 (2002), 755.
27. Mary C. Potter, Brad Wyble, Carl Erick Hagmann, and Emily S. McCourt. 2013. Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception, & Psychophysics* 76, 2 (Dec. 2013), 270–279. DOI: http://dx.doi.org/10.3758/s13414-013-0605-z
28. Rick Raymer. 2011. Gamification: Using Game Mechanics to Enhance eLearning. *eLearn* 2011, 9 (Sept. 2011). DOI: http://dx.doi.org/10.1145/2025356.2031772
29. Paul Rodway. 2005. The modality shift effect and the effectiveness of warning signals in different modalities. *Acta Psychologica* 120, 2 (2005), 199–226.
30. Richard A Schmidt. 1991. Frequent augmented feedback can degrade learning: Evidence and interpretations. In *Tutorials in motor neuroscience*. Springer, 59–75.
31. Ladan Shams, Yukiyasu Kamitani, and Shinsuke Shimojo. 2000. Illusions: What you see is what you hear. *Nature* 408, 6814 (Dec. 2000), 788–788. DOI: http://dx.doi.org/10.1038/35048669
32. Roland Sigrist, Georg Rauter, Robert Riener, and Peter Wolf. 2013. Augmented visual, auditory, haptic, and multimodal feedback in motor learning: A review. *Psychonomic bulletin & review* 20, 1 (2013), 21–53.
33. Charles Spence. 2010. Crossmodal spatial attention. *Annals of the New York Academy of Sciences* 1191, 1 (2010), 182–200.
34. Charles Spence. 2011. Crossmodal correspondences: A tutorial review. *Attention and Perceptual Psychophysics* (2011). DOI: http://dx.doi.org/DOI10.3758/s13414-010-0073-7
35. Charles Spence and Jon Driver. 1997. Cross-modal links in attention between audition, vision, and touch: Implications for interface design. *International Journal of Cognitive Ergonomics* (1997).
36. Charles Spence and Jon Driver. 2004. *Crossmodal space and crossmodal attention*. Oxford University Press.
37. Charles Spence and Cristy Ho. 2015. Multisensory information processing. In *APA handbook of human systems integration*, D. A. Boehm-Davis, F. T. Durso, and J. D. Lee (Eds.). American Psychological Association, Washington, DC, Ch. 27.
38. JK Stefanucci and DR Proffitt. 2005. Multimodal interfaces improve memory. In *Proceedings of the 11th International Conference on Human-Computer Interaction (HCI'05)*.
39. Penelope Sweetser and Peta Wyeth. 2005. GameFlow: a model for evaluating player enjoyment in games. *Computers in Entertainment (CIE)* 3 (July 2005), 3–3. DOI: http://dx.doi.org/10.1145/1077246.1077253 ACM ID: 1077253.
40. Andrew Webster. 2013. Gaming in darkness: 'Papa Sangre II' is a terrifying world made entirely of sound. (Oct. 2013). http://www.theverge.com/2013/10/31/5048298/papa-sangre-ii-is-a-terrifying-world-made-of-sound
41. Eric N. Wiebe, Allison Lamb, Megan Hardy, and David Sharek. 2014. Measuring engagement in video game-based environments: Investigation of the User Engagement Scale. *Computers in Human Behavior* 32 (March 2014), 123–132. DOI: http://dx.doi.org/10.1016/j.chb.2013.12.001
42. Fredrik Winberg. 2006. Supporting cross-modal collaboration: adding a social dimension to accessibility. In *Haptic and Audio Interaction Design*. Springer, 102–110.
43. Peter Wright, Jayne Wallace, and John McCarthy. 2008. Aesthetics and experience-centered design. *ACM Transactions on Computer-Human Interaction* 15, 4 (Dec. 2008), 18:1–18:21. DOI: http://dx.doi.org/10.1145/1460355.1460360