Perception, 2012, volume 41, pages 339 – 353

doi:10.1068/p7161

Cross-modality correspondence between pitch and spatial location modulates attentional orienting

Rocco Chiou, Anina N Rich

Department of Cognitive Science, and ARC Centre of Excellence in Cognition and its Disorders, Macquarie University, NSW 2019, Australia; e-mail: roccochiou@gmail.com; anina.rich@mq.edu.au

Received 26 October 2011, in revised form 2 February 2012

Abstract. The brain constantly integrates incoming signals across the senses to form a cohesive view of the world. Most studies on multisensory integration concern the roles of spatial and temporal parameters. However, recent findings suggest cross-modal correspondences (eg high-pitched sounds associated with bright, small objects located high up) also affect multisensory integration. Here, we focus on the association between auditory pitch and spatial location. Surprisingly little is known about the cognitive and perceptual roots of this phenomenon, despite its long use in ergonomic design. In a series of experiments, we explore how this cross-modal mapping affects the allocation of attention with an attentional cuing paradigm. Our results demonstrate that high and low tones induce attention shifts to upper or lower locations, depending on pitch height. Furthermore, this pitch-induced cuing effect is susceptible to contextual manipulations and volitional control. These findings suggest the cross-modal interaction between pitch and location originates from an attentional level rather than from response mapping alone. The flexible contextual mapping between pitch and location, as well as its susceptibility to top – down control, suggests the pitch-induced cuing effect is primarily mediated by cognitive processes after initial sensory encoding and occurs at a relatively late stage of voluntary attention orienting.

Keywords: cross-modal correspondence, auditory – visual interactions, exogenous attention, endogenous attention, polarity correspondence

1 Introduction

In everyday life, we effortlessly integrate information across our senses, combining disparate signals to gain a seamless conscious percept of events in the world. Traditionally, most researchers focus on the roles of spatial and temporal parameters in multisensory integration (Frens et al 1995; Jones and Jarick 2006). In recent years, researchers have started to explore how multisensory integration is modulated by high-level cognitive factors, such as semantic congruency (Doehrmann and Naumer 2008) and cross-modal associations (Parise and Spence 2008, 2009). In the case of cross-modal associations, people show a systematic tendency to map high-pitched sounds with small, bright objects located high up (Spence 2011). Such tendencies have been exploited in ergonomic design. For instance, the mapping between auditory pitch and vertical location has been shown to make visual and acoustic signals integrate more effectively in environments such as auditoria (Cabrera and Morimoto 2007; Roffler and Butler 1968).

The ubiquity of cross-modal associations and the functional significance of successfully integrating information across our senses make their cognitive roots important and intriguing to study. The correspondences between audition and vision may result from an intrinsic tendency of our perceptual system to map stimuli systematically across senses. Alternatively, it may depend upon semantic representations and have nothing to do with perceptual processing. Previous studies have shown that the cross-modal correspondences between pitch and visual attributes may occur involuntarily, in the sense that they affect behaviour when there is no benefit to the participant in making the associations (Spence 2011).

As with other cross-modal mappings that have been identified (eg smaller, brighter, and more angular shapes being congruent with high-pitched sounds—Gallace and Spence

2006; Marks 1987), the cross-modal correspondence between pitch and spatial location (high versus low) has been measured by speeded classification tasks. The results of several studies show that the pitch of a sound can influence the speed at which one discriminates the location of a visual target. That is, responding to an upper (versus lower) visual stimulus is faster when preceded or accompanied by a high (versus low) tone, compared with the opposite (incongruent) pairing (Ben-Artzi and Marks 1995; Bernstein and Edelstein 1971; Evans and Treisman 2010; Patching and Quinlan 2002). Two stimulus–response compatibility studies also found a similar effect when the response buttons were positioned in an upper or lower location: responding to a high (low) tone was faster with a button at an upper (lower) location (Lidji et al 2007; Rusconi et al 2006). In addition, a recent study reported that 3- to 4-month-old preverbal infants showed longer preferential looking at cross-modally congruent bimodal displays (eg pitch went up/down as the visual stimulus rose/fell) than at cross-modally incongruent displays (Walker et al 2010). Together, these studies suggest a fundamental correspondence between auditory pitch and spatial location.

The source of the pitch effect on visual tasks may be due to cross-modal mappings affecting response-selection processes. According to the polarity-correspondence principle (Proctor and Cho 2006), stimuli (eg a high tone and an upper visual target) and response alternatives (eg an upper response button) are coded as positive and negative polarity along physical or conceptual dimensions. Thus, a conflict in polarity codes means stimuli and responses activate opposite ends of a polarity spectrum, which slows down response-selection processes. Another possibility is that auditory pitch has a direct impact on the perceived visual height, such that a concomitant high tone makes the visual stimulus appear higher than it actually is. Maeda and colleagues (2004) tested this perceptual account and showed that, when viewing ambiguous moving gratings, participants tended to report upward/downward visual motion if the concomitant auditory stimuli ascended/descended in pitch. However, the task used in this study is not criterion free and thus is open to the criticism that decisional bias may affect participants' judgments. In addition, the effects of other cross-modal mappings may be based on shifts in decision criteria rather than perceptual phenomena (the pitch–size mapping—Keetels and Vroomen 2011; the pitch–brightness mapping—Marks et al 2003), although some studies with criterion-free measures suggest a perceptual origin (Ernst 2007; Parise and Spence 2009). The third plausible interpretation is that pitch biases spatial attention towards an upper or lower location in space, suggesting the cross-modal correspondence occurs at an earlier processing stage than does response selection.

Speeded classification tasks can quantify the cross-modal correspondences. The need for response selection, however, means such designs cannot identify any earlier stage at which the cross-modal interaction might arise. Evans and Treisman (2010) aimed to circumvent the problem of response selection by counterbalancing all stimulus–response mappings (see appendix 1 of Evans and Treisman 2010) and using an indirect task (discrimination of grating orientation). They observed cross-modal congruency effects with the indirect task and suggested the effect occurs at a perceptual level. However, a conflict between pitch and location may still interfere with other response-related processes (eg response readiness—see Hommel 1996), which is hard to gauge and avoid with their design. Moreover, a discrimination task does not clarify the stage of perceptual processing, from initial sensory encoding to late semantically mediated attention orienting, at which the mapping occurs. Here, we examined the stage in which the cross-modal mapping originates using a speeded detection task that required participants to simply press the same button whenever they detected a visual target, thus effectively removing the response-selection component.

If the cross-modal correspondence between pitch and spatial location is underpinned by attentional mechanisms, we should see effects of pitch on simple detection

of visual targets in upper and lower spatial locations. If, by contrast, the effect occurs primarily at a response-related level, it should not be evident in a task where response selection is not required. To assess the attentional hypothesis, we used a modified version of Posner's attention-cuing paradigm (Posner 1980). Non-predictive and non-lateralised (ie coming from both sides of the computer monitor) tones of high and low pitch were presented as cues. Participants were asked to detect low-contrast visual targets. If pitch causes a shift of attention rather than a response conflict, we should see congruency effects. That is, a high-pitched tone should shift attention to an upper location, and a low-pitched tone should bias attention to a lower location, and responses should be quicker when the target appears at the corresponding location. On the other hand, responses should be slower when the target appears at the opposite side to the pitch – location mapping.

In experiment 1 we tested whether pitch influences the allocation of visual attention in space. We had two spatial frames: in the vertical condition, the visual target appeared above or below the central fixation; in the horizontal condition, the target appeared to the left or right of the fixation. In experiment 2 we explored the influence of pitch difference on the cuing effect by manipulating the magnitude of the difference between high and low tones. In experiment 3 we tested whether the cuing effect reflects an intrinsic mapping between absolute pitch and particular location or a contextual mapping between relative pitch and location. Finally, in experiment 4 we tested whether the cuing effect is susceptible to volitional control by making the tones predictive such that targets were more likely to be in the opposite location from the natural cross-modal mappings.

To pre-empt our results, detection of visual targets was affected by the cross-modal mapping between pitch and vertical location, even though the sound was not spatially lateralised and not predictive of target location. In contrast, auditory pitch caused no cuing effects in the horizontal condition (experiment 1). In addition, frequency difference potently modulated the pitch-induced attention cuing effect. Reducing the difference between high and low tones abolished the effect (experiment 2). We also found that manipulating the contextual range of pitches modulated the direction of attention shifts. Depending on the relative pitch within a context, attention was shifted upwards or downwards in response to sounds of identical pitch height (experiment 3). Finally, we found that the cross-modal mapping could be overwhelmed by voluntary attention: predictive cues could reverse the direction of the cuing effect (experiment 4). Taken together, our results show that the cross-modal mapping occurs during attention orienting, an earlier stage than response-related processes. Furthermore, its susceptibility to contexts and volition suggests that, rather than occurring at an early stage of multisensory processing, this effect occurs after the recognition of contextual relative pitch and reflects deployment of voluntary attention.

2 Experiment 1

In experiment 1 we used non-predictive, non-lateralised sound cues prior to a visual target appearing in one of two locations, to test for a shift in visual attention caused by pitch (figure 1). In separate blocks, the locations could be above/below fixation ('vertical') or to the left/right of fixation ('horizontal'). This allows us to test whether there is a generic mapping between pitch and any spatial axis, as predicted by the polarity account that any congruency effects simply reflect a binary assignment of opposite polarities (Proctor and Cho 2006), or whether it is specific to the vertical axis. There is a clear congruency mapping for vertical location from the literature—high pitch with upper location and vice versa (Evans and Treisman 2010). We defined the congruency of the horizontal condition based on two studies with participants with substantial musical training in which low/high pitch mapped to the left/right (Lidji et al 2007; Rusconi et al 2006). As our participants were not musicians, the horizontal congruency mappings are unlikely to be strong and should act as a control to test the polarity mapping hypothesis.

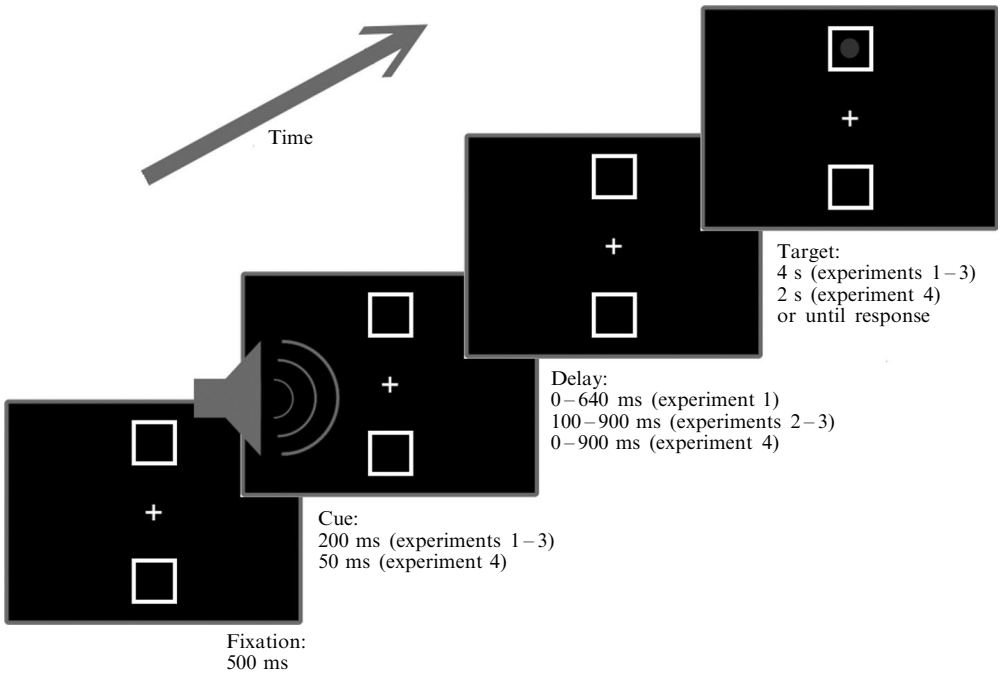

Figure 1. Example of the sequence of events in a trial. Note that the auditory cue was presented simultaneously from both sides of the monitor. In the horizontal condition of experiment 1, the two boxes were situated to the left and right of the central cross. Not to scale.

2.1 *Methods*

2.1.1 *Participants.* Twelve participants (four male, eight female; aged 20 – 25 years) from the Macquarie University subject pool took part after giving informed consent. All reported normal or corrected-to-normal vision and normal hearing and were naive to the purpose of the experiment. None were musicians.

2.1.2 *Apparatus, stimuli, and design.* A Pentium III computer was used for stimulus presentation and response collection, and the stimuli were displayed on a 17-inch CRT monitor. The experimental procedure was controlled by Matlab 7.5 with Psychophysics Toolbox (Brainard 1997; Pelli 1997). The participants sat at a viewing distance of 57 cm with a chin rest to stabilise head position. The trial procedure is shown in figure 1. The stimuli appeared in white against a black background, with the exception of the low-contrast visual target, which was a dark grey circle (RGB triplet: 25, 25, 25; 1° in diameter). Trials began with the fixation display consisting of a central cross (1 deg $\times$ 1 deg) flanked by two placeholders. The two placeholders subtended 3.5 deg and were positioned 7.4 deg either to the left and right of (in the horizontal condition) or above and below (in the vertical condition) the central cross. After 500 ms, a sinusoidal tone of either high (1500 Hz) or low (300 Hz) pitch was presented for 200 ms, while the fixation display remained on the screen. Sounds were presented at a comfortable hearing level. The sounds were played through two loudspeakers positioned to the left and right of the computer screen so there was no lateral spatial information. After various delay times (0, 80, 160, 320, or 640 ms), the low-contrast visual target appeared in one of the two placeholders. the target was equally likely to appear in either of the two locations, and the tone did not predict the target location. This display remained visible until a response was made or 4000 ms had elapsed.

Participants were instructed to fixate centrally and press the spacebar with their preferred hand as quickly as possible when they saw the grey circle. On 8% of trials,

no target appeared (catch trials), to discourage participants from making anticipatory responses. The participants received a warning signal on the screen ('ERROR!' in red) if they responded on a catch trial. The inter-trial interval was 750 ms. On another 8% of trials, no sound was played before the target (no-sound trials), to discourage participants from using the sound to predict target onset, although note the variable delay made this unpredictable even on cued trials.

There were two conditions randomly intermingled for each of the vertical and horizontal block types. In the vertical condition, the cross-modally congruent condition included high tone – upper target and low tone – lower target trials. The cross-modally incongruent condition had the reverse mappings (high tone – lower target; low tone – upper target). The congruency in the horizontal condition was based on two studies on the pitch – location mapping of professional musicians (Lidji et al 2007; Rusconi et al 2006). Namely, the congruent condition was low tone – left target and high tone – right target, and the incongruent condition had the reverse mappings (low – right; high – left).

2.1.3 *Procedure.* In every block, there were two congruency types (congruent or incongruent) and five delay times (0, 80, 160, 320, and 640 ms) randomly intermingled. The whole experiment consisted of two separate sections (vertical and horizontal). Each section was composed of a practice block of twenty-four trials and five experimental blocks of ninety-six trials, giving a total of forty experimental trials per condition, as well as eight catch trials and eight no-cue trials per block. The order (vertical or horizontal first) was counterbalanced across participants. Participants were informed that the auditory stimuli provided no information about the target location.

2.2 *Results*

Outliers (defined as RTs > 3 SD above an individual's mean for each condition) and anticipatory responses (RTs faster than 100 ms) were removed prior to statistical analyses. With these criteria, 4% of trials were discarded. Catch-trial errors and missed responses were less than 1% of total data and were removed from further analyses.

Mean RTs of each condition are shown in figure 2. A repeated-measures ANOVA was conducted on mean RTs, with factors of Delay Time (0, 80, 160, 320, and 640 ms), Spatial Axis (horizontal or vertical), and Cross-modal Congruency (congruent or incongruent). The Greenhouse – Geisser adjustments were used for correcting violations of sphericity, where necessary, and the Bonferroni correction was used in a posteriori comparisons.

As is evident in figure 2, RTs gradually decreased as the delay time between cue and target became longer until the longest delay time. The repeated-measures ANOVA confirmed a significant main effect of Delay Time ($F_{1.63,\,17.95} = 26.63$, $p < 0.0001$, $\eta^2 = 0.68$). This pattern illustrates the foreperiod effect (Bertelson 1967), with the

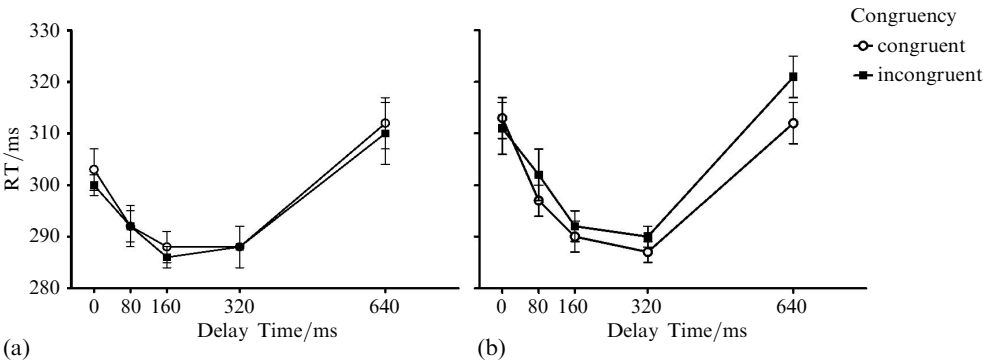

Figure 2. Mean reaction times from experiment 1 as a function of Delay Time and Cross-modal Congruency in the (a) horizontal and (b) vertical condition. Error bars represent ±1 SEM.

readiness of making a response increasing with prolonging delay time to a point (here 320 ms) then decreasing when the delay is long (640 ms). Crucial to the aim of our study, there was a significant interaction between Spatial Axis and Cross-modal Congruency ($F_{1,11} = 9.06$, $p = 0.01$, $\eta^2 = 0.45$—see figure 2). As figure 2 also shows, incongruent trials were slower than congruent trials in the vertical but not the horizontal condition. A posteriori pair-wise comparisons by Spatial Axis showed that the Cross-modal Congruency effect was significant only in the vertical condition ($p = 0.04$—see figure 2b), not in the horizontal condition ($p = 0.18$—see figure 2a). No other statistics reached significance (all ps > 0.10).

2.3 *Discussion*

The results of experiment 1 showed that non-predictive and non-lateralised tones elicit a congruency effect between pitch and spatial location in the vertical axis. As the task did not involve response selection, the significant congruency effect suggests that the cross-modal mapping between pitch and location originates at an attention-orienting stage. Moreover, the significant effect of Delay Time, reflecting the response readiness (foreperiod) effect, did not interact with the Congruency effect. The absence of interaction suggests that the two effects may be independent of one another, reflecting distinct cognitive processes (Sternberg 1969).

We found no evidence for an analogous cuing effect in the horizontal plane. As our participants were not musicians, this is consistent with the notion that the tendency to associate pitch with the left or right relies on musical training (Lidji et al 2007; Rusconi et al 2006). Importantly, this demonstrates that the vertical cuing effect reflects an implicit mapping between pitch and spatial height, rather than this being indicative of a general tendency to assign auditory pitches to binary stimuli/responses.

3 Experiment 2

In experiment 2 we explored whether the pitch-induced attention cuing effect was modulated by the magnitude of frequency difference between high and low tones. A previous study on the intra-modal mapping of pitch and loudness (faster RTs to high-pitched loud/low-pitched soft sounds than to the opposite) found that increasing frequency difference led to an increase in the size of the congruency effect when participants perform judgments on loudness (Melara and Mounts 1994). The authors argued that increasing frequency difference made variations in pitch less negligible and more likely to interact with task-relevant processes on loudness. In experiment 1 the frequency difference was 1200 Hz, which is relatively large and perceptually salient (around 27 semitones). We hypothesised that a large frequency difference makes it more probable for participants to notice whether a task-relevant sound is high- or low-pitched, which biases attention to a cross-modally congruent location. In contrast, a difference that is less obvious might be more effectively ignored when it is not relevant to the task. Based on this logic, we tested the effect of manipulating frequency difference on the pitch-induced cuing effect.

3.1 *Methods*

3.1.1 *Participants.* Twelve participants (six male, six female; aged 20 – 25 years) from Macquarie University were paid to take part. All reported normal vision and hearing and were naive to the purpose of the study. None of these participants had participated in experiment 1.

3.1.2 *Stimuli and design.* The apparatus and stimuli were as in experiment 1. The design was identical to experiment 1, with the following exceptions: we included only the vertical condition, manipulated the magnitude of the pitch difference, and used slightly different cue – target delays. We used a large frequency difference (1200 Hz, equivalent

to ~27 semitones; tones are identical to experiment 1) and a small frequency difference (in which the low and high tone were 300 Hz and 400 Hz, respectively; ~4.98 semitones). The two frequency difference conditions were separated into different sections, and the order of the two sections was counterbalanced across participants. Each section was composed of a practice block of twenty-four trials and five experimental blocks of ninety-six trials, giving a total of forty trials per condition. The delay times between cue and target were 100, 300, 500, 700, and 900 ms. Instructions were the same as in experiment 1.

3.2 *Results*

Outliers (3%) and catch-trial errors (0.8%) were removed prior to analysis. Figure 3 shows the mean RTs for this experiment. We conducted a repeated-measures ANOVA with the factors of Delay Time (100, 300, 500, 700, or 900 ms), Frequency Difference (100 or 1200 Hz), and Cross-modal Congruency (congruent or incongruent). The analyses revealed significant main effects of Delay Time ($F_{1.96,\,21.60} = 25.97$, $p < 0.0001$, $\eta^2 = 0.70$) and Cross-modal Congruency ($F_{1,\,11} = 5.15$, $p = 0.04$, $\eta^2 = 0.31$) and a significant interaction between Frequency Difference and Cross-modal Congruency ($F_{1,\,11} = 12.97$, $p = 0.004$, $\eta^2 = 0.54$). The difference between cross-modally congruent and incongruent conditions was significant when the frequency difference was large ($p = 0.001$—see figure 3b) but not when the difference was small ($p = 0.51$—see figure 3a). It is possible that participants may have shown a stronger effect if they did the small difference condition first, owing to a within-session contextual effect. We therefore conducted additional planned analyses to assess whether a congruency effect occurred in the small difference condition depending on the order. As the order of the two frequency-difference sessions was counterbalanced across participants, we separated them into those performing the small difference condition first and those doing it second and then tested for the presence of a congruency effect. We found no evidence of an effect of congruency in the small difference condition either in the participants doing this condition first ($t_5 = 1.70$, $p = 0.15$) or in those doing it second ($t_5 = -0.95$, $p = 0.38$). In contrast, the congruency effect was significant in the large difference condition regardless of whether it was first or second (both ps < 0.05).

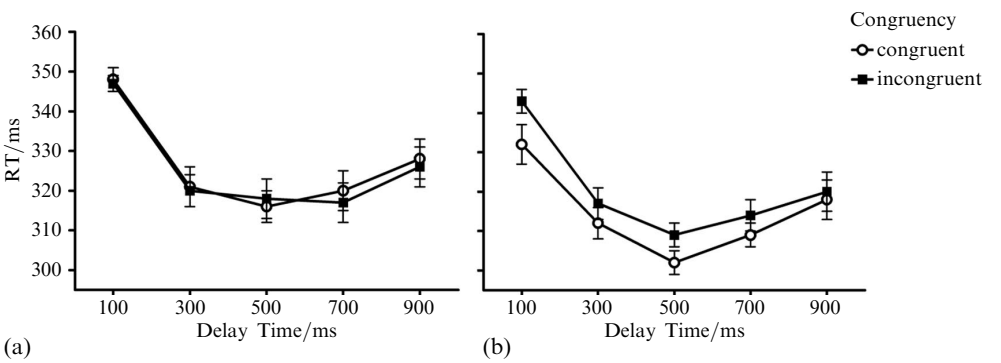

Figure 3. Mean reaction times from experiment 2 as a function of Delay Time and Cross-modal Congruency in the (a) 100 Hz difference and the (b) 1200 Hz difference condition. Error bars represent ±1 SEM.

3.3 *Discussion*

The attention-cuing effect in the large difference condition replicated the vertical condition of experiment 1 with different cue–target delay times. In contrast, decreasing the frequency difference between high and low tone to 100 Hz eliminated the cuing effect. The results suggest decreasing the salience of perceived difference in pitch makes the tones less likely to be coded as one end of polar opposites (ie high/low tone), which

have cross-modal associations with vertical locations (ie upper/lower position). Hence, the pitch cuing effect seems to depend on a perceptually salient difference. The elimination of the cuing effect cannot be due to an inability to discern a 100 Hz pitch difference, as previous studies have shown that frequency discrimination threshold for a 300 Hz tone is approximately 1.8 Hz ($\sim$0.6% of the tone), much smaller than 100 Hz (Sek and Moore 1995). Alternatively, deciding the pitch and coding the tone as one end of the two extremes might take more time in the small difference condition. This latter interpretation does not seem plausible, as we found no evidence of a cuing effect even at the longest delay time (ie 900 ms). Another interpretation is that the magnitude of the attention shifts in space is modulated by pitch difference. Thus, a small pitch difference would cause a spatially small attention shift, which can be revealed only when the distance between upper and lower targets is small. This argument necessitates an absolute scale on which pitch and visual location are mapped. In experiment 3 we test for such an intrinsic set mapping between pitch and location by manipulating the pitch context.

4 Experiment 3

In experiment 3 we explored whether the cross-modal mapping between pitch and location is contingent on an absolute mapping or the relative pitch within the context of the range used in an experimental block. If the cross-modal mapping is context dependent, the same sound will elicit attention shifts to opposite directions depending on the context. If, in contrast, the mapping reflects an association between absolute pitch and specific location, we should not be able to change the direction cued by a particular tone. To this end, we manipulated the range of high and low tones in separate experimental sections such that the same tone was the low sound in one section and the high sound in the other.

4.1 *Methods*

4.1.1 *Participants.* Eighteen participants (eight male, ten female; aged 20 – 25 years) from Macquarie University were paid to take part. All reported normal vision and hearing and were naive to the purpose of the study. None had participated in either of the previous two experiments.

4.1.2 *Stimuli and design.* The apparatus, stimuli, trial procedure, and design were identical to those of experiment 2, except for changes specified below. We used two frequency ranges. In one section, the low tone was 100 Hz, and the high tone was 900 Hz. In the other section, the low tone was 900 Hz, and the high tone was 1700 Hz. The order of the two sections was counterbalanced across participants. Each section was composed of a practice block of twenty-four trials and five experimental blocks of ninety-six trials, giving a total of forty experimental trials per condition. Instructions were the same as in previous experiments.

4.2 *Results*

Outliers (3%) and catch-trial errors (1%) were excluded from further analyses. A repeated-measures ANOVA was conducted on mean RTs, with factors of Delay Time (100, 300, 500, 700, or 900 ms), Context (100 – 900 Hz or 900 – 1700 Hz), and Cross-modal Congruency (congruent or incongruent).

As is clear in figure 4a, there were significant main effects of Delay Time ($F_{2.17,\,36.95} = 24.05$, $p < 0.0001$, $\eta^2 = 0.58$) and Cross-modal Congruency ($F_{1,\,17} = 14.78$, $p < 0.01$, $\eta^2 = 0.46$), once again replicating our primary effect of pitch-induced attentional cuing. No other effects were significant in the omnibus ANOVA (all ps > 0.15). To directly assess whether the relative pitch under a context determined the direction of attention orienting, we submitted the RTs of the 900 Hz condition in the two

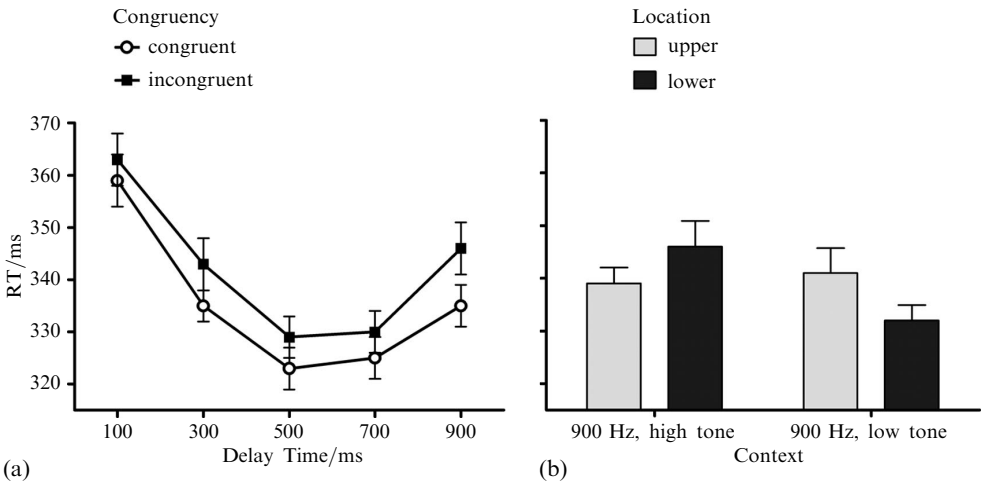

Figure 4. (a) Mean reaction times (±1 SEM) from experiment 3 as a function of Delay Time and Cross-modal Congruency. (b) Mean reaction times (+1 SEM) from the 900 Hz condition of experiment 3 as a function of Context and Target Location, collapsed across delays. Note that all four bars in the graph shown in (b) are from the 900 Hz condition, separated by whether the context was for this tone to be the high (left bars) or low (right bars) pitch and the location of the target (upper or lower). Error bars represent ±1 SEM.

different contexts into a repeated-measures ANOVA with the factors of Delay Time, Context, and Spatial Location (upper or lower). We found a significant effect of Delay Time ($F_{2.18,37.06} = 23.32$, $p < 0.0001$, $\eta^2 = 0.57$). Importantly, as is evident in figure 4b, there was a significant interaction between Context and Spatial Location ($F_{1,17} = 53.14$, $p < 0.0001$, $\eta^2 = 0.75$). A posteriori comparisons (one-tailed) revealed that, when the 900 Hz sound served as the low tone, detecting lower targets was significantly faster than detecting upper targets ($t_{17} = 2.16$, $p = 0.02$). Conversely, when the 900 Hz sound served as the high tone, there was a trend for detecting upper targets to be faster than lower targets ($t_{17} = 1.61$, $p = 0.06$).[(1)]

4.3 *Discussion*

Our results demonstrate that manipulating the context of the frequency range modulates the direction of the resulting attention shifts. This shows the pitch effect is context dependent rather than an absolute mapping between pitch and location. The contextual effect also suggests the pitch cuing does not arise from early sensory encoding, as the 900 Hz sound is identical in both contexts. Instead, the cuing effect seems to rely on the recognition of contextual relative pitch, implying a relatively late locus in attentional processes.

Note that, despite the same distance in the frequency dimension, the perceived pitch difference between 100 and 900 Hz (~38 semitones) is larger than that between 900 and 1700 Hz (~11 semitones), as auditory perception for pitch is not linear but logarithmic with respect to frequency (Sek and Moore 1995). Based on the results of experiment 2 in which only the large frequency difference caused a congruency effect, one might expect this effect to be stronger for the ~38 semitone difference than for the ~11 semitone difference. However, the absence of a Context by Congruency interaction in the omnibus ANOVA suggests that, once the pitch difference is obvious enough for participants to readily notice, the perceived pitch difference does not affect the magnitude of the effect.

[(1)] Two-tailed tests also show a significant difference between high and low targets in the 900 Hz low-tone condition ($p = 0.04$), but the 900 Hz high-tone condition did not reach significance ($p = 0.12$; numerically the difference was in the expected direction, detecting upper targets was numerically faster than detecting lower ones).

These results are consistent with a previous study that demonstrated the contextual effect on the cross-modal mapping between pitch and lightness (Marks 1987). In one experiment of that study, 220 Hz and 360 Hz were presented as the low and high tones, respectively. The author reported a significant interaction between pitch and lightness, with faster responses to a white stimulus when participants heard a 360 Hz sound than a 220 Hz sound, and vice versa when responding to a black stimulus. However, in another experiment where tones of 100, 220, 360, and 800 Hz were used within the same block the interaction between the same two tones (ie 220 Hz and 360 Hz) and lightness was greatly reduced, presumably because the sounds became the intermediate pitches under this context. These and our results indicate that the mapping between auditory pitch and visual features is relative in nature so that what is 'high-pitched' or 'bright' in one context may be 'low-pitched' or 'dark' in another.

5 Experiment 4

The finding that contextual relative pitch determines the direction of attention shifts implies that substantial mental processes after early auditory encoding mediate the pitch cuing effect. This leads us to suspect, despite pitch causing a shift in attention when it is non-predictive, the effect may be underpinned by volitional attention. It has been well-documented that volitional orienting is sensitive to top–down control, whereas involuntary attention is relatively unaffected (Jonides 1981). Based on this, we conducted two experiments. In experiment 4a we shorten the cue duration to 50 ms to be in keeping with the brief duration conventionally used in attention orienting tasks and tested whether the effect replicates with a brief cue. In experiment 4b we made the tones predictive of the opposite location to the mapping used in experiments 1 through 3. On 80% of the trials, targets appeared on the location opposite to the location of the natural cross-modal mappings (eg an upper target following a low tone). Participants were explicitly informed of these probabilities. In our previous experiments, the pitch effect usually reached significance at around 300 to 600 ms after cue onset. If volitional attention mediates the pitch effect, we would expect the cuing effect to be reversed at SOAs of 300 and 600 ms, when participants are motivated to adopt an inverse mapping. By contrast, if the pitch effect resists top–down control, like the involuntary cuing effects caused by a peripheral flash (Wright and Richard 2000), we should observe similar effects to those found in previous experiments and replicated in experiment 4a.

5.1 *Methods*

5.1.1 *Participants.* Twelve and ten students (aged 20–25 years) from Macquarie University participated in experiment 4a (five male, seven female) and 4b (six male, four female), respectively. All reported normal vision and hearing and were naive about our research aim. None participated in the previous three experiments.

5.1.2 *Stimuli and design.* The methods were identical to previous experiments, except for changes specified below. In experiment 4a we shortened the tone duration to 50 ms, as well as the target duration to a maximum of 2 s. Low and high tones were 250 Hz and 2500 Hz, respectively. The cue–target delay times used in experiment 4a were 0, 150, 300, 450, and 600 ms.

Experiment 4b was identical to 4a, with the exception that on 80% of the trials a high/low tone preceded a lower/upper target (incongruent), whereas on 20% of the trials targets appeared on the location associated with pitch (congruent). Thus, incongruent trials were four times more common than congruent trials. Participants were explicitly informed about this contingency. The delay times between cue and target were 0, 300, 600, and 900 ms. In each block, there were two congruency types (incongruent/expected or congruent/unexpected) and four delay times. The eight possible conditions

were presented in a randomly intermingled sequence. Each block included eighty-eight trials, including eight catch trials per block. The experiment started with a practice block of eighty-eight trials, followed by nine experimental blocks, giving a total of one hundred and forty-four trials in each incongruent (expected) condition and thirty-six trials in each congruent (unexpected) condition. All participants reported they were aware of the contingency between pitch and location after practice.

5.2 *Results*

5.2.1 *Experiment 4a.* Outliers (2%) and catch-trial errors (1%) were excluded prior to analysis. A repeated-measures ANOVA was conducted on mean RTs of each condition, with factors of Delay Time (0, 150, 300, 450, or 600 ms) and Cross-modal Congruency (congruent or incongruent). Figure 5a shows the mean RTs for each condition. The main effect of Delay Time was significant ($F_{1.91,21.03} = 21.12$, $p < 0.0001$, $\eta^2 = 0.65$). Importantly, we replicated the main effect of Congruency with the brief tone duration ($F_{1,11} = 4.70$, $p = 0.05$, $\eta^2 = 0.29$), and, as with previous experiments, there was no significant interaction between Delay and Congruency ($F_{2.99,32.94} = 1.33$, $p > 0.28$).

5.2.2 *Experiment 4b.* Outliers (2%) and catch-trial errors (0.9%) were excluded. Figure 5b shows the mean RTs for each condition. A repeated-measures ANOVA, with the factors of Delay Time (0, 300, 600, or 900 ms) and Cross-modal Congruency (congruent or incongruent), revealed significant main effects of delay time ($F_{1.84,16.61} = 14.34$, $p < 0.0001$, $\eta^2 = 0.61$) and Congruency ($F_{1,9} = 8.28$, $p = 0.01$, $\eta^2 = 0.49$). As evident in figure 5, RTs were overall faster for targets on the incongruent (expected) locations than on the congruent (unexpected) locations, indicating volitional control overrode the pitch effect. There was no significant interaction between Delay Time and Congruency ($F_{1.96,17.67} = 2.61$, $p = 0.10$, $\eta^2 = 0.22$). The pattern of data in figure 5 does hint, however, that the volitional effect took time to emerge but was sustained across a long range of delay times.[(2)]

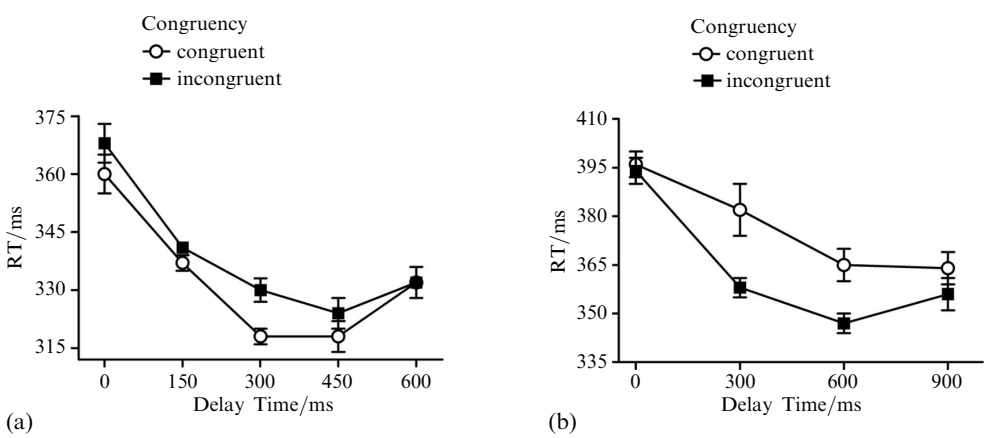

Figure 5. (a) Mean reaction times (±1 SEM) from experiment 4a as a function of Delay Time and Cross-modal Congruency. (b) Mean reaction times (±1 SEM) from experiment 4b as a function of Delay Time and Cross-modal Congruency (Expectancy). Error bars represent standard error of the mean. Note that the cue duration was 50 ms in experiment 4.

[(2)] A posteriori pair-wise tests show that at 0 ms Delay Time there is no reliable difference ($p > 0.69$), consistent with previous research that voluntary attention takes at least 200 ms to have a measurable effect (Horowitz et al 2009). At 300, 600, and 900 ms Delay Times the pitch effect was significantly reversed by volition (all ps < 0.05). Note, however, that these a posteriori tests are less conservative than strictly appropriate given the lack of an interaction in the omnibus ANOVA.

5.3 *Discussion*

The results of experiment 4a and 4b were clear-cut—the pitch effect on attention orienting was reversed when participants expected targets to appear at the location opposite to the natural mapping of pitch and location. Jonides (1981) proposes that the orienting of attention needs to meet the criterion of resisting volitional control to be considered involuntary. Although non-predictive auditory pitch orients attention, it is unable to act against strategic deployment of attention to maximise performance. Thus, our finding suggests the pitch cuing effect reflects more voluntary processes than involuntary orienting.

6 General discussion

The cross-modal correspondence between auditory pitch and spatial location has been demonstrated to affect reaction times in psychophysical experiments with binary responses (Ben-Artzi and Marks 1995; Evans and Treisman 2010) and is used by ergonomists to design multimodal displays (Cabrera and Morimoto 2007). We aimed to explore the cognitive processes underpinning this cross-modal mapping, specifically the effect of irrelevant pitch on visual attention. Experiment 1 showed that non-predictive and non-lateralised high- and low-pitched tones bias visual attention to locations in the vertical plane. Such effects did not occur in the horizontal plane, suggesting that the pitch-induced cuing effect is not simply a bias to assign polar opposites to any binary stimulus, unlike the process suggested to occur for binary judgments (Proctor and Cho 2006). In experiment 2 we found that reducing the frequency difference between the high tone and low tone could eliminate this attention cuing effect. Experiment 3 showed that the effect relies on relative pitch within a context. Finally, experiment 4 demonstrated that volitional control can override the effect of pitch on attention. Taken together, our results suggest that the cross-modal interaction of irrelevant pitch and location occurs at the stage of attention orienting, but that voluntary attention and the recognition of relative pitch height (rather than an intrinsic cross-modal scale between pitch and location) are critical.

A possible contributing factor to the present findings is the loudness of the auditory stimuli. We equalised the amplitude of tones, not their loudness, in all experiments. As a result, within the frequency range we used (100 – 2500 Hz), high tones would be perceived louder than would low tones, owing to the loudness threshold of the human auditory system decreasing with frequency (cf equal loudness contours—Schneider et al 1972). However, we believe that our results are unlikely to be fully explained by a possible loudness – space mapping, as the cross-modal effects of pitch can be observed regardless of whether loudness is specifically controlled. For instance, Patching and Quinlan (2002) demonstrated a congruency effect of the pitch – space mapping with the loudness of high and low tones matched, suggesting it cannot be explained by participants associating loud tones with high locations. Parise and Spence (2009) also showed effect of the pitch – size and pitch – shape mappings when loudness is controlled. Therefore, although loudness covaries with auditory frequency, it seems that this factor alone cannot entirely explain the present results.

The influences of cross-modal congruency and response readiness (indexed by the main effect of delay times) both significantly account for the variation in reaction times. However, these two factors do not interact in any of the four experiments. According to the influential additive factor method (Sternberg 1969), an interaction between two factors means either the influences of the two factors converge at a cognitive processing stage or a mental process is affected jointly by the two factors. In contrast, the absence of an interaction between two factors suggests the two factors selectively affect two individual cognitive processes in series. As always, we need to be cautious in interpreting the lack of an effect; however, the repeated lack of an interaction between Delay Time and Congruency across all of our experiments implies

that the influence of cross-modal congruency and that of response readiness occur in different stages of cognitive processing.

If we had observed an absolute pitch – location mapping, it would suggest that the pitch-induced attention shifts arise at an inter-sensory stage prior to cognitive processing. However, the significant contextual effect of relative pitch suggests instead that the process of mapping pitch to location occurs only after participants recognise whether the sound represents a high or low tone. This also means that, although the cross-modal mapping affects attention, this effect is cognitively mediated and does not seem to be based on purely perceptual processes. Our explanation fits well with previous studies suggesting that post-sensory processing (whether it is semantic labeling or polarity mapping) is necessary for cross-modal interactions between auditory pitch and visual features to occur (Gallace and Spence 2006; Patching and Quinlan 2002).

Our results could be explained by the polarity correspondence principle if one assumes that both pitch and location have some intrinsic polarity mapping such that high pitch and upper location are coded similarly (and vice versa for low tone and lower location). The polarity correspondence principle was originally proposed to explain the stimulus – response congruency effect in binary classification tasks, such as the mapping between numbers and left/right responses (Proctor and Cho 2006). Unlike a binary classification task, our detection task does not require response selection. Therefore, a polarity explanation would have to include a step at which auditory pitch orients visual attention to the location that shares the same polarity coding which, in turn, affects the performance on target detection. In addition, our lack of effect in the horizontal plane (in non-musicians) suggests either the polarity account requires experience or that it is not an adequate explanation of our vertical cuing effects. It is certainly not the most parsimonious explanation.

Numerical cuing effects bear a similarity to our auditory pitch cueing effect in the sense that they both map to space via conceptual mediation. Fischer and colleagues (2003) reported that spatially non-predictive numerals can also induce attention shifts to the left or right, as if visual attention is reflexively biased by an invisible mental number line. Subsequent studies then reported that the numeral-induced attention cuing effect could be easily modulated by contextual manipulation (Dodd 2010; Galfano et al 2006; Ristic et al 2006), leading to conclusions that it is underpinned by volitional attentional processes. Other attentional cuing effects, such as those induced by eye gaze and arrows, are unaffected by probability manipulations at short SOAs (Friesen et al 2004; Kuhn and Kingstone 2009). Our pitch-induced attention cuing effect occurs involuntarily in the sense that the sound is task-irrelevant and spatially non-predictive. Nevertheless, its susceptibility to contextual modulation and volitional control implies it is more akin to the numeral-induced cuing effect rather than the less volitional cuing effect caused by eye gaze and arrows.

The results of our study clearly demonstrate that spatially non-lateralised and non-predictive sounds of different pitches can induce attention shifts, and this pitch effect can be flexibly modulated by contextual factors such as frequency range and top – down control, indicating volitional attention is the underlying mechanism. This cross-modal mapping may subserve unusual cross-modal phenomena such as auditory – visual synaesthesia, in which auditory stimuli elicit conscious visual experience (Ward et al 2006). In particular, recent findings suggest high-pitched sounds elicit synaesthetic images that appear in higher locations, leading researchers to suggest that auditory – visual synaesthesia may share mechanisms with cross-modal mappings of non-synaesthetes (Chiou et al 2012; Fernay et al 2011). In a broader sense, our findings also imply that our perceptual and attentional systems have an intrinsic tendency towards cross-modal associations, consistent with the critical role multisensory integration has in our conscious perception.

Acknowledgments. RC is supported by a Macquarie University Research Excellence Scholarship and the Educational Ministry of Taiwanese Government. ANR is supported by the Australian Research Council (DP0984494). We thank Matthew Finkbeiner and Samantha Baggot for helpful comments on the manuscript.

References

Ben-Artzi E, Marks L E, 1995 "Visual-auditory interaction in speeded classification: role of stimulus difference" *Perception & Psychophysics* **57** 1151 – 1162

Bernstein I H, Edelstein B A, 1971 "Effects of some variations in auditory input upon visual choice reaction time" *Journal of Experimental Psychology* **87** 241 – 247

Bertelson P, 1967 "The time course of preparation" *Quarterly Journal of Experimental Psychology* **19** 272 – 279

Brainard D H, 1997 "The Psychophysics Toolbox" *Spatial Vision* **10** 433 – 436

Cabrera D, Morimoto M, 2007 "Influence of fundamental frequency and source elevation on the vertical localization of complex tones and complex tone pairs" *Journal of the Acoustical Society of America* **122** 478 – 488

Chiou R, Stelter M, Rich A N, 2012 "Beyond colour perception: Auditory – visual synaesthesia induces experiences of geometric objects in specific locations" *Cortex* http://dx.doi.org/10.1016/j.cortex2012.04.006

Dodd M D, 2010 "Negative numbers eliminate, but do not reverse, the attentional SNARC effect" *Psychological Research* **75** 2 – 9

Doehrmann O, Naumer M J, 2008 "Semantics and the multisensory brain: how meaning modulates processes of audio-visual integration" *Brain Research* **1242** 136 – 150

Ernst M O, 2007 "Learning to integrate arbitrary signals from vision and touch" *Journal of Vision* **7**(5):7, 1 – 14

Evans K K, Treisman A, 2010 "Natural cross-modal mappings between visual and auditory features" *Journal of Vision* **10**(1):6, 1 – 12

Fernay L, Reby D, Ward J, 2011 "Visualized voices: A case study of audio-visual synesthesia" *Neurocase* **18** 50 – 56

Fischer M H, Castel A D, Dodd M D, Pratt J, 2003 "Perceiving numbers causes spatial shifts of attention" *Nature Neuroscience* **6** 555 – 556

Frens M A, Van Opstal A J, Van der Willigen R F, 1995 "Spatial and temporal factors determine auditory – visual interactions in human saccadic eye movements" *Perception & Psychophysics* **57** 802 – 816

Friesen C K, Ristic J, Kingstone A, 2004 "Attentional effects of counterpredictive gaze and arrow cues" *Journal of Experimental Psychology: Human Perception and Performance* **30** 319 – 329

Galfano G, Rusconi E, Umilta C, 2006 "Number magnitude orients attention, but not against one's will" *Psychonomic Bulletin Review* **13** 869 – 874

Gallace A, Spence C, 2006 "Multisensory synesthetic interactions in the speeded classification of visual size" *Perception & Psychophysics* **68** 1191 – 1203

Hommel B, 1996 "S-R compatibility effects without response uncertainty" *Quarterly Journal of Experimental Psychology A* **49** 546 – 571

Horowitz T S, Wolfe J M, Alvarez G A, Cohen M A, Kuzmova Y I, 2009 "The speed of free will" *Quarterly Journal of Experimental Psychology* **62** 2262 – 2288

Jones J A, Jarick M, 2006 "Multisensory integration of speech signals: the relationship between space and time" *Experimental Brain Research* **174** 588 – 594

Jonides J, 1981 "Voluntary versus automatic control over mind's eye's movement", in *Attention and Performance IX* Eds J B Long, A D Baddeley (Hillsdale, NJ: Lawrence Erlbaum Associates)

Keetels M, Vroomen J, 2011 "No effect of synesthetic congruency on temporal ventriloquism" *Attention, Perception, & Psychophysics* **73** 209 – 218

Kuhn G, Kingstone A, 2009 "Look away! Eyes and arrows engage oculomotor responses automatically" *Attention, Perception, & Psychophysics* **71** 314 – 327

Lidji P, Kolinsky R, Lochy A, Morais J, 2007 "Spatial associations for musical stimuli: a piano in the head?" *Journal of Experimental Psychology: Human Perception and Performance* **33** 1189 – 1207

Maeda F, Kanai R, Shimojo S, 2004 "Changing pitch induced visual motion illusion" *Current Biology* **14** R990 – R991

Marks L E, 1987 "On cross-modal similarity: auditory-visual interactions in speeded discrimination" *Journal of Experimental Psychology: Human Perception and Performance* **13** 384 – 394

Marks L E, Ben-Artzi E, Lakatos S, 2003 "Cross-modal interactions in auditory and visual discrimination" *International Journal of Psychophysiology* **50** 125 – 145

Melara R D, Mounts J R, 1994 "Contextual influences on interactive processing: effects of discriminability, quantity, and uncertainty" *Perception & Psychophysics* **56** 73 – 90
Parise C, Spence C, 2008 "Synesthetic congruency modulates the temporal ventriloquism effect" *Neuroscience Letters* **442** 257 – 261
Parise C V, Spence C, 2009 "'When birds of a feather flock together': synesthetic correspondences modulate audiovisual integration in non-synesthetes" *PLoS One* **4** e5664
Patching G R, Quinlan P T, 2002 "Garner and congruence effects in the speeded classification of bimodal signals" *Journal of Experimental Psychology: Human Perception and Performance* **28** 755 – 775
Pelli D G, 1997 "The VideoToolbox software for visual psychophysics: transforming numbers into movies" *Spatial Vision* **10** 437 – 442
Posner M I, 1980 "Orienting of attention" *Quarterly Journal of Experimental Psychology* **32** 3 – 25
Proctor R W, Cho Y S, 2006 "Polarity correspondence: A general principle for performance of speeded binary classification tasks" *Psychological Bulletin* **132** 416 – 442
Ristic J, Wright A, Kingstone A, 2006 "The number line effect reflects top – down control" *Psychonomic Bulletin & Review* **13** 862 – 868
Roffler S K, Butler R A, 1968 "Localization of tonal stimuli in the vertical plane" *Journal of the Acoustical Society of America* **43** 1260 – 1266
Rusconi E, Kwan B, Giordano B L, Umilta C, Butterworth B, 2006 "Spatial representation of pitch height: the SMARC effect" *Cognition* **99** 113 – 129
Schneider B, Wright A A, Edelheit W, Hock P, Humphrey C, 1972 "Equal loudness contours derived from sensory magnitude judgments" *Journal of the Acoustical Society of America* **51** 1951 – 1959
Sek A, Moore B C, 1995 "Frequency discrimination as a function of frequency, measured in several ways" *Journal of the Acoustical Society of America* **97** 2479 – 2486
Spence C, 2011 "Crossmodal correspondences: a tutorial review" *Attention, Perception, & Psychophysics* **73** 971 – 995
Sternberg S, 1969 "The discovery of processing stages: Extensions of Donders' method" *Attention and Performance II. Acta Psychologica* **30** 276 – 315
Walker P, Bremner J G, Mason U, Spring J, Mattock K, Slater A, Johnson S P, 2010 "Preverbal infants' sensitivity to synaesthetic cross-modality correspondences" *Psychological Science* **21** 21 – 25
Ward J, Huckstep B, Tsakanikos E, 2006 "Sound – colour synaesthesia: to what extent does it use cross-modal mechanisms common to us all?" *Cortex* **42** 264 – 280
Wright R D, Richard C M, 2000 "Location cue validity affects inhibition of return of visual processing" *Vision Research* **40** 2351 – 2358

© 2012 a Pion publication

ISSN 0301-0066 (print) ISSN 1468-4233 (electronic)

PERCEPTION
VOLUME 41 2012

www.perceptionweb.com

Conditions of use. This article may be downloaded from the Perception website for personal research by members of subscribing organisations. Authors are entitled to distribute their own article (in printed form or by e-mail) to up to 50 people. This PDF may not be placed on any website (or other online distribution system) without permission of the publisher.