Cognition 196 (2020) 104119

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

The Role of Selective Attention in Cross-modal Interactions between Auditory and Visual Features

Karla K. Evans

University of York, Department of Psychology, Heslington, York, YO10 5DD, United Kingdom

ARTICLE INFO

Keywords:
Cross-modal integration
Attention
Vision
Audition

ABSTRACT

Evans and Treisman (2010) showed systematic interactions between audition and vision when participants made speeded classifications in one modality while supposedly ignoring another. We found perceptual facilitation between high pitch and high visual position, high spatial frequency and small size, and interference between high pitch and low position, low spatial frequency and large size, while the converse was the case between low pitch and the same visual features. The present study examined the role of selective attention in these cross-modal interactions. Participants performed speeded classification or search tasks of low or high load while attempting to ignore irrelevant stimuli in a different modality. In both paradigms, congruency between the visual and the irrelevant auditory stimulus had an equal effect in the low and in the high perceptual load conditions. A third experiment tested divided attention, requiring participants to compare stimuli across modalities and respond to the visual-auditory compound. The congruency effect was as large with attention focused on one modality as when it was divided across both. These findings offer converging evidence that cross-modal interactions between corresponding basic features are independent of selective attention.

1. Introduction

On hearing a high-pitched sound, we tend to look upwards and to visualize a small rather than a large object. On seeing a small object, we expect a high-pitched sound. We may be surprised when two features in different modalities seem incongruous, for example when a small person speaks in a deep voice. The existence of cross-modal relationships between auditory (pitch and loudness) and visual (vertical position, size, brightness, shape) features have been investigated using primarily psychophysical methods such as matching and speeded classification tasks (Marks, 2004; Spence, 2011). These links seem involuntary and cross-cultural; some have even been shown by very small children (Dolscheid, Hunnius, Casasanto, & Majid, 2014; Marks, Hammeal, & Bornstein, 1987; Mondloch & Maurer, 2004; Walker et al., 2009) and non-human primates (Ludwig, Adachi, & Matsuzawa, 2011). An interaction between different cross-modal correspondences does not necessarily mean that the signals from two modalities are integrated into one percept. Yet interactions can be strong enough that it is hard to dissociate them in any clear terms from an integrated percept. For example, an auditory stimulus associated with high reward can trigger a cross-modal interaction increasing visual sensitivity that even correlates with changes in stimulus representation the visual cortex (Pooresmaeili et al., 2014). Therefore the robust empirical evidence that supports the view that cross-modal interactions of above mentioned features produce an integrated percept of component signals (for review see Marks, 2004; Spence, 2011; Spence & Deroy, 2013) often leads to these two terms being used interchangeably in literature for afore mentioned cross-modal correspondences.

However, there has been relatively little study of the conditions under which these cross-modal correspondences occur. In particular, the role of attention has not been much studied. Are they truly automatic, as suggested by the fact that the interactions are sometimes involuntary and unconscious?

Attention is not one unitary system but rather a series of mechanisms or rather important properties of multiple perceptual and cognitive operations allowing for control of information processing. Selection is one of those core properties by means of which the cognitive system chooses the information relevant for current behavior. Focused spatial attention as one of the selective attention mechanisms has been shown to be needed for many perceptual tasks, one of which is feature-binding to form objects within a modality (Treisman, 1998; Treisman & Gelade, 1980). But is the selective attention mechanisms necessary for integration across modalities? Does increasing the selective attentional demand by increasing perceptual load or by dividing attention between concurrent tasks change the outcome of cross-modal interaction? It may be misleading to equate the mechanism of spatially focused attention

E-mail address: Karla.evans@york.ac.uk.

https://doi.org/10.1016/j.cognition.2019.104119
Received 2 March 2016; Received in revised form 2 October 2019; Accepted 25 October 2019

(used in binding features according to feature integration theory) with that of other attentional mechanisms. But it is of interest anyway to explore the effects of load in competing tasks with those of spatial focus in vision.

The role of attention in cross-modal integration in general is still not well understood and remains the object of fierce debate. Some studies investigating the interaction between attention and the integration of multimodal sensory inputs have shown clear effects of attention while others have argued that attention has no effect (for review see Talsma, Senkowski, Soto-Faraco, & Woldorff, 2010; Ten Oever et al., 2016). The lack of or need for attention in cross-modal integration is often associated with either calling the process of integration as automatic or not automatic. Here too there is no clear consensus. Some studies have argued for non-automatic nature of cross-modal correspondences (Chiou & Rich, 2012; Klapetek, Ngo, & Spence, 2012) while others contend that there is evidence suggesting automaticity (Evans & Treisman, 2010; Parise & Spence, 2009, 2012; Peiffer-Smadja, 2010). Spence & Deroy in their 2013 review on cross-modal correspondences point out that there are different criteria by which one might define an automatic process and that it would be better to consider automaticity in cross-modal integration as a general term referring to a number of distinct features. They cite Moors and De Houwer (2006) four critical features by which automaticity might be judged: the goal-independence, the non-conscious, the load-insensitivity and the speed criteria.

In previous studies Evans and Treisman (2010) compared performance when a tone and the visual position of a simultaneously presented grating were congruent (high pitch with high visual position and low pitch with low visual position) and when they were incongruent (high pitch with low visual position or low pitch with high visual position). We found that the congruence was detected unconsciously and unintentionally, since it affected reaction times even when the task was to discriminate the stimuli on an orthogonal dimension (the left or right orientation of the visual grating and whether the tone was played on a piano or on a violin) thus automatically. However, it is possible that the interaction occurred because the task did not take all the available attentional resources and there were enough remaining to process the irrelevant dimensions as well. This brings in the question of automaticity criteria of load-insensitivity. According to the Load Theory of Attention and Cognitive Control (Lavie & Dalton, 2014), when the amount of resources required to perform a cognitive operation does not exceed the capacity of the system, no attentional effects on perception should be observed.

I attempt to review the earlier studies and our findings in the General Discussion to see whether any generalizations about the causes and conditions are possible. First, I will describe three experiments suggesting that selective attention has no effect on cross-modal integration, at least for simple auditory and visual stimuli for which there are correspondent matching relations. The aim of these experiments was to see whether the cross-modal congruency effect we observed in our earlier experiments depends on the selective attention if we increase demands on selective attention by either increasing perceptual load or the need to divide attentional focus in the concurrent primary task. More specifically I examine if the increase in the perceptual load would deplete attentional resource. I varied the demands on selective attention in three different ways: by varying the perceptual discrimination difficulty (i.e. perceptual load) of the attended modality in the speeded classification paradigm, by varying the search difficulty in a visual search task and by increasing demand going from a single to a dual task requirement. Would the cross-modal congruency effect between the relevant and irrelevant stimuli disappear or would it persist under a higher perceptual load or divided attention and thus bigger attentional demands. For a process to be automatic it must satisfy, among other things, the load–insensitivity criterion (Jonides & Irwin, 1981), which states that automatic processes are insensitive to the demands of a concurrent task. If attention is not needed for cross-modal interaction between correspondent features to occur, then we should observe no significant difference in the magnitude of the congruency effect between low and high load conditions.

2. Modulating Perceptual Load

2.1. Experiment 1

In Experiment 1 I used the same speeded classification paradigm as in our earlier experiment, testing the effect of congruence relations between pitch of sound and visual position in the vertical plane. This was an indirect effect since the primary task was to discriminate the stimuli on an orthogonal dimension other than the dimensions whose congruence effect was being tested. In the primary task, participants were asked to discriminate between the sounds of two instruments or between two grating orientations, while the stimuli were still high or low in pitch for the instruments and high or low in visual position for the oriented gratings. In order to vary the perceptual load of the primary tasks, I compared an easy to a difficult discrimination on each dimension. The aim of the experiment was to test whether increasing the difficulty of the discrimination in the task-relevant modality would differentially affect the congruency effect of the stimuli in the irrelevant modality. - I hypothesized that since noise is not intorduced into the system we most likely will not observe inverse effectiveness principle (Stein & Meredith, 1993) with more cross-modal interaction in the task irrelevant dimensions under high perceptual load. But rather with higher perceptual load in the primary task, there will be less resources to select irrelevant dimension of pitch and position of the stimuli (i.e. distractors) for processing and potentially a diminished cross-modal interaction of those correspondent features. Given that I am not manipulating working memory or task coordination in the two conditions the task difficulty increase does not manipulate cognitive control load but just perceptual load.

2.2. Method

2.2.1. Participants

Nineteen Princeton University undergraduates (9 males, 11 female) participated in the experiment after giving informed consent, as one option to fulfill a course requirement. The sample size was based on a power analysis of previous published work (Evans & Treisman, 2010) investigating the same congruency effect, yielding an estimated effect size of $\eta_p^2 = .57$. An effect size of .57, $\alpha = .05$ and power = 0.95 returns a minimum sample size of 10. The chosen sample size of 20 has a power of 0.99 with $\eta_p^2 = .57$, $\alpha = .05$. All had normal or corrected-to-normal vision and hearing. Every aspect of this study was carried out in accordance with the regulations of the Princeton University's Institutional Review Panel for Human Subjects.

2.2.2. Apparatus and Stimuli

The stimuli were presented using MATLAB (Mathworks Inc., Natick, MA) with Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997) running on a Macintosh G3. The visual displays were presented on a 17" Apple screen at a viewing distance of 57 cm and monitor refresh rate of 75 Hz. The sound wave files generated by GarageBand were played through speakers positioned to the left and right of the computer screen and with center-to-center distance of 20 cm between the speakers.

The feature discriminations, which participants performed in separate experiments in both the auditory and visual modalities, (instrument for the tones and orientation for the visual gratings) were orthogonal to the features whose correspondence was varied (i.e. pitch of tones and vertical position of the visual stimuli).

In a pilot study, we identified two pairs of visual and two pairs of auditory stimuli for which one pair was easier to discriminate than the other. We then matched the discriminability of the low and high

discriminability pairs across the modalities, using both accuracy and a response latency measure in a speeded classification task. I tested twelve pairs of sounds in a speeded classification task and chose two pairs that matched the performance of the same participants on an easy and a hard visual discrimination task. Error rates averaged 9% for the hard tasks and 5% for the easy tasks in both auditory and visual discriminations. Response times averaged 553 ms (542 ms for auditory task and 564 for visual task) for the hard tasks and 509 ms (530 ms for auditory task and 482 ms for visual task) for the easy tasks.

The visual stimuli were black and white sinusoidal gratings (luminance of 8 for black and 320 candels/m^2 for white) presented on a gray background (80 cd/m2) and subtending 2.5 degrees visual angle. The gratings were presented 4.5° either above or below a central fixation cross. The feature of the gratings that was to be discriminated was their orientation, with discrimination between 45 and 75 degrees to the right in the difficult, high load condition and 45 degrees to the left and 45 degrees to the right in the easy, low load condition. The auditory stimuli were tones of two different pitches (C3 and C4) presented at an intensity of 75 dB (A) and played by different instruments. In the difficult, high load condition, participants discriminated between a piano and an electric piano, and in the easy, low load condition they discriminated between a piano and a harpsichord.

2.2.3. Experimental Design and Procedure

The task was auditory in half of the blocks and visual in the other half. The auditory task required the classification of the instrument that produced the tone and the visual task required the classification of the orientation of the grating. In half of the blocks the relevant discrimination was hard (high load) and in the other half it was easy (low load). The stimulus presentation was bimodal in all trials. (Fig. 1) When they were task irrelevant, the stimulus values were fixed (e.g. only the piano tone of two different pitches was presented when participants did the visual task).

Each trial started with a fixation cross for 500 ms, followed by a simultaneous visual and auditory stimulus presentation for 120 ms. A black and white grating appeared in either the upper or lower part of the display and at the same time a tone, either high or low in pitch, was played from the speakers. The task of the participants and the difficulty of discrimination varied depending on the block of trials they were doing. During visual task blocks they pressed as fast as possible one key with the left hand when they detected a left oriented (easy blocks) or sharply tilted grating (difficult blocks) and another key with the right hand when they detected a right oriented (easy blocks) or shallow tilted grating (difficult blocks). During auditory task blocks they pressed one key using the left hand when they detected a piano tone and another with the right hand when they detected a harpsichord (easy blocks) or electric piano tone (difficult blocks). I changed these response hand and key mapping in a counterbalanced way across participants. Both response keys and hands were used equally often in the congruent and in the incongruent conditions, thus the cross-modal effects that were observed were independent of the response key assignment. Each participant completed two short practice blocks (one visual and one auditory) and eight experimental blocks (4 visual and 4 auditory classification tasks). There were 160 trials per condition (in each of 8 conditions: congruence (2) x perceptual load (2) x relevant modality (2)) in the experiment and 40 trials in the practice block. The order of trials within each block was randomly selected under the constraint that each condition was presented equally often.

The possible combinations of a high or low pitch tone with a grating in the upper or lower half of the display were presented equally often. Both congruency between auditory and visual stimuli and difficulty of task were varied within participants. Accuracy and reaction time in classifying either the visual or auditory stimuli were used as the dependent measures.

2.3. Results

The goal of the study was to test if the congruency effect observed in a speeded classification paradigm would significantly change when the perceptual load was increased. If there was a change it would suggest that attention is needed and modulates integration of the correspondent cross-modal features. A three-way repeated measures ANOVA (load x task x type of pairing) was calculated separately for percent correct and reaction times. In the present and in the following experiments, trials with incorrect responses or on which response times were shorter than 150 ms or greater than 3 standard deviations from the mean were treated as outliers and removed prior to the analysis (225 trials across all participants or 1% of all correct trials). As predicted from the pilot study, participants made more errors (14% vs. 9%; F (1, 18) = 23.37, $p < 0.00013$, $\eta_p^2 = .57$) and were slower (573 vs. 519; F (1, 18) = 60.27, $p < 0.00001$, $\eta_p^2 = .77$) in the high load condition than in the low load condition, indicating that the high load condition was more demanding. There was no difference, however, due to task modality, neither in accuracy (87% vs. 90%; F (1, 18) = 2.37,

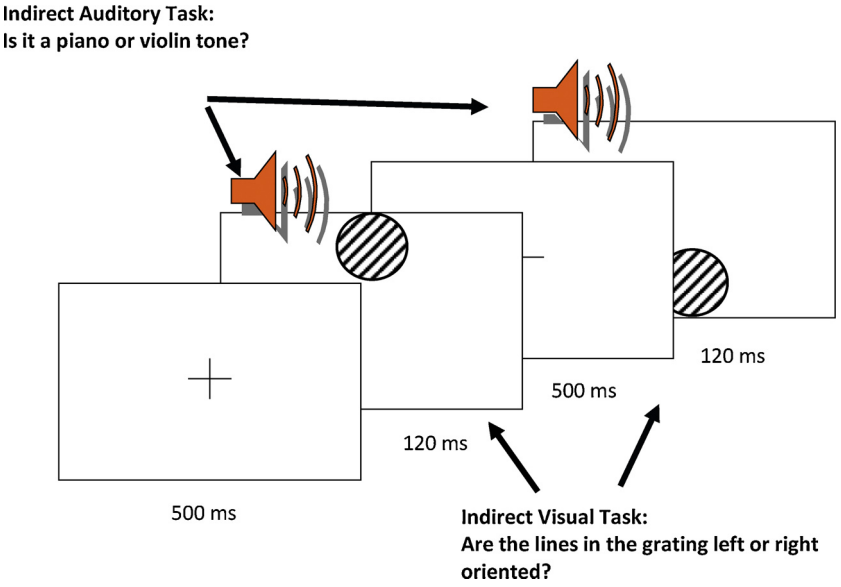

Fig. 1. Example time course of the speeded classification paradigm in Experiment 1.

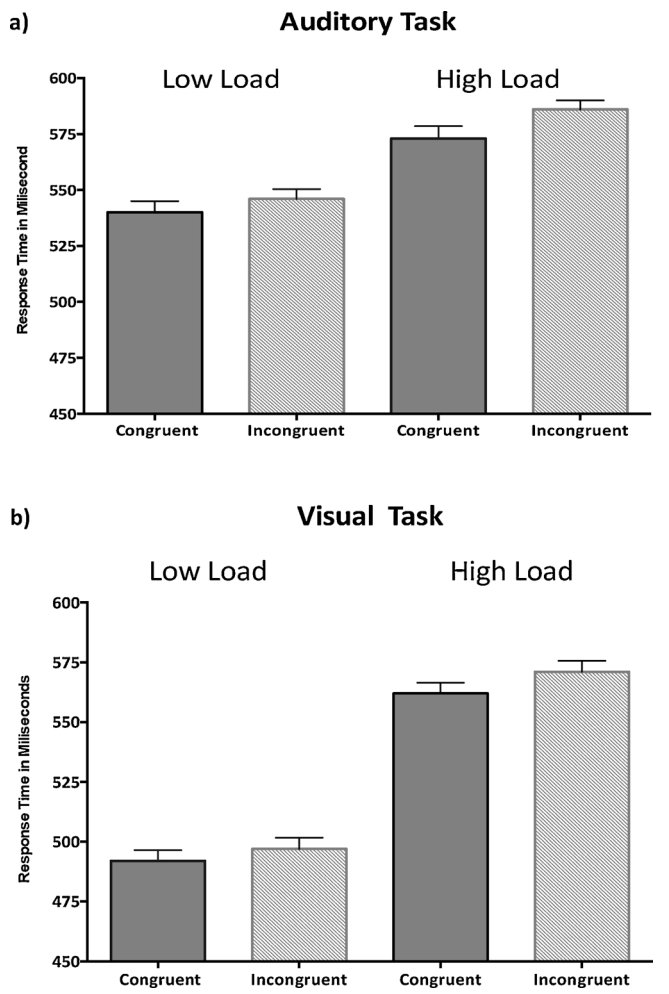

Fig. 2. Mean reaction times during **a)** auditory speeded classification (Low Load: Congruent 540 ms, s.e.m. 5 ms vs. Incongruent 546 ms, s.e.m. 4.4 ms; High Load: Congruent 573 ms, s.e.m. 4.5 ms vs. Incongruent 586 ms, s.e.m. 4.1 ms) and **b)** visual speeded classification task for two loads (low and high) (Low Load: Congruent 492 ms, s.e.m. 4.5 ms vs. Incongruent 497 ms, s.e.m. 4.6 ms; High Load: Congruent 562 ms, s.e.m. 4.5 ms vs. Incongruent 571 ms, s.e.m. 4.6 ms).

$p < 0.095$, $\eta_p^2 = .152$), nor in response time (531 vs. 561; F (1, 18) = 3.36, $p < 0.083$, $\eta_p^2 = .157$). However, unlike the results of the pilot study, load and task modality interacted and the low load visual task (495 ms, s.e.m. 16 ms) had significantly faster response times (F (1,18) = 9.17, $p < 0.007$, $\eta_p^2 = .334$) than the low load auditory task (543 ms, s.e.m. 19 ms), but this did not interact with congruency in any way ($F < 1$, $\eta_p^2 = .031$). There was no significant difference in accuracy between congruent and incongruent pairings (89% both; F (1, 18) < 1, $\eta_p^2 = .039$) but the response times were significantly faster for congruent than incongruent pairings (542 vs. 550; F (1, 18) = 51.60, $p < 0.000001$, $\eta_p^2 = .75$) replicating my previous studies (Evans & Treisman, 2010). The critical question concerned the two-way interaction between load and type of pairing in the reaction time data. The interaction between load (high vs. low) and type of pairing (congruent vs. incongruent) was not significant (F (1, 18) = 1.99, $p < 0.175$, $\eta_p^2 = .102$). The congruency effect between the auditory and visual stimuli was not significantly different in the low and in the high load conditions, signifying that the effects of congruence and load were additive (Fig. 2). This means that cross-modal interaction between pitch and visual position in the vertical plane happened automatically, unaffected by the load. When pitch and position of the visual stimulus were congruent, processing of both the visual and the auditory stimulus was facilitated, independent of the attentional demands of the concurrent primary task.

One problem is that there is some interaction, which does not reach significance. It is hard to prove a null effect because noise is going to reduce the significance. But what we can say is that the difference is in the opposite direction from that predicted if load reduced the effect of congruence. If anything, the congruence effect is slightly larger with high than with low load.

Another way one might analyze the data that would account for the individual mean differences is to normalize RT's of high and low load by calculating the congruency benefit score for different tasks. The congruency benefit score was computed extracting the difference between incongruent and congruent condition trials and then dividing the difference by the individual means RTs on incongruent trials (RT(incongruent) − RT(congruent)]/ RT(incongruent) (see Störmer, Eppinger, & Li, 2014). When these scores were entered into a two-way (task x load) repeated measure ANOVA there still was no significant change in the congruency effect ($F(1, 18) = 2.99$, $p = .101$, $\eta_p^2 = .143$) and no interaction with task ($F(1, 18) = .194$, $p = .665$, $\eta_p^2 = .011$).

There is no significant interaction between load and congruency in either auditory or visual task conditions.

2.4. Experiment 2

Load can be varied in a number of different ways. To look for converging evidence with the results of the previous experiment, I manipulated perceptual load using a visual search paradigm in Experiment 2. I asked whether congruence between the pitch of an irrelevant sound and the position of a search target in a visual display could increase the likelihood and speed of detecting the target, and if so, whether the difficulty of the search task due to increase in perceptual discriminability would affect the benefit of congruence.

2.5. Method

2.5.1. Participants

Twenty Princeton University undergraduates (6 female, 14 male) participated in the experiment after giving informed consent, as one option to fulfill a course requirement. The sample size was based on a power analysis of previous published work (Evans & Treisman, 2010) investigating the same congruency effect, yielding an estimated effect size of $\eta_p^2 = .57$. An effect size of .57, $\alpha = .05$ and power = 0.95 returns a minimum sample size of 10. The chosen sample size of 20 has a power of 0.99 with $\eta_p^2 = .57$, $\alpha = .05$. All had normal or corrected-to-normal vision and hearing. Every aspect of this study was carried out in accordance with the regulations of the Princeton University's Institutional Review Panel for Human Subjects.

2.5.2. Apparatus and Stimuli

The same apparatus and settings were used as in experiment one. The visual stimulus consisted of an array of 5 letters chosen from the English alphabet presented centrally in a column subtending 20 degrees visual angle from top to bottom of the display. The letters were black on a white background, clearly visible, subtending 3 degrees of visual angle each. The sounds were played through speakers positioned left and right of the screen. They had the same onset and duration as the visual displays. The sound was one of two piano tones, a high pitch G2 and a low pitch C2 tone presented at an intensity of 75 dB (A).

2.5.3. Experimental Design and Procedure

The task was to identify which of two visual targets was present (either a letter X or a letter S). A target appeared in every trial with equal probability in any of the four positions except the central one. Each trial started with a fixation cross for 500 ms, followed by a target visual array and tone for 1000 ms. A display consisting of five letters, all of the same size, appeared in a column at the center of the display. The

task of the participants was to press as fast as possible one key when they detected the letter X and another when they detected the letter S with the same hand. The response key mapping was changed in a counterbalanced way across participants. Both response keys were used equally often in the congruent and in the incongruent conditions, thus the cross-modal effects that we observed were independent of the response key assignment. Each participant completed four experimental blocks, and one short practice block composed of 20 trials. There were 120 trials for each of 4 conditions (congruence (2) x perceptual load (2)) in a given experimental block (480 trials per block). The order of trials within each block was randomly selected under the constraint that each condition was tested equally often.

The independent variables were the congruence (between the pitch of a sound presented simultaneously with the visual array and the location of the target) and the perceptual load (i.e. difficulty in discrimination between target and distractors) determined by the variety of letters in the array. The possible combinations of a high or low pitch tone with a target in the upper or lower half of the display were presented equally often. The pairing was congruent when the high pitch was paired with targets appearing in the upper two visual positions, and incongruent when paired with the lower two visual positions. A visual array containing four different distractor letters was a difficult task condition and an array containing four identical distractor letters constituted an easy task condition. The distractor letters were selected randomly from the alphabet, omitting the two target letters, X and S. Both factors were varied within participants. (Fig. 3) The dependent variable was the speed of response on correct target detection in different conditions.

2.6. Results

The purpose of Experiment 2 was to investigate whether the congruence of pitch in a task-irrelevant modality could influence a search for a target in the task-relevant modality. If the pitch of the tone is automatically integrated with the visual dimension of vertical location then the sound may increase sensitivity to the target in the congruent positions, increasing the accuracy and reducing the latency of the response.

Performance in both load conditions was well over chance with significantly better accuracy in the low load condition 97% (s.e.m. 1%) in comparison to the high load 95% (s.e.m. 1%) (F (1, 19) = 20.74, $p < 0.002$, $\eta_p^2 = .517$). There was no significant difference in accuracy between the congruent and incongruent pairings ($F < 1$, $\eta_p^2 = .009$) with both averaging 96% correct (s.e.m. 1%) and no interaction with the difficulty of task ($F < 1$, $\eta_p^2 = .001$).

The two-way repeated measures ANOVA (task difficulty x congruence of pairing) performed on reaction times showed a main effect

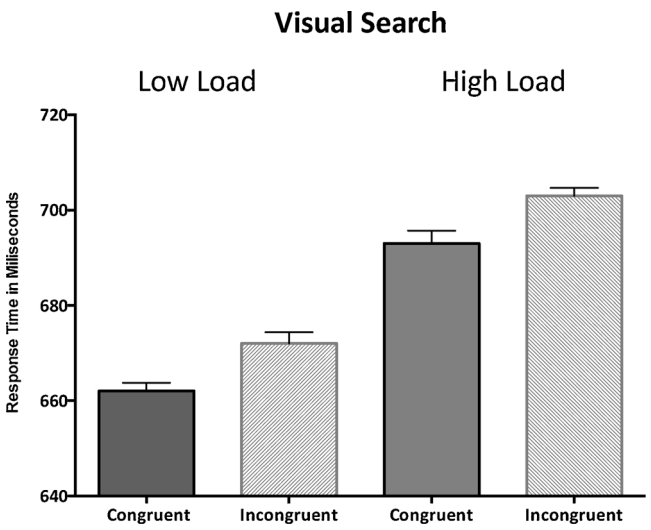

Fig. 4. Mean reaction times during correct target identification in low and high load visual search trials (Low Load: Congruent 662 ms, s.e.m. 1.7 ms vs. Incongruent 672 ms, s.e.m. 2.3 ms; High Load: Congruent 693 ms, s.e.m. 2.7 ms vs. Incongruent 703 ms, s.e.m. 1.7 ms). The response times showed a significant effect of congruency, a significant effect of load and no significant interaction between load and congruence.

of task difficulty (F (1, 19) = 70.94, $p < 0.000001$, $\eta_p^2 = .785$) as well as a main effect of congruence (F (1, 19) = 30.35, $p < 0.00001$, $\eta_p^2 = .611$) but no significant interaction between the two ($F < 1$, $\eta_p^2 = .001$) (Fig. 4). When converting the data to normalized RT's by calculating the congruency effect score using the method outlined in Exp. 1, here too there is no significant difference between the congruency effect in two task difficulties (t(19) = .359, p = .723). During the high load condition the participant's average reaction time was 698 ms (s.e.m. 2 ms) and in the low load condition 667 ms (s.e.m. 2 ms). Response time with the congruent pairings was 677 ms (s.e.m. 2 ms) whereas with the incongruent pairings it was 687 ms (s.e.m. 2 ms). The reaction time data reveal how the pitch of a behaviorally irrelevant stimulus, affects the efficiency of visual search, with high pitch facilitating search when the visual targets were in the upper two positions in the visual display and low pitch facilitating search for targets in the lower part of visual display. There was no difference between the extreme and intermediate target positions in how fast they were searched when a congruent or incongruent pitch was presented. These findings further support the idea that the association between pitch and position is automatic and perceptual. One possibility was that when the perceptual load is high, (e.g. in our experiment when the visual search was

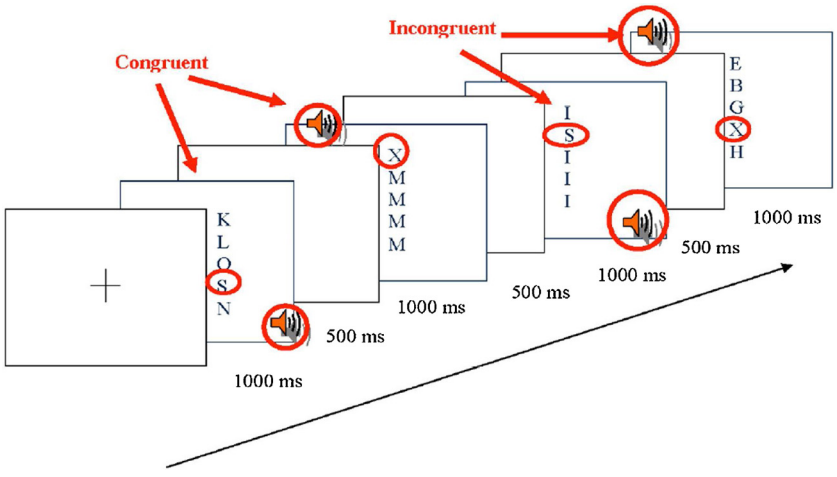

Fig. 3. Example time course of the visual search paradigm in Experiment 2. Visual displays composed of a column of letters (high load - distractor letters all different letters; low load – distractor letters all the same). The task was to respond as fast as possible whether there was a target letter X or a target letter S. Target letters were equally likely to appear in any of the four positions above or below the center, omitting the central position. At the same time as the letter display, a tone was played, which was either high or low in pitch.

difficult), then the saliency of the task-irrelevant auditory stimulus would be reduced and in turn its interaction with the visual stimulus would be diminished (e.g. no or reduced congruency effect). That did not happen suggesting that the interaction of pitch and visual position is independent of attention.

2.7. Discussion

In both Experiments 1 and 2 the congruent pairings of pitch of sound and vertical position of the visual stimulus speeded responses in classification and search for targets relative to the incongruent pairings. However, neither showed any interaction with the increased attentional demand due to perceptual load, suggesting that the cross-modal interaction between pitch and position is the result of a fully automatic process.

Experiment 1 replicated the basic results obtained in previous experiment (Evans & Treisman, 2010), showing that the speed of stimulus classification in one modality is dependent on the association of the features of that stimulus with the feature of a behaviorally irrelevant stimulus in another modality simultaneously presented. Experiment 2 presented further evidence of the cross-modal interaction using the visual search paradigm. The goal of the two experiments was to test if the observed cross-modal interaction is subject to attentional demands. The results provide converging evidence that raising the attentional demand with increasing perceptual load does not significantly alter the congruency effect. Neither Experiment 1 nor Experiment 2 showed any evidence that attention plays a crucial role in the cross-modal interaction between pitch and vertical position. There are similar studies that have shown location non-informative auditory cues guided attention and eye–movements toward congruous visual stimuli (Iordanescu, Guzman-Martinez, Grabowecky, & Suzuki, 2008; Iordanescu, Grabowecky, Franconeri, Theeuwes, & Suzuki, 2010; Mossbridge, Grabowecky, & Suzuki, 2011) however not under any load. For example, Mossbridge et al. (2011) report that the direction of sound frequency change guided visual-spatial attention resulting in more accurate and faster matching between a central probe color and the peripheral colored circle that was congruent to the frequency sweep (e.g. ascending sweep and top display position).

3. Dividing attention between modalities

So far, experiments on perceptual load effects have used selective attention tasks. In the final experiment, I tested the effect of demand on selective attention in a task requiring divided attention to the two modalities. Would an increased demand on attention affect the congruency effect? Several studies that investigated divided attention between two independent visual and auditory inputs have shown that attending to one modality had no effect on concurrent identification of a second target in another modality (Duncan, Martens, & Ward, 1997; Hein, Parr, & Duncan, 2006; Massaro & Warner, 1977; Potter, Chun, Banks, & Muckenhoupt, 1998; Soto-Faraco & Spence, 2002; Taylor, Lindsay, & Forbes, 1967; Treisman & Davies, 1973). However, opposite results have been found when the demand is increased (Arnell & Duncan, 2002). These results suggest that the major source of attentional restriction lies in modality-specific sensory systems and not between modalities. If this were the case, it would imply that dividing attention across modalities need not disrupt cross-modal integration.

In view of these conflicting findings, the question that I hoped to address in Experiment 3 is whether dividing attention between the modalities would affect the interaction between basic cross-modal features that share a correspondent relationship, such as pitch and vertical position. Based on my previous findings that show congruency effects even when observers are selectively attending to one modality, I predicted that we would not see a change in the congruency effect even when attention is depleted by performing two concurrent tasks.

3.1. Experiment 3

The aim of Experiment 3 was to test whether different demands on attention modulates the magnitude of the congruency effect in a dual–task paradigm. On the one hand, it could reduce the congruency effect if attention is needed to integrate stimuli in different modalities. On the other hand, it could actually increase the congruency effect by ensuring that both modalities are attended and thus increasing the salience of their congruency.

3.2. Method

3.2.1. Participants

Twenty-one Princeton University undergraduates (5 male, 16 female) participated in the experiment after giving informed consent, as one option to fulfill a course requirement. The sample size was based on a power analysis of previous published work (Evans & Treisman, 2010) investigating the same congruency effect, yielding an estimated effect size of $\eta_p^2 = .57$. An effect size of .57, $\alpha = .05$ and power = 0.95 returns a minimum sample size of 10. The chosen sample size of 19 has a power of 0.99 with $\eta_p^2 = .57$, $\alpha = .05$. All had normal or corrected-to-normal vision and hearing. Every aspect of this study was carried out in accordance with the regulations of the Princeton University's Institutional Review Panel for Human Subjects.

3.2.2. Apparatus and Stimuli

The same apparatus, setting and stimuli were used as in experiment one.

3.2.3. Experimental Design and Procedure

The task was to identify either auditory only, visual only or both visual and auditory stimuli, depending on the block. The auditory task was to classify the instrument that produced the tone (piano or violin) and the visual task was to classify the orientation of the grating (left or right). The dual task required participants to detect the presence or absence of a target in either modality. The two possible targets were a left oriented grating (right oriented grating for half of the participants) and a piano tone (violin tone for half of the participants). During the dual task blocks there were no instances of a left oriented grating accompanied by a piano tone (or a right oriented grating by a violin tone) since then the participants would have to give two responses. The stimulus presentation was bimodal in all trials.

Each trial started with a fixation cross for 500 ms, followed by a simultaneous visual and auditory stimulus for 120 ms. A display consisting of a black and white grating appeared either in the upper or lower part of the display and at the same time a tone either high or low in pitch was played from the speakers. The task of the participants varied depending on the block of trials they were doing. During visual task blocks they pressed as fast as possible one key when they detected a left oriented grating and another when they detected a right oriented grating. During auditory task blocks their task was to press one key when they detected a piano tone and another when they detected a violin tone. During the dual task they pressed one key when the target they detected was a left oriented grating (right oriented grating for half of participants) or when the target was a piano tone (violin tone for half of the participants) with left hand and another with the right hand when neither of the two targets was present. The response hand and key mapping were changed in all the conditions in a counterbalanced way across participants. Both response keys were used equally often in the congruent and in the incongruent conditions, thus the cross-modal effects that we observed were independent of the response key assignment. Each participant did three experimental blocks, and one short practice block. There were 120 trials per condition (congruence (2) x task (3)) in the experimental block and 40 trials in the practice block. The order of trials within each block was randomly selected under the constraint that each condition was tested equally often.

The independent variables were the congruence of the pitch of the sound and the vertical position of the grating and the task difficulty. The possible combinations of a high or low pitch tone with a grating in the upper or lower half of the display were presented equally often. The second factor was the single or dual modality task. The discrimination between two stimuli in one modality (visual or auditory) was considered a single modality condition and the discrimination between the presence or absence of either a visual or an auditory target constituted the dual modality condition. The dependent variable was the speed of response on correct discriminations in the different conditions. I compared the difference between congruent and incongruent audio-visual pairs both in the single modality tasks and in the dual modality tasks.

3.3. Results

Accuracy in all three task conditions was well over chance with significantly better accuracy in the single task conditions (visual 93%, s.e.m. 2% and auditory 90%, s.e.m. 1%) than in the dual task condition 80% (s.e.m. 2%), (F (1, 20) = 51.79, $p < 0.000001$, $\eta_p^2 = .729$; F (1, 20) = 31.36, $p < 0.00001$, $\eta_p^2 = .612$). There was no significant difference in accuracy between the congruent and incongruent pairings ($F < 1$, $\eta_p^2 = .053$) with both averaging 87% correct (s.e.m. 2%) and no interaction with the task type.

Fig. 5**a** shows the RTs in the Dual task and Single task conditions, and Fig. 5**b** shows the RTs in the single task conditions separately for the visual and auditory task. The two-way repeated measures ANOVA (Dual or 2 single tasks x type of pairing) performed on reaction times showed a main effect of task type (F (1, 20) = 60.71, $p < 0.000001$, $\eta_p^2 = .752$) and a main effect of congruence (F (1, 20) = 23.77, $p < 0.000001$, $\eta_p^2 = .541$), but no significant interaction between the two ($F < 1$, $\eta_p^2 = .016$). (Fig. 5). The two single tasks (visual 513 ms, s.e.m. 11 ms and auditory 590 ms, s.e.m. 10 ms) both show significantly faster response times in comparison to the dual task 706 ms (s.e.m. 10 ms) (F (1, 20) = 105.77, $p < 0.000001$, $\eta_p^2 = .841$; F (1, 20) = 50.05, $p < 0.000001$, $\eta_p^2 = .715$). The visual discrimination was significantly faster than the auditory one (F (1, 20) = 19.08, $p < 0.00029$, $\eta_p^2 = .488$) (Fig. 5**b**). The critical interaction between task (unimodal auditory, unimodal visual and dual task) and congruency was not significant, suggesting that the interaction between pitch and position is not modulated by attentional demand.

4. Discussion

This third experiment in the series of studies on the role of selective attention tested the effect of the correspondence between pitch and spatial position under conditions of divided attention. Earlier studies using a dual-task paradigm in cross-modal processing of sensory properties suggest that the major source of attentional limits lies within modality-specific sensory systems and not between modalities. This suggests that dividing attention across modalities should not disrupt cross-modal integration. But the question was whether this would still hold when there is a correspondent relationship between the two modalities. Alsius and colleagues (Alsius, Navarra, Campbell, & Soto-Faraco, 2005) have shown by measuring the participant's susceptibility to the McGurk illusion (McGurk & MacDonald, 1976) under dual-task conditions that audiovisual speech integration is modulated by the amount of available attentional resources. This task involves the processing of verbal material rather than simple sensory properties. The results of Experiment 3 suggest, however, that that is not the case in cross-modal interaction between basic audiovisual correspondent features. When dividing their attention between two modalities while engaged in a dual task, participants performed overall less well than during the single task but the behavioral gain when congruent pairings of features were present did not differ significantly between the two tasks. The increased demand on attention here probably arises in the identification and response to the targets and not at the sensory integration level.

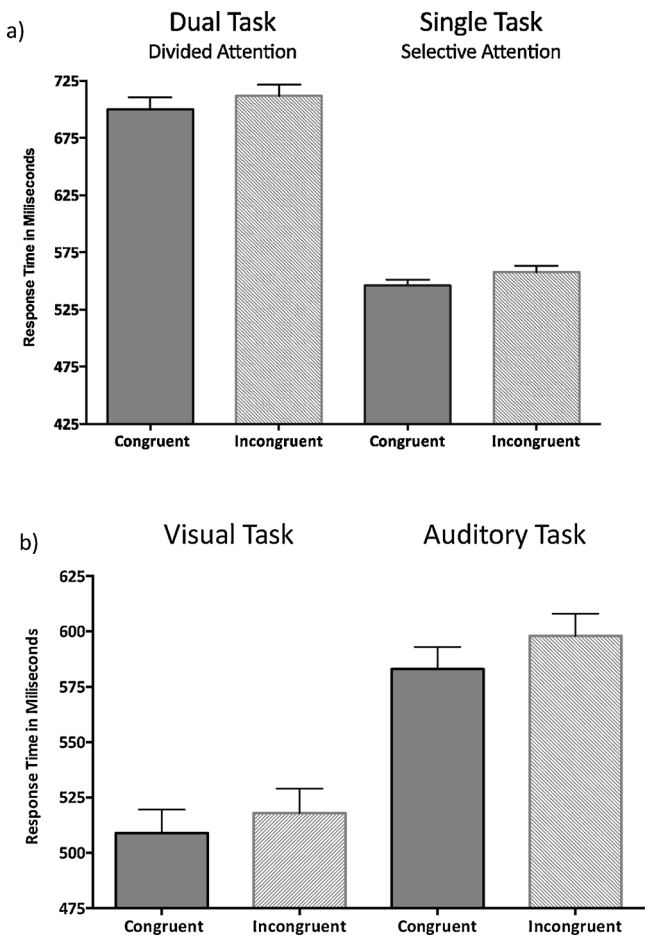

Fig. 5. a) Mean reaction time in dual-task (Congruent 700 ms, s.e.m. 10.6 ms vs. Incongruent 712 ms, s.e.m. 9.7 ms) and single task conditions during speeded classification. There is no significant interaction between task type (dual or single) and congruency (faster response time to congruent audiovisual pairings than incongruent). Response times in single modality tasks were faster than in the dual task.
b) Mean response time for two different single task conditions (visual only and auditory only) in cross-modal audiovisual pairings during speeded classification of stimuli. Observers were fastest when classifying only visual stimuli (Congruent 509 ms, s.e.m. 10 ms vs. Incongruent 518 ms, s.e.m. 11 ms) than auditory (Congruent 583 ms, s.e.m. 10 ms vs. Incongruent 598 ms, s.e.m. 10 ms).

4.1. General Discussion

Selective attention is an important cognitive function that allows one to process relevant and ignore irrelevant information, thus selectively enhancing perception. The question I addressed in this paper is whether selective attention plays a role in cross-modal interactions. Treisman and Gelade (1980) have suggested that focused selective attention is needed to bind features within the visual modality. Would the same need for selection apply across modalities? Is it the case that when both auditory and visual features are attended, an integrated cross-modal event is perceived but when they are unattended it is not?

The literature on multisensory integration is deeply divided as to the role of attention in this perceptual process. There are about as many demonstrations supporting as opposing the claim that interactions across modalities occur in an automatic fashion, independent of focused attention. Some of the factors which vary across earlier experiments that have probed the role of attention in cross-modal integration are the type of task demands loading attentional resources, whether the stimuli

are related and if so, how, and the level of processing required, for example whether the tasks require processing at the sensory or the semantic level, or whether they are speech stimuli with auditory and lip-reading components.

Talsma et al. (2010) recently proposed a framework that argues that the extent to which attention plays a role in multimodal integration depends on the conflict between the interacting modalities. That is, multisensory integration will occur preattentively when there is no or low competition between multimodal stimuli, for example the tasks are easy or one of the stimulus modalities has a stimulus of high saliency automatically capturing attention. They argue that behavioral and neuroimaging studies that do not exhaust attentional resources in multimodal speech perception (Bernstein, Auer, & Moore, 2004; Colin et al., 2002; Dekle, Fowler, & Funnell, 1992; Soto-Faraco, Navarra, & Alsius, 2004) or in which a part of the cross-modal stimulus is particularly salient (e.g. due to abrupt onset) show no susceptibility to the manipulation of attentional resources. Conversely, when investigators purposefully depleted attentional resources, as was done in a series of studies examining multimodal speech perception (Alsius et al., 2005; Alsius, Navarra, & Soto-Faraco, 2007; Tiippana, Andersen, & Sams, 2004), cross-modal interactions were reduced or absent. Similarly, when the stimuli from different modalities were of equal saliency (Degerman et al., 2007; Talsma, Doty, & Woldorff, 2007) integration between the multimodal stimuli was susceptible to attentional modulation and occurred only when observers' selective attention was directed to the multimodal event.

Present studies may help us to separate the effects of depletion of selective attention from the nature of the stimuli. I varied the demand on selective attention in several different ways and found no effect of load on cross-modal interaction with our simple visual and auditory stimuli, suggesting that the relevant issue may be the type of stimuli we used. In the current studies congruency effects remained intact despite large increases in selective attentional demand in three different manipulations: in a speeded classification task, in visual search and in a divided attention task. The findings suggest that the congruency effect observed between simple sensory features of pitch and vertical position reflects an interaction that is automatic and independent of selective attention demands. Moreover, the visual search experiment showed that the cross-modal interaction might influence the direction of attention in space, as when pitch modulates the visual input so that visual attention is drawn toward the location of the visual target. It should also be noted that in these studies I have referred to the cross-modal interaction between corresponding features of auditory pitch and visual spatial position as an integrated percept based on finding consistent congruency effects across different paradigms (see also Dolscheid et al., 2014; Evans & Treisman, 2010). What is more there is evidence showing sound localization and ear anatomy is tuned to frequency-depended biases that follow statistics of natural auditory scenes showing congruency mapping between frequency and elevation (e.g. high frequency high elevation) (Parise, Knorre, & Ernst, 2014). However it is important to stress that congruency effects alone cannot make a strong case for a necessary integration into a cross-modal percept such as is the McGurk effect in audiovisual speech integration (Soto-Faraco et al., 2004).

One possibility is that the interaction of audiovisual correspondent features in afore presented experiments is resistant to modulation by attention because they are an innate aspect of perception (Walker et al., 2009). I argued that when the responses are orthogonal to the priming stimuli, which is the case in all three of the presented experiments, and the only association is between irrelevant aspects of the two stimuli, then the interpretation is likely to be a perceptual priming. Another plausible explanation, less extreme to the idea of innate correspondences, is that this perceived correspondence is the results of observers attributing the perceived correlation between frequency and spatial elevation inherent in auditory statistics of natural environment (Parise et al., 2014) to one common cause and automatically integrating the two cross-modal signals (Parise et al., 2014). A similar independence of attentional manipulations was found with the integration of visual and auditory information about emotions (visible face and heard voice) (Vroomen, Driver et al., 2001). Perhaps the correspondence between cues to emotional states could also be innately recognized or the result of natural scene statistics. On the other hand, cross-modal speech perception or arbitrary multimodal conjunctions do seem to be subject to attentional depletion effects, as described above. These must clearly be learned through experience rather than innately specified or arising from natural visual and auditory scene statistics.

Based on the current debate on the role of attention and by extension the question of automaticity in multimodal processing (Spence & Deroy, 2013; Talsma et al., 2010), it is evident that there is no one definitive answer or a clear dichotomy of yes or no valid for all. Different attentional mechanisms maybe relevant for some multimodal processes while not needed for others. The degree to which a process is automatic will depend on the extent it meets the goal-independence, non-conscious, load-insensitivity and speed criteria. The findings we present here suggest that selective attention is not needed for cross-modal integration of simple basic features such as pitch of sound and spatial potion of a visual stimulus. The findings of load-insensitivity further support the notion that this integration has a high degree of automaticity. This is but one of multiple cross-modal interactions that have been observed and they differ in their nature. Some are perceptual but others are due to verbal or semantic mediation. Further research may test the postulated hypothesis that selective attention is needed to integrate or permit interaction of cross-modal features only when they are not innately connected or naturally occurring mapping and the associations are learned through effortful experience.

Acknowledgments

I would like to knowledge Prof. Anne. M. Treisman for her significant contribution to the design of the studies and invaluable discussions of the results and the topic at large. Anne Treisman fell ill and passed away before these studies were prepared for the paper. We had intended to write jointly but this was overshadowed by her illness and she asked not to be listed as a co-author. In sorrow, I dedicate this work to her memory.

Appendix A. Supplementary data

The raw data for the three experiments discussed in this article can be found at the following link: https://osf.io/kfh56/?view_only=e921bfe8846b44969297f615446129e0.

References

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Curr Biol, 15*(9), 839–843.
Alsius, A., Navarra, J., & Soto-Faraco, S. (2007). Attention to touch weakens audiovisual speech integration. *Exp Brain Res, 183*(3), 399–404.
Arnell, K. M., & Duncan, J. (2002). Separate and shared sources of dual-task cost in stimulus identification and response selection. *Cognit Psychol, 44*(2), 105–147.
Bernstein, L. E., Auer, E. T., Jr., & Moore, J. K. (2004). Audiovisual speech binding: Convergence or association. In G. Calvert, C. Spence, & B. E. Stein (Eds.). *The handbook of multisensory processes*. Cambridge, Mass: MIT Press.
Brainard, D. H. (1997). The Psychophysics Toolbox. *Spat Vis, 10*(4), 433–436.
Chiou, & Rich (2012). Cross-modality correspondence between pitch and spatial location modulates attentional orienting. *Perception, 41*(3), 339–353.
Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk-MacDonald effect: a phonetic representation within short-term memory. *Clin Neurophysiol, 113*(4), 495–506.
Degerman, A., Rinne, T., Pekkola, J., Autti, T., Jaaskelainen, I. P., Sams, M., et al. (2007). Human brain activity associated with audiovisual perception and attention. *Neuroimage, 34*(4), 1683–1691.
Dekle, D. J., Fowler, C. A., & Funnell, M. G. (1992). Audiovisual integration in perception of real words. *Percept Psychophys, 51*(4), 355–362.
Dolscheid, S., Hunnius, S., Casasanto, D., & Majid, A. (2014). Prelinguistic infants are sensitive to space-pitch associations found across cultures. *Psychological Science, 25*(6), 1256–1261.

Duncan, J., Martens, S., & Ward, R. (1997). Restricted attentional capacity within but not between sensory modalities. *Nature, 387*(6635), 808–810.
Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision, 10*(1), 1–12 6.
Hein, G., Parr, A., & Duncan, J. (2006). Within-modality and cross-modality attentional blinks in a simple discrimination task. *Percept Psychophys, 68*(1), 54–61.
Jonides, J., & Irwin, D. E. (1981). Capturing attention. *Cognition, 10*(1–3), 145–150.
Iordanescu, L., Guzman-Martinez, E., Grabowecky, M., & Suzuki, S. (2008). Characteristic sounds facilitate visual search. *Psychonomic Bulletin & Review, 15*(3), 548–554.
Iordanescu, L., Grabowecky, M., Franconeri, S., Theeuwes, J., & Suzuki, S. (2010). Characteristic sounds make you look at target objects more quickly. *Attention, Perception, & Psychophysics, 72*(7), 1736–1741.
Klapetek, Ngo, & Spence (2012). Does crossmodal correspondence modulate the facilitatory effect of auditory cues on visual search? *Attention, Perception, & Psychophysics, 74*(6), 1154–1167.
Lavie, & Dalton (2014). *Load theory of attention and cognitive control. The Oxford Handbook of Attention*. Oxford University Press56–75.
Ludwig, V. U., Adachi, I., & Matsuzawa, T. (2011). Visuoauditory mappings between high luminance and high pitch are shared by chimpanzees (Pan troglodytes) and humans. *Proceedings of the National Academy of Sciences of the United States of America, 108*, 20661–20665.
Marks, L. E., Hammeal, R. J., & Bornstein, M. H. (1987). Perceiving similarity and comprehending metaphor. *Monogr Soc Res Child Dev, 52*(1), 1–102.
Marks, L. E. (2004). *Speeded Classification. The handbook of multisensory processes*85–92.
Massaro, D. W., & Warner, D. S. (1977). Dividing attention between auditory and visual perception. *Perception & Psychophysics, 21*(6), 569–574.
McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746–748.
Mondloch, C. J., & Maurer, D. (2004). Do small white balls squeak? Pitch-object correspondences in young children. *Cogn Affect Behav Neurosci, 4*(2), 133–136.
Moors, A., & De Houwer, J. (2006). Automaticity: a theoretical and conceptual analysis. *Psychological bulletin, 132*(2), 297.
Mossbridge, J. A., Grabowecky, M., & Suzuki, S. (2011). Changes in auditory frequency guide visual–spatial attention. *Cognition, 121*(1), 133–139.
Ten Oever, S., Romei, V., van Atteveldt, N., Soto-Faraco, S., Murray, M. M., & Matusz, P. J. (2016). The COGs (context, object, and goals) in multisensory processing. *Experimental brain research, 234*(5), 1307–1323.
Parise, C. V., & Spence, C. (2009). 'When birds of a feather flock together': synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS One, 4*(5), e5664.
Parise, C. V., & Spence, C. (2012). Audiovisual crossmodal correspondences and sound symbolism: a study using the implicit association test. *Experimental Brain Research, 220*(3–4), 319–333.
Parise, C. V., Knorre, K., & Ernst, M. O. (2014). Natural auditory scene statistics shapes human spatial hearing. *Proceedings of the National Academy of Sciences, 111*(16), 6104–6108.
Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat Vis, 10*(4), 437–442.
Peiffer-Smadja, N. (2010). *Exploring the bouba/kiki effect: A behavioral and Fmri study*Unpublished Ms Thesis. France: Universite Paris V–Descartes.
Pooresmaeili, A., FitzGerald, T. H., Bach, D. R., Toelch, U., Ostendorf, F., & Dolan, R. J. (2014). Cross-modal effects of value on perceptual acuity and stimulus encoding. *Proceedings of the National Academy of Sciences, 111*(42), 15244–15249.
Potter, M. C., Chun, M. M., Banks, B. S., & Muckenhoupt, M. (1998). Two attentional deficits in serial target search: the visual attentional blink and an amodal task-switch deficit. *J Exp Psychol Learn Mem Cogn, 24*(4), 979–992.
Soto-Faraco, S., Navarra, J., & Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. *Cognition, 92*(3), B13–23.
Soto-Faraco, S., & Spence, C. (2002). Modality-specific auditory and visual temporal processing deficits. *Q J Exp Psychol A, 55*(1), 23–40.
Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics, 73*(4), 971–995.
Spence, C., & Deroy, O. (2013). How automatic are crossmodal correspondences? *Consciousness and cognition, 22*(1), 245–260.
Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses.* The MIT Press.
Störmer, V., Eppinger, B., & Li, S. C. (2014). Reward speeds up and increases consistency of visual selective attention: A lifespan comparison. *Cognitive, Affective, & Behavioral Neuroscience, 14*(2), 659–671.
Talsma, D., Doty, T. J., & Woldorff, M. G. (2007). Selective attention and audiovisual integration: is attending to both modalities a prerequisite for early integration? *Cereb Cortex, 17*(3), 679–690.
Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in cognitive sciences, 14*, 400–410.
Taylor, M. M., Lindsay, P. H., & Forbes, S. M. (1967). Quantification of shared capacity processing in auditory and visual discrimination. *Acta Psychol (Amst), 27*, 223–229.
Tiippana, K., Andersen, T. S., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology, 16*, 457–472.
Treisman, A. M. (1998). Feature binding, attention and object perception. *Philos Trans R Soc Lond B Biol Sci, 353*(1373), 1295–1306.
Treisman, A. M., & Davies, A. (1973). Divided attention to ear and eye. In S. Kornblum (Ed.). *Attention and Performance IV* (pp. 101–117). London: Academic Press.
Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognit Psychol, 12*(1), 97–136.
Vroomen, J., Driver, J., & de Gelder, B. (2001). Is cross-modal integration of emotional expressions independent of attentional resources? *Cogn Affect Behav Neurosci, 1*(4), 382–387.
Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., et al. (2009). Preverbal Infants' Sensitivity to Synaesthetic Cross-Modality Correspondences. *Psychological Science*.