

**Crossmodal Attentional Control Sets**

**Between Vision and Audition**

Frank Mast<sup>1</sup>, Christian Frings<sup>1</sup>, & Charles Spence<sup>2</sup>

<sup>1</sup>*Trier University*

<sup>2</sup>*Oxford University*

Correspondence:

Christian Frings  
Trier University  
Faculty 1, Department of Psychology,  
Germany  
chfrings@uni-trier.de  
Tel.: +49 (0)651 201 2958

## ABSTRACT

The interplay between top-down and bottom-up factors in attentional selection has been a topic of extensive research and controversy amongst scientists over the past two decades. According to the influential contingent capture hypothesis, a *visual* stimulus needs to match the feature(s) implemented into the current attentional control sets in order to be automatically selected. Recently, however, evidence has been presented that attentional control sets affect not only visual but also crossmodal selection. The aim of the present study was therefore to establish contingent capture as a general principle of multisensory selection. A non-spatial interference task with bimodal (visual and auditory) distractors and bimodal targets was used. The target and the distractors were presented in close temporal succession. In order to perform the task correctly, the participants only had to process a predefined target feature in either of the two modalities (e.g., color when vision was the primary modality). Note that the additional crossmodal stimulation (e.g., a specific sound when hearing was the secondary modality) was not relevant for the selection of the correct response. Nevertheless, larger interference effects were observed when the distractor matched both the stimulus of the primary as well as the secondary modality and this pattern was even stronger if vision was the primary modality than if audition was the primary modality. These results are therefore in line with the crossmodal contingent capture hypothesis. Both visual and auditory early processing seem to be affected by top-down control sets even beyond the spatial dimension.

Keywords: Contingent capture; Crossmodal attention; Multisensory; Vision; Audition

## Introduction

During every waking moment, we are flooded by a vast amount of sensory information as a result of the many ongoing events in our environment. Due to our limited cognitive capacities, however, only a minority of this information can be processed consciously. Thus, there is a constant competition amongst the different stimuli in order to be selected for further processing (Pashler, 1998). One of the central roles for attention here is thought to be to bias the processing of early incoming perceptual information. According to prominent theories of *visual* attention (e.g., the Feature Integration Theory, Treisman & Gelade, 1980; the Guided Search Model, Wolfe, 1994, 2007; the Theory of Visual Attention, Bundesen, 1990), selection is affected by both, top-down and bottom-up mechanisms. In everyday life, however, visual input is rarely processed in isolation, but rather together with information from the other senses, in order to enhance behavioural efficiency. For example, when searching for a friend at a lively party, it seems plausible to search the visual scene for his face but also to listen out for the booming sound of his voice. It is reasonable, therefore, to assume that top-down mechanisms (e.g., knowledge about the appearance and voice of one's friend) not only play an important role in unisensory but also in crossmodal attentional selection (that is selection between different senses) as well as in the selection and integration of multisensory events (see e.g., Talsma, Senkowski, Soto-Faraco, Woldorff, 2010; Tang, Wu, & Shen, 2016, for reviews). In this paper we therefore enhance and generalize the idea of contingent capture to crossmodal selection as research on this particular attention phenomenon has only recently started to be examined in multisensory contexts (Mast, Frings, & Spence, 2015; Matusz & Eimer, 2013).

## AUDIOVISUAL CONTINGENT CAPTURE

One paradigm that has frequently been used in research on visual selective attention in order to dissociate top-down from bottom-up mechanisms is the exogenous spatial cuing task (see Posner, 1980; Yantis & Jonides, 1984). Over the last couple of decades, researchers have investigated the exogenous control of crossmodal spatial attention for all possible combinations of visual, auditory, and tactile cue and target stimuli (Spence & Driver, 1997; Spence, Nicholls, Gillespie, & Driver, 1997).

Originally, the assumption was that certain stimulus events, no matter whether they were unisensory or multisensory, had the potential to capture attention and reflect a purely stimulus-driven mechanism of selective attention (e.g., Jonides, 1980; Theeuwes, 1991; Yantis & Jonides, 1984). This notion was challenged by Folk and colleagues' (1992; see also Folk, Remington, & Wright, 1994) notion of *contingent* capture. According to the latter account, participants can set up attentional control sets for a specific task-relevant feature. As a consequence of attentional control sets, only those stimuli that match the current attentional control set have the potential to automatically capture a participant's spatial attention. When, for example, the task involves localizing a red target stimulus, participants are assumed to set-up their attentional control sets for the color 'red' (see also Ansorge & Becker, 2014; Goller & Ansorge, 2015). Accordingly, a stimulus (cue or target) needs to be red in order to be selected automatically. One might think of attentional control sets as an abstract inner representation of the searched-for target stimuli. Since Folk et al. (1992) published their influential first study, *contingent attentional capture* has been replicated in various studies (see Awh, Belopolsky, & Theeuwes, 2012; Burnham, 2007; Theeuwes, 2010, for reviews; but see also Lamy & Kristjánsson, 2013).

Recent studies have indicated that attentional control sets can also be compiled of multiple features from different sensory modalities, namely from vision and touch (Mast,

## AUDIOVISUAL CONTINGENT CAPTURE

Frings, & Spence, 2015; see also Matusz & Eimer, 2013). Mast and his colleagues combined a non-spatial visual response compatibility task with additional (response irrelevant) tactile stimulation. Each trial consisted of two visual stimuli that were presented from the same location in close temporal succession. The participants were instructed to try and ignore the identity of the first stimulus (the distractor) and to respond to the identity of the second stimulus (the target). In response compatible trials, the distractor was mapped on to the same response as the subsequent target. In the response incompatible trials, by contrast, the distractor was mapped on to the opposing response instead. In order to examine the compilation of crossmodal attentional control sets, the visual primary task was combined with additional tactile stimulation. That is, the visual target stimulus was always accompanied by a simultaneously-presented additional tactile stimulus. It is important to stress that the tactile stimulation itself was not mapped on to either response. The co-occurrence of the visual target and the tactile stimulus was assumed to result in participants establishing a bimodal attentional control set (incorporating both visual and tactile components). Intriguingly, bimodal distractors caused more pronounced interference effects than unimodal distractors. It was argued that the difference in the size of the interference effects was due to differences in the feature-overlap between the features of the distractor (unimodal vs. bimodal) and the features implemented into the participants' top-down sets. Therefore, the results suggest multisensory top-down sets having both visual and tactile features.

The aim of the present study was therefore to further support the *crossmodal* contingent capture hypothesis and to underline the importance of contingent capture in selection in general. Studying crossmodal attentional control sets across the different senses and in different experimental paradigms, is important since the interplay between top-down and

bottom-up mechanisms has been found to differ between the different senses modalities (e.g., see Bundesen, Kyllingsbaek, Houmann, & Jensen, 1997; Moray, 1959, for differences in the potential of one's own name to capture attention depending on whether the name is spoken or written). What is more, the top-down influence on automatic distractor processing has been found to vary as a function of the modalities combined in a given task (e.g., see Mast, Frings, & Spence, 2014, where participants were able to ignore tactile distractor information when attending to a visual target but not vice versa). Therefore, we examined the compilation of audiovisual attentional control sets in a non-spatial interference task that was derived from the typical contingent capture task (see Matusz & Eimer, 2013, for a crossmodal exogenous spatial cuing task). All of the published studies that previously examined crossmodal contingent capture (Mast et al., 2015; Matusz & Eimer, 2013) combined a *visual* primary task with additional crossmodal (auditory or tactile) stimulation. Thus, we went beyond the previous research in this area by analysing audio-visual contingent capture effects and further by varying whether vision or audition was the response-relevant dimension. Note that previous research on audio-visual integration/ selection have shown differences in dependence of whether vision or audition was the task-relevant modality (e.g., Yuval-Greenberg & Deouell, 2009; Thelen, Matusz, & Murray, 2014; van der Burg, Olivers, Bronkhorst, & Theeuwes, 2008). Thus, our study can answer the question whether crossmodal contingent capture affects selection also if participants respond to non-visual targets and hence reflects a modality-unspecific attention mechanism.

### Overview of the present study

## AUDIOVISUAL CONTINGENT CAPTURE

Following Mast et al. (2014, 2015), a non-spatial response compatibility task was used. In Experiment 1, the presentation of the visual target was always accompanied by a sound. It can be argued that participants set-up their attentional control sets for a visual and an auditory feature. On the one hand, the visual feature (i.e., color) should be implemented into the top-down sets because its identity indicates the response that should be executed (the response feature). On the other hand, the auditory feature should be implemented into the participant's attentional control set because it indicates the presence of the target stimulus (the selection feature). While the targets were always accompanied by the same auditory stimulus, the distractors were combined with either a target congruent sound or else with a target incongruent sound (see Fig. 1). Note that the distractor sound was not correlated with the identity of the subsequent visual target (i.e., it was non-predictive).

In Experiment 2, the relevant modality was audition while a particular visual feature (a coloured circle) always accompanied the targets but only half of the distractors displayed this particular target feature. Once again, the accompanying feature was not correlated with the response feature.

In both experiments, the strength of attentional capture effects for the distractors was assumed to vary as a function of the feature-overlap between the features of the distractor and the features of the attentional control sets. That is, more pronounced attentional capture effects were expected as the feature overlap between the distractor and the multisensory top-down set increases.

## EXPERIMENT 1

As outlined above, Experiment 1 combined stimuli from both vision (red and green circles) and audition (with either 200 or 700 Hz pure sounds). In line with the previous research (e.g., Mast et al., 2015; Matusz & Eimer, 2013), the hypothesis was that the participants compile their top-down sets for both the visual and auditory features. Thus, the participants' top-down sets should contain at least two crossmodal features; color as the response feature (given that color indicates the correct response) and the pitch of the target sound as the selection feature (a feature that indicates the presence of the target). The featural-overlap of the distractor and the participants' top-down sets is assumed to vary as a function of the congruency of the auditory stimulus. The proposed differences in feature overlap between the distractor and the top-down set should be reflected in less pronounced attentional capture effects for the auditory incongruent condition as compared to the auditory congruent condition (see Fig. 2 for an explanation).

### Methods

*Participants.* Twenty-one students (3 male; mean age of 22 years ranging from 19 to 27 years) from Trier University served as participants in Experiment 1. All of the participants reported normal or corrected-to-normal vision and no impairments of their auditory perception.

*Design.* The participants were tested in a  $2 \times 2$  experimental design with response compatibility (compatible vs. incompatible) and auditory congruency (congruent vs. incongruent) as factors. Auditory congruency and response compatibility were tested as within-participants factors.



## AUDIOVISUAL CONTINGENT CAPTURE

*Apparatus and materials.* The participants were tested in individual testing sessions. The data collection took place in a completely dark soundproofed room. The instructions as well as the visual stimuli were presented on a 7'' monitor (Model FT0070TM, Faytech Ltd., Henzen, China). The refresh rate of the monitor was 60 Hz, which was placed approximately 12 cm in front of the participant's body midline. The participant's responses were detected with a standard PC mouse connected via USB 2.0 port. The auditory stimulus was presented by means of a loudspeaker cone placed at the rear of the screen.

The visual stimuli used in Experiment 1 were green (CIE L\*a\*b\*-value: 46, -52, 50) and red circles (CIE L\*a\*b\*-value: 53, 80, 67) with a diameter of approximately 1.72° of visual angle. The targets and distractors were always presented from the same central location on the screen which was indicated by a fixation cross at the start of each trial. Two different auditory stimuli were created with the open source software Audacity Vs. 2.0.2. Both sound files consisted of sinusoidal waves, differing only in terms of their frequency. The frequency of one of the stimuli was 200 Hz and its loudness was 46 dB whereas the frequency of the other sound was 700 Hz and its loudness was 60 dB.

*Procedure.* During each trial, two audiovisual stimuli were presented. in close temporal succession from the same central location. The participants' task was to try to ignore the first stimulus (the distractor) and to respond to the colour of the second stimulus (the target) by pressing the left (for the colour green) or right (for the colour red) mouse button. Each trial was initiated by the presentation of a central fixation cross for 600 ms. The fixation cross was immediately followed by the distractor stimulus for 33 ms. The inter-stimulus interval (ISI) varied randomly on a trial-by-trial basis (90 ms, 130 ms, or 170 ms). Finally, the target was presented for 33 ms. The participants had to respond to the

## AUDIOVISUAL CONTINGENT CAPTURE

target within 2,000 ms of the offset of the target. After having responded to the color of the target, a 600 ms blank-screen was presented before the next trial started automatically.

The colour of the distractor was uncorrelated with the colour of the subsequent target, i.e. in half of the trials the distractor colour matched the target colour (response compatible trials) while in the other half it did not match the target colour (response incompatible trials). Both targets and distractors were accompanied by an additional irrelevant auditory stimulus. The sound that accompanied the presentation of the target sound was kept constant for each participant, but was counterbalanced across participants (i.e. for a particular participant, all targets were accompanied by a 200 Hz tone). While the target sound was kept constant, the distractor sound varied matching the target sound in 50% of trials (congruent trials) and not matching the target sound in the other 50% of trials (incongruent trials). The auditory distractor features were also orthogonal to the colour feature of the target (i.e. the target colour could neither be predicted by the distractor colour nor the distractor sound).

At the beginning of the experiment, the participants had to work their way through two short training phases. In the first training block, only the target was presented in order to facilitate the participants' learning of the stimulus-response mapping (16 trials). The second training phase (48 trials), was almost identical with the experimental trials, with the sole exception of feedback concerning the participants' responses which was provided after each trial. The experimental block comprised 432 trials (108 compatible with congruent auditory stimulation, 108 compatible with incongruent auditory stimulation, 108 incompatible with congruent auditory stimulation, 108 incompatible with incongruent auditory stimulation). After every 40th trial, the participants were offered a short break. Should the participant make three errors in a row then they were offered another break.

## Results

Only those trials in which the participant responded to the target correctly were considered for the analyses of reaction times (RTs). Additionally, all of those trials in which the RT was below 200 ms, as well as those trials in which it was 1.5 interquartile ranges above the third quartile of each participant's individual RT distribution (Tukey, 1977) were excluded from the data analyses. 8.2% of all trials were excluded from the analysis due to these restrictions. Mean RTs and error rates (ERs) are highlighted in Table 1.

*RTs.* The RTs from Experiment 1 were submitted to a  $2$  (response compatibility: compatible vs. incompatible)  $\times$   $2$  (auditory congruency: congruent  $\times$  incongruent) MANOVA with Pillai's trace as criterion (see O'Brien & Kaiser, 1985, for a comparison of MANOVA and ANOVA in repeated measure designs). The main effect of auditory congruency just failed to reach the conventional criterion for statistical significance,  $F(1, 20) = 4.23, p = .053, \eta_p^2 = .174$ . No significant difference was observed for the mean RTs between congruent and incongruent auditory trials. Most important in terms of our main hypothesis, the MANOVA revealed a significant response compatibility  $\times$  auditory congruency interaction,  $F(1, 20) = 13.09, p = .002, \eta_p^2 = .396$  (see Fig. 3). That is, larger compatibility effects were observed in the congruent than in the incongruent auditory condition ( $M = 143$  ms,  $SD = 57$  ms vs.  $M = 119$  ms,  $SD = 61$  ms, respectively).

*ERs.* The error rates were submitted to the same  $2 \times 2$  MANOVA. The analyses of the ERs only revealed a significant main effect of response compatibility,  $F(1, 20) = 19.06, p < .001, \eta_p^2 = .488$ . The participants made fewer errors in the response compatible trials

than in the response incompatible trials. However, neither the main effect of auditory congruency,  $F(1, 20) < 1$ , nor the response compatibility  $\times$  auditory congruency interaction reached significance,  $F(1, 20) < 1$ .

### Discussion

In Experiment 1, a visual response compatibility task was systematically combined with auditory stimulation in order to examine the compilation of bimodal attentional control sets. For each participant, the visual target stimulus was consistently presented together with a specific target sound (200 or 700 Hz). The idea was that due to the constant co-occurrence of the visual target and a specific sound, the participants would use visual and auditory features in order to set-up their attentional control sets. While the visual targets were consistently presented together with the same auditory stimulus, the auditory stimulus (200 or 700 Hz) that was presented during the visual distractors varied randomly between trials. In line with our main hypothesis, larger compatibility effects were observed for the congruent than for the incongruent auditory condition. These results support the proposed feature overlap between the different distractors and the attentional control sets (see Fig. 2). That is, the congruent auditory distractors match both features of the participants' top-down sets (color and frequency). By contrast, the incongruent auditory distractors only matched the visual but did not match the auditory feature of the top-down sets. Thus, the results support the crossmodal contingent capture hypothesis with attentional capture effects nicely mirroring the proposed feature overlap between the distractor and the multisensory top-down set.

## EXPERIMENT 2

### Methods

*Participants.* Another group of 26 students (14 women, 12 men; mean age 25 years ranging from 19 to 32 years) from the University of Trier took part in the second experiment. All of the participants reported normal or corrected-to normal vision, and none of them reported any impairment of their sense of hearing.

*Design.* The participants were tested in a 2 (response compatibility: compatible vs. incompatible)  $\times$  2 (visual congruency: congruent vs. incongruent) experimental design. Once again, the two experimental factors were tested as within-participants factors.

*Apparatus and materials.* The data collection was conducted in 1 of 3 soundproofed cabins. The instructions as well as the visual stimuli were presented on a 22'' monitor (Model FlexScan S2202 W, EIZO Europe GmbH, Mönchengladbach, Germany). The refresh rate of the monitor was 60 Hz, it was placed approximately 50 cm in front of the participant's body midline and the distance was held constant by means of a chinrest. The auditory stimuli were presented via headphones. The participant's responses were detected with a standard PC mouse connected via a USB 2.0 port.

The visual stimuli were the same as in the previous experiment. However, due to changes in the experimental setup because of the change in the testing cabins (display size) the size of the visual stimuli slightly changed. The diameter measured for the visual stimuli was 2.52°. The two sound files again consisted of sinusoidal waves but differed in terms of their frequency; that is, stimulus one was of 400 Hz (its loudness was 47 dB) whereas stimulus two was of 700 Hz (its loudness was 60 dB).

## AUDIOVISUAL CONTINGENT CAPTURE

*Procedure.* The procedure closely followed that of Experiment 1 except for the following details: The participants' primary task was to respond to the pitch of the auditory targets (the second stimulus) and to try and ignore the pitch of the distractors (the first stimulus). The appearance of the auditory target was always accompanied by the same visual stimulus whereas the presentation of the auditory distractor was accompanied by target congruent visual stimulus in 50% of trials or by an incongruent visual stimulus in the remaining 50% of trials.

The trial procedure was very similar to that used in Experiment 1. The duration of target and distractor presentation was slightly increased to 50 ms. The ISI between the two distractor and the target varied randomly (50, 70, or 90 ms) on a trial-by-trial basis. Furthermore, catch trials were introduced in Experiment 2 in order to prevent the participants from using strategies to avoid the response irrelevant visual stimuli (such as closing their eyes to prevent distraction from the visual stimuli). The catch trials followed the same temporal rules as the congruent or incongruent visual trials. However, during the catch trials, a blue instead of a red or green visual stimulus accompanied the auditory distractor. The participants were instructed not to respond to the target at all when the preceding distractor was accompanied by the blue visual stimulus.

Once again, the experiment was divided into three consecutive blocks of trials, from which the first two blocks were considered as training and the third and final block was the experimental block. The two training blocks were identical to those in Experiment 1. The experimental block comprised 360 trials in total (72 catch trials, 144 visual congruent, 144 visual incongruent; with 72 response compatible and incompatible trials per congruency condition). As in the previous experiment, a break was offered to the participants after

every 40th trial. When three errors were recorded in consecutive trials, an additional break was offered.

## Results

The same criteria as in the previous experiments were used for data trimming. The data from three of the participants had to be discarded due to their extremely high error rates in the catch trials (100% errors in the catch trials for all three participants; 6.7% errors for catch trials as sample mean). Due to these restrictions, 11% of all trials were not considered in the RT analyses (8% errors). Mean RTs and error rates are depicted in Table 2.

*RTs.* A 2 (response compatibility: compatible vs. incompatible)  $\times$  2 (visual congruency: congruent vs. incongruent) MANOVA was conducted with Pillai's trace as criterion. The MANOVA revealed a significant main effect for response compatibility,  $F(1, 22) = 65.83, p < .001, \eta_p^2 = .750$ . The participants responded more rapidly in the response compatible trials than in the response incompatible trials, as expected. The main effect of visual congruency also reached statistical significance,  $F(1, 22) = 12.63, p = .002, \eta_p^2 = .365$ , indicating that the participants' responses were more rapid in the congruent visual trials ( $M = 631$  ms) than in the incongruent visual trials ( $M = 649$  ms). In line with our main hypothesis, the interaction between response compatibility and visual congruency was once again significant,  $F(1, 22) = 12.17, p = .002, \eta_p^2 = .356$ . That is, larger compatibility effects were observed in the congruent than in the incongruent auditory condition ( $M = 155$  ms,  $SD = 87$  ms vs.  $M = 116$  ms,  $SD = 82$  ms, respectively; see Fig. 4).

*ERs.* The same 2 (response compatibility: compatible vs. incompatible)  $\times$  2 (visual congruency: congruent vs. incongruent) MANOVA as for the RTs was also applied to examine the ERs. The MANOVA revealed only a significant main effect for response compatibility,  $F(1, 22) = 14.83$ ,  $p = .001$ ,  $\eta_p^2 = .403$ , with participants making more errors in response incompatible trials than in response compatible trials. Neither the main effect of visual congruency,  $F(1, 22) < 1$ , nor the response compatibility  $\times$  visual congruency interaction reached significance  $F(1, 22) < 1$ .

### Discussion

The main aim of Experiment 2 was to increase the generalisability of crossmodal contingent capture by utilizing a modality other than vision for the primary task. Therefore, audition was chosen for the response compatibility task and then systematically combined with additional visual stimuli. The participant had to respond to the pitch of the auditory target which was accompanied by a specific visual figure (e.g., a green circle). Thus, the participants were supposed to set-up their attentional control sets for both auditory (the sound pitch as the response feature) and visual features (colour as the selection feature). The auditory distractor was accompanied by either a colour congruent (a green circle) or a colour incongruent visual stimulus (a red circle). The data from Experiment 2 fully confirm those from our first experiment. That is, larger compatibility effects were observed for congruent visual trials than for incongruent visual trials. Thus, the results support crossmodal attentional control sets whereby the feature overlap between the distractors and the top-down sets determine the strength of attentional capture effects. In the congruent



## AUDIOVISUAL CONTINGENT CAPTURE

visual condition, the distractor matched both features of the crossmodal top-down set; the auditory response feature and the visual selection feature. By contrast, for the incongruent visual trials, the distractor only matched the auditory response feature but the distractor did not match the visual selection feature. Accordingly, the differences in the size of the compatibility effects nicely matched the predicted differences in the feature overlap between the distractors and the top-down sets for the congruent visual and the incongruent visual condition.

### General Discussion

The aim of the present study was to establish *crossmodal contingent capture* in crossmodal and multisensory selection. Across two experiments, the compilation of attentional control sets with features from vision and audition was examined. In Experiment 1, vision was used for the response compatibility task and additional auditory stimulation was presented in order to examine crossmodal attentional control sets. By contrast, in Experiment 2, audition was used in the response compatibility task and was combined with additional, non-irrelevant visual stimuli. In both experiments, the main empirical prediction was that the size of the distractor effects varies as function of the feature-overlap between the distractors and the top-down compiled attentional control sets. In both experiments, larger compatibility effects were reported in the congruent condition (auditory or visual) than in the incongruent condition (auditory or visual). The difference in the size of the compatibility effects indicates more pronounced attentional capture for congruent distractors as compared to incongruent distractors. Thus, evidence for crossmodal

## AUDIOVISUAL CONTINGENT CAPTURE

contingent capture was presented from a visuo-auditory as well as from an audiovisual non-spatial response compatibility task, thus underpinning the importance of multisensory attentional control sets when it comes to crossmodal selection.

The data presented here are in line with recent studies on crossmodal contingent capture (Mast et al., 2015; Matusz & Eimer, 2013). Importantly, the present study provides a new, somewhat more general understanding of the mechanisms involved in crossmodal contingent capture. In particular, Matusz and Eimer applied an audio-visual *spatial* cuing task and found evidence for the impact of crossmodal attentional control sets when it comes to the control of visual spatial attention. Here, we applied a very similar manipulation (the combination of a unisensory primary task with response irrelevant crossmodal stimulation) but instead of an exogenous cuing study, a non-spatial response compatibility task was used. Intriguingly, the same pattern of results was observed for both studies. Thus, the present results indicate that multisensory attentional control sets play an important role in both spatial and non-spatial crossmodal selective attention.

Mast et al. (2015) applied a comparable *non-spatial* task as in the present study but with additional tactile (instead of auditory) stimulation. The results of the two audio-visual experiments reported here nicely match the results of our earlier visuo-tactile study (). Taken together, these studies demonstrate that crossmodal attentional control sets can be compiled with features from vision, audition, and touch. Importantly, the results of Experiment 2 support the suggestion that this mechanism is not restricted to visual primary tasks but can be applied for the auditory and tactile primary tasks as well. Still, these findings should not be taken to suggest that multisensory processes are controlled in a purely top-down fashion. For instance, Matusz and Eimer (2011) demonstrated that visual top-down templates are ineffective in preventing capture from audiovisual distractors

(thereby boosting spatial cuing). Other studies suggested that the detection of simultaneity across the senses and its influence on brain and cognitive processes is independent of the particular task, population, or state of the individual (e.g., conscious awareness) also speaking for strong bottom-up influences of multisensory processing occurring early in the brain (less than 100ms post-stimulus onset) and within low-level cortices (e.g., Murray et al. 2016).

One issue that might be criticized here is the fact that a response compatibility task was used to examine the compilation of attentional control sets. The interpretation of response compatibility effects in favour of contingent capture (Folk et al., 1992) might appear doubtful at first glance because the results are strongly confounded by response priming effects (see Neumann & Klotz, 1994; for a review see Kiesel, Kunde, & Hoffmann, 2007) which researchers in exogenous cuing tasks typically explicitly try to avoid. It seems reasonable to assume that the compatibility effects observed here reflect a combination of response priming and attentional capture effects. However, our line of argument refers exclusively to the difference in the size of the compatibility effects between the crossmodally congruent and incongruent conditions. Note that the additional crossmodal stimulation during distractor presentation was not mapped on to a specific response. Consequently, the crossmodal stimulation could not *prime* either of the two response alternatives.

With respect to purely perceptual repetition priming one might argue that in the congruent-compatible condition target processing is enhanced without referring to top-down sets: still perceptual priming could not easily explain why response selection in one modality should be enhanced by perceptually priming an irrelevant modality (see Mast & Frings, 2014 for a detailed discussion of this issue). Nevertheless, in order to completely

disentangle the influences of perceptual priming versus attentional control sets one would have to vary whether participants could predict the irrelevant target features. Here it is safe to conclude that distractors that match target features in a setting with predictable target features lead to enhanced processing.

We argue that the congruent crossmodal stimulation during the presentation of the distractor increased the similarity between the distractor and the searched-for target. Thus, a congruent distractor might receive more attentional resources than an incongruent distractor because of its increased similarity to the search template. In various visual studies (e.g., Ansorge & Heumann, 2003, 2004), target-distractor similarity has been found to be a significant factor when it comes to predicting attentional capture. Ansorge and Heumann had their participants complete a visual exogenous cuing task in which the participants' task involved localizing a target defined by a specific visual feature (e.g., blue). Intriguingly, more pronounced attentional capture effects were reported in those trials with target similar distractors (e.g., blueish-green cues) than in those trials with target dissimilar distractors (e.g., yellowish-red cues). The feature-overlap between distractors and top-down sets presented here (see also Mast et al., 2014, 2015) is very similar to the idea of target-distractor similarity, however extended to similarity based on multiple features from multiple sensory modalities.

Further support for the notion of audio-visual top-down sets comes from those studies that have combined visual search tasks (Iordanescu et al., 2008, 2010) with additional auditory stimuli. In the studies of Iordanescu and colleagues, the participants had to perform a visual search task with naturalistic objects (e.g., a dog) as the target. As the central experimental manipulation, auditory stimulation was varied during the presentation of the visual search display. The search displays were either accompanied by target

## AUDIOVISUAL CONTINGENT CAPTURE

consistent (a barking sound when searching for the picture of a dog) sound, a distractor consistent (a meowing sound when the picture of a cat is presented as a distractor) sound, or as a control, a no-sound condition. The main finding to emerge from these studies was that visual search performance was enhanced by the presentation of target-consistent sounds. The authors argued that target-consistent sounds increased the saliency of the visual target object. Intriguingly, no difference in search performance was observed between the no-sound condition and the distractor-consistent condition. Thus, the potential of a distractor stimulus to capture attention was not affected by the presentation of a distractor consistent auditory stimulus. Therefore, auditory enhancement in visual search tasks seems to be goal directed or, to use the vocabulary of the present study, driven by crossmodal attentional control sets.

To summarize, the results of the present study underpin the importance of contingent capture (Folk et al., 1992) as a mechanism of selection across the senses. That is, participants can set-up their attentional control sets for features from different modalities in order to optimize visual but also auditory selection. Many models of attentional selection have focussed on visual attention. Only in recent years has the interplay between multisensory integration and endogenous and exogenous attention been researched (for reviews see e.g., Chen & Spence, in press; Talsma et al., 2010; Tang et al., 2016). The contingent capture approach as one paradigm to analyse top-down attention on the bottom-up processing of stimuli is an intriguing way also to analyse multisensory integration. Humans act in multisensory contexts and thus can set their attentional control sets to ‘expect’ multisensory events.

REFERENCES

- Ansorge, U., & Becker, S. I. (2014). Contingent capture in cueing: The role of color search templates and cue-target color relations. *Psychological Research*, 78, 209-221.
- Ansorge, U., & Heumann, M. (2003). Top-down contingencies in peripheral cuing: The roles of color and location. *Journal of Experimental Psychology: Human Perception and Performance*, 29, 937-948.
- Ansorge, U., & Heumann, M. (2004). Peripheral cuing by abrupt-onset cues: The influence of color in S-R corresponding conditions. *Acta Psychologica*, 116, 115-143.
- Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in Cognitive Sciences*, 16, 437-443.
- Bundesen, C. (1990). A theory of visual attention. *Psychological Review*, 97, 523-547.
- Bundesen, C., Kyllingsbaek, S., Houmann, K. J., & Jensen, R. M. (1997). Is visual attention automatically attracted by one's own name? *Perception & Psychophysics*, 59, 714-720.
- Burnham, B. R. (2007). Displaywide visual features associated with a search display's appearance can mediate attentional capture. *Psychonomic Bulletin & Review*, 14, 392-422.

- Chen, Y.-C., & Spence, C. (2016). Hemispheric asymmetry: A novel signature of attention's role in multisensory integration. *Psychonomic Bulletin & Review*. DOI: [10.3758/s13423-016-1154-y](https://doi.org/10.3758/s13423-016-1154-y)
- Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Human Perception & Performance*, 18, 1030-1044.
- Folk, C. L., Remington, R. W., & Wright, J. H. (1994). The structure of attentional control: Contingent attentional capture by apparent motion, abrupt onset, and color. *Journal of Experimental Psychology: Human Perception & Performance*, 20, 317-329.
- Gibson, B. S., & Kelsey, E. M. (1998). Stimulus-driven attentional capture is contingent on attentional set for displaywide visual features. *Journal of Experimental Psychology: Human Perception and Performance*, 24, 699-706.
- Goller, F., & Ansorge, U. (2015). There is more to trial history than priming in attentional capture experiments. *Attention, Perception, & Psychophysics*, 77, 1574-1584.
- 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, 1736-1741.
- Iordanescu, L., Guzman-Martinez, E., Grabowecky, M., & Suzuki, S. (2008). Characteristic sounds facilitate visual search. *Psychonomic Bulletin & Review*, 15, 548-554.
- Jonides, J. (1980). Towards a model of the mind's eye's movement. *Canadian Journal of Psychology*, 34, 103-112.

## AUDIOVISUAL CONTINGENT CAPTURE

- Kiesel, A., Kunde, W., & Hoffmann, J. (2007). Mechanisms of subliminal response priming. *Advances in Cognitive Psychology*, 3, 307-315.
- Lamy, D. F., & Kristjánsson, Á. (2013). Is goal-directed attentional guidance just intertrial priming? A review. *Journal of Vision*, 13, 1-19.
- Mast, F., & Frings, C. (2014). The impact of the irrelevant: The task environment modulates the impact of irrelevant features in response selection. *Journal of Experimental Psychology: Human Perception & Performance*, 40, 2198-2213.
- Mast, F., Frings, C., & Spence, C. (2014). Response interference in touch, vision, and crossmodally: Beyond the spatial dimension. *Experimental Brain Research*, 232, 2325-2336.
- Mast, F., Frings, C., & Spence, C. (2015). Multisensory top-down sets: Evidence for contingent crossmodal capture. *Attention, Perception, & Psychophysics*, 77, 1970-1985.
- Matusz, P. J., & Eimer, M. (2011). Multisensory enhancement of attentional capture in visual search. *Psychonomic Bulletin & Review*, 18, 904-909.
- Matusz, P. J., & Eimer, M. (2013). Top-down control of audiovisual search by bimodal search templates. *Psychophysiology*, 50, 996-1009.
- Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, 11, 56-60.



## AUDIOVISUAL CONTINGENT CAPTURE

- Murray, M. M., Thelen, A., Thut, G., Romei, V., Martuzzi, R., & Matusz, P. J. (2016). The multisensory function of the human primary visual cortex. *Neuropsychologica*, 83, 161-169.
- Neumann, O., & Klotz, W. (1994). Motor responses to nonreportable, masked stimuli: Where is the limit of direct parameter specification? In C. Umiltà & M. Moscovitch (Eds.), *Attention and performance 15: Conscious and nonconscious information processing* (pp. 123-150). Cambridge, MA: MIT Press.
- O'Brien, R. G., & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin*, 97, 316-333.
- Pashler, H. E. (1998). *The psychology of attention*. Cambridge, MA: MIT Press.
- Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32, 3-25.
- Spence, C. (2010). Crossmodal spatial attention. *Annals of the New York Academy of Sciences*, 1191, 182-200.
- Spence, C. (2013). Just how important is spatial coincidence to multisensory integration? Evaluating the spatial rule. *Annals of the New York Academy of Sciences*, 1296, 31-49.
- Spence, C. [J.], & Driver, J. (1994). Covert spatial orienting in audition: Exogenous and endogenous mechanisms. *Journal of Experimental Psychology: Human Perception and Performance*, 20, 555-574.

## AUDIOVISUAL CONTINGENT CAPTURE

Spence, C., & Driver, J. (1997). Audiovisual links in exogenous covert spatial orienting. *Perception & Psychophysics*, 59, 1-22.

Spence, C., & Driver, J. (Eds.). (2004). *Crossmodal space and crossmodal attention*. Oxford, UK: Oxford University Press.

Spence, C., Nicholls, M. E., Gillespie, N., & Driver, J. (1998). Cross-modal links in exogenous covert spatial orienting between touch, audition, and vision. *Perception & Psychophysics*, 60, 544-557.

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.

Tang, X., Wu, J., & Shen, Y. (2016). The interactions of multisensory integration with endogenous and exogenous attention. *Neuroscience and Behavioral Reviews*, 61, 208-224.

Theeuwes, J. (1991). Exogenous and endogenous control of attention: The effect of visual onsets and offsets. *Perception & Psychophysics*, 49, 83-90.

Theeuwes, J. (2010). Top-down and bottom-up control of visual selection. *Acta Psychologica*, 135, 77-99.

Thelen A., Matusz P. J., & Murray M. M. (2014). Multisensory context portends object memory. *Current Biology*, 24, 734-735.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12, 97-136.

## AUDIOVISUAL CONTINGENT CAPTURE

Tukey, J. W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.

van der Burg E., Olivers C. N., Bronkhorst A. W., Theeuwes J. (2008). Pip and pop: Nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 34, 1053-1065.

Wolfe, J. M. (1994). Guided search 2.0. A revised model of visual search. *Psychonomic Bulletin & Review*, 1, 202-238.

Wolfe, J. M. (2007). Guided search 4.0: Current progress with a model of visual search. In W. Gray (Ed.), *Integrated models of cognitive systems* (pp. 99-119). Oxford, UK: Oxford University Press.

Yantis, S. & Jonides, J. (1984). Abrupt visual onsets and selective attention: Evidence from visual search. *Journal of Experimental Psychology: Human Perception & Performance*, 10, 601-621.

Yuval-Greenberg, S. and Deouell L.Y. The Dog's Meow: Asymmetrical Interaction in Cross-Modal Object Recognition (2009). *Experimental Brain Research*, 193, 603-614.

## AUDIOVISUAL CONTINGENT CAPTURE

*Table 1.* Mean RTs (Reaction Times; in ms) and error rates (in brackets; in percentages) as a function of the response compatibility (incompatible vs. compatible) and auditory congruency (congruent vs. incongruent) in Experiment 1.

	Auditory congruency	
	Congruent sound	Incongruent sound
Incompatible	609 (11.5)	606 (11.0)
Compatible	466 (3.9)	486 (4.0)
Compatibility effect	142 (7.5)	119 (7.0)

## AUDIOVISUAL CONTINGENT CAPTURE

*Table 2.* Mean RTs (Reaction Times; in ms) and error rates (in brackets; in percentages) as a function of response compatibility (incompatible vs. compatible) and auditory congruency (congruent vs. incongruent) in Experiment 2.

	Visual congruency	
	Congruent figure	Incongruent figure
Incompatible	709 (11.8)	707 (11.7)
Compatible	554 (2.7)	591 (2.5)
Compatibility effect	155 (9.1)	116 (9.2)

## FIGURE LEGENDS

*Figure 1.* Sequence of events for trials in Experiment 1. In the response compatible trials, the target and distractor were both presented in the same color whereas, in the response incompatible trials, the visual stimuli were presented in different colors. The visual target was always accompanied by the same auditory stimulus whereas the auditory stimulus presented during the visual distractor varied randomly between trials. Thus, in the auditory congruent trials, both visual stimuli were accompanied by the same sound whereas, in the auditory incongruent trials, they were accompanied by different sounds. The stimuli are not drawn to scale.

*Figure 2.* The figure depicts the logic of why a difference in the size of the compatibility effects for auditory congruent versus incongruent distractors should be found. It is assumed that the participants set-up their attentional control sets for target color and the accompanying sound stimulus (the sound frequency). The feature match between the distractor and the top-down sets varies as a function of the distractor sound. A visual distractor with a congruent auditory stimulus matches both the color and frequency of the attentional control set. By contrast, a visual distractor accompanied by an incongruent sound only matches the color of the attentional control set but not its frequency. Finally, the feature match predicts the size of the resulting compatibility effects. See the text for further details.

## AUDIOVISUAL CONTINGENT CAPTURE

*Figure 3.* Mean reaction time (in ms; on the left side) and error rate (in %; on the right side) compatibility effects in Experiment 1 as a function of auditory congruency. The error bars depict the standard error of the means. Note that the comparison of the compatibility effects in the congruent versus incongruent conditions reflects the interaction of compatibility x congruency reported in the text.

*Figure 4.* Mean reaction time (in ms) compatibility effects in Experiments 2 as a function of auditory congruency (congruent vs. incongruent vs. no-sound). The error bars depict the standard error of the means. Note that the comparison of the compatibility effects in the congruent versus incongruent conditions reflects the interaction of compatibility x congruency reported in the text.

Figure 1.

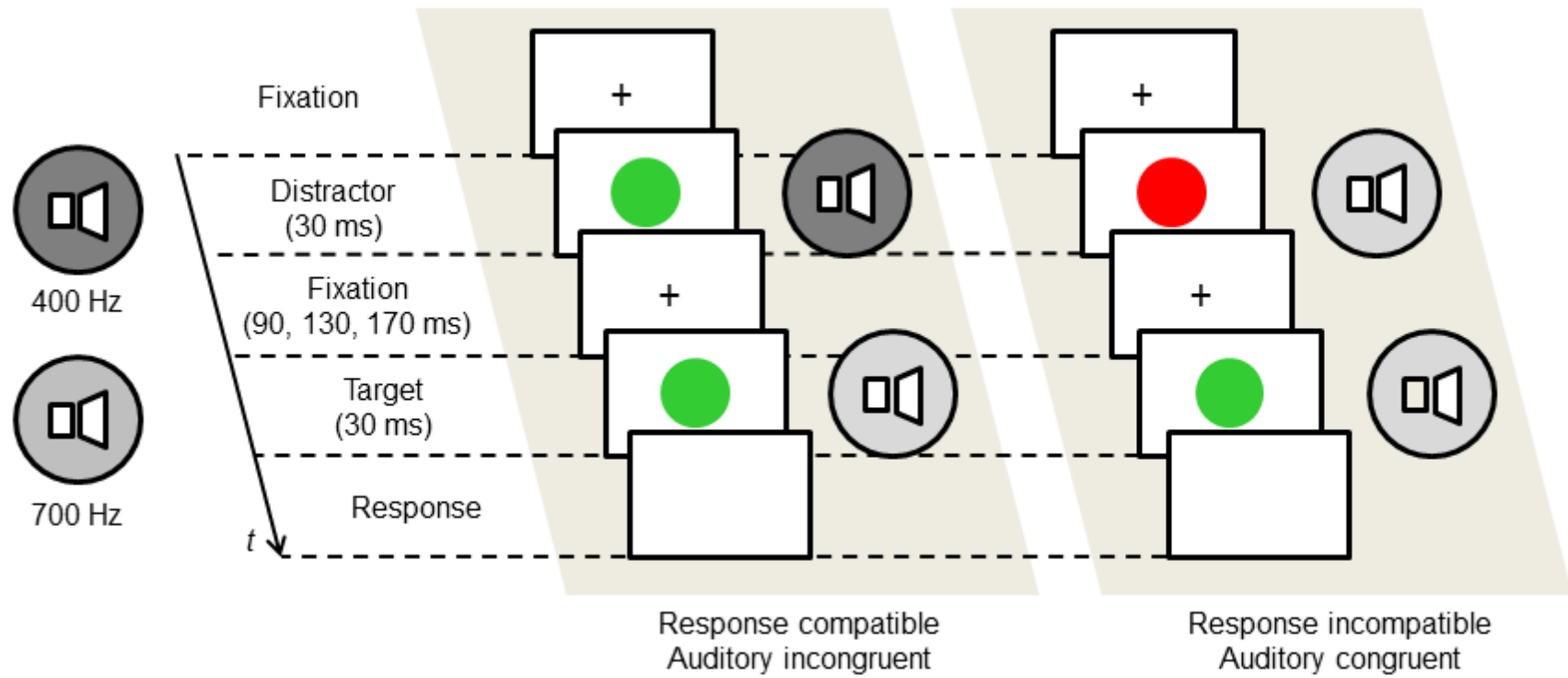




Figure 2.

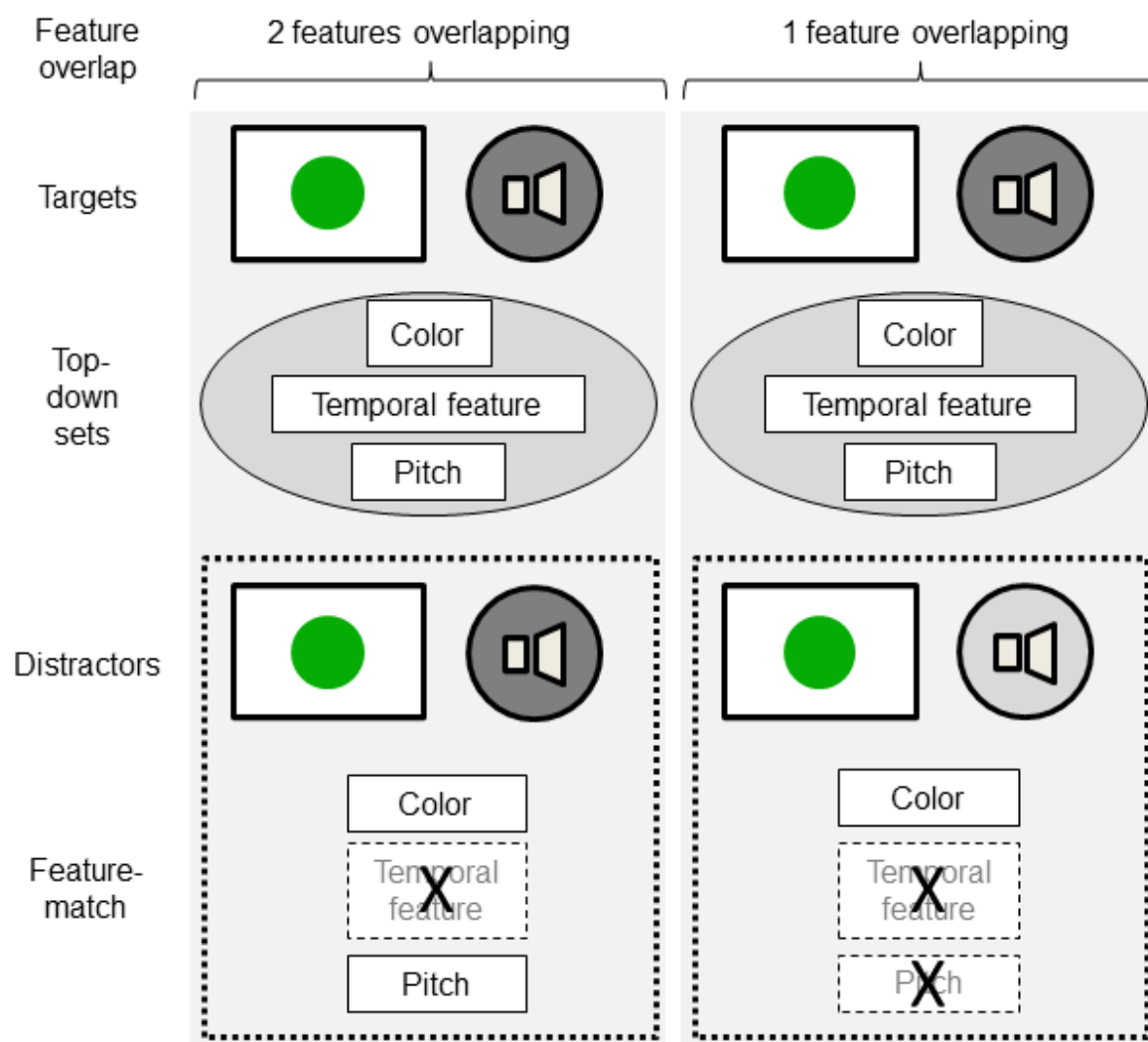


Figure 3.

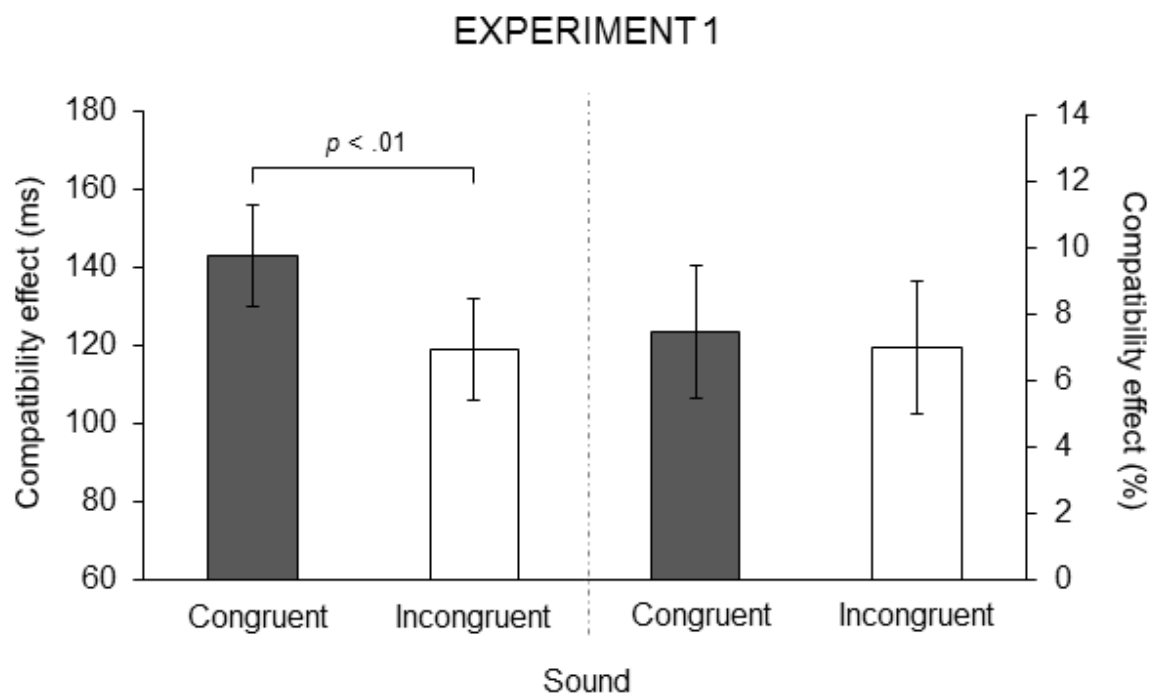


Figure 4.

