A MATTER OF BOTTOM-UP OR TOP-DOWN PROCESSES: THE ROLE OF ATTENTION IN MULTISENSORY INTEGRATION

EDITED BY: Jess Hartcher-O'Brien, Salvador Soto-Faraco and Ruth Adam
PUBLISHED IN: Frontiers in Integrative Neuroscience

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88945-193-7
DOI 10.3389/978-2-88945-193-7

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

A MATTER OF BOTTOM-UP OR TOP-DOWN PROCESSES: THE ROLE OF ATTENTION IN MULTISENSORY INTEGRATION

Topic Editors:
Jess Hartcher-O'Brien, Delft University of Technology, Netherlands
Salvador Soto-Faraco, Institució Catalana de Recerca i Estudis Avançats & Universitat Pompeu Fabra, Spain
Ruth Adam, Ludwig-Maximilians-Universität (LMU), Germany

The integration of information from various sensory modalities influences behaviour. It can induce behavioural benefits such as faster reaction times and enhanced detection of noisy signals but may also produce illusions, all of which have been characterized by specific neuronal signatures. Yet, while these effects of multisensory integration are largely accepted, the role of attention in this process is still the object of intense debate. On the one hand, it has been suggested that attention may guide multisensory integration in a top-down fashion by selection of specific inputs to be integrated out of the plethora of information in our environment. On the other hand, there is evidence that integration could occur in a bottom-up manner, based on temporal and spatial correlations, and outside the focus of attention. An extreme example is the multisensory enhancement of neural responses in anesthetised animals.

Attention itself is not a unitary construct, and may refer to a range of different selection mechanisms. Therefore, the interplay between attention and multisensory integration can take many forms which explain, in part, the diversity of findings and the disputes in the literature.

The goal of this Research Topic is to help clarify the picture by trying to answer the following questions from various perspectives: Under which circumstances does multisensory integration take place without attention?, and, When does attention determine the fate of multisensory integration?

Citation: Hartcher-O'Brien, J., Soto-Faraco, S., Adam, R., eds. (2017). A Matter of Bottom-Up or Top-Down Processes: The Role of Attention in Multisensory Integration. Lausanne: Frontiers Media. doi: 10.3389/978-2-88945-193-7

Table of Contents

frontiers in Integrative Neuroscience

EDITORIAL
published: 28 February 2017
doi: 10.3389/fnint.2017.00005

Editorial: A Matter of Bottom-Up or Top-Down Processes: The Role of Attention in Multisensory Integration

Jess Hartcher-O'Brien[1], *Salvador Soto-Faraco*[2,3] *and Ruth Adam*[4]*

[1] *Perceptual Intelligence Lab, Industrial Design Engineering, Delft University of Technology, Delft, Netherlands,* [2] *Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain,* [3] *Departament de Tecnologies de la Informació i les Comunicacions, Universitat Pompeu Fabra, Barcelona, Spain,* [4] *Institute for Stroke and Dementia Research, Klinikum der Universität München, Ludwig-Maximilians-Universität (LMU), Munich, Germany*

Keywords: attention, bottom-up, congruency, cross-modal, multisensory integration, top-down

Editorial on the Research Topic

A Matter of Bottom-Up or Top-Down Processes: The Role of Attention in Multisensory Integration

Our everyday environments are multisensory and our brains handle this rich information in an extremely efficient way. Yet, attention's role in the process of multisensory integration (MSI) is still the object of intense debate. Whilst some evidence supports that attention guides MSI via top-down selection of inputs, others suggest that bottom-up integration can occur pre-attentively capitalizing on temporal and spatial correlations. Understanding the role of attention in MSI is further complicated by the fact that attention itself refers to a variety of different selection mechanisms. Thus, the interplay between attention and MSI can take many forms and lead, as evident in the literature, to mixed findings and apparent contradictions (e.g., Driver, 1996; Talsma et al., 2010; Jack et al., 2013; Macaluso et al., 2016). This *Frontiers in Integrative Neuroscience* Research Topic aims at helping clarify the nature of this interplay by posing a specific and narrow question. The reader will find a collection of 10 empirical papers plus an opinion and a review article which can broadly be classified into those addressing the contribution of bottom-up processing in MSI, and those exploring top-down modulations.

One framework for exploring the interplay between attention and MSI is to assume that attention leads to reweighting of sensory information (Bresciani and Ernst, 2007). Focusing on *bottom-up* contributions, Vercillo and Gori address such potential reweighting using the Maximum Likelihood Estimate model. The effect of attention on the weighting of sensory information could be disentangled by measuring observer's audio-tactile spatial estimates, showing that bottom-up attention increases precision and alters sensory weighting in MSI. These findings corroborate selective attention's role in adjusting the brain's computations for achieving an integrated multisensory percept. Keeping the bottom-up perspective, Hazan et al. address visual search in the barn owl. Similar to ventriloquism (e.g., Pick et al., 1969) and visual search in humans (Onat et al., 2007), the owls' visual search behavior was modulated by sound, demonstrating that audio-visual interactions guided visual attention. Visual search mechanisms might be similar among mammalian and non-mammalian species, owing to correlations between visual and auditory events in nature. An ultimate demonstration of the effect of bottom-up processes would consist in showing MSI for sub-threshold stimuli, in the absence of top-down biases. Aller et al. take a step in this direction showing that the visibility of a visual event under continuous flash suppression (CFS) increases when a sound is congruent (instead of incongruent). Albeit, as the authors argue, possible top-down processes may still exert an influence, the CFS framework provides a

Edited and reviewed by:
Sidney A. Simon,
Duke University, USA

***Correspondence:**
Ruth Adam
ruth.adam@med.uni-muenchen.de

Received: *27 January 2017*
Accepted: *13 February 2017*
Published: *28 February 2017*

Citation:
Hartcher-O'Brien J, Soto-Faraco S and Adam R (2017) Editorial: A Matter of Bottom-Up or Top-Down Processes: The Role of Attention in Multisensory Integration. Front. Integr. Neurosci. 11:5. doi: 10.3389/fnint.2017.00005

clear conceptualization of the question of bottom-up versus top-down processes. Jones' study explores both attentional cuing via bottom-up temporal entrainment and spatial cuing of attention in unisensory and cross-modal events. Both temporal and spatial attention-MSI interactions facilitated behavioral responses: attention produced a response advantage when deployed in a bottom-up temporal-cuing fashion and via top-down spatial attention manipulations. However, there was no measurable interaction between the bottom-up and top-down processes observed.

This research topic also includes two review/opinion papers with different views on bottom-up MSI (De Meo et al.; Talsma). De Meo et al. interpret expressions of early multisensory interaction as *integration*. Such that integration phenomena are irreducible, albeit top-down control processes can regulate their expression. Talsma et al. instead argue that cross-modal interactions that take place early, requiring no role of attention, do not result in *integration*. This controversy suggests that we may be missing crucial evidence, or are looking at extant evidence from incongruous angles. Defining what is meant by "integration" would already be an important step in the right direction.

Despite the attempts to find core, bottom-up MSI interactions, top-down attentional components may also determine the outcome of MSI (e.g., Aller et al.). Whether these influences are general, or confined to specific contexts, is still a matter of debate. This Research Topic includes five articles that have identified *top-down* influences employing various manipulations of multisensory *congruency*.

In an attempt to disentangle bottom-up versus top-down contributions Donohue et al. manipulate attentional load and observer goals. Audio-visual binding in the bounce-stream paradigm was modulated by spatial cueing, suggesting that attention alters temporal binding of audio-visual signals in this task. Attention produced a response advantage when deployed in a bottom-up temporal-cuing approach and via the top-down spatial attention manipulation. However, similar to Jones' conclusion (Jones), there was no measurable interaction between bottom-up and top-down processes. Employing an audio-visual congruency manipulation with the attentional blink paradigm, Adam and Noppeney could show that task-irrelevant sounds influence detection of, and awareness to, a visual target. Increased awareness of visual inputs was based not only on the congruency of current sensory evidence but also on prior knowledge, hinting that top-down expectations affect decisions regarding multisensory events and enhance integration. Mastroberardino et al. addresses whether task-irrelevant stimuli modulate cross-modal processing of semantically-congruent cues, by neutralizing low-level contributions. Consistent with the idea of extensive processing of cross-modal semantic relations, their fMRI results reveal that semantic-congruency engages fronto-parietal networks related to visuo-spatial control. Consequently, one could think of semantic congruency as providing a bias signal that exerts influence (yet not dominance) on the competitive interplay between bottom-up and top-down processes for the control of processing resources. Once one accepts that top-down influences are pervasive in MSI, the question of content-dependency arises. Su's study explores to what extent content congruency will determine low levels of information processing in MSI and illustrates that audio-visual correspondence relations derived from human movements exert an important influence on auditory deviant detection and even on cross-modal synchrony perception.

The relation between attention and MSI further increases in complexity when manipulating stimulus-elicited *emotions*. Only a few studies have investigated multisensory emotion processing, despite the importance of both emotions and MSI to adaptive behavior. Takagi et al. establish that attentional instructions and audio-visual congruency modulate sensory dominance in emotion processing. This study highlights how important it is to provide participants with detailed and clear instructions when characterizing MSI-attention interactions. Finally, Doose-Grünefeld et al. find no direct relationship between MSI and attention. Their study also investigates MSI of emotional signals, yet in patients with depression. The patients rated faces as more fearful when displayed with happy sounds and appeared impaired in processing positive auditory information even when task-irrelevant. Neurocognitive tests revealed that those patients had impaired attention, which was not related to their emotion perception. Thus, impaired attention cannot directly explain deficits in multisensory (emotional) processing.

CONCLUSION AND WHERE DO WE GO FROM HERE

The work presented in this Research Topic demonstrates that the relation between MSI and attention is complex and unlikely to be answered by one single study. By bringing together these diverse works we observe that stimulus context effects, such as spatial/temporal co-location (e.g., Hazan et al.) or semantic (e.g., Mastroberardino et al.) and emotional congruency (e.g., Takagi et al.), as well as the goal of the observer, such as changing task for similar stimuli (e.g., Donohue et al.; Jones) tend to characterize whether MSI will be modulated by top-down attentional effects (e.g., Adam and Noppeney; Mastroberardino et al.; Talsma) or will seem to occur preattentively (e.g., Aller et al.; De Meo et al.; Hazan et al.; Su; Vercillo and Gori). It is fair to say that the interplay depends on many factors and, in some situations, involves no direct relation between attention and MSI (e.g., Doose-Grünefeld et al.). Clearer and universally agreed definitions would limit the same results being used for different perspectives on the debate of attention's role in MSI. Future research using standardized instructions and experimental designs, e.g. CFS, controlling for either bottom-up or top-down influences (or both) across different contexts and observer goals would help get closer to a resolution of this ongoing debate.

AUTHOR CONTRIBUTIONS

JH, SS, and RA co-editored the Research Topic and wrote the editorial.

REFERENCES

Bresciani, J. P., and Ernst, M. O. (2007). Signal reliability modulates auditory–tactile integration for event counting. *Neuroreport* 18, 1157–1161. doi: 10.1097/WNR.0b013e3281ace0ca

Driver, J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature* 381, 66–68. doi: 10.1038/381066a0

Jack, B. N., O'Shea, R. P., Cottrell, D., and Ritter, W. (2013). Does the ventriloquist illusion assist selective listening? *J. Exp. Psychol. Hum. Percept. Perform.* 39, 1496–1502. doi: 10.1037/a0033594

Macaluso, E., Noppeney, U., Talsma, D., Vercillo, T., Hartcher-O'Brien, J., and Adam, R. (2016). The curious incident of attention in multisensory integration: bottom-up vs. top-down. *Multisens. Res.* 29, 557–583. doi: 10.1163/22134808-00002528

Onat, S., Libertus, K., and König, P. (2007). Integrating audiovisual information for the control of overt attention. *J. Vis.* 7:11. doi: 10.1167/7.10.11

Pick, H. L. Jr., Warren, D. H., and Hay, J. C. (1969). Sensory conflict in judgments of spatial direction. *Percept. Psychophys.* 6, 203–205. doi: 10.3758/BF03207017

Talsma, D., Senkowski, D., Soto-Faraco, S., and Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends Cogn. Sci.* 14, 400–410. doi: 10.1016/j.tics.2010.06.008

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Integrative Neuroscience

ORIGINAL RESEARCH
published: 07 May 2015
doi: 10.3389/fnint.2015.00034

Attention to sound improves auditory reliability in audio-tactile spatial optimal integration

Tiziana Vercillo and Monica Gori*

Robotics, Brain, and Cognitive Sciences Department, Fondazione Istituto Italiano di Tecnologia, Genoa, Italy

The role of attention on multisensory processing is still poorly understood. In particular, it is unclear whether directing attention toward a sensory cue dynamically reweights cue reliability during integration of multiple sensory signals. In this study, we investigated the impact of attention in combining audio-tactile signals in an optimal fashion. We used the Maximum Likelihood Estimation (MLE) model to predict audio-tactile spatial localization on the body surface. We developed a new audio-tactile device composed by several small units, each one consisting of a speaker and a tactile vibrator independently controllable by external software. We tested participants in an attentional and a non-attentional condition. In the attentional experiment, participants performed a dual task paradigm: they were required to evaluate the duration of a sound while performing an audio-tactile spatial task. Three unisensory or multisensory stimuli, conflictual or not conflictual sounds and vibrations arranged along the horizontal axis, were presented sequentially. In the primary task participants had to evaluate in a space bisection task the position of the second stimulus (the probe) with respect to the others (the standards). In the secondary task they had to report occasionally changes in duration of the second auditory stimulus. In the non-attentional task participants had only to perform the primary task (space bisection). Our results showed an enhanced auditory precision (and auditory weights) in the auditory attentional condition with respect to the control non-attentional condition. The results of this study support the idea that modality-specific attention modulates multisensory integration.

Keywords: attention, multisensory integration, auditory, bayes theorem, sensory cue

OPEN ACCESS

Edited by:
Jessica Hartcher-O'Brien, l'Ecole Normale Superieuer, France

Reviewed by:
John S. Butler, Trinity College Dublin, Ireland
Yoshiyuki Sato, University of Electro-Communications, Japan
Kielan Yarrow, City University London, UK

****Correspondence:***
Tiziana Vercillo, Robotics, Brain, and Cognitive Sciences Department, Fondazione Istituto Italiano di Tecnologia, Via Morego 30, 16163, Genoa, Italy
tiziana.vercillo@iit.it

Received: *30 September 2014*
Accepted: *22 April 2015*
Published: *07 May 2015*

Citation:
Vercillo T and Gori M (2015) Attention to sound improves auditory reliability in audio-tactile spatial optimal integration. Front. Integr. Neurosci. 9:34. doi: 10.3389/fnint.2015.00034

Introduction

Spatio-temporal coincident sensory signals are combined together to generate multisensory percepts. Sensory information is weighted accordingly to its reliability and integrated in a statistically optimal fashion (Clarke and Yuille, 1990; Ghahramani et al., 1997; Ernst and Banks, 2002; Alais and Burr, 2004; Landy et al., 2011). Although years of intensive studies have produced a wide body of research on the topic of multisensory integration, it is still unclear whether or not attended stimuli are integrated differently from those that are not attended. Specifically, it is not clear whether the mechanism of multisensory integration occurs automatically and pre-attentively or whether attention affects the sensory binding. Several studies support the first idea, reporting differences in the perceptual estimates when people attend to one or another sensory modality in a multisensory task (Bertelson and Radeau, 1981; Warren et al., 1981). For example, Oruc et al. (2008)

demonstrated that crossmodal dynamic ventriloquism (Soto-Faraco et al., 2002), the illusory reversal in the perceived direction of motion of a target modality induced by the opposite motion direction of a distractor modality, can be affected by modality-specific attention. Similarly in another study, Alsius et al. (2005) reported that the audio-visual McGurk illusion is powerfully reduced when participants perform a concurrent auditory or visual task, suggesting that the high attentional load precludes multisensory processing.

Differently, other studies found that attention has no effect on multisensory integration, supporting the idea that sensory cues are combined pre-attentively. For example, Driver (1996) showed that the ventriloquist cross-modal illusion can enhance selective spatial attention to speech sounds, suggesting that the multisensory binding has to occur before the auditory attentive selection. Furthermore, other studies suggest that there are no effects of endogenous (Bertelson et al., 2000) and automatic visual attention (Vroomen et al., 2001) on audio-visual ventriloquism. Bertelson et al. (2000) reported no effect of attention when participants had to localize the apparent source of a sound presented with a synchronous peripheral flash while monitoring occasional slight changes in shape of a visual target in a central or in a peripheral position, supporting the idea that multisensory integration is a pre-attentive process (Driver, 1996; Vroomen et al., 2001).

Although a great deal of consideration has been paid to the effect of attention on multisensory processing, there is much less effort directed to quantify such effects with the Maximum Likelihood Estimation (MLE) model. Helbig and Ernst (2008) have recently investigated the effects of modality-specific attention on multisensory optimal integration, adopting a dual task paradigm. Participants were asked to evaluate similarities or differences between two sequences of letters while performing a visual-haptic size discrimination task. Participants' performance was later compared to an ideal observer (MLE model) to test for optimal integration. Results showed no effect of modality-specific attention on visual-haptic optimal integration, sustaining the hypothesis that the mechanism of integration is pre-attentive. Visual and tactile weights were untouched by the distractor task. Furthermore, the bimodal JNDs, although increased in the dual task condition, were still lower than both of the unisensory JNDs, as predicted by the MLE model.

Interestingly, the distractor task used by Helbig and Ernst (2008), and by several other studies to date (Alsius et al., 2005) involved the use of stimuli with qualitatively different properties from those used in the primary task. A possible reason for the absence of attentional effects on multisensory integration could be that the simultaneous encoding of qualitatively different stimuli (e.g., size vs. letter) increases the attentional load, rather than focusing attention on a sensory modality.

Here we examined the attentional modulation of multisensory integration in a dual task where the same stimulus had to be evaluated twice. Participants were asked to execute an acoustic temporal discrimination task while performing an audio-tactile spatial bisection task. Recent researches reported that audition and touch can interact pre-attentively. Butler et al. (2012) demonstrated audio-tactile pre-attentive interaction at the cortical level during frequency processing. Yau et al. (2009) reported that auditory stimuli can interfere with tactile frequency perception when auditory and tactile stimuli share similar frequencies. Of greater interest for our study is that audio-tactile integration seems to vary according to the perceptual task that participants have to perform. Yau et al. (2010) reported separate integration mechanisms for audio-tactile interactions in frequency and intensity perception. While the effects of sensory capture appear to be stronger and pre-attentive for frequency perception, suggesting shared processing for spectral analysis, audio-tactile interactions for intensity discrimination depend on the attended modality.

We investigated the effect of attention on auditory precision and multisensory optimal processing when participants had to simultaneously evaluate an auditory stimulus in two different domains (temporal and spatial) while integrating it with a tactile signal in the spatial (and not in the temporal) domain. We expected that the simultaneous estimation of multiple characteristics of the same stimulus may affect its reliability during multisensory integration. Moreover we compared the performance of all the participants with an optimal estimator.

Methods and Procedures

Participants

Ten adults (28 ± 1 years of age) participated at experiment. All of them had normal hearing. Participants were blindfolded before entering the room, so they had no notion of the experimental setup. All participants signed informed consents before starting the experiment. Testing procedures were approved by the ASL3 of Genoa (Italy).

Stimuli

For the audio-tactile stimulation we developed a device composed by 9 units which could be controlled individually. Units were separated by 3.5 cm (11° of visual angle). Each unit was composed by a speaker producing a 2978 Hz pure tone associated to a 2V vibrating motor (**Figure 1**). The vibrotactile motors produced tactile stimulation of 120 Hz, with vibration amplitude of 0.55 G.

Procedures

Participants placed their right arm on a support at the eyes' level, at a distance of 18 cm from their eyes. The device was positioned on the forearm, with the 5th unit (the middle of the array) aligned with the nose, the 1st unit close to the hand (the left side of participants' head) and the 9th unit close to elbow (at the right side of participants' head). Participants wore acoustic earmuffs (Howard leight, Viking™ V1) during all the experiment, to attenuate the noise emitted by the vibrotactile stimulator while hearing sounds at ordinary volumes and frequencies normally.

Two tests were performed. In the non-attentional condition we measured discrimination thresholds and PSEs in a spatial bisection task. Only-audio, only-tactile and audio-tactile stimuli were provided. For each trial, we presented a sequence of three stimuli (auditory, tactile or both) for a total duration of 1.7 s, with the second and the third stimuli occurring always 600 ms after

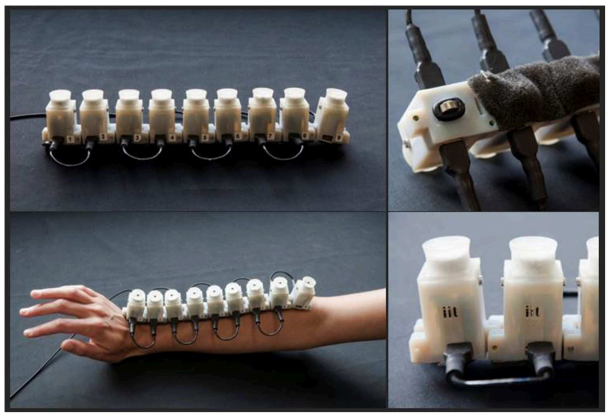


FIGURE 1 | Images from the device used during the test. It is composed by 9 units constituted by a speaker and a vibrating motor individually controllable. Units were separated by 3.5 cm. The device was located on the forearm, with the the 1st unit close to the hand and the 9th unit close to elbow.

the onset of the previous stimulus. The duration of the auditory and the tactile stimuli was 500 ms. The locations of the first and the third stimulus (standard stimuli) of the sequence were fixed at the 1st (−14 cm) and the 9th (+14 cm) units, respectively, while the location of the second stimulus was controlled by the adaptive QUEST algorithm (Watson and Pelli, 1983). The QUEST algorithm estimates PSE after each response and places the next trial near that estimate. To ensure that a wide range of positions was sampled, that estimate was jittered by a random amount, drawn from a Gaussian distribution of space constant 10 cm, and the nearest unit to that estimate chosen. We will refer to the position estimated by the QUEST algorithm as the "probe."

In the unisensory tasks participants were presented with a sequence of three vibrotactile stimulations or sounds and participants had to report whether the second stimulus appeared closer in space to the first or the third stimulus. The second auditory or tactile stimulus was placed at the position estimated by the QUEST algorithm (the probe). In the bimodal task, the sequence of three vibrotactile signals was associated to three sounds. In this last condition, the second stimulus could have been presented in conflict with auditory and tactile stimuli located in different positions and at different distances from the probe. The audio-tactile conflict (Δ) was calculated as S_A–S_T, with S_A and S_T representing the spatial distance of the auditory and the tactile stimuli with respect to the probe (see Alais and Burr, 2004; Gori et al., 2012). In the no-conflict condition ($\Delta = 0$ cm), the location of the auditory and the tactile stimulus corresponded to the probe. In the conflict conditions ($\Delta = \pm 7$ cm), auditory and tactile stimuli were presented at ± 3.5 cm from the probe. For example if the probe was 0 (the fifth unit, the center of the device), in the $\Delta = +7$ cm condition the sound was located at +3.5 cm and the vibration at −3.5 cm [3.5 − (−3.5) = 7 cm]; conversely, in the $\Delta = -7$ condition the sound was located at −3.5 cm while the vibration at +3.5 cm [−3.5 − (3.5) = −7]. In the case that the probe was estimated in a position outside of the stimulus array, the closest unit to the extreme position was selected. Therefore, the second auditory or tactile stimuli could have been presented also in the two extremes locations. In the first and the third audio-tactile stimulus, the auditory and tactile components were presented aligned, with no spatial conflict.

Participants performed 90 trials for both the unisensory conditions and 90 trials for each conflict in the bimodal condition. Conditions were mixed within each block, and presented in a random order. Data for each condition were fitted with cumulative Gaussians. The proportion of rightward responses was plotted as a function of the speaker position, and the data fitted with a Cumulative Gaussian function by means of the Maximum Likelihood method to estimate both PSE (point of subjective equality, given by the mean) and threshold (standard deviation). The space constant (σ) of the fit was taken as the estimate of threshold indicating precision for the bisection task. Standard errors in the threshold and PSEs were computed with bootstrap simulation (Efron and Tibshirani, 1993). All conflict conditions were used to obtain the bimodal threshold estimates. Despite the audio-tactile conflict the stimulation appeared as a single stimulus; participants did not notice the conflict even when asked. Unimodal and bimodal (conflictual or not) audio-tactile thresholds and PSEs were compared with the prediction of the MLE model.

In the attentional condition we introduced an auditory dual task to focus participants' attention only in the auditory stream. This time in addition to the spatial bisection task, participants were also asked to identify occasionally changes in duration of the second auditory stimulus. The duration of the second sound was manipulated only in the 30% of the trials (catch trials) for each block. The task was extremely easy to perform since the second sound might have been 150 ms longer or shorter than its normal duration and than the other two sounds of the sequence. All these catch trials were excluded from the data analysis. The remaining data for each condition were fitted with cumulative Gaussians. Unimodal and bimodal audio-tactile thresholds and PSEs were compared with the prediction of the MLE model. Participants performed the same amount of trials as they did in the non-attentional condition. The order of the two attentional conditions was counterbalanced across participants.

Maximum Likelihood Model

The MLE calculation assumes that the optimal bimodal estimate of PSE ($\hat{S}_{AT}$) is given by the weighted sum of the independent audio and tactile estimates ($\hat{S}_A$ and $\hat{S}_T$).

$$\hat{S}_{AT} = w_A \hat{S}_A + w_T \hat{S}_T \tag{1}$$

Where weights w_A and w_T sum to unity and are inversely proportional to the variance (σ^2) of the underlying noise distribution, assessed from the standard deviation σ of the Gaussian fit of the psychometric functions for audio and tactile judgments:

$$w_A = \sigma_T^2 / \left(\sigma_T^2 + \sigma_A^2\right),\ w_T = \sigma_A^2 / \left(\sigma_T^2 + \sigma_A^2\right) \tag{2}$$

The MLE prediction for the audio-tactile threshold (σ_{AT}) is given by:

$$\sigma_{AT}^2 = \frac{\sigma_A^2 \sigma_T^2}{\sigma_A^2 + \sigma_T^2} \leq \min\left(\sigma_A^2, \sigma_T^2\right) \quad (3)$$

where σ_A and σ_T are the audio and tactile unimodal thresholds. The improvement is greatest ($\sqrt{2}$) when $\sigma_A = \sigma_T$.

To calculate the audio and tactile weights from the PSEs, we substituted the actual second sound position (relative to standard) into Equation (1):

$$\hat{S}(\Delta) = (w_A \Delta - w_T \Delta) = (1 - 2w_T)\,\Delta \quad (4)$$

The slope of the function is given by the first derivative:

$$\hat{S}(\Delta)' = 1 - 2w_T \quad (5)$$

Rearranging:

$$w_T = \left(1 - \hat{S}(\Delta)'\right)/2 \quad (6)$$

The slope $\hat{S}(\Delta)'$ was calculated by linear regression of PSEs for all values of Δ, separately for each subject and each condition.

Results

Unisensory Tasks

Figures 2A,B show psychometric functions from one representative subject in the only-audio and in the only-tactile condition. Each function describes the proportion of trials where the second stimulus was perceived more on the right for all its spatial locations. Light green and light red curves represent the performance of the subject in the non-attentional task and dark green and dark red curves the performance in the attentional task. The point of subjective equality (PSE) represents the stimulus position that participants judged as more on the right in 50% of the trials. The slopes of the psychometric functions, given by the standard deviations of the best-fitting Gaussian error function, provide an estimate of the precision in the spatial task. The steeper the curve, the higher the precision. We mainly based the statistical analysis on these two measures, as described below. Looking at the thresholds (the slopes of the psychometric functions in **Figures 2A,B**, that are also reported in **Figures 4C,D**) it is clear that in the non-attentional condition, participants performed the tactile task with higher precision than the auditory one [one tailed paired t-test; $t_{(9)} = 2.08$; $P = 0.03$]. Interestingly, we found that the temporal auditory task improves auditory precision in the spatial task [one tailed paired t-test; $t_{(9)} = 1.88$; $P = 0.04$] and declines the tactile precision [one tailed paired t-test;$t_{(9)} = -2.17$; $P = 0.02$]. The improved auditory precision, and the lack of significant difference between auditory and tactile thresholds in the attentional condition [one tailed paired t-test; $t_{(9)} = 1.18$; $P = 0.86$] result in a large enhancement of the predicted auditory weights in the attentional condition with respect to the non-attentional condition [one tailed paired t-test; $t_{(9)} = 5.55$; $P = 0.001$]. **Figure 2C** shows predicted average audio and tactile weights, calculated from all the individual thresholds in the two experimental conditions. Predicted auditory weights in the non-attentional condition were equal to 0.32 ± 0.07 and become equal to 0.57 ± 0.07 in the attentional condition. Interestingly, predicted tactile weights vary from 0.68 ± 0.07, in the non-attentional condition, to 0.43 ± 0.21 in the attentional condition. Following the MLE model (Equation 1) we should expect tactile dominance in the non-attentional condition and auditory dominance or no dominance in the attentional condition.

Bimodal Tasks

Figure 3 reports bimodal psychometric functions from the same representative subject for the three audio-tactile conflicts: $\Delta = 0$, $\Delta = -7$ (dark gray curve), $\Delta = +7$ (light gray curve). The proportion of the "stimulus more on the right" responses

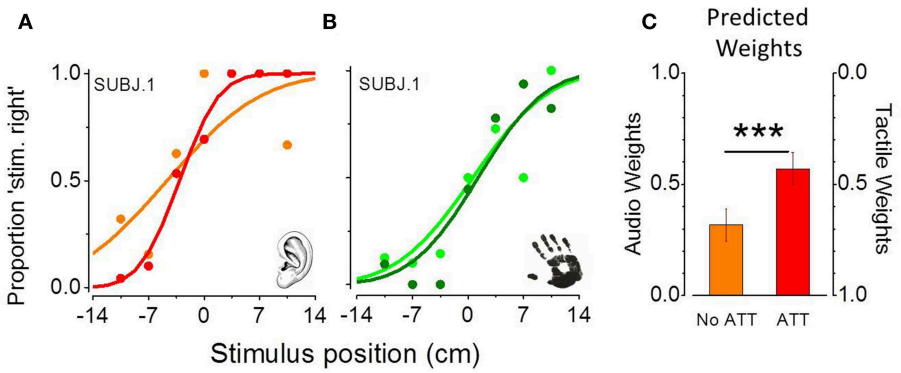


FIGURE 2 | Psychometric functions from one representative subject in the only-audio (A) and the only-tactile condition (B). The curves plot proportion of trials where the second stimulus was perceived more on the right for all its positions. Light green and light red curves represent subject's performance for the non-attentional condition and dark green and dark red curves the performance for the attentional condition. Data were fitted with cumulative Gaussians, to obtain PSE (50%) and threshold (standard deviation of the best-fitting function). Thresholds decrease in the attentional task only for the auditory condition. **(C)** Shows predicted average audio and tactile weights, calculated from individual thresholds in the two experimental conditions. Auditory weights are higher in the attentional condition. (***p-value < 0.001).

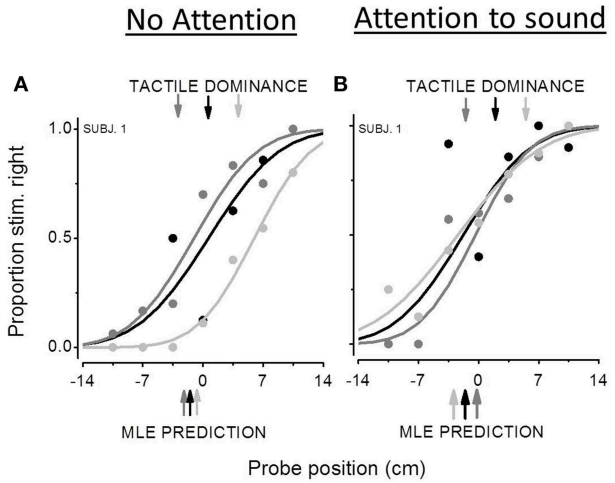


FIGURE 3 | Bimodal psychometric functions from the same participant in the non attentional condition (A) and in the attentional condition (B). The black curve represents subject's performance in the $\Delta = 0$ audio-tactile conflict, the dark gray curve the performance in the $\Delta = -7$ conflict and the light gray curve the $\Delta = +7$ conflict. The x-axis represents the estimated position of the second stimulus by the QUEST algorithm (cm). Positive values of PSE mean a spatial bias to the left, while negative values a bias to the right. The lower color-coded arrows show the MLE prediction. The upper color-coded arrows indicate the predicted PSEs in the case of tactile dominance.

is plotted as a function of the estimated position of the probe (the location calculated by the QUEST algorithm). Positive values of the PSE mean that participants are following the modality presented at −3.5 cm from the probe. For example in the conflict condition of $\Delta = +7$, a positive PSE means that subject are founding their perceptual estimates on the tactile modality. Indeed when the probe is higher than 0, the auditory stimulus is located at +3.5 cm, then closer to the extreme right, while the tactile stimulus is closer to the 0 position. Conversely, negative values of the PSE mean that participants are founding their perceptual estimates on the sensory modality presented at +3.5 cm with respect to the probe. The lower color-coded arrows show the MLE prediction. The upper color-coded arrows indicate the predicted PSEs in the case of tactile dominance. Results from the first experiment (**Figure 3A**) showed poor audio-tactile integration with tactile dominance. In the $\Delta = -7$ condition, when the auditory stimulus is located more on the left than the tactile one, the psychometric curve is shifted toward negative value, denoting a bias to the right in the direction of the tactile stimulus. In this condition, as well as the $\Delta = 0$ condition, measured PSEs are very similar to PSEs predicted by the MLE model. Conversely, in the $\Delta = +7$ condition, the psychometric function is shifted toward positive values implying a bias to the left in the direction of the tactile stimulus and the measured PSE is closer to the one predicted in the case of tactile dominance. In the auditory attentional condition, the bimodal psychometric functions are in the inverted position; however, they are all fairly centered on the 0 confirming that the two sensory modalities share similar weights.

Figures 4A,B show average PSEs for the three bimodal conflicts plotted as a function of the audio-tactile conflict (Δ) in the non-attentional (**Figure 4A**) and attentional (**Figure 4B**) conditions. The two dashed lines describe the ideal performance in the case of auditory (light and dark red) or tactile (light and dark green) dominance. Black line and symbols represent observed PSEs data, gray line and symbols represent the model prediction. As predicted by the model, in the non-attentional condition bimodal PSEs follow the tactile conflict suggesting a tactile dominance. Average PSEs are equal to 0.45 ± 1.26 for the $\Delta = +7$ conflict, 0.33 ± 0.79 for the $\Delta = 0$ conflict and -2.05 ± 0.79 for the $\Delta = -7$ conflict. Predicted PSEs were 2 ± 1.17 for the $\Delta = +7$ conflict, 0.8 ± 1.08 for the $\Delta = 0$ conflict and -0.5 ± 1.22 for the $\Delta = -7$ conflict. We ran a Two-Way ANOVA to study differences between predicted and observed PSEs and between PSEs measured in different conflict conditions. In the non-attentional condition, we found a significant effect of the conflict [$F_{(2, 54)} = 3.39$; $P = 0.04$], but no significant differences between predicted and observed PSEs [$F_{(1, 54)} = 2.11$; $P = 0.15$] and no interaction between the two factors [$F_{(2, 54)} = 0.20$; $P = 0.81$]. In the attentional condition, we found no differences between PSEs across conflicts [$F_{(2, 54)} = 0.11$; $P = 0.88$], between predicted and observed PSEs [$F_{(1, 54)} = 1.55$; $P = 0.21$] and no interaction between the two factors [$F_{(2, 54)} = 0.53$; $P = 0.58$]. The effect of the conflict that we have found in the non-attentional condition confirms that participants founded their perceptual judgment mainly on one sensory modality. Additionally, the lack of differences between predicted and observed PSEs implies a good prediction from the MLE model. **Figures 4C,D** show the average thresholds for the audio (A), tactile (T) and audio-tactile (AT) estimates as well as the predicted bimodal thresholds. Since individual bimodal thresholds were similar across the three AT conflicts, for both the non-attentional [repeated measure ANOVA; $F_{(2, 27)} = 1.03$; $P = 0.37$] and the attentional condition [repeated measure ANOVA; $F_{(2, 27)} = 2.37$; $P = 0.11$], we calculated average bimodal thresholds for each participant.

For the non-attentional condition (**Figure 4C**), we compared unimodal and bimodal observed and predicted thresholds in a One-Way ANOVA and found significant difference [$F_{(3, 36)} = 4.72$; $P = 0.007$]. However, the Tukey HSD correction for multiple comparisons revealed a significant difference between auditory and bimodal thresholds ($P = 0.04$) but not between tactile and bimodal thresholds ($P = 0.99$) or between predicted and observed bimodal thresholds ($P = 0.87$). In the attentional condition (**Figure 4D**), the One-Way ANOVA reported a significant difference between unimodal and bimodal observed and predicted thresholds [$F_{(3, 36)} = 3.22$; $P = 0.03$]. The Tukey HSD correction showed no significant difference between tactile and bimodal thresholds ($P = 0.07$) between auditory and bimodal thresholds ($P = 0.5$) or between predicted and observed bimodal thresholds ($P = 0.99$). These results suggest that both optimal integration and sensory dominance are possible.

Predicted and observed auditory weights are similar for all the participants in all the experimental conditions. **Figure 5A** shows observed individual audio weights plotted as a function of predicted individual audio weights in the non-attentional (light

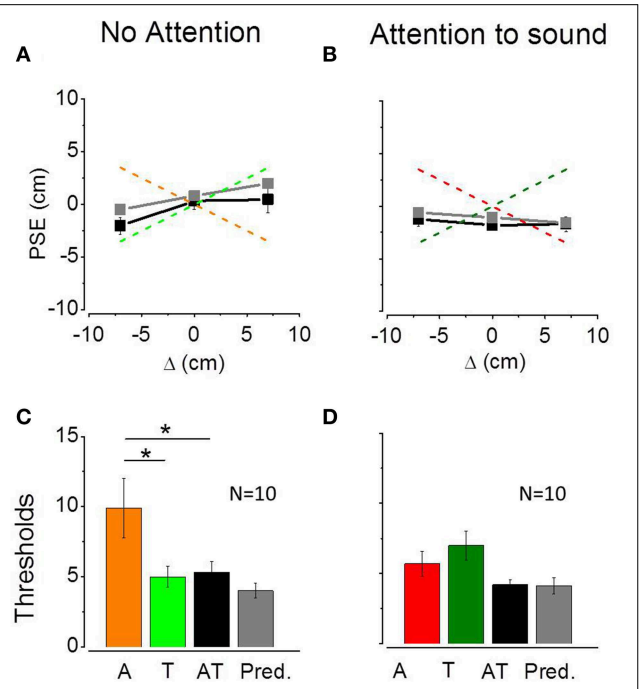


FIGURE 4 | (A,B) Average PSEs ($N = 10$) for the three bimodal conflicts plotted as a function of the audio-tactile conflict (Δ) in the non-attentional condition **(A)** and in the attentional condition **(B)**. Error bars represent standard errors of the mean. Red and green lines describe the ideal performance in the case of auditory or tactile dominance, respectively. Black line and symbols show measured data, while gray line and symbols describe the model prediction. **(C,D)** Average thresholds for the audio (A), tactile (T), and audio-tactile (AT) estimates and predicted bimodal thresholds in the non-attentional **(C)** and attentional condition **(D)**. Error bars represent standard errors of the mean. In the non attentional condition **(C)** auditory thresholds are different from tactile and bimodal thresholds (*p-value < 0.05).

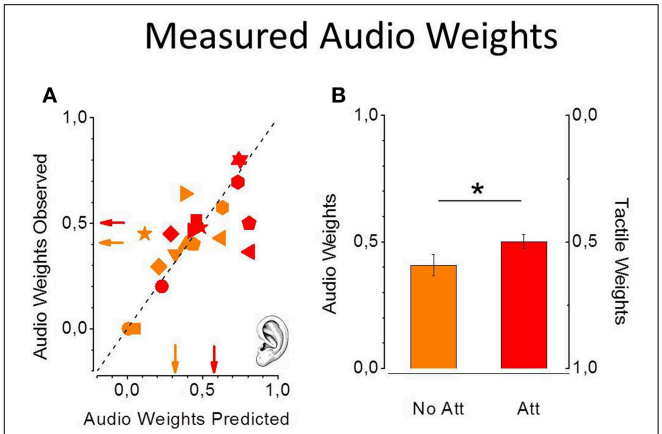


FIGURE 5 | Predicted and observed individual auditory weights from the 10 participants in the non-attentional (light red) and the attentional (dark red) condition (A). Standard errors were computed with bootstrap simulation. Data are all scattered on the equality line meaning perfect prediction from the MLE model. Measured average auditory weights are higher in the attentional condition **(B)**. (*p-value < 0.05).

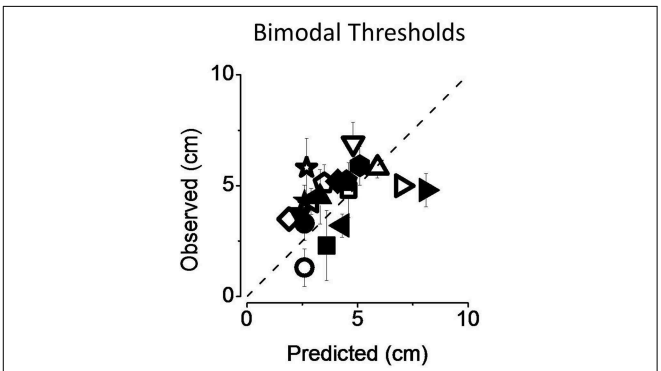


FIGURE 6 | Predicted and observed individual thresholds (A) for both the non-attentional (empty symbols) and the attentional condition (filled symbols). Standard errors were computed with bootstrap simulation. Data are well-predicted by the MLE model.

symbols) and attentional (dark symbols) conditions. All the data are scattered on the equality line suggesting that the model successfully predicted participants' performance. Moreover, observed and predicted audio weights are not statistically different [non-attentional condition: one tailed paired t-test, $t_{(9)} = -0.75$, $P = 0.46$; attentional condition: $t_{(9)} = 0.81$, $P = 0.43$]. More important, the average auditory weights are significantly higher in the attentional (dark red) condition with respect to the the non-attentional (light red) condition [one tailed paired t-test, $t_{(9)} = -2.23$, $P = 0.02$, see **Figure 5B**], as predicted by the model.

In **Figure 6** we reported individual bimodal thresholds for both attentional (filled symbols) and non-attentional condition (empty symbols). Individual data are all scattered on the equality line and average thresholds are similar to those predicted by the MLE model.

Discussion

In this study we investigated the effects of modality-specific attention on sensory reliability and on multisensory integration. We compared an attentional and non-attentional condition. In the attentional condition we found increased precision in the attended modality and a collateral change in cue weighting in the bimodal estimate. Our results support the idea that attention to a sensory modality can affect multisensory processing.

In the non-attentional condition participants performed the tactile task with higher precision than the auditory one. As predicted by the MLE model the tactile modality directed the final multisensory estimates. The dual auditory task (attentional condition) significantly improved auditory precision and increased unimodal weights. Also in this condition the bimodal estimates were successfully predicted by the MLE model, suggesting optimal integration. These results show that attention to sounds reduces auditory thresholds and that the improved auditory precision affects multisensory perceptual judgments and accuracy. However, in all the conditions we have tested, bimodal thresholds were not significantly different from either the best unimodal threshold or the MLE prediction, therefore,

neither account could be rejected. Similarly, Alais and Burr (2004) and Gori et al. (2012) reported near-optimal integration with bimodal thresholds sometimes comparable to unisensory thresholds. The lack of significant improvement on precision may be due to several factors. As Alais and Burr (2004) have suggested, there may be an additional noise source at the level of bimodal combination not considered in the model or there may be correlations between the noise sources of the sensory modalities. The lack of statistical power might be another possible reason for failing to find strong support for MLE integration.

Our study appears to be in conflict with previous results from Helbig and Ernst (2008) that have recently examined the effect of a dual visual task on visuo-tactile integration. The distractor task used in Helbig and Ernst's experiment (2008) involved visual stimuli different from those used in the primary task. Participants had to evaluate the similarity between two sequences of letters presented just above the position of the visual stimulus of the primary size discrimination task. Authors reported that performing a dual visual task impaired precision in the visual modality but did not affect visual and tactile weights in visuo-tactile integration. Combining these two tasks might require extremely high cognitive resources. Indeed, authors found that also the tactile modality was slightly affected by the distractor task. Contrarily, in our task we asked participants to evaluate two different characteristics of the same stimulus: one spatial and the other temporal. Probably, the double-task that we used increases attention to the stimulus rather than withdrawing attention from it. Moreover, we presented the dual task randomly only in the 30% of the trials and analyzed the remaining 70% to be sure that participants were focused on the stimulus and not distracted by the secondary task.

Another possible explanation for the difference between our results and those from Helbig and Ernst (2008) is that spatio-temporal features of a stimulus may be encoded together in the brain. In both the studies participants were engaged in a double-task, a paradigm that generally increases the attentional load and results in a lower performance. Surprisingly, we found higher precision in the attentional condition than in the non-attentional one. Previous studies demonstrated that space and time are not processed separately but probably share similar neural mechanisms and similar cortical circuits (Burr and Morrone, 2006; Johnston et al., 2006). Under this perspective, performing a spatio-temporal dual task could not result in a reduction of spatial precision, but rather in an increased reliability of the attended stimulus.

Researchers have found that directing attention toward a particular region of space or to a sensory modality improves performance in several tasks. Yeshurun and Carrasco (1998) explored the effect of spatial attention on a texture segregation visual task and found attentional facilitation reflecting signal enhancement. Moreover, a neurophysiological study from Spitzer et al. (1988) reported that increasing the amount of effort required to perform a perceptual task, such as orientation or color discrimination, can affect information processing in the visual stream. When the task is more difficult the performance improves and neuronal responses to stimuli are larger and more selective. In a similar way, the attentional effort required by the dual task on the auditory stimulus used in our experiment might have improved the discriminative ability of the participants.

Our results are in line with several studies showing that attention to a sensory modality might affect perceptual estimates in multisensory tasks (Bertelson and Radeau, 1981; Warren et al., 1981; Alsius et al., 2005). For example Oruc et al. (2008) demonstrated that in crossmodal dynamic ventriloquism the motion signals from different sensory modalities are combined differently depending on modality-specific attention, but only when the susceptibility for capture between the two signals is comparable. Alsius et al. (2005) also showed that the McGurk illusion is severely reduced when participants are concurrently performing an unrelated visual or auditory task.

Yau et al. (2010) showed that auditory and tactile signals can be combined differently based on the perceptual task. Here we report a strong attentional modulation of AT integration. The current study adds an interesting contribution to the large body of empirical research supporting the idea that attention to modality can affect the process of multisensory integration. Moreover, although previous studies investigated AT integration with the MLE model in the temporal judgments (Ley et al., 2009) we explored optimal integration also in the spatial domain. Further studies might investigate whether this attentional effect can also reduce the visual "dominance" in an audio-visual or visuo-tactile spatial integration.

References

Alais, D., and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Curr. Biol.* 14, 257–262. doi: 10.1016/j.cub.2004.01.029

Alsius, A., Navarra, J., Campbell, R., and Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Curr. Biol.* 15, 839–843. doi: 10.1016/j.cub.2005.03.046

Bertelson, P., and Radeau, M. (1981). Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Percept. Psychophys.* 29, 578–584. doi: 10.3758/BF03207374

Bertelson, P., Vroomen, J., de Gelder, B., and Driver, J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. *Percept. Psychophys.* 62, 321–332. doi: 10.3758/BF03205552

Burr, D., and Morrone, C. (2006). Time perception: space-time in the brain. *Curr. Biol.* 16, R171–R173. doi: 10.1016/j.cub.2006.02.038

Butler, J. S., Foxe, J. J., Fiebelkorn, I. C., Mercier, M. R., and Molholm, S. (2012). Multisensory representation of frequency across audition and touch: high density electrical mapping reveals early sensory-perceptual coupling. *J. Neurosci.* 32, 15338–15344. doi: 10.1523/JNEUROSCI.1796-12.2012

Clarke, J. J., and Yuille, A. L. (1990). *Data Fusion for Sensory Information Processing*. Boston, MA: Kluwer Academic. doi: 10.1007/978-1-4757-2076-1

Driver, J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature* 381, 66–68. doi: 10.1038/381066a0

Efron, B., and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall. doi: 10.1007/978-1-4899-4541-9

Ernst, M. O., and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415, 429–433. doi: 10.1038/415429a

Ghahramani, Z., Wolpert, D. M., and Jordan, M. I. (1997). "Computational models of sensorimotor integration," in *Self-Organization, Computational Maps and*

Motor Control, eds P. G. Morasso and V. Sanguineti (Amsterdam: Elsevier Science Publishers), 117–147.

Gori, M., Sandini, G., and Burr, D. (2012). Development of visuo-auditory integration in space and time. *Front. Integr. Neurosci.* 6:77. doi: 10.3389/fnint.2012.00077

Helbig, H. B., and Ernst, M. O. (2008). Visual-haptic cue weighting is independent of modality-specific attention. *J. Vis.* 8, 21–16. doi: 10.1167/8.1.21

Johnston, A., Arnold, D. H., and Nishida, S. (2006). Spatially localized distortions of event time. *Curr. Biol.* 16, 472–479. doi: 10.1016/j.cub.2006.01.032

Landy, M. S., Banks, M. S., and Knill, D. C. (2011). "Ideal-observer models of cue integration," in *Book of Sensory Cue Integration,* eds J. Trommershauser, K. Körding, and M. S. Landy (New York, NY: Oxford University Press), 5–30.

Ley, I., Haggard, P., and Yarrow, K. (2009). Optimal integration of auditory and vibrotactile information for judgments of temporal order. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 1005–1019. doi: 10.1037/a0015021

Oruc, I., Sinnett, S., Bischof, W. F., Soto-Faraco, S., Lock, K., and Kingstone, A. (2008). The effect of attention on the illusory capture of motion in bimodal stimuli. *Brain Res.* 1242, 200–208. doi: 10.1016/j.brainres.2008.04.014

Soto-Faraco, S., Lyons, J., Gazzaniga, M., Spence, C., and Kingstone, A. (2002). The ventriloquist in motion: illusory capture of dynamic information across sensory modalities. *Brain Res. Cogn. Brain Res.* 14, 139–146. doi: 10.1016/S0926-6410(02)00068-X

Spitzer, H., Desimone, R., and Moran, J. (1988). Increased attention enhances both behavioral and neuronal performance. *Science* 240, 338–340. doi: 10.1126/science.3353728

Vroomen, J., Bertelson, P., and de Gelder, B. (2001). The ventriloquist effect does not depend on the direction of automatic visual attention. *Percept. Psychophys.* 63, 651–659. doi: 10.3758/BF03194427

Warren, D. H., Welch, R. B., and McCarthy, T. J. (1981). The role of visual-auditory "compellingness" in the ventriloquism effect: implications for transitivity among the spatial senses. *Percept. Psychophys.* 30, 557–564. doi: 10.3758/BF03202010

Watson, A. B., and Pelli, D. G. (1983). QUEST: a Bayesian adaptive psychometric method. *Percept. Psychophys.* 33, 113–120. doi: 10.3758/BF03202828

Yau, J. M., Olenczak, J. B., Dammann, J. F., and Bensmaia, S. J. (2009). Temporal frequency channels are linked across audition and touch. *Curr. Biol.* 19, 561–566. doi: 10.1016/j.cub.2009.02.013

Yau, J. M., Weber, A. I., and Bensmaia, S. J. (2010). Separate mechanisms for audio-tactile pitch and loudness interactions. *Front. Psychol.* 1:160. doi: 10.3389/fpsyg.2010.00160

Yeshurun, Y., and Carrasco, M. (1998). Attention improves or impairs visual performance by enhancing spatial resolution. *Nature* 396, 72–75. doi: 10.1038/23936

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in
INTEGRATIVE NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 13 February 2015
doi: 10.3389/fnint.2015.00011

Visual-auditory integration for visual search: a behavioral study in barn owls

Yael Hazan[1], Yonatan Kra[1], Inna Yarin[1], Hermann Wagner[2] and Yoram Gutfreund[1]*

[1] Department of Neuroscience, The Ruth and Bruce Rappaport Faculty of Medicine and Research Institute, Technion, Haifa, Israel
[2] Department of Zoology and Animal Physiology, Institute for Biology II, RWTH Aachen University, Aachen, Germany

Edited by:
Ruth Adam,
Ludwig-Maximilian-University,
Germany

Reviewed by:
Christopher I. Petkov, Newcastle University, UK
Maik Christopher Stüttgen, University Medical Center Mainz, Germany

***Correspondence:**
Yoram Gutfreund, Department of Neuroscience, The Ruth and Bruce Rappaport Faculty of Medicine and Research Institute, Technion, Bat-Galim, Haifa 31096, Israel
e-mail: yoramg@tx.technion.ac.il

Barn owls are nocturnal predators that rely on both vision and hearing for survival. The optic tectum of barn owls, a midbrain structure involved in selective attention, has been used as a model for studying visual-auditory integration at the neuronal level. However, behavioral data on visual-auditory integration in barn owls are lacking. The goal of this study was to examine if the integration of visual and auditory signals contributes to the process of guiding attention toward salient stimuli. We attached miniature wireless video cameras on barn owls' heads (OwlCam) to track their target of gaze. We first provide evidence that the area centralis (a retinal area with a maximal density of photoreceptors) is used as a functional fovea in barn owls. Thus, by mapping the projection of the area centralis on the OwlCam's video frame, it is possible to extract the target of gaze. For the experiment, owls were positioned on a high perch and four food items were scattered in a large arena on the floor. In addition, a hidden loudspeaker was positioned in the arena. The positions of the food items and speaker were changed every session. Video sequences from the OwlCam were saved for offline analysis while the owls spontaneously scanned the room and the food items with abrupt gaze shifts (head saccades). From time to time during the experiment, a brief sound was emitted from the speaker. The fixation points immediately following the sounds were extracted and the distances between the gaze position and the nearest items and loudspeaker were measured. The head saccades were rarely toward the location of the sound source but to salient visual features in the room, such as the door knob or the food items. However, among the food items, the one closest to the loudspeaker had the highest probability of attracting a gaze shift. This result supports the notion that auditory signals are integrated with visual information for the selection of the next visual search target.

Keywords: saliency, saccades, multisensory, visual search, barn owls, selective attention, sound localization

INTRODUCTION

An animal in its environment is constantly bombarded by sensory input, while the animal can only orient and react to one object or event at a time. Therefore, mechanisms have evolved to select the most behaviorally relevant stimulus at any particular time. This brain process is called saliency mapping (Itti and Koch, 2000) and it lies at the base of selective attention. Animals tend to respond and attend to the stimulus which they perceive as the most salient (Luck and Ford, 1998).

It is widely accepted that animals compute a dynamic saliency value to different locations in space based on a combination of external factors, such as stimulus intensity, stimulus history, spatial context, etc., and internal factors, such as cognitive biases, behavioral tasks, reward history, motivations, etc. (Fecteau and Munoz, 2006). Therefore, the saliency of a stimulus is not a physical feature but rather a perceived feature dependent strongly on the context, history, surroundings, and internal state of the animal (Dutta and Gutfreund, 2014). It has been shown that the relationship between visual and auditory signals is an important external factor determining the saliency of stimuli in cats, primates, and humans (Stein et al., 1988; Driver and Spence, 1998; Recanzone, 2009). Congruent visual and auditory stimuli (same location and same time) are more likely to attract the animal's gaze and attention, i.e., more salient, compared to unimodal stimuli or incongruent stimuli (Stein and Meredith, 1993; Frassinetti et al., 2002; Stein and Stanford, 2008). This process of combining visual and auditory signals is called visual-auditory integration. Thus, an animal is said to integrate visual and auditory information if the response to a combined stimulus is different from the response to each stimulus alone (Stein et al., 2014).

At the single neuron level, neurons that respond to both visual and auditory signals (bimodal neurons) have been identified in numerous levels of the nervous system (Beauchamp, 2005; Sugihara et al., 2006; Wallace et al., 2006; Kayser et al., 2007). Among these, the superior colliculus (SC), a mid-brain structure believed to be involved in selective attention, has been the most studied (Wallace et al., 1996). In cats and primates, many of the neurons in the SC have been shown to integrate auditory and visual signals in ways that mirror the behavioral observations, i.e., neurons respond maximally to visual and auditory signals that are congruent in time and space (Meredith and Stein, 1986; Meredith et al.,

1987). These findings support the notion that the SC circuitry combines visual and auditory signals to perform saliency mapping in bimodal environments (Boehnke and Munoz, 2008; Mysore and Knudsen, 2011; Dutta and Gutfreund, 2014).

In non-mammalian species, physiological studies of visual-auditory integration for selective attention have been carried out mostly in the optic tectum (OT), the homolog of the SC. Among these, the OT of the barn owl consists of numerous bimodal neurons, integrating signals to enhance responses to congruent bimodal events, particularly if such events are surprising (Zahar et al., 2009). It has been suggested that this type of response facilitates the detection of salient stimuli (Reches et al., 2010; Gutfreund, 2012). The OT of barn owls possesses the most robust and accurate map of auditory space in any animal species studied so far (Knudsen, 1987). This auditory map is aligned with a precise retinotopic visual map (Knudsen, 1982). In addition, mechanisms of stimulus selection in the OT have been studied extensively in barn owls (Reches and Gutfreund, 2008; Mysore et al., 2010; Mysore and Knudsen, 2013). Therefore, this species has a great potential of being used as a research model for the study of visual-auditory integration for saliency mapping. Despite previous studies in barn owls on this subject at the neuronal level (Gutfreund et al., 2002; Reches and Gutfreund, 2009; Zahar et al., 2009; Reches et al., 2010), behavioral characterization of visual-auditory integration at the behavioral level is scarce (Whitchurch and Takahashi, 2006). The goal of this study is to contribute to such characterization and to develop new ways of studying visual-auditory integration at the behavioral level.

The perceived saliency of objects in the environment is manifested, in many species, in their visual search behavior (Wolfe and Horowitz, 2004; Hayhoe and Ballard, 2005; Berman and Colby, 2009). Visual search is the process of actively scanning the environment. Many animal species possess a small retinal area with a higher density of photoreceptors known as area centralis (in some species, this area is accompanied by an anatomical dipping in the retina, in which case it is called fovea). Such animals tend to shift their gaze so that points of interest will be acquired by this specialized retinal area. Animals exhibiting such a behavior are called foveating animals. It is widely accepted that the target at the retinal center in foveating animals is correlated with the focus of attention (Eckstein, 2011). Thus, by tracking the scan path of an animal in its environment, it is possible to obtain information on what objects and conditions are likely to attract the animal's attention, i.e., are perceived as salient. This experimental procedure of gaze tracking has been used widely to study attention in humans and other species (Reinagel and Zador, 1999; Harmening et al., 2011; MacInnes et al., 2014). Gaze tracking can be technically difficult particularly when performed in freely moving animals, since it requires the exact measurement of both eye and head orientation as well as its relationship with the structure of the environment.

Barn owls possess a tubular eye structure that limits eye movement. Thus, in contrast to most other foveating animals, barn owls do not move their eyes in the orbits, maintaining a mostly fixed eye position relative to the head (Steinbach and Money, 1973; du Lac and Knudsen, 1990). Instead, they compensate for this lack by prominent head motions (Masino and Knudsen, 1990; Ohayon et al., 2006). This makes them an attractive animal model for the study of attention and visual search because it is not necessary to measure eye relative to head movements. It has been demonstrated unequivocally that barn owls are foveating targets, i.e., they use a single retinal location to acquire targets of interest (Ohayon et al., 2008; Harmening et al., 2011). The retina of barn owls contains a single area centralis but no visible fovea (Wathey and Pettigrew, 1989). It remains an open question whether the functional fovea in barn owls corresponds with the anatomically defined area centralis. In this study, we took advantage of the lack of eye movement and the spontaneous visual search behavior of barn owls. We attached miniature video cameras to the heads of barn owls in order to track the scan path and points of interests in the environment. In the first part, we show that the functional fovea corresponds with the area centralis. In the second part, we show that sounds influence visual search behavior in ways that support visual-auditory integration for saliency mapping.

MATERIALS AND METHODS

ANIMALS

Three adult barn owls (*Tyto alba*) were used in multiple test sessions. The owls were hatched in the breeding colony in the Technion's Rappaport Faculty of Medicine and were hand-raised by lab members. Before the experiments, the owls were accustomed to the experimental room by being maintained in the room, on a perch, for about 2 h a day for a period of about 2 weeks. To increase the owls' motivation to search spontaneously during the experiments, they were only fed on the perch at the end of each experiment from the food items on the floor. However, body weight was maintained normal. All procedures were approved by the Technion's Committee for the Ethical Use of Laboratory Animals.

VIDEO CAMERA ATTACHMENT AND RECORDINGS

The OwlCam used was similar to the OwlCam developed and used by Harmening et al. (2011). It is devised from a miniature microcamera combined with a video broadcasting chip (900 MHz) and a rechargeable lithium-polymer battery [for more details on the camera, see Harmening et al. (2011)]. The OwlCam was attached to the head using a 3D printed attachment unit designed in the lab. One part of the unit was attached permanently to the skull bone with dental cement in a single surgical procedure. The other part, glued to the camera, was screwed to the permanent part at the beginning of each experiment and removed at the end of the experiment. The attachment unit was designed to allow the adjustment of the OwlCam orientation with respect to the head and to maintain a locked rigid positioning of the camera. The total weight of the device including the battery and attachment unit was 10.5 g and the dimensions of the camera case was 10 mm × 10 mm × 20 mm (**Figures 1B,C**). Owls wearing the mount showed no apparent behavioral changes including normal posture and flying.

The OwlCam delivered a wireless black and white video signal at 30 frames per second with an effective vertical resolution of about 380 lines and a view angle of about 60°. The video signal was collected with an off-the-shelf tunable video receiver (positioned about 2 m from the owl) and digitized at a resolution of 480 × 720 pixels for offline analysis.

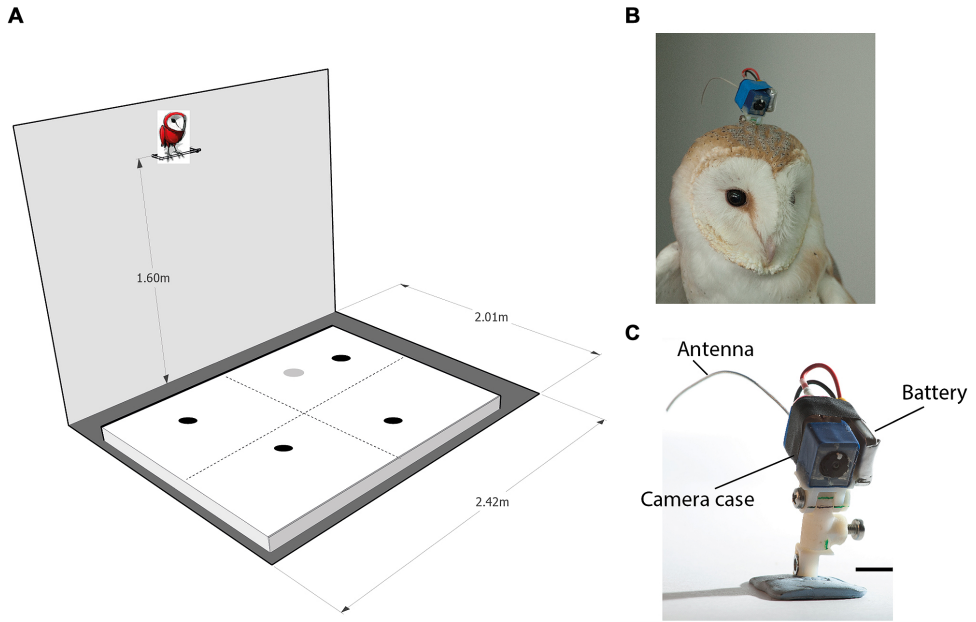


FIGURE 1 | The experimental apparatus. (A) A scheme of the experimental room. An arena 200 cm × 160 cm is shown positioned on the floor. The perch, positioned 160 cm above the ground adjacent to the long wall, is shown with a cartoon owl on top of it. The dashed lines on the arena surface designate imaginary lines that divide the arena to four equal quadrants. The dark spots on the arena designate possible positions of four food items. Items were spread so that each quadrant will contain one item. The gray spot on the arena designates a possible location of the loud speaker. **(B)** An owl with the OwlCam attached to its head. **(C)** A close view of the OwlCam with the attachment unit and the battery in place. The scale bar designates 10 mm.

EXPERIMENTAL SET-UP

The experiments were performed in a 200 cm × 240 cm room. Although the room was stripped of any furniture, it was a regular lab room with some salient features on its walls such as electrical outlets, a dark door and a window covered with black paper. Moderate illumination was provided by a ceiling-mounted bulb. A 30 cm long perch was mounted to the wall opposite the door at a height of 160 and 10 cm away from the wall (see **Figure 1A** for a sketch of the experimental room). During the experiments, the owls were attached to the perch with a leg leash that was long enough to allow free walking and turning on the perch but prevented the owls from flying off the perch. A 160 cm × 200 cm wooden frame covered with a white cloth (the arena) was positioned on the floor centered under the perch. Before the beginning of the experiment, a curtain was drawn to block the owl's view of the floor. Then, several food items (dead black lab mice on petri dishes) were scattered on the arena. At this point, the experimenter would leave the room and the curtain was drawn from outside the room with a string to reveal the arena to the owl.

In several of the experiments, a loudspeaker was positioned under the cloth, hidden from the owls. To register the position of the loudspeaker in the video frames, we attached an infrared LED to the speaker. The light emitted by the LED was clearly visible in the video (see arrow in **Figure 2**) but invisible to humans and barn owls (Netser et al., 2010). The loudspeaker was used to generate short and unexpected auditory stimuli. The auditory stimuli were stored on a PC connected to a Tucker-Davies Technologies (TDT) system III (∼100 kHz sampling rate; 24 bit A/D), running custom Matlab programs. In order to reduce habituation to the sound, we used a library of 18 playbacks of natural sounds such as rustling leaves, animal sounds, etc. The different sounds varied in amplitude, frequency, temporal structure and duration (300–800 ms). Sounds were generated manually by a button pressed about once every 2 min. Each button press generated one randomly chosen sound from the library. The volume of all sounds was adjusted so that all were clearly audible to human listeners in the room. To synchronize the video recordings with the times of the auditory stimulation, an electronic switch was triggered from the sound-generating system to temporarily switch off the power supply to the OwlCam receiver. This resulted in about 5–6 disrupted frames in the video sequence, signaling the onset of the auditory stimulation. The disrupted frames were detected offline.

DATA ANALYSIS

The behavior of the owls was mostly characterized by abrupt head movements and prolonged fixation period where the image was stable (see **Figure 2** and Ohayon et al., 2008). From each fixation period a single frame was extracted for further analysis. Fixation periods were identified manually by viewing the video sequence frame by frame and identifying stable periods in which the point of view doesn't change. For visualization of the behavior (**Figures 2** and **6**) we used a frame by frame correlation function. Each frame in the video sequence was first passed through an edge detection filter, creating a reduced black and white image, the frames where then divided into an array of 8 × 10 rectangulars; the average value of all pixels in each rectangular was measured, creating 80 pixels

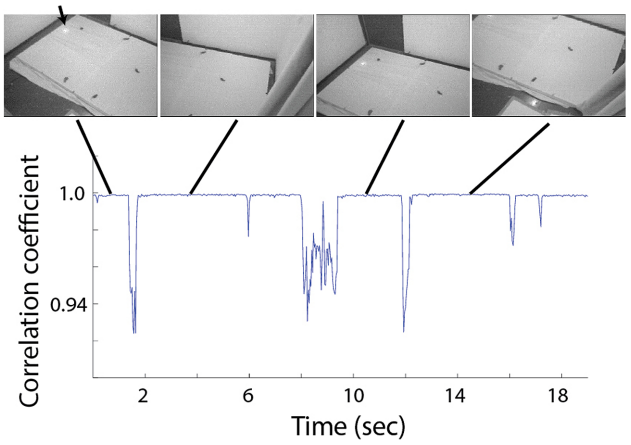


FIGURE 2 | Correlation between adjacent frames during 19 s of typical visual search behavior. Four video frames extracted from periods of fixation are shown above. Each shows a different view of the experimental arena and the scattered food items. The arrow points to the position of the loudspeaker, visible via the infra-red light source attached to it.

images, each of which was correlated with its preceding frame and the correlation coefficient as a function of time was obtained.

Fixation frames were analyzed manually. All observable targets and speaker locations were marked with a mouse curser, and the coordinates in the frame were saved (**Figure 3A**). To estimate the coordinates of the functional fovea, all target locations from one experiment were collapsed in one graph (**Figure 3B**), and a density function (the density of the points in each pixel) was calculated to estimate the probability for targets in pixel space (**Figure 3C**). To map the projection of area centralis onto the video frame, a prominent retinal landmark in birds called pecten oculus (Wathey and Pettigrew, 1989) was used. The pecten oculus is a pigmented structure covering the entrance of the optic nerve to the retina. Therefore it is easily viewable with an ophthalmoscope as a dark structure on the highly reflective background of the retina (Netser et al., 2010). We mounted the camera on the owl's head. Then the owl was held by hand, with its head fixed pointing straight ahead. Another experimenter, who was standing a meter away from the owl, viewed the eye of the barn owl through an ophthalmoscope, and adjusted the position of the ophthalmoscope relative to the owl's head until viewing the superior tip of the pectin (see **Figure 4B** for an illustration of the pecten). When this was achieved the experimenter marked the video frame by snapping his fingers and moved to the second eye. The relevant video images were extracted offline, and the positions of the ophthalmoscope beam viewing the two pectens were marked manually (see **Figure 4A**) to obtain the coordinates of the projection of the retinal landmarks on the video frame. Since the owls do not move their eyes considerably, the coordinates of the retinal locations are fixed throughout the experiment.

Distances between targets were measured in pixels. Conversion to distance in centimeters was estimated based on the number of pixels recorded in a 10 cm line at the center of the arena and the center of the video frame (30 pixels per 10 cm). Distortion errors

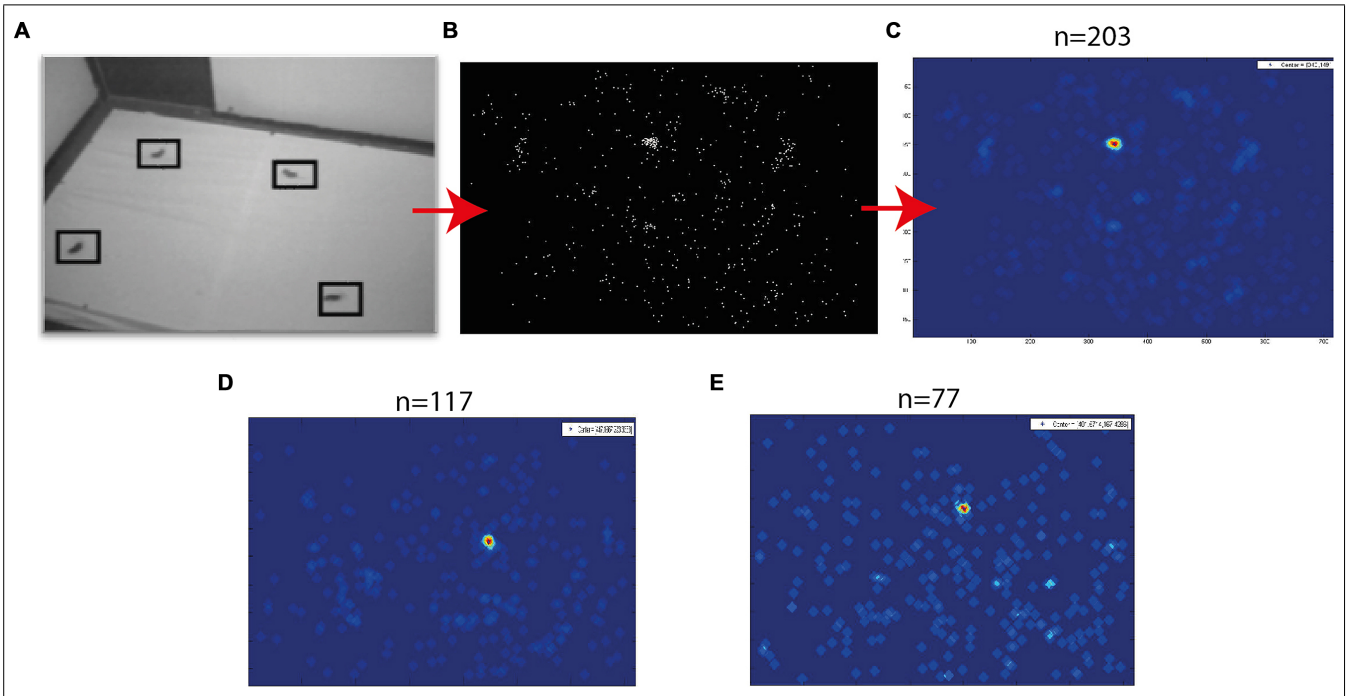


FIGURE 3 | Target probability maps. (A) An example of a single fixation frame. Four targets are marked by squares. **(B)** A plot showing all marked targets collected from a single experimental day. **(C)** The density function of the points in **(B)**. Red colors indicate high density and blue low density. **(D,E)** Two additional examples of density functions obtained in different experimental days. The numbers above panels **(C–E)** designate the number of fixation frames analyzed to obtain this image.

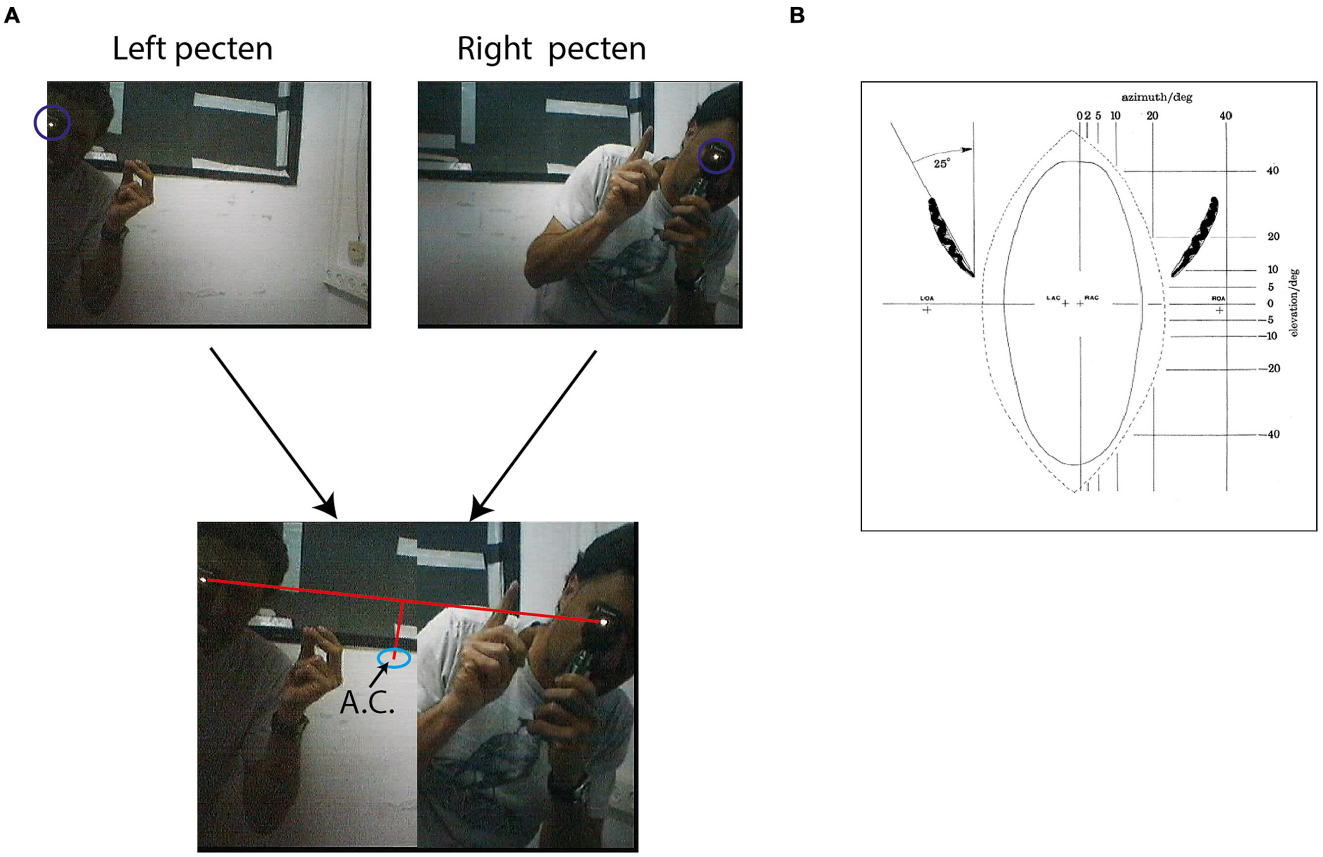


FIGURE 4 | Mapping of area centralis. (A) Two frames from the video sequence are shown above. One frame shows an experimenter standing about 1 m from the owl viewing with an ophthalmoscope the superior tip of the left pected and the second from the experimenter viewing the superior tip of the right pecten. The beam of the ophthalmoscope in both cases is marked by a circle. Below the two half frames are combined to one image. A line is drawn between the two beams and the point that is 7° below the mid-point of the line is calculated. This point, marked by a circle, is the estimation of area centralis. **(B)** A scheme showing the projection of the pectens (left and right) on a calibrated screen in front of the owls. The plus marks at the center show the calculated projections of left area centralis (LAC) and right area centralis (RAC). The plus marks at the sides show the projections of the left optical axis (LOA) and the right optical axis (ROA). Positions of LAC, RAC, LOA, and ROA were estimated based on average distances from pectens obtained by analysis of whole-mounted retinas. Figure with permission from Pettigrew (1979).

due to camera optics and non-equal distances to positions on the arena (keystone distortions) were estimated to be mostly less than 25% and were ignored in this analysis. Since the positions of the targets and speaker were shifted in the arena between different experimental sessions, ignoring these distortions is not expected to create any systemic biases in the results.

Conventional statistical methods were used to assess significance of results. A p-value smaller than 5% was considered significant.

RESULTS

ANATOMICAL STRUCTURE OF FUNCTIONAL FOVEA

In the first part of this study, we ask whether owls shift their gaze toward targets of interest so as to view the target image on the retinal area centralis. On each experimental day, the OwlCam was first attached and the coordinates for the right and left pectens were mapped as described in the Section "Materials and Methods." After mapping the pecten tips, the owls were free to stand on the perch. Three to five food items were scattered on the arena, and the owl was left in the room for a period of about half an hour. During this time, the video signal was saved continuously for offline analysis.

A typical video sequence consisted of a series of abrupt head motions (head saccades), each head saccade terminating in a stable period (fixation period). **Figure 2** shows the frame-to-frame correlation function (see Materials and Methods) over a period of 20 s in one such typical experiment. It can be seen that the sequence is composed mostly of stable periods (correlation indices close to 1), but every few seconds, the correlation indices drop abruptly below one, indicating a gaze shift to a new location. For analysis, a single frame was extracted from each stable period (fixation period). Fixation frames in which at least one food item was observed were taken for further analysis while the others were discarded. The positions in the frames of all observed food items were marked manually (**Figure 3A**) and pooled together to create a distribution of food items on the video frame (**Figure 3B**). The probability of each pixel to contain a food item was thus estimated by the density function (**Figure 3C**). It can be seen that the density function resulted in a single most probable point.

This pattern was observed consistently (see **Figures 3D,E** for two more examples). In most cases, 10–20 min of recording of spontaneous scan behavior (corresponding to about 80–200 fixation periods) were sufficient to expose such clear, single, most probable point. These results verify the results from previous studies (Ohayon et al., 2008; Harmening et al., 2011) that during a search task, owls scan their environment in a serial manner and repeatedly direct their gaze in a way that brings targets of interest to a specific retinal position, which we refer to as the functional fovea. Since the OwlCam moves with the owl's head and the eyes are fixed in the head, the functional fovea does not change its position in the video sequences regardless of the owl's movement.

It is hypothesized that the functional fovea corresponds with the area centralis. To demonstrate that this is indeed the case, we measured both the functional fovea and the area centralis in the same video frame. This dual measurement was performed 15 times in three birds. The functional fovea coordinates on the frame were extracted as the point of maximal probability of an average of at least 60 fixation frames (as demonstrated in **Figure 3**). The coordinates of the area centralis on the video image were mapped based on the histological study of Pettigrew (1979), which analyzed whole-mounted retinas and concluded that the retinal area centralis is located, on average, 25° temporal and 7° above the superior tip of the pecten oculus (**Figure 4B**). We therefore measured the position of the superior tip of the pecten with an ophthalmoscope for both the left and right eyes (see Materials and Methods). A line was drawn between the two positions to designate the horizontal plane. The point 7° perpendicularly below the mid-point was registered as the area centralis (**Figure 4A**). The distances between the left and right pectens were estimated to be between 52 and 55°. These numbers are consistent with the measurements of Pettigrew (1979; 25° per each eye) and indicate that the projections of the area centrales of both eyes tend to converge at a single point directly in front of the owl. Thus, a single functional fovea corresponds with both the left and right area centrales. To assess the reliability of the measurement of the projection of area centralis we repeated in seven cases the measurement twice, once before and once after the release of the owl on the perch for half an hour. In each case we measured the distance in pixels between the two measurements. The average distance was 12.8 pixels and the STD 8 pixels.

Figure 5A shows the location of the functional fovea on the video frame together with the corresponding area centralis. It can be seen that generally there is good agreement between the two independent measurements. The errors are shown in **Figure 5B**, where, for each experiment, the position of the area centralis relative to the position of the functional fovea (0,0 point in the graph) is depicted. The smallest error was 4 pixels and the maximal error was 50 pixels. However, note that the median of the area centralis was 7 pixels biased to the left of the functional fovea (sign test, $n = 15$; $p < 0.05$). This deviation of the area centralis relative to the functional fovea, which corresponds to 2–3°, may arise from small differences in eye positions when the head is immobilized and visible light is shone onto the eye (conditions for measuring the area centralis) compared to active viewing of the environment (conditions for measuring the functional fovea). Interestingly, in the data from Pettigrew (1979), the average mid-position between the two pectens is also shifted to the left of the gaze point by 2–3° [**Figure 4B**, the midpoint between right area centralis (RAC) and left area centralis (LAC)]. Therefore, it seems that when passive and head-fixed, the owls have a tendency to drift eye positions slightly to the left. In the subsequent experiments, where possible, we estimated the point of gaze based on the functional fovea. However, in several cases, not enough spontaneous fixation points

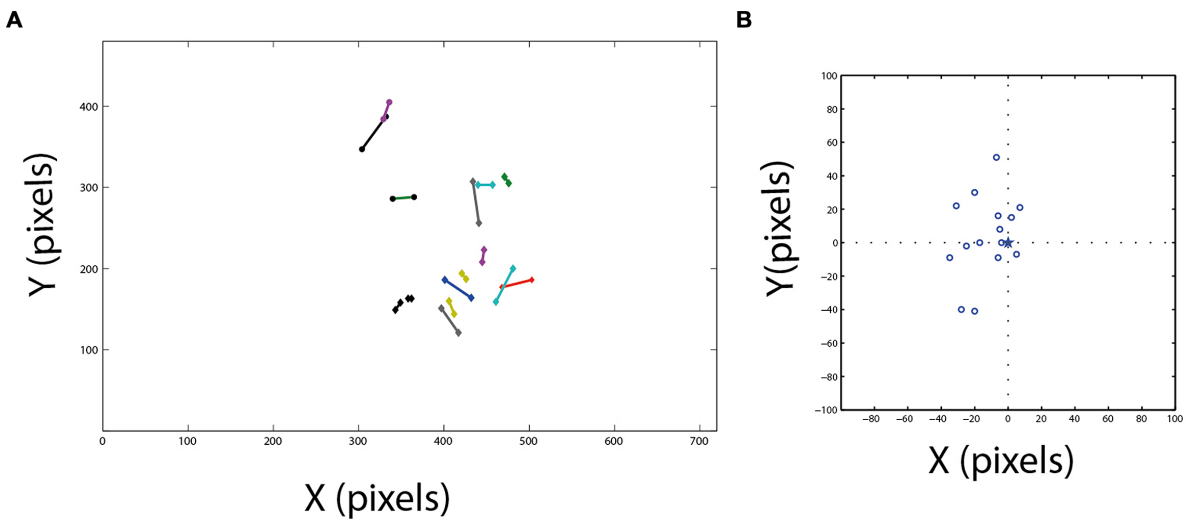


FIGURE 5 | Comparison between area centralis and the functional fovea. (A) Fifteen estimated projections of area centralis on the video frame are shown together with the corresponding estimations of the functional fovea. Corresponding pairs are connected with lines. **(B)** The projections of area centralis from panel **(A)** are shown in relation to the position of the corresponding functional fovea. Dash lines mark the zero axes and star marks the position of the functional fovea at the center.

were obtained to give a good estimation. In these cases, the area centralis was used to estimate the point of gaze.

AUDITORY EFFECTS ON VISUAL SEARCH

We used the same owls in the second part of the research. The owls performed an active visual search task with food items as before, however, in this case four food items were scattered in the arena. Each food item was positioned in one quadrant of the arena, and a loudspeaker was hidden under the cloth in one of the quadrants so that the food item in the same quadrant was closer to the loudspeaker but never closer than 10 cm (30 pixels). During each session which was 10–15 min long, a sound stimulus was occasionally emitted from the speaker (see Materials and Methods). About 6–10 sounds were emitted during a single session. After each session, the curtain was drawn, the speaker moved to another quadrant and the positions of the food items were moved slightly for the next session. Each owl performed 2–3 sessions during one experimental day. We analyzed the responses to about 800 sounds in 85 sessions.

The responses were analyzed offline and first divided into two groups, one in which the fixation point did not change in the 1.5 s after the onset of the sound stimulus (considered as no response and omitted from further analysis) and a second in which a head saccade was elicited in the 1.5 s after the stimulus (considered as a response to the auditory stimulus). Overall in about 35% of stimulus presentations, a head response was induced after stimulation. However, as mentioned earlier, the owls searched the room spontaneously by head saccades even without the sound. Therefore, to examine whether the presence of the sound increased the probability of a head saccade, we averaged the frame to frame correlation function in the 1.5 s following an auditory stimulus and compared it with the average correlation function in the 1.5 s before the auditory stimulus. The average correlation function is an estimation of the saccade probability; the smaller the average correlation, the higher the probability of a saccade. **Figure 6A** shows the average correlation function from 172 stimulus presentations in two owls. A typical delay to an acoustical evoked saccade is 150–300 ms (Whitchurch and Takahashi, 2006), therefore, effects happening in the first six frames unlikely to be attributed to the sound. Examining the probabilities from frame 6 onward, no apparent difference can be seen between the probability of head saccades before and after the stimulus. The lower correlations in the first six frames are attributed to the jitter of the synchronization signal in the video (see Materials and Methods) and do not reflect a head motion. Thus, it seems that the auditory stimulation did not affect the rate of spontaneous head saccades.

Nevertheless, we continued analyzing the targets of the head saccades following the stimuli to explore the possibility that even though the probability of eliciting a saccade is not affected by the auditory stimulus, the choice of the next target is. Among the trials in which a head response was measured in the 1.5 s following the stimulus ($n = 305$), in about half, the gaze changes landed outside the arena ($n = 161$), in many cases attracted toward salient features in the room such as the door knob or the window frame. Only 144 saccades were directed toward the arena. Those latter saccade end points were the subject of subsequent analysis.

The histogram in **Figure 6B** shows the distances between the speaker location and the fixation points on the arena from all three owls. We define a radius of 50 pixels (about 15 cm) from the center of the target as a hit response. This relatively broad circle around the target was chosen to encompass the errors from mapping the area centralis and the errors expected due to different view angles. Only 6% of all responses were considered hits to the speaker. Thus, it seems that the speaker location was a poor attracting point for the final gaze position. This is despite the fact that owls can pinpoint sound sources accurately (Knudsen et al., 1979) and that the sounds were clearly audible and consisted of sound playbacks of objects from the owl's natural environment. **Figures 6C,D** show the distribution of the gaze end-points of 108 head saccades in which the speaker location was observable at the end-point frame. The same data are plotted in panel C with relation to the location of the speaker and in panel D with relation to the location of the visual target closest to the speaker. It is apparent that the position of the visual target tends to attract gaze while the position of the speaker does not.

To examine the tendency of choosing the different targets, we ranked the targets according to distance from the speaker: the target closest to the speaker was ranked 1 and the target furthest from the speaker was ranked 4. Note that in each session, the speaker location was shifted and therefore the targets ranks were updated accordingly. In the 144 saccades following a stimulus that were directed to the arena, we measured the distance in pixels between the gaze point and the closest target. The data are shown in the histogram in **Figure 7**. Sixty-six of the saccades ended in a position closest to target 1, 32 to target 2, 21 to target 3, and 25 to target 4. This distribution significantly favored the target closest to the speaker location [$\chi^2(3) = 35.05$; $p < 0.05$]. In addition, the mean distance of the head saccades closest to target 1 (58.2 pixels) was significantly smaller [$t(142) = -2.75$; $p < 0.05$] from the mean distances to targets 2, 3, and 4 (82.4, 110.2 and 97.2 pixels, respectively). Finally, the number of trials considered hits (below 50 pixels) was 39 for target 1 compared to 14 for target 2, 8 for target 3, and 12 for target 4. These correspond to hit rates of 0.59 (39/66) to target 1, 0.44 (14/32) to target 2, 0.38 (8/21) to target 3 and 0.48 (12/25) to target 4, not significantly different from an equal hit rate to all targets [$\chi^2(3) = 1.91$; $p > 0.5$].

One concern is that the bias we observed toward the target closest to the speaker may arise from prior behavioral biases that the owls might have (for example the owls spontaneously preferring targets near the door). However, this may only pose a concern if the speaker locations were not equally distributed among the four quadrants of the arena. In our experiments out of 85 recording sessions in 21 sessions the speaker was in the upper left quadrant (close to the door), 27 sessions in the lower left, 14 sessions in the upper right, and 23 sessions in the lower right. To address to what extent this can pose a problem we performed a probability simulation by assuming that the owls' behavior is independent of the speaker location and simply picking for each trial a speaker location based on the above distribution and a fixation location based on hypothetical distributions of the owls' behavior. First we simulated the very unlikely but the worst case scenario that the owls only look at the lower left quadrant, which is the quadrant in which the speaker happened to be the most. In this case, out of

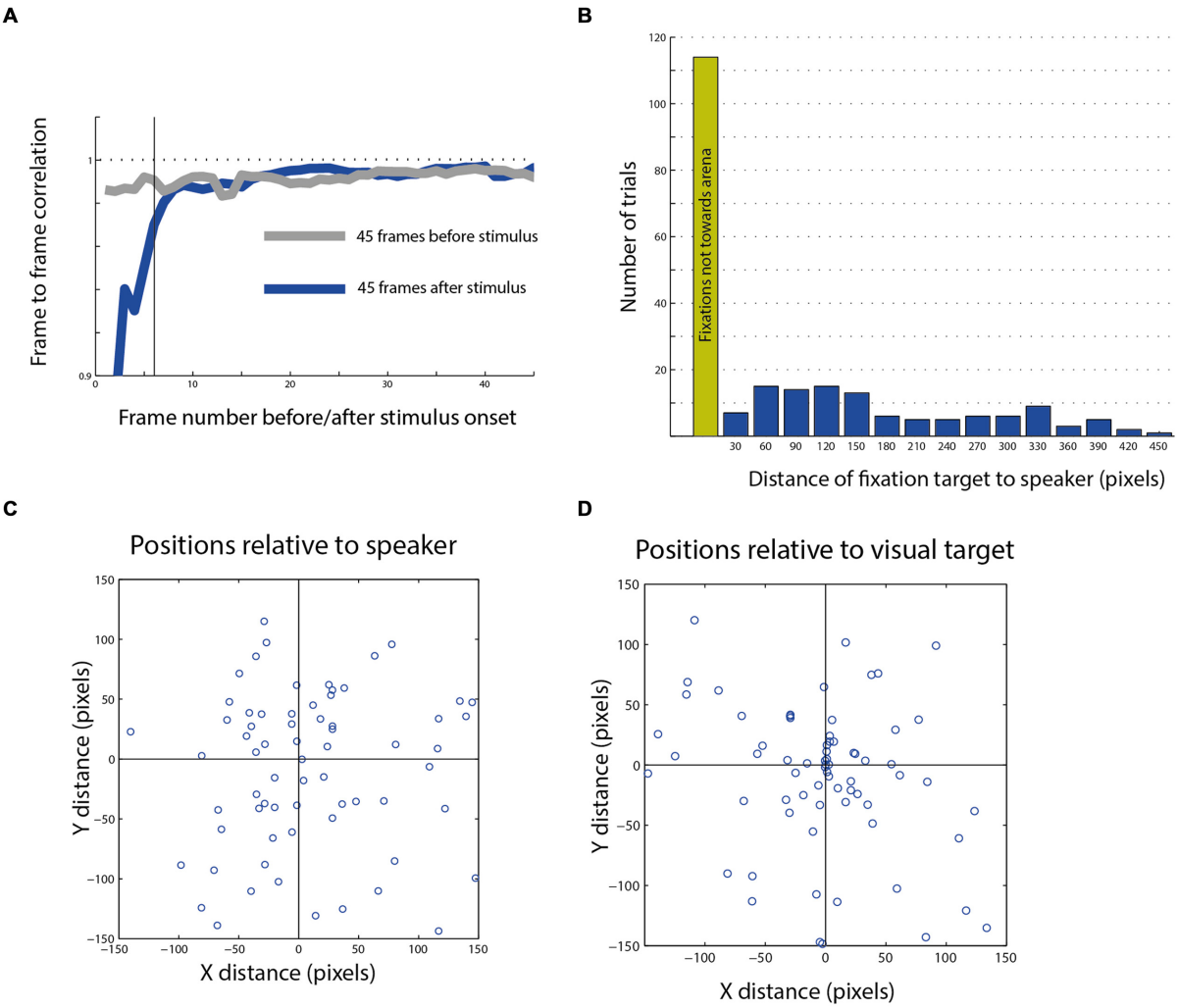


FIGURE 6 | Post stimulus head saccades end-points. (A) Average frame to frame correlation graph from 45 frames after the stimulus compared with the average graph from the 45 frames before the stimulus. The dashed horizontal line indicates a correlation of 1. The dark vertical line indicates the sixth frame after stimulus onset, after which the effect of the trigger signal on the correlation function disappears. **(B)** The distribution of the distances of fixation points from speaker location of all movements toward the arena (blue columns). The green column shows the number of fixation points that landed outside of the arena. **(C)** A scatterplot showing the population of fixation end-points which landed in the general area of the speaker plotted with relation to the speaker location (0,0 point). **(D)** The same population as in **(C)**, plotted with relation to the visual target closest to the speaker (0,0 point).

10,000 simulations we find the mean number of fixations on the quadrant close to the target to be 45 out of 144 trials with only 5 out of 10000 simulations giving a number above 66 (the value that we measured in our experiments). Any other distribution of the owls' spontaneous movements resulted with a less mean value and smaller percentage above 66. Thus, it is highly unlikely that the preference we see toward target 1 is due to mere probabilities and uneven distributions.

DISCUSSION

In this study, we investigated visual-auditory integration during spontaneous visual search behaviors in barn owls. The basic assumption in visual search experiments is that by measuring the probability of a target to attract gaze it is possible to estimate the perceived saliency of the target (Treue, 2003). By measuring gaze probabilities of barn owls in various environments Ohayon et al. (2008) have shown that this assumption holds true for barn owls as well.

Most previous behavioral studies of visual-auditory integration in animals used operant conditioning techniques to train the animals to pinpoint the location of a single modality stimulus and then measure the reaction time and accuracy of responses to unimodal and bimodal stimuli (Stein et al., 1988; Whitchurch and Takahashi, 2006; Schiller et al., 2012). In barn owls, it was shown that the reaction time and accuracy of responses to bimodal stimuli were not better than those of responses to unimodal stimuli. Yet, the response as a whole could still benefit from a bimodal stimulus by enjoying both the faster reaction time of an auditory response as well as the better accuracy of a visual response (Whitchurch and Takahashi, 2006). Thus, this previous study does not provide

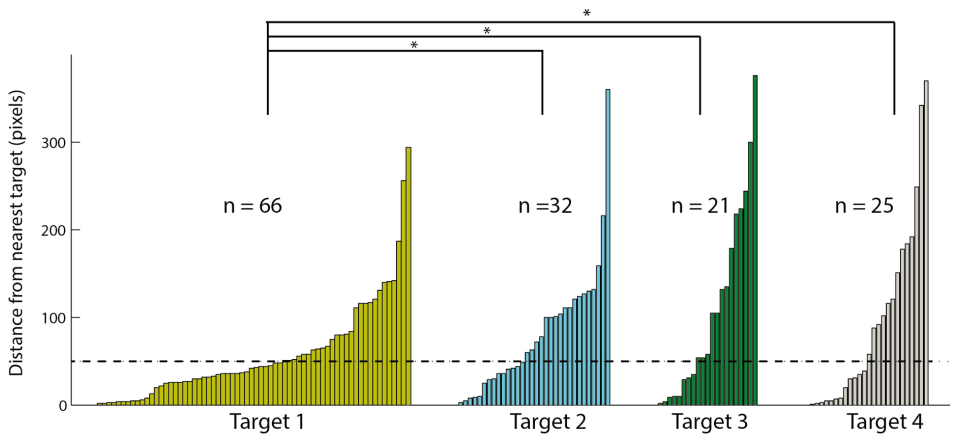


FIGURE 7 | The distances to the nearest visual target of all fixation points toward the arena. Data is divided to four targets based on the distance of the target to the speaker, target one is the closest to the speaker and target four the furthest. Within each target the results are sorted according to the distance, smaller distances on the left and larger distances on the right. Asterisks designate a significant ($p < 0.05$) difference in the average distance. The dashed line indicates the 50 pixels criterion line.

evidence that owls can improve accuracy or speed by integrating visual and auditory information. It is, however, possible that visual and auditory information is integrated, not for localizing single targets, but rather for enhanced saliency mapping (Talsma et al., 2010). To investigate this possibility, we adopted a different approach: we took advantage of the owl's natural tendency to scan the environment by abrupt shifts of gaze every few seconds (Ohayon et al., 2008). We first show that by doing so, the owls tend to bring targets of interest (i.e., high saliency) to project onto the retinal area centralis. Thus, by tracking the projection of the area centralis onto the scene, it is possible to estimate the perceived saliency of targets. We then estimated the saliency of targets as a function of their proximity to a hidden loudspeaker that occasionally produced a sound stimulus. Three possible outcomes are envisioned: one, that the owls will respond by turning their gaze to localize the sound-source; two, that the owls will ignore the sounds and continue scanning the environment independent of the sound source; and three, that the owls will not localize the sound source but rather adjust their visual search behavior according to the location of the sound source. The first two possibilities would indicate not a visual-auditory integration but rather a dominance of one modality over the other or competition between the two modalities. The latter possibility would indicate that information from the sound stimulus is integrated with visual information to create a joint functional saliency map (Onat et al., 2007). We have found that the owls tend to look more at the vicinity of the target close to the speaker. This observation is consistent with the third possibility above, but it can also be explained by the first possibility, i.e., that the owls from time to time localized the sound source independent of the visual target. However, if this was the case we would have expected that the mean distance of the head saccades closest to target 1 would be larger compared to the same values for targets 2–4 and that the hit rates on target will be smaller for target 1. In fact, the mean distance to the visual target was smaller for target 1 and the hit rates were not significantly different between the targets. Thus, our results support the third possibility: the visual target closest to the sound source attracted more attention compared to the other targets. These results therefore are consistent with the notion that visual-auditory integration is used for saliency mapping in barn owls.

A similar experimental approach was taken in a study of human visual search (Onat et al., 2007). Human subjects were exposed to images of natural scenes, and the probability of elements in the image to attract eye fixations were measured under three conditions: when no sounds were generated; when a speaker on the right side of the image was playing sounds; and when a speaker on the left side of the image was playing sounds. Using this apparatus, the authors showed that similar to results in this study the saliency of elements in the image are modulated according to their distance from the sound source. Fixations on visual elements on the side close to the sound source were more probable. A common notion in saliency mapping is that saliency is first mapped for individual features such as shape, color, modality, etc., and then the individual computations are combined into a global saliency map that integrates different modalities and features (Itti and Koch, 2000; Treue, 2003). According to this notion, saliency is biased toward the stimulus closest to the loudspeaker since in the global saliency map information is integrated with spatial information from the auditory sense. The similarity between the results of this study and the study of Onat et al. (2007) suggest that similar multisensory strategies are used by humans and barn owls to determine the next target for overt attention. This study, therefore, is in line with several recent behavioral studies in barn owls that point to similarities between humans and barn owls in visual search and attention allocation behaviors (Johnen et al., 2001; Ohayon et al., 2008; Harmening et al., 2011).

One clear observation in our results was that the sound stimulus itself seemed to be of low saliency relative to the visual targets. In most cases, it did not elicit a gaze shift, and when it did, it was seldom toward its source. In fact, it seemed that the probability of

changing a gaze shift after the sound was not elevated compared to spontaneous responses (**Figure 6A**). This is somewhat surprising, particularly since barn owls are well known for their accurate sound localization behaviors (Knudsen et al., 1979; Konishi, 2000), in some cases using hearing alone to capture mice in complete darkness (Payne, 1971). This lack of interest in the sound was despite the fact that the sounds were clearly audible, were composed of natural elements, and were very restricted in time. One likely explanation for lack of interest may be that the owls were used repeatedly in the same task. Barn owls habituate dramatically to repeating sounds (Netser et al., 2011). Therefore their responses to the auditory stimulus in this experiment are expected to habituate over time. It is plausible that in natural unfamiliar cases where sounds may carry behaviorally relevant information, the owls will respond considerably more to sudden sounds. In the conditions of the current experiments, the owls learned that the sounds carry no behavioral meaning and were of low saliency. However, although this low saliency was not enough to compete with the saliency of the visual targets, it was still enough to bias their perceived saliency.

In the cases where, following the sounds, the owls changed their gaze toward the general direction of the speaker, the gaze shifts tended to acquire the visual target and not localize the sound source (**Figures 6C,D**). The tendency to shift gaze toward close visual targets and not directly to the auditory source is consistent with the notion of visual capture or ventriloquism (Recanzone, 2009). It is well documented that humans and primates, when confronted with visual and auditory stimuli that are incongruent in space, tend to localize the sound as if coming from the location of the visual stimulus (Recanzone, 1998; Woods and Recanzone, 2004), and accordingly increase the saliency of the visual stimulus location (Spence and Driver, 2000). Visual capture makes sense because in most conditions the reliability of visual localization is larger than the reliability of auditory localization. Although no direct evidence for visual capture in barn owls has been reported, experiments with prismatic spectacles suggest that it does take place (Knudsen and Knudsen, 1989; Gutfreund and Knudsen, 2004). We therefore hypothesize that in clutter environments where visual features are abundant, the owls will show a tendency to acquire visual targets rather than the location of the auditory stimulus. It would be interesting to find out whether the owls will continue to be attracted to the salient visual targets close to the sound source more than to the sound source itself, if the experimental conditions could be varied to make the sound more salient to the owl, for example, by associating the sound with a food reward or by creating more natural unpredicted conditions.

An open question is where in the brain the interactions between visual and auditory signals for saliency mapping take place? In mammals, focus has been drawn mostly to the lateral intraparietal cortex [LIP; Bisley and Goldberg (2003)] and to the SC (Stanford et al., 2005; Fecteau and Munoz, 2006). In birds, the analog forebrain area to LIP is unknown, however, the homolog region to the SC is the OT (Jarvis et al., 2005). The OT in barn owls has been studied extensively, and a series of studies recently point to the notion that activity in the OT reflects the saliency mapping necessary for overt selective attention (Winkowski and Knudsen, 2007; Reches and Gutfreund, 2008; Zahar et al., 2009; Mysore et al., 2010, 2011; Netser et al., 2011). Thus, it is likely that the behavioral results observed in this study reflect activity of tectal neurons. In barn owls, tectal neurons that integrate visual and auditory signals (multisensory neurons) are highly abundant (Knudsen, 1982; Zahar et al., 2009), more common compared to the OT of other avian species (Wang, 2003) or to the SC of mammals (King and Palmer, 1985; Populin and Yin, 2002). It was shown that multisensory neurons in the barn owl's OT tend to combine visual and auditory signals in a supra-linear manner, if the stimuli are congruent in space and time and if the stimuli are surprising (Zahar et al., 2009). These findings resemble rules of visual-auditory integration found in the SC (Meredith and Stein, 1983). However, the behavioral paradigm of the current study requires a different type of interaction. Here, the auditory stimuli were abrupt and scarce in time, while the visual scene was fixed (in room coordinates). Thus, if the saliency of the scene is expressed in the activity of neurons in the tectal map, we expect the auditory stimulus to modulate the activity in the map over a time span larger than the duration of the stimulus itself. Short-term auditory memory in tectal neurons has been reported for periods of up to a minute and possibly more (Reches and Gutfreund, 2008; Netser et al., 2011). Thus, the substrate for modulating visual activity by short localized auditory stimuli exists in the OT.

CONCLUSION

In this study we report that auditory information biases spontaneous visual search in barn owls. This suggests that auditory-visual integration takes place at early pre-attentive stages in order to guide spatial attention. These findings are consistent with a model proposed by Itti and Koch (2000) to explain bottom–up mechanisms of visual search in primates. Our findings therefore suggests that similar, pre-attentive, visual-auditory integration takes place in non-mammalian species as well, pointing to the generality of such integration. In nature the visual and the auditory scenes are highly dependent. Thus, it makes sense to integrate the two modalities at early stages to enhance the important behavior of identifying the most salient target for attentional responses.

ACKNOWLEDGMENTS

We thank Wolf Harmening and Alex Porotskiy for developing and assembling the OwlCam and Alon Gutfreund for producing **Figure 1**. This work was supported by grants from the Israel Science Foundation (ISF) and the German-Israeli Foundation (GIF).

REFERENCES

Beauchamp, M. S. (2005). See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex. *Curr. Opin. Neurobiol.* 15, 145–153. doi: 10.1016/j.conb.2005.03.011

Berman, R., and Colby, C. (2009). Attention and active vision. *Vision Res.* 49, 1233–1248. doi: 10.1016/j.visres.2008.06.017

Bisley, J. W., and Goldberg, M. E. (2003). Neuronal activity in the lateral intraparietal area and spatial attention. *Science* 299, 81–86. doi: 10.1126/science.1077395

Boehnke, S. E., and Munoz, D. P. (2008). On the importance of the transient visual response in the superior colliculus. *Curr. Opin. Neurobiol.* 18, 544–551. doi: 10.1016/j.conb.2008.11.004

Driver, J., and Spence, C. (1998). Cross-modal links in spatial attention. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 353, 1319–1331. doi: 10.1098/rstb.1998.0286

du Lac, S., and Knudsen, E. I. (1990). Neural maps of head movement vector and speed in the optic tectum of the barn owl. *J. Neurophysiol.* 63, 131–146.

Dutta, A., and Gutfreund, Y. (2014). Saliency mapping in the optic tectum and its relationship to habituation. *Front. Integr. Neurosci.* 8:1. doi: 10.3389/fnint.2014.00001

Eckstein, M. P. (2011). Visual search: a retrospective. *J. Vis.* 11:14. doi: 10.1167/11.5.14

Fecteau, J. H., and Munoz, D. P. (2006). Salience, relevance, and firing: a priority map for target selection. *Trends Cogn. Sci.* 10, 382–390. doi: 10.1016/j.tics.2006.06.011

Frassinetti, F., Bolognini, N., and Ladavas, E. (2002). Enhancement of visual perception by crossmodal visuo-auditory interaction. *Exp. Brain Res.* 147, 332–343. doi: 10.1007/s00221-002-1262-y

Gutfreund, Y. (2012). Stimulus-specific adaptation, habituation and change detection in the gaze control system. *Biol. Cybern.* 106, 657–668. doi: 10.1007/s00422-012-0497-3

Gutfreund, Y., and Knudsen, E. I. (2004). "Visual instruction of the auditory space map in the midbrain," in *The Handbook of Multisensory Processes*, eds G. Calvert, C. Spence, and B. E. Stein (Cambridge, MA: The MIT press), 613–624.

Gutfreund, Y., Zheng, W., and Knudsen, E. I. (2002). Gated visual input to the central auditory system. *Science* 297, 1556–1559. doi: 10.1126/science.1073712

Harmening, W. M., Orlowski, J., Ben-Shahar, O., and Wagner, H. (2011). Overt attention toward oriented objects in free-viewing barn owls. *Proc. Natl. Acad. Sci. U.S.A.* 108, 8461–8466. doi: 10.1073/pnas.1101582108

Hayhoe, M., and Ballard, D. (2005). Eye movements in natural behavior. *Trends Cogn. Sci.* 9, 188–194. doi: 10.1016/j.tics.2005.02.009

Itti, L., and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res.* 40, 1489–1506. doi: 10.1016/S0042-6989(99)00163-7

Jarvis, E. D., Gunturkun, O., Bruce, L., Csillag, A., Karten, H., Kuenzel, W., et al. (2005). Avian brains and a new understanding of vertebrate brain evolution. *Nat. Rev. Neurosci.* 6, 151–159. doi: 10.1038/nrn1606

Johnen, A., Wagner, H., and Gaese, B. H. (2001). Spatial attention modulates sound localization in barn owls. *J. Neurophysiol.* 85, 1009–1012.

Kayser, C., Petkov, C. I., Augath, M., and Logothetis, N. K. (2007). Functional imaging reveals visual modulation of specific fields in auditory cortex. *J. Neurosci.* 27, 1824–1835. doi: 10.1523/JNEUROSCI.4737-06.2007

King, A. J., and Palmer, A. R. (1985). Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus. *Exp. Brain Res.* 60, 492–500. doi: 10.1007/BF00236934

Knudsen, E. I. (1982). Auditory and visual maps of space in the optic tectum of the owl. *J. Neurosci.* 2, 1177–1194.

Knudsen, E. I. (1987). Neural derivation of sound source location in the barn owl. An example of a computational map. *Ann. N. Y. Acad. Sci.* 510, 33–38. doi: 10.1111/j.1749-6632.1987.tb43463.x

Knudsen, E. I., Blasdel, G. G., and Konishi, M. (1979). Sound localization by the barn owl (*Tyto alba*) measured with the search coil technique. *J. Comp. Physiol.* 133, 1–11. doi: 10.1007/BF00663105

Knudsen, E. I., and Knudsen, P. F. (1989). Vision calibrates sound localization in developing barn owls. *J. Neurosci.* 9, 3306–3313.

Konishi, M. (2000). Study of sound localization by owls and its relevance to humans. *Comp. Biochem. Physiol. A. Mol. Integr. Physiol.* 126, 459–469. doi: 10.1016/S1095-6433(00)00232-4

Luck, S. J., and Ford, M. A. (1998). On the role of selective attention in visual perception. *Proc. Natl. Acad. Sci. U.S.A.* 95, 825–830. doi: 10.1073/pnas.95.3.825

MacInnes, W. J., Hunt, A. R., Hilchey, M. D., and Klein, R. M. (2014). Driving forces in free visual search: an ethology. *Atten. Percept. Psychophys.* 76, 280–295. doi: 10.3758/s13414-013-0608-9

Masino, T., and Knudsen, E. I. (1990). Horizontal and vertical components of head movement are controlled by distinct neural circuits in the barn owl. *Nature* 345, 434–437. doi: 10.1038/345434a0

Meredith, M. A., Nemitz, J. W., and Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *J. Neurosci.* 7, 3215–3229.

Meredith, M. A., and Stein, B. E. (1983). Interactions among converging sensory inputs in the superior colliculus. *Science* 221, 389–391. doi: 10.1126/science.6867718

Meredith, M. A., and Stein, B. E. (1986). Spatial factors determine the activity of multisensory neurons in cat superior colliculus. *Brain Res.* 365, 350–354. doi: 10.1016/0006-8993(86)91648-3

Mysore, S. P., Asadollahi, A., and Knudsen, E. I. (2010). Global inhibition and stimulus competition in the owl optic tectum. *J. Neurosci.* 30, 1727–1738. doi: 10.1523/JNEUROSCI.3740-09.2010

Mysore, S. P., Asadollahi, A., and Knudsen, E. I. (2011). Signaling of the strongest stimulus in the owl optic tectum. *J. Neurosci.* 31, 5186–5196. doi: 10.1523/JNEUROSCI.4592-10.2011

Mysore, S. P., and Knudsen, E. I. (2011). The role of a midbrain network in competitive stimulus selection. *Curr. Opin. Neurobiol.* 21, 653–660. doi: 10.1016/j.conb.2011.05.024

Mysore, S. P., and Knudsen, E. I. (2013). A shared inhibitory circuit for both exogenous and endogenous control of stimulus selection. *Nat. Neurosci.* 16, 473–478. doi: 10.1038/nn.3352

Netser, S., Ohayon, S., and Gutfreund, Y. (2010). Multiple manifestations of microstimulation in the optic tectum: eye movements, pupil dilations, and sensory priming. *J. Neurophysiol.* 104, 108–118. doi: 10.1152/jn.01142.2009

Netser, S., Zahar, Y., and Gutfreund, Y. (2011). Stimulus-specific adaptation: can it be a neural correlate of behavioral habituation? *J. Neurosci.* 31, 17811–17820. doi: 10.1523/JNEUROSCI.4790-11.2011

Ohayon, S., Harmening, W., Wagner, H., and Rivlin, E. (2008). Through a barn owl's eyes: interactions between scene content and visual attention. *Biol. Cybern.* 98, 115–132. doi: 10.1007/s00422-007-0199-4

Ohayon, S., van der Willigen, R. F., Wagner, H., Katsman, I., and Rivlin, E. (2006). On the barn owl's visual pre-attack behavior: I. Structure of head movements and motion patterns. *J. Comp. Physiol. A Neuroethol. Sens. Neural Behav. Physiol.* 192, 927–940. doi: 10.1007/s00359-006-0130-9

Onat, S., Libertus, K., and Konig, P. (2007). Integrating audiovisual information for the control of overt attention. *J. Vis.* 7, 11.1–11.16.

Payne, R. S. (1971). Acoustic location of prey by barn owls (*Tyto alba*). *J. Exp. Biol.* 54, 535–573.

Pettigrew, J. D. (1979). Binocular visual processing in the owl's telencephalon. *Proc. R. Soc. Lond. B Biol. Sci.* 204, 435–454. doi: 10.1098/rspb.1979.0038

Populin, L. C., and Yin, T. C. (2002). Bimodal interactions in the superior colliculus of the behaving cat. *J. Neurosci.* 22, 2826–2834.

Recanzone, G. H. (1998). Rapidly induced auditory plasticity: the ventriloquism aftereffect. *Proc. Natl. Acad. Sci. U.S.A.* 95, 869–875. doi: 10.1073/pnas.95.3.869

Recanzone, G. H. (2009). Interactions of auditory and visual stimuli in space and time. *Hear. Res.* 258, 89–99. doi: 10.1016/j.heares.2009.04.009

Reches, A., and Gutfreund, Y. (2008). Stimulus-specific adaptations in the gaze control system of the barn owl. *J. Neurosci.* 28, 1523–1533. doi: 10.1523/JNEUROSCI.3785-07.2008

Reches, A., and Gutfreund, Y. (2009). Auditory and multisensory responses in the tectofugal pathway of the barn owl. *J. Neurosci.* 29, 9602–9613. doi: 10.1523/JNEUROSCI.6117-08.2009

Reches, A., Netser, S., and Gutfreund, Y. (2010). Interactions between stimulus-specific adaptation and visual auditory integration in the forebrain of the barn owl. *J. Neurosci.* 30, 6991–6998. doi: 10.1523/JNEUROSCI.5723-09.2010

Reinagel, P., and Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network* 10, 341–350. doi: 10.1088/0954-898X/10/4/304

Schiller, P. H., Kwak, M. C., and Slocum, W. M. (2012). Visual and auditory cue integration for the generation of saccadic eye movements in monkeys and lever pressing in humans. *Eur. J. Neurosci.* 36, 2500–2504. doi: 10.1111/j.1460-9568.2012.08133.x

Spence, C., and Driver, J. (2000). Attracting attention to the illusory location of a sound: reflexive crossmodal orienting and ventriloquism. *Neuroreport* 11, 2057–2061. doi: 10.1097/00001756-200006260-00049

Stanford, T. R., Quessy, S., and Stein, B. E. (2005). Evaluating the operations underlying multisensory integration in the cat superior colliculus. *J. Neurosci.* 25, 6499–6508. doi: 10.1523/JNEUROSCI.5095-04.2005

Stein, B. E., Huneycutt, W. S., and Meredith, M. A. (1988). Neurons and behavior: the same rules of multisensory integration apply. *Brain Res.* 448, 355–358. doi: 10.1016/0006-8993(88)91276-0

Stein, B. E., and Meredith, M. A. (1993). *The Merging of the Senses.* Cambridge, MA: MIT Press.

Stein, B. E., and Stanford, T. R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nat. Rev. Neurosci.* 9, 255–266. doi: 10.1038/nrn2331

Stein, B. E., Stanford, T. R., and Rowland, B. A. (2014). Development of multisensory integration from the perspective of the individual neuron. *Nat. Rev. Neurosci.* 15, 520–535. doi: 10.1038/nrn3742

Steinbach, M. J., and Money, K. E. (1973). Eye movements of the owl. *Vision Res.* 13, 889–891. doi: 10.1016/0042-6989(73)90055-2

Sugihara, T., Diltz, M. D., Averbeck, B. B., and Romanski, L. M. (2006). Integration of auditory and visual communication information in the primate ventrolateral prefrontal cortex. *J. Neurosci.* 26, 11138–11147. doi: 10.1523/JNEUROSCI.3550-06.2006

Talsma, D., Senkowski, D., Soto-Faraco, S., and Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends Cogn. Sci.* 14, 400–410. doi: 10.1016/j.tics.2010.06.008

Treue, S. (2003). Visual attention: the where, what, how and why of saliency. *Curr. Opin. Neurobiol.* 13, 428–432. doi: 10.1016/S0959-4388(03)00105-3

Wallace, M. T., Carriere, B. N., Perrault, T. J. Jr., Vaughan, J. W., and Stein, B. E. (2006). The development of cortical multisensory integration. *J. Neurosci.* 26, 11844–11849. doi: 10.1523/JNEUROSCI.3295-06.2006

Wallace, M. T., Wilkinson, L. K., and Stein, B. E. (1996). Representation and integration of multiple sensory inputs in primate superior colliculus. *J. Neurophysiol.* 76, 1246–1266.

Wang, S. R. (2003). The nucleus isthmi and dual modulation of the receptive field of tectal neurons in non-mammals. *Brain Res. Brain Res. Rev.* 41, 13–25. doi: 10.1016/S0165-0173(02)00217-5

Wathey, J. C., and Pettigrew, J. D. (1989). Quantitative analysis of the retinal ganglion cell layer and optic nerve of the barn owl *Tyto alba. Brain Behav. Evol.* 33, 279–292. doi: 10.1159/000115936

Whitchurch, E. A., and Takahashi, T. T. (2006). Combined auditory and visual stimuli facilitate head saccades in the barn owl (*Tyto alba*). *J. Neurophysiol.* 96, 730–745. doi: 10.1152/jn.00072.2006

Winkowski, D. E., and Knudsen, E. I. (2007). Top-down control of multimodal sensitivity in the barn owl optic tectum. *J. Neurosci.* 27, 13279–13291. doi: 10.1523/JNEUROSCI.3937-07.2007

Wolfe, J. M., and Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nat. Rev. Neurosci.* 5, 495–501. doi: 10.1038/nrn1411

Woods, T. M., and Recanzone, G. H. (2004). Visually induced plasticity of auditory spatial perception in macaques. *Curr. Biol.* 14, 1559–1564. doi: 10.1016/j.cub.2004.08.059

Zahar, Y., Reches, A., and Gutfreund, Y. (2009). Multisensory enhancement in the optic tectum of the barn owl: spike count and spike timing. *J. Neurophysiol.* 101, 2380–2394. doi: 10.1152/jn.91193.2008

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 11 November 2014; accepted: 28 January 2015; published online: 13 February 2015.

Citation: Hazan Y, Kra Y, Yarin I, Wagner H and Gutfreund Y (2015) Visual-auditory integration for visual search: a behavioral study in barn owls. Front. Integr. Neurosci. ***9****:11. doi: 10.3389/fnint.2015.00011*

This article was submitted to the journal Frontiers in Integrative Neuroscience.

frontiers in
INTEGRATIVE NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 27 February 2015
doi: 10.3389/fnint.2015.00016

A spatially collocated sound thrusts a flash into awareness

*Máté Aller[1], Anette Giani[2], Verena Conrad[2], Masataka Watanabe[2] and Uta Noppeney[1,2]**

[1] Computational Cognitive Neuroimaging Laboratory, Computational Neuroscience and Cognitive Robotics Centre, University of Birmingham, Birmingham, UK
[2] Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Edited by:
Salvador Soto-Faraco, Universitat Pompeu Fabra, Spain

Reviewed by:
Elliot D. Freeman, City University, UK
Yi-Chuan Chen, Lancaster University, UK
Claudia Lunghi, Università degli Studi di Firenze, Italy

***Correspondence:**
Uta Noppeney, Computational Cognitive Neuroimaging Laboratory, Computational Neuroscience and Cognitive Robotics Centre, University of Birmingham, B15 2TT Birmingham, UK
e-mail: U.Noppeney@bham.ac.uk

To interact effectively with the environment the brain integrates signals from multiple senses. It is currently unclear to what extent spatial information can be integrated across different senses in the absence of awareness. Combining dynamic continuous flash suppression (CFS) and spatial audiovisual stimulation, the current study investigated whether a sound facilitates a concurrent visual flash to elude flash suppression and enter perceptual awareness depending on audiovisual spatial congruency. Our results demonstrate that a concurrent sound boosts unaware visual signals into perceptual awareness. Critically, this process depended on the spatial congruency of the auditory and visual signals pointing towards low level mechanisms of audiovisual integration. Moreover, the concurrent sound biased the reported location of the flash as a function of flash visibility. The spatial bias of sounds on reported flash location was strongest for flashes that were judged invisible. Our results suggest that multisensory integration is a critical mechanism that enables signals to enter conscious perception.

Keywords: multisensory integration, awareness, attention, consciousness, audiovisual, perception, ventriloquism, perceptual illusion

INTRODUCTION

For effective interactions an organism needs to merge signals from different senses into a coherent and unified percept of the environment. A controversial question is to which extent multisensory integration is automatic or relies on higher cognitive resources such as attention or awareness (for review see Talsma et al., 2010). Even though recent studies have demonstrated that awareness and attention can be dissociated (Koch and Tsuchiya, 2007, 2012; Wyart and Tallon-Baudry, 2008; Watanabe et al., 2011), in many situations attention and awareness are closely intertwined. Hence, for the purpose of this study we do not yet intend to dissociate these aspects, but loosely define "automatic integration" as integration that is relatively immune to attention and awareness. According to the account of automatic integration multisensory co-stimulation increases the bottom-up stimulus saliency (Onat et al., 2007). Thus, signals that co-occur within a spatial and temporal window of integration can automatically amplify stimulus salience. Multisensory integration thereby enables multisensory events to enter perceptual awareness and capture an organism's attention.

In support of automatic integration a vast body of psychophysics and neurophysiological research has shown that multisensory integration is immune to attentional modulation (Bertelson et al., 2000; Vroomen et al., 2001; Stekelenburg et al., 2004; Bresciani et al., 2006), emerges prior to participants' awareness (Alsius and Munhall, 2013) and even persists in the anesthetized non-human primate brain (e.g., superior colliculus, primary sensory areas) (Kayser et al., 2005; Stanford et al., 2005). Yet, the account of "automatic" integration has more recently been challenged. For instance, the audiovisual McGurk illusion falters, when attention is diverted to a secondary task (Alsius et al., 2005) or when subjects are unaware of the visual speech gestures (Munhall et al., 2009). Moreover, neuroimaging studies have shown profound attentional modulation of neural multisensory integration indices. Thus, attention modulated the amplification of the BOLD response for congruent audiovisual speech signals in superior colliculi, primary sensory and association cortices (Fairhall and Macaluso, 2009). Likewise, EEG studies showed attentional influences on audiovisual interactions already at $\leq$100 ms poststimulus (Talsma et al., 2007). With respect to perceptual awareness, the role of primary sensory areas is still debated. While numerous studies have demonstrated that activations in primary sensory areas correlate with participants' awareness (Tong, 2003), others have suggested that these activations may be mediated by concurrent attentional effects (Watanabe et al., 2011). Collectively, this body of research suggests a multifaceted and not yet completely understood interplay between multisensory integration and higher cognitive processes such as attention or awareness (Talsma et al., 2010).

This intricate relationship partly results from the hierarchical nature of multisensory perception where different types of information (e.g., temporal, spatial, semantic, phonological) are integrated at distinct cortical levels (Bonath et al., 2007; Driver and Noesselt, 2008; Lewis and Noppeney, 2010; Werner and Noppeney, 2010; Lee and Noppeney, 2011, 2014). Conversely, perceptual awareness and attentional capture rely on a cascade of neural processes. Thus, experiments using masking (Chen and Spence, 2011), attentional blink (Soto-Faraco and Spence, 2002; Olivers and Van der Burg, 2008; Adam and Noppeney, 2014), binocular/perceptual rivalry (Hupé et al., 2008; van Ee et al., 2009; Alais et al., 2010; Conrad et al., 2010, 2012, 2013; Lunghi et al.,

2010, 2014; Zhou et al., 2010; Guzman-Martinez et al., 2012; Klink et al., 2012; Lunghi and Alais, 2013; Lunghi and Morrone, 2013) or flash suppression (Palmer and Ramsey, 2012; Alsius and Munhall, 2013) are likely to perturb the interplay between perceptual awareness and multisensory integration at different processing stages (for related discussion focusing on visual context, see Fogelson et al., 2014; Peremen and Lamy, 2014; for a recent review see Deroy et al., 2014). In particular, using binocular rivalry numerous studies have demonstrated that a concurrent non-visual signal increases the dominance and decreases the suppression times of the congruent visual percept. Yet, because of the presence of two rivaling percepts, these binocular rivalry experiments make it more difficult to unambiguously determine that the rivalry dynamics was shaped by interactions between the non-visual signals with the suppressed rather than the dominant percept (for further discussion, please see Conrad et al., 2010).

Continuous flash suppression (CFS) is a powerful technique to manipulate participants' perceptual awareness (Tsuchiya and Koch, 2005). Flashing a mask to one eye can render even a salient stimulus presented to the other eye invisible. Critically, CFS is thought to affect cortical activity already at the primary cortical level via a gain control mechanism (Yuval-Greenberg and Heeger, 2013). CFS thus provides a very useful paradigm to investigate whether a concurrent non-visual signal can counteract the effect of flash suppression at the primary cortical level. Indeed, a previous study has demonstrated that an auditory speech signal makes participants more likely to detect a congruent relative to an incongruent speech video under CFS (Alsius and Munhall, 2013; see also Palmer and Ramsey, 2012). These results suggest that audiovisual synchrony and temporal correlations are important determinants for audiovisual interactions prior to participants' awareness. Moreover, as natural speech signals evolve continuously over time, temporal expectations may also play an important role in enabling participants to detect visual speech signals.

Yet, as this previous study has presented auditory and visual signals only in a spatially congruent fashion, it could not evaluate the role of spatial congruency, which is another critical cue for multisensory binding. Spatial congruency may enable multisensory interactions via at least two mechanisms. First, spatial congruency may act as a bottom-up cue informing the brain that two signals are likely to come from a common source and should hence be bound into a coherent percept. Second, a spatially collocated sound may reduce the spatial uncertainty about a concurrent flash. Even though spatial congruency affects detection performance only rarely in redundant target paradigm (Forster et al., 2002; Bertini et al., 2008) the second mechanism may be more important in paradigms where the visual signal has been strongly attenuated by various experimental manipulations such as flash suppression or masking. Spatial uncertainty may be reduced via bottom-up mechanisms that enable the formation of more precise audiovisual spatial salience maps. Alternatively, a co-located sound may reduce spatial uncertainty even via top-down expectations that stabilize visual representations potentially even after they have accessed awareness.

Previous studies have demonstrated that a sound increases the detectability of a collocated yet masked visual flash at threshold visibility (Frassinetti et al., 2002; Bolognini et al., 2005). Yet, as these masking studies reduced flash detectability only to threshold performance of 70%, the suppression of awareness for the undetected stimuli was rather shallow. Moreover, it is still unknown whether masking and dynamic CFS reduce visual awareness via similar neural mechanisms (Fogelson et al., 2014; Peremen and Lamy, 2014).

To further investigate the role of spatial congruency in multisensory integration prior to perceptual awareness, the current study combined spatial audiovisual stimulation with dynamic CFS (Tsuchiya and Koch, 2005; Maruya et al., 2008). On each trial, participants were presented with a single flash in the center, their left or right hemifield together with a sound that was spatially congruent or incongruent. Participants located the flash (i.e., flash localization) and judged its visibility (i.e., visual detection task). First, we investigated whether participants were better at detecting the flash when the sound was spatially collocated. We hypothesized that spatial constraints are critical for audiovisual integration processes prior to participants' awareness. Second, we investigated whether the concurrent sound biased participants' perceived flash location and whether this bias depended on flash visibility. Importantly, as CFS obliterated visual awareness only in a fraction of trials, we were able to compare the audiovisual spatial bias for physically identical flashes that were visible or invisible.

MATERIALS AND METHODS

PARTICIPANTS

After giving informed consent, 24 healthy young adults with normal or corrected-to-normal vision participated in this study (14 females, mean age: 26.7 years, standard deviation: 5.3, range: 18–40; 22 right-handed). One subject was excluded because she did not follow task instructions properly as she located the visual stimuli almost exclusively in the center (98.5%, (group mean $\pm$ SD): 35.7% $\pm$ 17.5%). The study was approved by the local ethics review board of the University of Tübingen.

STIMULI AND APPARATUS

Participants sat in a dimly lit room in front of a computer monitor at a viewing distance of 1 m. They viewed one half of the monitor with each eye using a custom-built mirror stereoscope. Visual stimuli were composed of targets and masks that were presented on a gray, uniform background with a mean luminance of 15.5 cd/m^2. One eye viewed the target stimuli, the other eye the masks.

The target stimuli were three gray discs (Ø 0.29°, mean luminance: 25.4 cd/m^2), located in the center and 5.72° visual angle to the left and right. On each trial, the color of exactly one of the targets changed to white (mean luminance: 224.2 cd/m^2) for a duration of 100 ms. This change in brightness will be referred to as "flash". To suppress the flash's perceptual visibility, the other eye was shown three dynamic Mondrians (Ø 2°, mean luminance: 35.6 cd/m^2) (Tsuchiya and Koch, 2005; Maruya et al., 2008). We employed dynamic CFS, as this proved a powerful and reliable method to suppress perceptual awareness of a brief and hence relatively salient flash. To match the target's

location the Mondrians were also located in the center or 5.72° to the left and right of the fixation dot. Each Mondrian consisted of sinusoidal gratings (Ø 0.57°) which changed their color and position randomly at a frequency of 10 Hz. Each grating's texture was shifted every 16.6 ms to generate apparent motion. Visual stimuli were presented with a fixation spot in the center of the screen and were framed by a gray, isoluminant square aperture of 8.58° × 13.69° in diameter to aid binocular fusion.

Auditory stimuli were pure tones with a carrier frequency of 1 kHz and a duration of 100 ms. They were presented via four external speakers, placed above and below the monitor. Upper and lower speakers were aligned vertically and located 2.3° to the left and 2.3° to the right of the monitor's center. Speakers' location was chosen by trading off physical alignment of visual and auditory stimulus locations and sound localization performance. Moreover, it traded off optimization for the two research questions we addressed in this study: (i) the role of audiovisual localization; and (ii) auditory bias on perceived visual location. At a distance of 2.3° mean sound localization accuracy amounted to ∼70%.

Psychophysical stimuli were generated and presented on a PC running Windows XP using the Psychtoolbox version 3 (Brainard, 1997; Kleiner et al., 2007) running on Matlab 7 (Mathworks, Nantucket, Massachusetts). Visual stimuli were presented dichoptically using a gamma-corrected 30" LCD monitor with a resolution of 2560 × 1600 pixels at a frame rate of 60 Hz (GeForce 8600GT graphics card). Auditory stimuli were digitized at a sampling rate of 44.8 kHz via an M-Audio Delta 1010LT sound card and presented at a maximal amplitude of 73 dB sound pressure level. Exact audiovisual onset timing was confirmed by recording visual and auditory signals concurrently with a photo-diode and a microphone.

EXPERIMENTAL DESIGN

Participants were presented with an auditory beep emanating from either the left or right. In synchrony with the beep, one eye was presented with a brief flash in the center or participants' left or right hemifield. The visibility of the flash was suppressed by presenting masks to the other eye using the method of dynamic CFS (Maruya et al., 2008). Hence, the 3 × 2 factorial design manipulated (1) "flash location" (3 levels: left, center, right) and (2) "sound location" (2 levels: left, right) (**Figure 1A**). On each trial, participants located the flash (left, right or center). Moreover, they performed a graded detection task by judging the visibility of the flash (invisible, unsure, visible).

This experimental design enabled us to address two questions: First, we investigated whether participants were better at detecting the flash, when auditory and visual signals were approximately collocated. Second, as the flash was visible only in a fraction of trials, we were able to quantify the effect of sound on localizing physically identical flashes that were visible or invisible.

EXPERIMENTAL PROCEDURE

As seen in **Figure 1B**, each trial started with the presentation of the fixation dot for a duration of 1000 ms. Next, participants' one eye was presented with three gray discs, located in the center, 5.72° visual angle to the left and right. Participants' awareness of these discs was suppressed by showing dynamic Mondrians at the corresponding locations to the other eye (i.e., dynamic CFS). The Mondrian masks and the discs were presented on the screen until participants had responded to all questions. The assignment of eyes to disks or masks was changed after each trial, to enhance suppression. After a random interval of 500–1000 ms one of the three discs "flashed", i.e., changed its luminance for a duration of 100 ms. In synchrony with the flash, an auditory beep was played from the left or right. In addition, on 22.2% of the trials, the so-called catch trials, participants were also asked to locate the sound (left vs. right discrimination; in addition to the visibility judgment and flash localization). This allowed us to assess the spatial information that is available for sound localization. Moreover, it ensures that participants did not completely ignore the sound.

Participants responded by pressing one of three buttons on a keyboard. The button assignment was counterbalanced across participants as follows: Participants used three sets of buttons to respond to the three question types (flash localization, sound localization (on catch trials only) and visibility judgment). Each set contained three buttons, one central, one to the left and one to the right. One set of buttons was operated with one hand and the other two sets were operated with the other hand. The association of the hands to the button sets was counterbalanced across participants. Moreover, we also counterbalanced the button response assignment for the flash visibility question. Within subjects we counterbalanced the two possible question orders (i.e., (i) flash localization, (ii) sound localization (only on catch trials), (iii) visibility judgment; alternatively: (i) sound localization (only on catch trials), (ii) flash localization, (iii) visibility judgment).

Prior to the main experiment, participants were familiarized with stimuli and task. First, they completed 2–3 sessions of sound localization. Next, there were two short practice sessions of the main paradigm. During the main experiment participants completed a total of 24 experimental sessions distributed over two successive days, resulting in a total of 1296 trials (i.e., 216 trials per condition).

ANALYSIS

Our analysis addressed two questions:

Effect of spatial congruency on visibility judgment

We investigated whether a synchronous sound boosts "a suppressed visual signal" into participants' awareness depending on spatial congruency. In other words, we asked whether participants were better at detecting a flash, when the sound was approximately collocated with the flashing disc. Visibility judgment as the dependent variable was quantified as the percentage of non-catch trials judged as visible. As participants' visibility judgment depended on stimulus eccentricity, we limited this analysis only to those trials with left/right flashes and excluded trials with flashes in the center. Moreover, we pooled over the left and right hemifield as there was no significant difference between left and right hemifield in percentage judged visible. Hence, congruent conditions included flash left/sound left and flash right/sound right combination. Likewise,

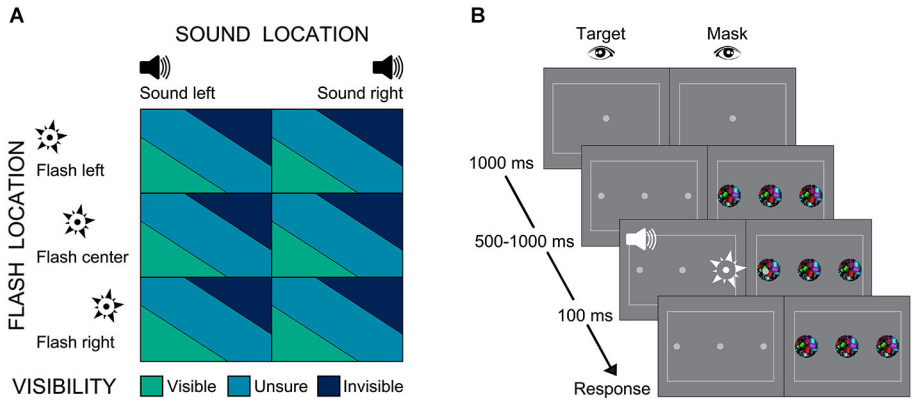


FIGURE 1 | Experiment paradigm and sample trial. **(A)** Experiment design 2 × 3 factorial design with factors: (i) Sound location: left, right; (ii) Flash location: left, center, right. **(B)** Example trial and procedure of dynamic flash suppression.

incongruent conditions included flash left/sound right and flash right/sound left combinations. We performed paired *t*-tests to compare participants' visibility judgment between congruent and incongruent conditions. However, to be consistent with the statistical analyses used for comparisons concerning the relative auditory weight (detailed in the next paragraph) we also performed a non-parametric bootstrap test based on the one-sample t-statistic for the congruent minus incongruent difference (Efron and Tibshirani, 1993).

Effect of sound location on perceived flash location as a function of visibility

We investigated whether the influence of the sound on flash localization depended on the visibility of the flash. Critically, the flash signal intensity was fine-tuned in several pilot studies, so that approximately 50% of the flashes were judged invisible across participants at the group level. Hence, the flash visibility varied across trials and participants because of internal systems noise and participant-specific effects rather than external signal strength. We hypothesized that the influence of the true sound location would be inversely related to flash visibility. In other words, we expected that the influence of the sound on perceived flash location should be maximal for trials where the flash was judged invisible.

To quantify the influence of true sound location on participants' perceived flash location, we first coded the perceived and true flash and sound locations as −1 for left, 0 for center and 1 for right. Separately for visible, unsure and invisible trials, we then estimated a general linear model where participants' perceived flash location as the dependent variable was predicted by the true flash and sound location on each trial:

$$V_p = \beta_0 + (\beta_V * V_t) + (\beta_A * A_t) + \varepsilon \quad (1)$$

with $\boldsymbol{V_p}$ = perceived/reported flash location, $\boldsymbol{V_t}$ = true flash location, $\boldsymbol{A_t}$ = true sound location, β_0 = intercept term, β_V = coefficient for true flash location, β_A = coefficient for true sound location, ε = error term. As the audiovisual spatial discrepancies in this experiment were smaller than 10° visual angle, we assumed that auditory and visual signals are combined linearly as assumed under the standard forced fusion model (Alais and Burr, 2004). In other words, the influence of the true sound location (as quantified by the regression coefficient β_A) is assumed not to vary with the spatial discrepancy. Hence, we did not include an interaction term $A_t \times V_t$ in the regression model.

We computed the relative auditory weight as an index of the influence of sound on perceived flash location according to:

$$\textit{Relative Auditory Weight} = \frac{\beta_A}{\beta_A + \beta_V} \quad (2)$$

We tested whether the relative auditory weight was greater than zero using one-sample *t*-tests. A positive auditory weight indicates that the perceived visual location is shifted towards the true auditory location as expected for a reverse ventriloquist illusion. A negative auditory weight suggests that the perceived visual location is shifted away from the true auditory location (i.e., repulsion effect). An auditory weight that is not significantly different from zero suggests that the location of the sound does not significantly influence the perceived location of the flash. For comparison across visibility levels a one-way repeated measures ANOVA was performed with factor visibility. Planned pairwise comparisons were performed using paired *t*-tests. Moreover, to refrain making any parametric assumptions (n.b. the relative auditory weight conforms to a ratio distribution) we repeated these comparisons using non-parametric bootstrap-based tests.

RESULTS

EFFECT OF SPATIAL CONGRUENCY ON VISIBILITY JUDGMENT

Figure 2A shows the percentage of trials judged visible, unsure and invisible. As expected we observed a significant increase in percentage judged visible, when the sound was presented in the same relative to the opposite hemifield (percentage judged visible: congruent − incongruent (mean ± SEM): 1.8 ± 0.51; Cohen's d: 0.73; paired-samples *t*-test, $t_{(22)} = 3.51$, $p = 0.002$, bootstrap-based $p < 0.001$) (see **Figure 2B** for individual differences). Conversely, we observed a significant decrease in percentage

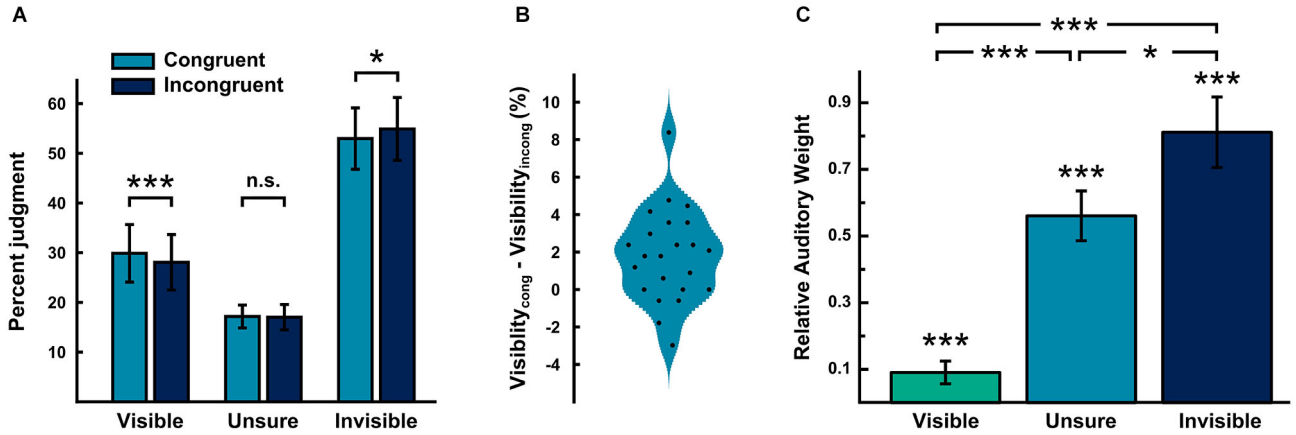


FIGURE 2 | Behavioral results. (A) Bar plots showing the percentage of flashes judged visible, unsure and invisible for audiovisual spatially congruent and incongruent conditions (across subjects mean ± SEM). Critically, the % judged visible was significantly higher for audiovisual spatially congruent relative to incongruent conditions. **(B)** Violin plot showing the distribution of the individual differences in percentage of flashes judged visible between the spatially congruent and incongruent conditions. The individual data points are overlaid. **(C)** Bar plots showing the relative auditory weights (across subjects mean ± SEM) obtained from the regression model separately for visible, unsure and invisible trials. As the regression model (specified in the methods) can only be estimated with at least three trials present for a particular visibility level, the number of subjects varies across the different visibility levels (visible: $n = 21$; unsure: $n = 22$; invisible: $n = 23$).

judged invisible for spatially congruent relative to incongruent trials (percentage judged invisible: congruent − incongruent (mean ± SEM): 1.94 ± 0.65; Cohen's d: −0.62; paired-samples t-test, $t_{(22)} = -2.98$, $p < 0.007$; bootstrap-based $p = 0.011$). This suggests that a sound influences whether visual signals reach perceptual awareness depending on audiovisual spatial congruency. As we did not include any trials where no flash was presented, we cannot compute the d-prime for the congruent and incongruent conditions or formally dissociate sensitivity and decisional bias. However, as the evaluation of audiovisual spatial congruency obviously entails spatial localization of both flash and sound, it is inconsistent to assume that audiovisual spatial congruency takes effect by influencing the decisional bias in the visibility judgment task. Moreover, had we included trials without a flash to estimate the false alarm rate, we would have still included the same false alarm rate for spatially congruent and incongruent conditions when computing the d-prime. In other words, the % judged visible directly corresponds to the d-primes for congruent and incongruent conditions.

EFFECT OF SOUND LOCATION ON PERCEIVED FLASH LOCATION AS A FUNCTION OF VISIBILITY

We quantified the influence of sound on perceived flash location across visibility levels in terms of the relative auditory weight obtained from the regression approach (see methods). As the regression model specified can only be estimated when at least three trials are present for a particular visibility level, the relative auditory weights are based on a different number of subjects across the different visibility levels (visible: $n = 21$; unsure: $n = 22$; invisible: $n = 23$). **Figure 2C** shows the relative auditory weights on the perceived location of a visible, unsure and invisible flash. We observed positive relative auditory weights for all three visibility levels. Critically, the relative auditory weights significantly differed across visibility levels (main effect of visibility: $F_{(1.6,29.8)} = 25.6$, MSE = 3.75, $p < 0.001$). More specifically, the relative auditory weight for visible trials was significantly different from that for unsure or invisible trials (paired-t test: unsure-visible $t_{(19)} = 6.54$, parametric $p < 0.001$, bootstrap-based $p < 0.001$; invisible-visible $t_{(20)} = 6.44$, parametric $p < 0.001$, bootstrap-based $p < 0.001$; n.b. the degrees of freedom vary as different numbers of subjects could be included, see above). As expected the auditory influence on perceived flash location was greatest when the flash was judged invisible.

DISCUSSION

Combining spatial audiovisual stimulation and CFS we investigated whether and how signals from different sensory modalities can interact prior to perceptual awareness. CFS is thought to affect visual perception by attenuating neural activity already in primary visual cortices similar to reducing the contrast of the stimulus (Yuval-Greenberg and Heeger, 2013). It is likely that this attenuation of neural activity destabilizes neural representations and prevents them from propagating up the cortical hierarchy thereby obliterating them from perceptual awareness. To measure the effect of a concurrent sound on participants' visual awareness, we tuned the strength of the visual flash such that it entered participants' awareness only on a fraction of trials. We then investigated whether the effect of a synchronous sound on participants' visibility judgment depended on audiovisual spatial congruency. Indeed, our results demonstrate that participants were more likely to detect the flash, when the sound was co-localized than non-collocated with the flash. In support of an "automatic" account of audiovisual integration these results suggest that an aware auditory signal can boost a weak visual signal into participants' awareness. Critically, the sound was brief and synchronous with the flash

across all conditions. Hence, the effects of spatial congruency are unlikely to be explained by a reduction in temporal uncertainty or more precise temporal expectations. Instead they suggest that audiovisual interactions prior to perceptual awareness are governed not only by temporal (as shown by Alsius and Munhall, 2013) but also by spatial constraints. There are at least two mechanisms by which a collocated sound may enhance flash visibility. First, a collocated sound may influence visual perception via bottom-up mechanisms that boost visual salience and enable the formation of spatially more precise salience maps. Second, a collocated sound may reduce visual spatial uncertainty via top-down mechanisms that enable more effective allocation of attentional resources and stabilize visual representations potentially even after they have accessed awareness. In the current paradigm, top-down mechanisms may be less likely because audiovisual signals were presented in synchrony and participants could respond immediately after the flash. Yet, future electrophysiological studies are needed to determine the role of bottom-up from top-down mechanisms in audiovisual interactions during flash suppression.

In sum, our results suggest that audiovisual interactions emerge largely prior to awareness governed by the classical principles of spatial congruency (Stein and Meredith, 1993; Wallace et al., 2004). These interactions in turn enhance stimulus salience and thereby enable a visual signal to elude flash suppression and enter participants' awareness. A controversial question is whether spatial congruency acts as a fundamental principle of multisensory integration or depends on stimulus characteristics and task-constraints (for excellent review see Spence, 2013). Accumulating evidence from behavioral research suggests that spatial congruency benefits performance predominantly in tasks where spatial information is relevant (e.g., overt or covert spatial orienting—Harrington and Peck, 1998; Arndt and Colonius, 2003; Diederich et al., 2003; Santangelo and Spence, 2008; Spence, 2010), but less so in detection (e.g., redundant target paradigms or identification tasks—Forster et al., 2002; Bertini et al., 2008; Girard et al., 2011). The current study cannot fully exclude that the role of spatial congruency emerges because subjects were engaged in both visibility judgment and spatial localization. Yet, as in previous masking studies (e.g., Frassinetti et al., 2002; Bolognini et al., 2005) an increase in detection performance was also observed in the absence of an additional localization task, spatial task demands do not seem absolutely critical. Instead, we would suggest that concurrent sounds automatically interact with visual signals as a function of spatial discrepancy in low level visual areas thereby amplifying the neural activity and boosting the flash into participants' awareness. Future studies are needed to further characterize the critical spatial integration window by systematically manipulating the spatial discrepancy of the audiovisual signals under flash suppression. Together with additional EEG and fMRI studies this research line would allow us to further pinpoint the cortical level at which sounds interact with visual processing under flash suppression.

In addition to judging the flash's visibility participants also located the flash on each trial. As the spatial discrepancy was approximately 8 degrees visual angle, we would expect that a concurrent, yet spatially discrepant sound biases the perceived visual location (Alais and Burr, 2004). The critical question of this study was whether participants' perceived flash location was influenced by the sound as a function of flash visibility. As expected we observed that the influence of sound location on perceived flash location increased gradually from visible to unsure and invisible trials. This audiovisual spatial bias profile is consistent with the principle of reliability-weighted integration where a stronger weight should be given to the more reliable signal. Indeed, numerous psychophysics and recent neurophysiological studies (Ernst and Banks, 2002; Alais and Burr, 2004; Morgan et al., 2008; Fetsch et al., 2011, 2013) have demonstrated that humans and non-human primates integrate signals weighted by their reliability approximately in accordance with predictions from Maximum Likelihood Estimation. In contrast to these previous studies we did not manipulate the reliability of the external signals. Instead, the flashes were physically identical across all visibility levels. Yet, identical physical signals will elicit neural representations that vary in their reliability across trials because of trial-specific internal systems noise (Faisal et al., 2008). Thus, as the brain does not have access to the true physical reliability of the sensory signals but only to the uncertainty of the internal representations, it is likely that the sensory weights in the integration process depend on both the noise in the environment and the trial-specific noise in the neural system. Thus, our findings suggest that the relative auditory weight in the integration process depends on the reliability of the trial-specific internal representation evoked by the visual signal. For example, if the visual signal is too weak to elude flash suppression and propagate to higher order association areas, "multisensory" representations for instance in parietal areas or response selection processes in frontal areas may be more strongly dominated by auditory inputs (Gottlieb et al., 1998; Macaluso et al., 2003; Macaluso and Driver, 2005; Bisley and Goldberg, 2010). As sensory noise also determines flash visibility, one may also argue that visible flashes bias participants' perceived sound location via higher order cognitive biasing mechanisms. In other words, if a flash elicits a noisy representation that does not enter participants' awareness, participants locate the sound purely based on the auditory input. By contrast, if a flash elicits a strong sensory representation that enters awareness, participants' perceptual decision is biased by the concurrent visual input. Future neurophysiological and neuroimaging studies are required to determine the neural mechanisms underlying this reliability weighting that emerges from internal noise rather than manipulation of external signal strength.

ACKNOWLEDGMENTS

This study was funded by the Max Planck Society and the European Research Council (ERC-multsens). We thank Mario Kleiner for help with stimulus generation and building the experimental set up, Beatrix Barth and Natalie Christner for help with data collection and Joana Leitão for support and helpful discussions.

REFERENCES

Adam, R., and Noppeney, U. (2014). A phonologically congruent sound boosts a visual target into perceptual awareness. *Front. Integr. Neurosci.* 8:70. doi: 10.3389/fnint.2014.00070

Alais, D., and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Curr. Biol.* 14, 257–262. doi: 10.1016/j.cub.2004.01.029

Alais, D., van Boxtel, J. J., Parker, A., and van Ee, R. (2010). Attending to auditory signals slows visual alternations in binocular rivalry. *Vision Res.* 50, 929–935. doi: 10.1016/j.visres.2010.03.010

Alsius, A., and Munhall, K. G. (2013). Detection of audiovisual speech correspondences without visual awareness. *Psychol. Sci.* 24, 423–431. doi: 10.1177/0956797612457378

Alsius, A., Navarra, J., Campbell, R., and Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Curr. Biol.* 15, 839–843. doi: 10.1016/j.cub.2005.03.046

Arndt, P. A., and Colonius, H. (2003). Two stages in crossmodal saccadic integration: evidence from a visual-auditory focused attention task. *Exp. Brain Res.* 150, 417–426. doi: 10.1007/s00221-003-1424-6

Bertelson, P., Vroomen, J., de Gelder, B., and Driver, J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. *Percept. Psychophys.* 62, 321–332. doi: 10.3758/bf03205552

Bertini, C., Leo, F., and Làdavas, E. (2008). Temporo-nasal asymmetry in multisensory integration mediated by the superior colliculus. *Brain Res.* 1242, 37–44. doi: 10.1016/j.brainres.2008.03.087

Bisley, J. W., and Goldberg, M. E. (2010). Attention, intention and priority in the parietal lobe. *Annu. Rev. Neurosci.* 33, 1–21. doi: 10.1146/annurev-neuro-060909-152823

Bolognini, N., Frassinetti, F., Serino, A., and Làdavas, E. (2005). "Acoustical vision" of below threshold stimuli: interaction among spatially converging audiovisual inputs. *Exp. Brain Res.* 160, 273–282. doi: 10.1007/s00221-004-2005-z

Bonath, B., Noesselt, T., Martinez, A., Mishra, J., Schwiecker, K., Heinze, H.-J., et al. (2007). Neural basis of the ventriloquist illusion. *Curr. Biol.* 17, 1697–1703. doi: 10.1016/j.cub.2007.08.050

Brainard, D. H. (1997). The psychophysics toolbox. *Spat. Vis.* 10, 433–436. doi: 10.1163/156856897x00357

Bresciani, J.-P., Dammeier, F., and Ernst, M. O. (2006). Vision and touch are automatically integrated for the perception of sequences of events. *J. Vis.* 6, 554–564. doi: 10.1167/6.5.2

Chen, Y.-C., and Spence, C. (2011). The crossmodal facilitation of visual object representations by sound: evidence from the backward masking paradigm. *J. Exp. Psychol. Hum. Percept. Perform.* 37, 1784–1802. doi: 10.1037/a0025638

Conrad, V., Bartels, A., Kleiner, M., and Noppeney, U. (2010). Audiovisual interactions in binocular rivalry. *J. Vis.* 10:27. doi: 10.1167/10.10.27

Conrad, V., Kleiner, M., Bartels, A., Hartcher O'Brien, J., Bülthoff, H. H., and Noppeney, U. (2013). Naturalistic stimulus structure determines the integration of audiovisual looming signals in binocular rivalry. *PLoS One* 8:e70710. doi: 10.1371/journal.pone.0070710

Conrad, V., Vitello, M. P., and Noppeney, U. (2012). Interactions between apparent motion rivalry in vision and touch. *Psychol. Sci.* 23, 940–948. doi: 10.1163/187847612x646497

Deroy, O., Chen, Y., and Spence, C. (2014). Multisensory constraints on awareness. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 369:20130207. doi: 10.1098/rstb.2013.0207

Diederich, A., Colonius, H., Bockhorst, D., and Tabeling, S. (2003). Visual-tactile spatial interaction in saccade generation. *Exp. Brain Res.* 148, 328–337. doi: 10.1007/s00221-002-1302-7

Driver, J., and Noesselt, T. (2008). Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses and judgments. *Neuron* 57, 11–23. doi: 10.1016/j.neuron.2007.12.013

Efron, B., and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Boston, MA: Springer US.

Ernst, M. O., and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415, 429–433. doi: 10.1038/415429a

Fairhall, S. L., and Macaluso, E. (2009). Spatial attention can modulate audiovisual integration at multiple cortical and subcortical sites. *Eur. J. Neurosci.* 29, 1247–1257. doi: 10.1111/j.1460-9568.2009.06688.x

Faisal, A. A., Selen, L. P. J., and Wolpert, D. M. (2008). Noise in the nervous system. *Nat. Rev. Neurosci.* 9, 292–303. doi: 10.1038/nrn2258

Fetsch, C. R., DeAngelis, G. C., and Angelaki, D. E. (2013). Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons. *Nat. Rev. Neurosci.* 14, 429–442. doi: 10.1038/nrn3503

Fetsch, C. R., Pouget, A., DeAngelis, G. C., and Angelaki, D. E. (2011). Neural correlates of reliability-based cue weighting during multisensory integration. *Nat. Neurosci.* 15, 146–154. doi: 10.1038/nn.2983

Fogelson, S. V., Kohler, P. J., Miller, K. J., Granger, R., and Tse, P. U. (2014). Unconscious neural processing differs with method used to render stimuli invisible. *Front. Psychol.* 5:601. doi: 10.3389/fpsyg.2014.00601

Forster, B., Cavina-Pratesi, C., Aglioti, S. M., and Berlucchi, G. (2002). Redundant target effect and intersensory facilitation from visual-tactile interactions in simple reaction time. *Exp. Brain Res.* 143, 480–487. doi: 10.1007/s00221-002-1017-9

Frassinetti, F., Bolognini, N., and Làdavas, E. (2002). Enhancement of visual perception by crossmodal visuo-auditory interaction. *Exp. Brain Res.* 147, 332–343. doi: 10.1007/s00221-002-1262-y

Girard, S., Collignon, O., and Lepore, F. (2011). Multisensory gain within and across hemispaces in simple and choice reaction time paradigms. *Exp. Brain Res.* 214, 1–8. doi: 10.1007/s00221-010-2515-9

Gottlieb, J. P., Kusunoki, M., and Goldberg, M. E. (1998). The representation of visual salience in monkey parietal cortex. *Nature* 391, 481–484. doi: 10.1038/35135

Guzman-Martinez, E., Ortega, L., Grabowecky, M., Mossbridge, J., and Suzuki, S. (2012). Interactive coding of visual spatial frequency and auditory amplitude-modulation rate. *Curr. Biol.* 22, 383–388. doi: 10.1016/j.cub.2012.01.004

Harrington, L. K., and Peck, C. K. (1998). Spatial disparity affects visual-auditory interactions in human sensorimotor processing. *Exp. Brain Res.* 122, 247–252. doi: 10.1007/s002210050512

Hupé, J., Joffo, L., and Pressnitzer, D. (2008). Bistability for audiovisual stimuli: perceptual decision is modality specific. *J. Vis.* 8, 1.1–1.15. doi: 10.1167/8.7.1

Kayser, C., Petkov, C. I., Augath, M., and Logothetis, N. K. (2005). Integration of touch and sound in auditory cortex. *Neuron* 48, 373–384. doi: 10.1016/j.neuron.2005.09.018

Kleiner, M., Brainard, D., and Pelli, D. (2007). "What's new in Psychtoolbox-3?," in *Perception* (Alezzo), 36 EVCP Abstract Supplement.

Klink, P. C., van Wezel, R. J. A., and van Ee, R. (2012). United we sense, divided we fail: context-driven perception of ambiguous visual stimuli. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 932–941. doi: 10.1098/rstb.2011.0358

Koch, C., and Tsuchiya, N. (2007). Attention and consciousness: two distinct brain processes. *Trends Cogn. Sci.* 11, 16–22. doi: 10.1016/j.tics.2006.10.012

Koch, C., and Tsuchiya, N. (2012). Attention and consciousness: related yet different. *Trends Cogn. Sci.* 16, 103–105. doi: 10.1016/j.tics.2011.11.012

Lee, H., and Noppeney, U. (2011). Physical and perceptual factors shape the neural mechanisms that integrate audiovisual signals in speech comprehension. *J. Neurosci.* 31, 11338–11350. doi: 10.1523/JNEUROSCI.6510-10.2011

Lee, H., and Noppeney, U. (2014). Temporal prediction errors in visual and auditory cortices. *Curr. Biol.* 24, R309–R310. doi: 10.1016/j.cub.2014.02.007

Lewis, R., and Noppeney, U. (2010). Audiovisual synchrony improves motion discrimination via enhanced connectivity between early visual and auditory areas. *J. Neurosci.* 30, 12329–12339. doi: 10.1523/JNEUROSCI.5745-09.2010

Lunghi, C., and Alais, D. (2013). Touch interacts with vision during binocular rivalry with a tight orientation tuning. *PLoS One* 8:e58754. doi: 10.1371/journal.pone.0058754

Lunghi, C., Binda, P., and Morrone, M. C. (2010). Touch disambiguates rivalrous perception at early stages of visual analysis. *Curr. Biol.* 20, R143–R144. doi: 10.1016/j.cub.2009.12.015

Lunghi, C., and Morrone, M. C. (2013). Early interaction between vision and touch during binocular rivalry. *Multisens. Res.* 26, 291–306. doi: 10.1163/22134808-00002411

Lunghi, C., Morrone, M. C., and Alais, D. (2014). Auditory and tactile signals combine to influence vision during binocular rivalry. *J. Neurosci.* 34, 784–792. doi: 10.1523/JNEUROSCI.2732-13.2014

Macaluso, E., and Driver, J. (2005). Multisensory spatial interactions: a window onto functional integration in the human brain. *Trends Neurosci.* 28, 264–271. doi: 10.1016/j.tins.2005.03.008

Macaluso, E., Driver, J., and Frith, C. D. (2003). Multimodal spatial representations engaged in human parietal cortex during both saccadic and manual spatial orienting. *Curr. Biol.* 13, 990–999. doi: 10.1016/s0960-9822(03)00377-4

Maruya, K., Watanabe, H., and Watanabe, M. (2008). Adaptation to invisible motion results in low-level but not high-level aftereffects. *J. Vis.* 8, 7.1–7.11. doi: 10.1167/8.11.7

Morgan, M. L., Deangelis, G. C., and Angelaki, D. E. (2008). Multisensory integration in macaque visual cortex depends on cue reliability. *Neuron* 59, 662–673. doi: 10.1016/j.neuron.2008.06.024

Munhall, K. G., ten Hove, M. W., Brammer, M., and Paré, M. (2009). Audiovisual integration of speech in a bistable illusion. *Curr. Biol.* 19, 735–739. doi: 10.1016/j.cub.2009.03.019

Olivers, C. N. L., and Van der Burg, E. (2008). Bleeping you out of the blink: sound saves vision from oblivion. *Brain Res.* 1242, 191–199. doi: 10.1016/j.brainres.2008.01.070

Onat, S., Libertus, K., and König, P. (2007). Integrating audiovisual information for the control of overt attention. *J. Vis.* 7, 11.1–11.16. doi: 10.1167/7.10.11

Palmer, T. D., and Ramsey, A. K. (2012). The function of consciousness in multisensory integration. *Cognition* 125, 353–364. doi: 10.1016/j.cognition.2012.08.003

Peremen, Z., and Lamy, D. (2014). Comparing unconscious processing during continuous flash suppression and meta-contrast masking just under the limen of consciousness. *Front. Psychol.* 5:969. doi: 10.3389/fpsyg.2014.00969

Santangelo, V., and Spence, C. (2008). Is the exogenous orienting of spatial attention truly automatic? Evidence from unimodal and multisensory studies. *Conscious. Cogn.* 17, 989–1015. doi: 10.1016/j.concog.2008.02.006

Soto-Faraco, S., and Spence, C. (2002). Modality-specific auditory and visual temporal processing deficits. *Q. J. Exp. Psychol. A* 55, 23–40. doi: 10.1080/02724980143000136

Spence, C. (2010). Crossmodal spatial attention. *Ann. N Y Acad. Sci.* 1191, 182–200. doi: 10.1111/j.1749-6632.2010.05440.x

Spence, C. (2013). Just how important is spatial coincidence to multisensory integration? Evaluating the spatial rule. *Ann. N Y Acad. Sci.* 1296, 31–49. doi: 10.1111/nyas.12121

Stanford, T. R., Quessy, S., and Stein, B. E. (2005). Evaluating the operations underlying multisensory integration in the cat superior colliculus. *J. Neurosci.* 25, 6499–6508. doi: 10.1523/jneurosci.5095-04.2005

Stein, B. E., and Meredith, M. A. (1993). *The Merging of the Senses.* (Cambridge, MA: The MIT Press), 211.

Stekelenburg, J. J., Vroomen, J., and de Gelder, B. (2004). Illusory sound shifts induced by the ventriloquist illusion evoke the mismatch negativity. *Neurosci. Lett.* 357, 163–166. doi: 10.1016/j.neulet.2003.12.085

Talsma, D., Doty, T. J., and Woldorff, M. G. (2007). Selective attention and audiovisual integration: is attending to both modalities a prerequisite for early integration? *Cereb. Cortex* 17, 679–690. doi: 10.1093/cercor/bhk016

Talsma, D., Senkowski, D., Soto-Faraco, S., and Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends Cogn. Sci.* 14, 400–410. doi: 10.1016/j.tics.2010.06.008

Tong, F. (2003). Primary visual cortex and visual awareness. *Nat. Rev. Neurosci.* 4, 219–229. doi: 10.1038/nrn1055

Tsuchiya, N., and Koch, C. (2005). Continuous flash suppression reduces negative afterimages. *Nat. Neurosci.* 8, 1096–1101. doi: 10.1038/nn1500

van Ee, R., van Boxtel, J. J. A., Parker, A. L., and Alais, D. (2009). Multisensory congruency as a mechanism for attentional control over perceptual selection. *J. Neurosci.* 29, 11641–11649. doi: 10.1523/JNEUROSCI.0873-09.2009

Vroomen, J., Bertelson, P., and de Gelder, B. (2001). The ventriloquist effect does not depend on the direction of automatic visual attention. *Percept. Psychophys.* 63, 651–659. doi: 10.3758/bf03194427

Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., and Schirillo, J. A. (2004). Unifying multisensory signals across time and space. *Exp. Brain Res.* 158, 252–258. doi: 10.1007/s00221-004-1899-9

Watanabe, M., Cheng, K., Murayama, Y., Ueno, K., Asamizuya, T., Tanaka, K., et al. (2011). Attention but not awareness modulates the BOLD signal in the human V1 during binocular suppression. *Science* 334, 829–831. doi: 10.1126/science.1203161

Werner, S., and Noppeney, U. (2010). Distinct functional contributions of primary sensory and association areas to audiovisual integration in object categorization. *J. Neurosci.* 30, 2662–2675. doi: 10.1523/JNEUROSCI.5091-09.2010

Wyart, V., and Tallon-Baudry, C. (2008). Neural dissociation between visual awareness and spatial attention. *J. Neurosci.* 28, 2667–2679. doi: 10.1523/JNEUROSCI.4748-07.2008

Yuval-Greenberg, S., and Heeger, D. J. (2013). Continuous flash suppression modulates cortical activity in early visual cortex. *J. Neurosci.* 33, 9635–9643. doi: 10.1523/JNEUROSCI.4612-12.2013

Zhou, W., Jiang, Y., He, S., and Chen, D. (2010). Olfaction modulates visual perception in binocular rivalry. *Curr. Biol.* 20, 1356–1358. doi: 10.1016/j.cub.2010.05.059

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 September 2014; accepted: 05 February 2015; published online: 27 February 2015.

*Citation: Aller M, Giani A, Conrad V, Watanabe M and Noppeney U (2015) A spatially collocated sound thrusts a flash into awareness. Front. Integr. Neurosci. **9**:16. doi: 10.3389/fnint.2015.00016*

This article was submitted to the journal Frontiers in Integrative Neuroscience.

frontiers in
INTEGRATIVE NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 06 January 2015
doi: 10.3389/fnint.2014.00096

Independent effects of bottom-up temporal expectancy and top-down spatial attention. An audiovisual study using rhythmic cueing

Alexander Jones*

Department of Psychology, Middlesex University, London, UK

Edited by:
Jessica Hartcher-O'Brien, l'Ecole Normale Superieuer, France

Reviewed by:
Jean Vroomen, University of Tilburg, Netherlands
Gustavo Rohenkohl, Ernst Strüngmann Institute (ESI) for Neuroscience in Cooperation with Max Planck Society, Germany

****Correspondence:***
Alexander Jones, Department of Psychology, Middlesex University, The Burroughs, London NW4 4BT, UK
e-mail: a.j.jones@mdx.ac.uk

Selective attention to a spatial location has shown enhanced perception and facilitate behavior for events at attended locations. However, selection relies not only on where but also when an event occurs. Recently, interest has turned to how intrinsic neural oscillations in the brain entrain to rhythms in our environment, and, stimuli appearing in or out of sync with a rhythm have shown to modulate perception and performance. Temporal expectations created by rhythms and spatial attention are two processes which have independently shown to affect stimulus processing but it remains largely unknown how, and if, they interact. In four separate tasks, this study investigated the effects of voluntary spatial attention and bottom-up temporal expectations created by rhythms in both unimodal and crossmodal conditions. In each task the participant used an informative cue, either color or pitch, to direct their covert spatial attention to the left or right, and respond as quickly as possible to a target. The lateralized target (visual or auditory) was then presented at the attended or unattended side. Importantly, although not task relevant, the cue was a rhythm of either flashes or beeps. The target was presented in or out of sync (early or late) with the rhythmic cue. Results showed participants were faster responding to spatially attended compared to unattended targets in all tasks. Moreover, there was an effect of rhythmic cueing upon response times in both unimodal and crossmodal conditions. Responses were faster to targets presented in sync with the rhythm compared to when they appeared too early in both crossmodal tasks. That is, rhythmic stimuli in one modality influenced the temporal expectancy in the other modality, suggesting temporal expectancies created by rhythms are crossmodal. Interestingly, there was no interaction between top-down spatial attention and rhythmic cueing in any task suggesting these two processes largely influenced behavior independently.

Keywords: entrainment, crossmodal, endogenous, exogenous, attention, expectancy, hazard function

INTRODUCTION

Our sensory system is constantly exposed to vast amounts of information. To efficiently deal with this information and guide behavior we need to select, prioritize and predict certain events and stimuli over others. The collective term for this selective mechanism is known as attention. There are many forms of attention and one of the most extensively researched is how we focus our attention towards different locations in space. Spatial attention research is typically divided into endogenous attention, which is under voluntary control and exogenous attention which is bottom-up and stimulus driven. The most common method to explore the behavioral effects of endogenous and exogenous attention has been using the Posner cueing task (Posner, 1980). The participant's task is to respond as quickly as possible to a target, usually presented peripherally to the left or right. In an endogenous version, the targets are preceded by a cue, usually centrally located, informing the most likely location of the target (70–80% likelihood). In an exogenous version, the cue, usually peripheral, does not give any indication of where the target may appear, however, the cue nevertheless typically elicits effects on target processing (Santangelo and Spence, 2008). Endogenous spatial attention has been studied extensively within and across modalities. Selective attention to a spatial location has shown to enhance perceptual processing (e.g., Mangun and Hillyard, 1990; Yeshurun and Carrasco, 1998) as well as facilitate response times (e.g., Posner et al., 1980) to visual stimuli at attended as compared to unattended locations (for a recent review see Carrasco, 2014).

Predictions about events in our environment rely not only on *where* something happens but also *when* an event occurs. Similar to spatial attention, focusing attention to a specific moment in time influences perception and biases our actions (Nobre, 2010). Temporal expectation can be generated in different ways and similar to voluntary spatial attention, instructive cues have been used to manipulate temporal expectations (e.g., Coull and Nobre, 1998; Naccache et al., 2002; Davranche et al., 2011; Zanto et al.,

2011; Rohenkohl et al., 2014; see Nobre and Rohenkohl, 2014 for a recent review). Coull and Nobre (1998) used a symbolic cue to indicate when an upcoming target would likely appear, either 300 ms or 1500 ms after cue onset. In this detection task they found behavioral benefits when the cue appeared at a temporally anticipated compared to an unexpected time interval. Zanto et al. (2011) extended these findings showing similar benefits of voluntary orienting to targets at a particular point in time using event-related potentials (ERPs) as well as behavioral discrimination and Go-NoGo tasks. Moreover, Correa et al. (2005) showed high temporal expectancies increased perceptual sensitivity (d') for detecting visual targets. Temporal cueing studies in the auditory modality have also shown effects of perceptual modulation by temporal cues. Several studies have observed a modulation of the early N1 component, suggested to originate from the primary auditory cortex, in response to temporally expected compared to unexpected tones (Lange and Röder, 2006; Lange et al., 2006; Lampar and Lange, 2011). There is thus mounting evidence showing that voluntary directing attention to a specific point in time influences both perception and modulates behavior.

Temporal expectancies can also be created by rhythms, something which commonly appears in our environment. For example the rhythm of breathing or our heartbeat, or the swaying of a tree, the sound and movement of walking or waves on a beach, the rhythmic structure of speech, or of course the rhythm in music. Rhythmic cueing has been used to investigate how external rhythms influences perception and performance. For example, Jones et al. (2002) presented participants with a standard tone which was followed by a sequence of tones presented in a rhythm. The participant's task was to judge whether a target tone had the same pitch as the standard tone. They found that performance accuracy was better when the target tone was presented *in* compared to *out of* sync with the preceding rhythm. Auditory perceptual discrimination has consistently shown to be better when stimuli coincide with the rhythm and perceptual performance deteriorates if the stimuli is presented too early or too late in relation to the rhythm (Jones et al., 2006; see Jones, 2010 for a review). Similarly, response times have also been reported to be improved for stimuli occurring on the beat of a particular rhythm compared to an asynchronous rhythm using both auditory (e.g., Sanabria et al., 2011) or visual stimuli (Doherty et al., 2005; Rohenkohl et al., 2012; Cravo et al., 2013).

More recently the concept of rhythmic cueing has seen an increased research interest from a more neuroscientific viewpoint in that intrinsic brain operations are profoundly rhythmic (Raichle, 2010). Groups of neurons in the brain fluctuate rhythmically together and create oscillations with different frequencies which can be measures using electroencephalogram (EEG). These self-generated brain oscillations have shown to modulate responses and influence motoric, perceptual and cognitive processes (Buzsaki, 2006; Thut and Miniussi, 2009). It has for example been shown that the threshold of detecting visual stimuli fluctuates over time along with the phase of ongoing EEG activity (Busch et al., 2009). Importantly, the neural oscillation can also entrain to external rhythms aligning the firing pattern according to rhythms in our environment (Arnal and Giraud, 2012). In other words, neurons start to fire in synchrony with external rhythms. Moreover, entrainment to particular rhythms has been suggested to underlie selective attention (Lakatos et al., 2013; Calderone et al., 2014). For example Lakatos et al. (2008) presented monkeys with auditory and visual interleaved rhythms and found selectively attention to one stream amplified neural responses to events in that stream. Moreover, entrainment has been shown to increase with participant effort (Lakatos et al., 2013). This further indicates entrainment can also be modulated by higher level processes, such as attention. Similar to spatial attention, temporal expectancies can be bottom-up or top-down. However, it remains to be fully established to what extent rhythmic cueing and entrainment occurs unintentionally, in a purely bottom-up fashion. Recent evidence suggests temporal expectancies do still occur as a result of a rhythm even though the rhythm is detrimental to the task, suggesting automatic effects of rhythmic stimuli in the absence of top-down processes (Breska and Deouell, 2014).

Predicting where or when an event will occur has independently been shown to influence perception and drive behavior. However, space and time are not dimensions which occur in isolation in our environment, yet only a handful of studies have explored these two types expectation together. Doherty et al. (2005) manipulated both temporal and spatial expectancies by presenting participants with a ball which moved from left to right across a screen. Towards the right side of the screen there was a section which occluded the ball before reappearing. Doherty et al. found that response times were faster when the ball reappeared behind the occluding band in sync with the preceding rhythm. Similarly, response times were faster when the ball reappeared in the spatial location which was predicted by the balls trajectory across the screen. The individual effects were also additive showing faster response times when both temporal and spatial expectancies matched, an additive effect also demonstrated on the visual P1 component. Recently Rohenkohl et al. (2014) also investigated the synergy between spatial and temporal expectancies. In their task a symbolic visual arrow simultaneously indicated the likely location of a target as well as the likely time point when to expect the target. Unlike Doherty et al. they found an interaction between spatial and temporal effects. Temporal expectations improved visual perception, but only at spatially attended and not unattended locations. Importantly, in both studies (see also Tang et al., 2013) participants were asked to use both types of expectancies to increase performance. That is, both temporal and spatial expectancies were generated top-down, or, in Doherty et al. (2005) study using a rhythmic cue, a likely mix of stimulus and voluntary temporal attention. What remains less clear is how stimulus driven temporal expectancies, created by external rhythms, are affected by top-down spatial attention.

Crossmodal spatial attention effects have been extensively reported and shown to enhance perceptual processing and facilitate behavior (Vroomen and de Gelder, 2000; Spence and Driver, 2004). However, less is known how entrainment operates across modalities. In a study by Lakatos et al. (2007) it was observed that somatosensory inputs can reset the phase of the neural oscillations in primary auditory cortex of macaque monkeys, and

in turn, auditory stimuli are enhanced or suppressed according to when in the oscillation they appear (see Kayser et al., 2008, for similar results in humans). This observation indicating that oscillations show crossmodal effects at a neural level. Moreover, recently Miller et al. (2013) also showed a crossmodal effect of entrainment whereby eye movements towards a visual target were faster if they occurred in sync with a preceding rhythm of tones.

The present study investigated how voluntary spatial attention affected automatic effects of rhythmic cueing using a simple detection task. Participants performed a typical Posner cueing task where an informative cue indicated to which side the target was most likely to appear. In addition, the cue consisted of four or five stimuli presented in a rhythm and the target was presented in or out of sync with this rhythm. Importantly this rhythm and the timing of the target was not task relevant. This novel paradigm allowed independent manipulation of top-down spatial attention and bottom-up temporal expectancies in order to investigate whether these represent dependent or independent mechanisms in driving behavior (as measured by response times). Furthermore, this study aimed to investigate whether rhythmic stimuli in one modality automatically influence the temporal expectancy in another modality. In separate tasks, participants were either presented with a visual cue and a visual target (VV), auditory cue and auditory target (AA) or in a crossmodal setting with a visual cue and an auditory targets (VA), or auditory cue and visual target (AV). Taken together, this study explored how endogenous spatial attention and stimulus-driven temporal expectancies, two processes which have independently shown to modulate perception and behavior, affected behavior in both a unimodal and crossmodal setting.

METHODS

PARTICIPANTS

The study consisted 16 participants in each task, 64 in total (16 males; 13 right-, 3 left-handed, and 48 females; 43 right-, 5 left-handed). The participants were naive to the study and participated voluntarily or in return for course credits. The participant number was based on similar behavioral studies (e.g., Lawrence and Klein, 2013). Each participant only took part in one of the following four tasks: AA, AV, VA, VV. Due to excessive responses to catch trials and/or an inability to perform the task two participants were removed and replaced in the AV task, one from the VA, and two from the VV. The study was approved by the Middlesex University ethics committee and all participants provided written informed consent.

STIMULI AND MATERIALS

The stimuli were presented and data collected using E-Prime v2 software (Psychology Software tools) run on a PC. Visual stimuli (fixation cross, visual cues and targets) were presented on a 17 inch monitor (1280 × 1024 pixels). A black fixation cross was presented in the middle of the screen. The visual cue consisted of an X above, below, to the left and right of the fixation cross creating the appearance of a larger cross in the center (see **Figure 1** for details). Visual targets (three black Xs) were presented to the left or right side of the monitor. The font was Courier new. The participant was seated with their eyes approximately in line with the fixation cross and approximately 400–500 mm away from the screen. The visual angle for the target typically ranged between 18.15 and 14.7°. Auditory stimuli were presented via headphones (Audio 355, Plantronics). Auditory stimuli were 100 ms in duration with a 5 ms rise and fall time. The cue consisted of either low tones (400 Hz) or high frequency tones (800 Hz) and always presented in stereo. Targets were presented to only one ear and were 600 Hz. A keyboard was used to collect response times. The *down arrow* key was positioned in a straight line behind the fixation cross.

DESIGN AND PROCEDURE

On each trial a rhythmic cue was presented. In the VA and VV task the color of the flashes (pink or blue) indicated whether the participant was to direct attention to the left or right. In the AV and AA tasks, the cue was either high or low tones, and this indicated which side attention was to be directed. A target then appeared at the attended (75%) or unattended side (25%) and the participant was to respond as quickly as possible by pressing the keyboard once a target appeared. In the AV and VV tasks, the target appeared to the left or right of the fixation cross and in the VA and AA tasks, the target was a tone presented to the left or right ear. The target appeared either in sync with the rhythm (the cue) or out of sync (early or late). The participant was not informed about this and it was not relevant to the task.

Each task consisted of five blocks with a total 260 trials (52 trials per block). Out of the 260 trials, 180 were attended (69%) and 60 unattended (23%), and 20 catch trials (8%). The weighting of targets, excluding catch trials was, 75/25 for attended/unattended targets. There was an even distribution of early, sync, and late trials. That is, for attended trials, there were 60 early, 60 sync, and 60 late trials per participant, and 20 unattended trials for each of the early, sync and late conditions. For half of the trials the cue consisted of four stimuli, and for half of the trials the rhythmic cue included five stimuli. Prior to the experiment the participant ran a practice block.

The participant was seated in an experimental booth in front of a PC monitor. In the tasks including auditory stimuli (all but the VV task) the participant wore headphones. Each trial started with the presentation of the rhythmic cue which consisted of four or five stimuli presented every 600 ms (see **Figure 1** for events in a trial). More specifically, the first of the rhythmic stimuli (the cue) was presented for 100 ms followed by an inter-stimulus interval (ISI) of 500 ms. The cue stimuli were presented four of five times creating a rhythm of 100 ms stimulus every 600 ms. After the last of the interpolated stimuli a target was presented. The critical ISI, preceding the target was 300 ms (early), 500 ms (sync), or 700 ms (late). The target was then presented for 100 ms. Participants responded with their dominant hand by pressing the *down arrow* key on the keyboard. If no response was recorded the trial terminated after 2000 ms. There was a random inter-trial interval of 2000–3000 ms. A centrally located fixation cross was presented throughout and participants were explicitly instructed to keep their gaze on this fixation cross at all times.

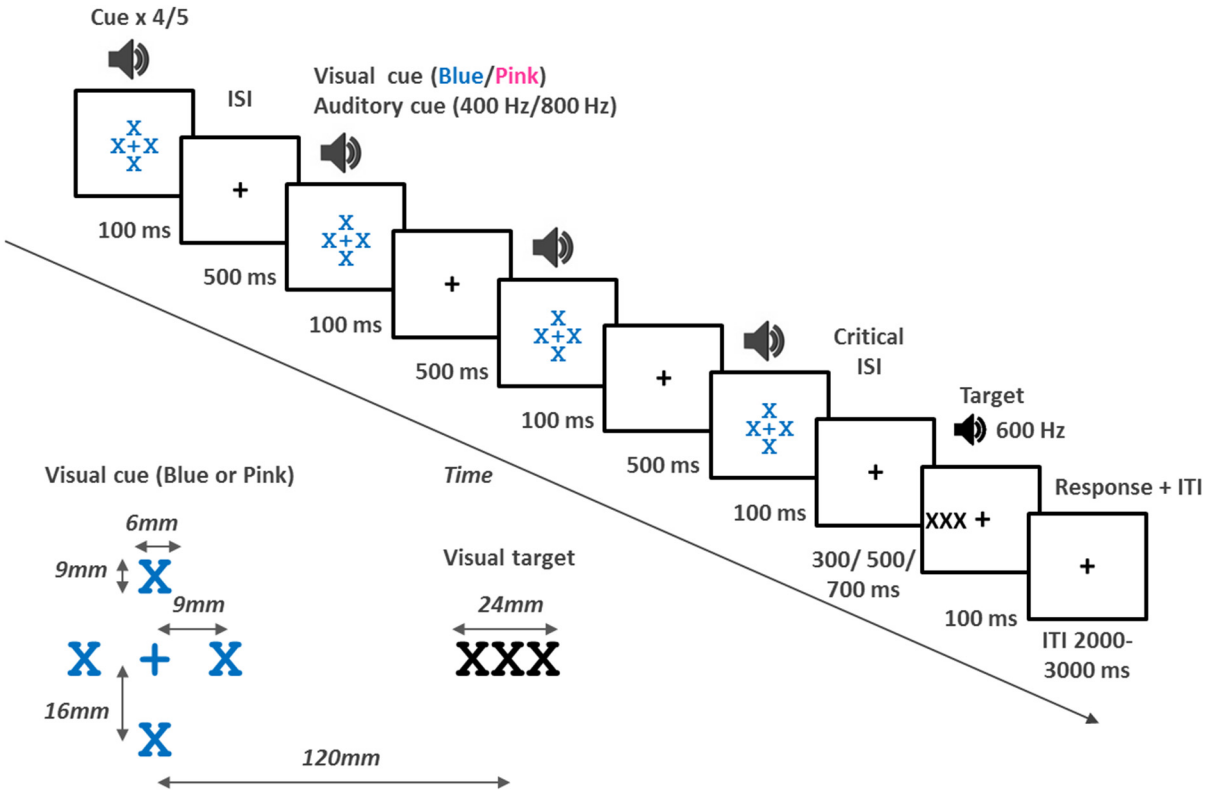


FIGURE 1 | *Top*: Schematic view of events in a trial. Each trial started with a cue which consisted of four or five interpolated stimuli, each 100 ms in duration, and presented with a 500 ms inter-stimulus interval (ISI) (SOA of 600 ms). In the VA and VV tasks the cue was either blue or pink Xs surrounding the fixation cross. The color of the cue indicated to which side to attend. In the AV and AA tasks the cue, which was either high or low frequency tones, indicated whether to attend to the left or right. The critical ISI was the interval between the last cue stimulus and the target. The target was 100 ms in duration. In the AV and VV tasks, the target was three X's which appeared to the left or right of the fixation cross. In the VA and AA tasks the cue was a 600 Hz tone presented to either the left or right ear. The participant responded by pressing a key on the keyboard. After a response an inter trial interval of between 2000–3000 ms followed. A fixation cross was presented throughout in all tasks. ***Bottom left*:** Schematic representation of the visual cue and visual target as it appeared on screen.

In the tasks with an auditory cue (AA and AV), for half the participants high frequency tones indicated to attend to the left and low tones indicated attend to the right. This allocation was reversed for the other half. Similarly, for half the participants who performed a task with a visual cue (VA and VV), a blue Xs indicated attend to the left and pink Xs attend to the right, and the reverse for the other half of participants.

The RT data was Log_{10}-transformed and submitted to a mix design ANOVA with the factors Task (AA, AV, VA, VV), Spatial attention (attended, unattended), Temporal expectancy (early, sync, late), and Rhythm count (four stimuli, five stimuli). Following the overall analysis, each task was analyzed separately.

RESULTS

SUMMARY

The results showed that participants responded faster to attended compared to unattended trials in all four tasks. There was also an effect of temporal expectancy in all tasks but the unimodal auditory task (AA). In the two cross-modal tasks (AV and VA) the targets were faster when in sync and late targets compared to early targets. In the visual task the late targets were faster than both in sync and early targets. Although clear effects of spatial attention and temporal expectancy were observed there was no evidence of an interaction between these two factors in any task. Thus suggesting temporal expectancy and spatial attention affected response times independently.

Overall analysis including task

A mixed design ANOVA, including Task as a factor, showed a main effect of Spatial attention ($F_{(1,60)} = 36.32$, $p < 0.001$, $\eta_p^2 = 0.38$) with faster RTs for attended (325.2 ms) compared to unattended trials (367.6 ms). There was a main effect of Temporal expectancy ($F_{(2,120)} = 27.27$, $p < 0.001$, $\eta_p^2 = 0.31$) and follow up pairwise-comparisons (Bonferroni corrected) showed that sync (342.8 ms) and late targets (339.7 ms) were significantly faster than early targets (356.6 ms) (both p's < 0.001). There was no difference between sync and late targets ($p = 0.18$). There was a main effect of Rhythm count ($F_{(1,60)} = 42.32$, $p < 0.001$, $\eta_p^2 = 0.41$) with targets preceded by four stimuli in the rhythm (352.8 ms) were slower compared to if the rhythmic cue contained five stimuli

(340.0 ms). There was a main effect of Task ($F_{(3,60)}$ = 6.24, p = 0.001, η_p^2 = 0.24) and Bonferroni *post hoc* test showed that the AV task was significantly faster (297.8 ms) compared to AA (380.5 ms) and VV task (369.1 ms) (p = 0.002 and p = 0. 005 respectively).

There was a Temporal expectancy*Task interaction ($F_{(6,120)}$ = 3.30, p = 0.005, η_p^2 = 0.14) and planned analysis for each task is presented below. There was no Task*Spatial attention interaction (p = 0.59, η_p^2 = 0.04). Important to note is there was no Spatial attention*Temporal expectancy interaction (p = 0.37, η_p^2 = 0.02). There was a Rhythm count*Temporal expectancy interaction ($F_{(2,120)}$ = 10.27, p < 0.001, η_p^2 = 0.15) suggesting the effect of Temporal expectancy was different according to the number of stimuli in the cue. This interaction will also be explored in the analysis of each task. No other main effects or interaction were significant.

VISUAL CUE—VISUAL TARGET (VV)

Overall participants missed less than 1% of targets and responded to 1.9% of catch trials in the VV task.

There was a significant effect of Spatial attention ($F_{(1,15)}$ = 5.74, p = 0.03, η_p^2 = 0.28) with attended trials being faster (342.6 ms) than unattended trials (395.5 ms). The purely visual task showed a main effect of Temporal expectancy ($F_{(2,30)}$ = 17.60, p < 0.001, η_p^2 = 0.54). Pairwise comparisons (Bonferroni corrected) demonstrated late targets (339.7 ms) to be faster (p = 0.002) than in sync targets (370.6 ms), and late targets were faster (p = 0.001) than early targets (382.4 ms), and in sync targets were also faster compared to early targets (p = 0.049; see **Figure 2**). There was a main effect of Rhythm count ($F_{(1,15)}$ = 24.03, p < 0.001, η_p^2 = 0.62) with on average faster RTs for visual targets preceded by five stimuli (356.9 ms) compared to four stimuli (381.2 ms). There was no Spatial attention*Temporal expectancy interaction (p = 0.18, η_p^2 = 0.11).

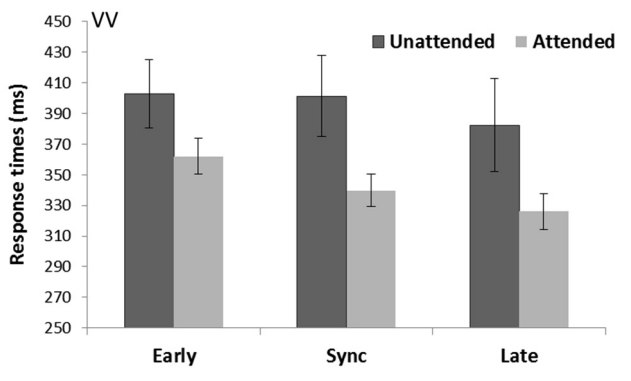


FIGURE 2 | Unimodal visual cue and visual targets task (VV). Mean response times (with standard error bars) for visual targets presented at spatially attended (light gray) and unattended side (dark gray), separately for early, sync and late conditions in relation to the visual rhythm (the cue). Response times were significantly faster for attended over unattended targets. There was a main effect of temporal expectancy where late targets were faster than both early and in sync targets, and in sync targets were also faster than early targets.

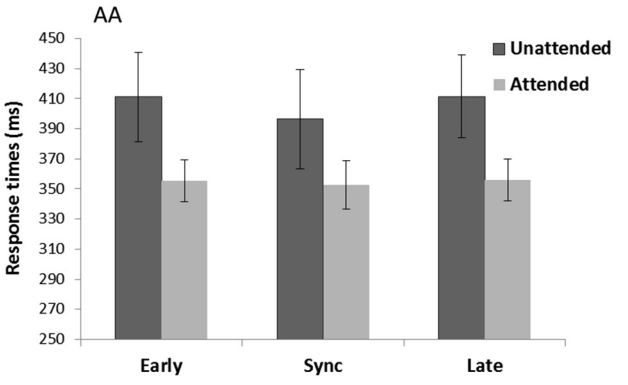


FIGURE 3 | Unimodal auditory cue and auditory targets task (AA). Mean response times (with standard error bars) for auditory targets presented at spatially attended (light gray) and unattended side (dark gray), separately for early, sync and late conditions in relation to the auditory rhythm (the cue). Response times were significantly faster for attended over unattended targets.

AUDITORY CUE—AUDITORY TARGET (AA)

Overall participants missed 1.4% of targets and responded to 4.4% of catch trials in the AA task.

There was a significant main effect of Spatial attention ($F_{(1,15)}$ = 8.99, p = 0.009, η_p^2 = 0.38) with faster RTs for attended (354.7 ms) compared to unattended trials (406.3 ms) (see **Figure 3**). There was also a main effect of Rhythm count ($F_{(1,15)}$ = 8.15, p = 0.012, η_p^2 = 0.35) with overall faster RTs for when the cue consisted of five (376.7 ms) compared to four tones (384.3 ms). There was no main effect of Temporal expectancy (p = 0.59, η_p^2 = 0.03) or Spatial attention*Temporal expectancy interaction (p = 0.36, η_p^2 = 0.07).

AUDITORY CUE—VISUAL TARGET (AV)

Participants missed 2% of targets and responded to 4.4% of catch trials in the AV task.

A main effect of Spatial attention ($F_{(1,15)}$ = 26.07, p < 0.001, η_p^2 = 0.64) showed attended trials were faster (276.6 s) compared to unattended trials (319.0 ms). There was also a main effect of Temporal expectancy ($F_{(2,30)}$ = 14.38, p < 0.001, η_p^2 = 0.49). Pairwise-comparisons (Bonferroni corrected) showed both sync (292.2 ms) and late targets (288.9 ms) were significantly faster than early targets (312.4 ms; p < 0.001 and p = 0.001 respectively; see **Figure 4**). There was no Spatial attention*Temporal expectancy interaction (p = 0.27, η_p^2 = 0.08).

There was a Rhythm count*Temporal expectancy interaction ($F_{(2,30)}$ = 10.74, p = 0.002, η_p^2 = 0.42). Follow-up analysis of trials with a rhythm of four tones preceding the visual target showed an effect of Temporal expectancy ($F_{(1,15)}$ = 20.56, p < 0.001, η_p^2 = 0.58). Pairwise-comparisons (Bonferroni corrected) showed a difference between early (327.9 ms) and in sync (296.4 ms) and early and late targets (288.8 ms) (both p's < 0.001). When there were five tones in the rhythm, no effect of target Temporal expectancy was present (p = 0.14).

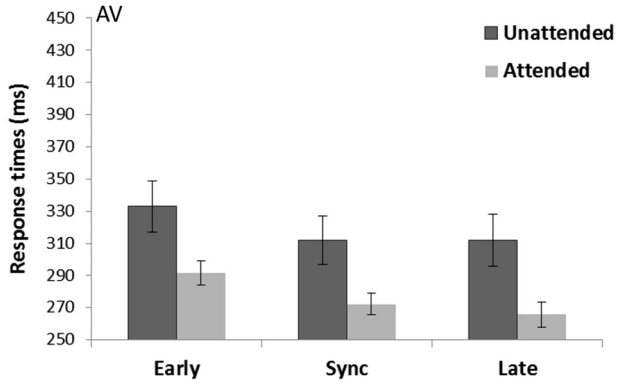


FIGURE 4 | Crossmodal auditory cue and visual targets (AV) task. Mean response times (with standard error bars) for targets presented at spatially attended (light gray) and unattended side (dark gray), separately for early, sync and late conditions in relation to the auditory rhythm (the cue). Response times were significantly faster for attended over unattended visual targets. There was a main effect of Temporal expectancy where in sync and late targets were faster than early targets.

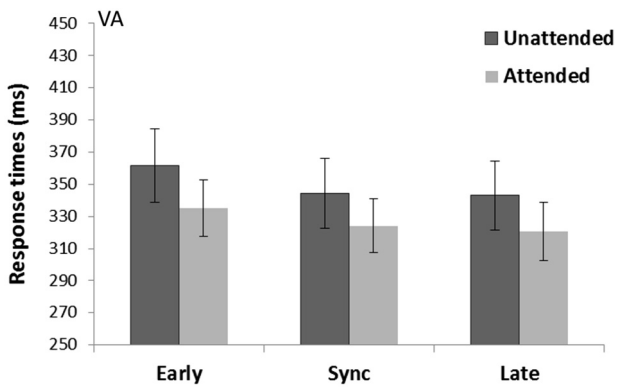


FIGURE 5 | Crossmodal visual cue and auditory targets task (VA). Mean response times (with standard error bars) for tones presented at spatially attended (light gray) and unattended side (dark gray), separately for early, sync and late conditions in relation to the visually presented rhythm (the cue). Response times were significantly faster for spatially attended over unattended targets and a main effect of Temporal expectancy showed in sync and late targets were faster than early targets.

VISUAL CUE—AUDITORY TARGET (VA)

Overall participants missed less than 1% of targets and responded to 1.6% of catch trials in the VA task.

A main effect of Spatial attention ($F_{(1,15)} = 6.83$, $p = 0.02$, $\eta_p^2 = 0.31$) revealed attended trials were faster (326.6 ms) compared to unattended trials (349.6 ms) (see **Figure 5**). There was also a main effect of Temporal expectancy ($F_{(2,30)} = 9.75$, $p = 0.001$, $\eta_p^2 = 0.39$) and pairwise-comparisons (Bonferroni corrected) showed in sync (370.6 ms) and late targets (354.3 ms) were faster than early target (382.4 ms) ($p = 0.02$ and $p < 0.001$ respectively). There was no Spatial attention*Temporal expectancy interaction ($p = 0.72$, $\eta_p^2 = 0.02$).

DISCUSSION

In four separate tasks, this study investigated the effects of voluntary spatial attention and bottom-up temporal expectancy in both unimodal and crossmodal conditions. In all tasks response times to targets were faster when they appeared at the attended compared to unattended location. This indicated that participants followed instructions and the results replicated what has previously been observed in unimodal visual (e.g., Posner, 1980; Wright and Ward, 1994) and auditory spatial attention tasks (Spence and Driver, 1994), as well as audiovisual crossmodal tasks (Spence and Driver, 1996; or Spence, 2010 for a review). The present study also demonstrated a main effect of temporal expectancy in both crossmodal tasks indicating that rhythmic stimuli in one modality automatically influenced the temporal expectancy in the other modality. An effect of rhythmic cueing was also observed in the unimodal visual task but not unimodal auditory task. Interestingly, there was no observed interaction between top-down spatial attention and temporal expectancy effects in any task suggesting temporal and spatial processing independently affected target response times.

That selectively attending to a spatial location enhances perceptual processing and facilitates behavior at the attended locations has been well documented (for a recent review see Carrasco, 2014). Although relatively less researched, voluntary temporal expectation has also been shown to influence perception and drive behavior (Nobre et al., 2011). Moreover, animal studies have demonstrated that temporal expectations can modulate neural processing in early sensory areas such as primary visual cortex (Lima et al., 2011) and primary auditory cortex (Jaramillo and Zador, 2011). As with spatial selective attention, temporal predictability can be divided into voluntary and stimulus-driven processes. Rhythmic cueing has been used to create temporal predictability, and the rhythm can be used to induce both voluntary or stimulus driven effects, or a combination of both, particularly depending on the instructions. In the present study the rhythmic structure of the cue was not task relevant, in order to investigate whether or not stimulus driven expectancies affected target processing. In other words, whether rhythmic cueing automatically influenced response times even when this temporal characteristic was not relevant or beneficial to the task.

In the unimodal visual task and the two crossmodal conditions the preceding rhythm influenced target detection times. In the visual task, responses to targets in sync with the rhythm were faster compared to early targets, and late targets were also faster compared to when the target was in sync with the rhythm. Similar effects of visual rhythmic cueing have been observed in a study by Rohenkohl et al. (2011) where participants attended to either the color or speed (rhythm) of a moving disc across the screen, to predict an upcoming target. They found both types of cue speeded up response times to targets. However, they found rhythmic cues facilitated response times, compared to an arrhythmic condition, regardless if the participant was instructed to use the rhythmic information or not, whilst the symbolic color cue was only effective when participants explicitly used this information. However, when using rhythmic stimuli to induce temporal expectancies,

it may still be difficult to tear apart the contribution of purely bottom-up effects caused by the rhythm, and any top-down influence such as directing attention to a specific point in time. In other words, are temporal expectancies created passively and purely unintentionally when we are exposed to rhythmic patterns? In some studies using a rhythmic cue, the target was always in sync with the rhythm in the rhythmic condition, as compared to a non-rhythmic condition (e.g., Doherty et al., 2005; Rohenkohl et al., 2012) and in others, the target was more likely to occur in sync compared to out of sync with a rhythm (e.g., Praamstra et al., 2006). In other words, crating conditions where it would be advantageous to use the rhythm to anticipate targets even though not explicitly instructed to do so. Breska and Deouell (2014) specifically investigated whether a rhythm automatically creates temporal expectancies. They included a condition where it was detrimental to use the rhythm to perform the task but still found that the rhythm affected target detection, concluding rhythms automatically exerts an effect on target processing. In the present study the rhythm was not task relevant, but, participants were not explicitly discouraged from using the rhythm. However, the probability of the target appearing at early, sync or late intervals was equally likely and therefore any strategy of expecting the target at a particular time point would not be advantageous. This together with participants concurrently performing another task, directing spatial attention, suggests the rhythmic cueing effects were mainly bottom-up and not involving higher level of processing. However, future studies may wish to specifically address the automaticity of rhythmic cueing and how systematically varying the automaticity is affected by top-down spatial attention. In any case, it can be concluded from the present study that rhythmic cueing effects were observed even though participants performed a concurrent spatial attention task.

In rhythmic cueing studies the perception of targets is typically best when the stimulus coincides with the rhythm and perceptual benefits decrease if the stimulus is presented too early or too late (e.g., Jones et al., 2002; Mathewson et al., 2010). In the current study the detrimental effect of asynchronous stimuli was only present for early, but not for late stimuli. The target was equally likely to appear at early, sync or late intervals meaning there was no strategic benefit in expecting the target at any particular time. However, the pattern of results can in part be explained by the hazard or foreperiod effect. The "hazard function" is an effect whereby an event is more likely to occur at a specific point in time if it has not yet occurred (Luce, 1986). In other words, if the stimulus has not appeared at the early time point it is then more likely to occur at the sync and subsequently the late time interval. This in turn can increase the anticipation and enhance motor readiness. Several steps were taken to account for and to minimize this potential bias. First, catch trials were used whereby no target was presented. This introduces the possibility that if a target has not occurred in sync with the rhythm, it will not necessarily occur at the late time interval. Moreover, to further reduce the hazard function effects and to increase temporal uncertainty of when the target may occur, the cue randomly consisted of either four or five stimuli. Finally, the participants were not informed about the temporal manipulation of the study. Nevertheless, the hazard function fits well with the pattern observed in the purely visual task (VV) whereby response times decreased orderly from early, sync and then late conditions. The expected pattern of results in terms of a model of entrainment would be that in sync targets would be faster than both early and late targets. In the two cross modal tasks there was no difference between late and in sync targets which may suggest both hazard function and entrainment effects influenced RTs. That is, in the early condition both entrainment and hazard effects predict slower RTs compared to the in sync condition. However, when the target is late, the hazard function predicts faster RTs compared to in sync targets, whilst an entrainment model would predict slower RTs. It is therefore possible that both entrainment and hazard function effects were present in this study. Future research may wish to use target discrimination tasks or detection of targets at perceptual threshold to further isolate entrainment effects from hazard functions.

Both crossmodal tasks showed similar effects of rhythmic cueing with a facilitation of response times for targets coinciding with the rhythm compared to when they appeared early. Importantly, this shows that stimulus driven rhythmic cueing is not limited to within a specific modality but effects can span across modalities. This is in line with a recent study by Miller et al. (2013) who found saccades to a visual target were faster when the target was preceded by a synchronous compared to an asynchronous auditory rhythm (see also Bolger et al., 2013 for similar results). The present study extends their findings by showing that audiovisual effects of rhythmic cueing are also found when the modalities are reversed, that is, a visual rhythm entrains auditory targets. This may suggest for a common mechanisms of temporal expectancy created by rhythms which is not modality specific. In line with this, Besle et al. (2011) observed large scaled entrainment of brain areas using intracranial electrocortical recordings in patients with epilepsy. They specifically found that the entrainment of visual stimuli was not confined to the primary visual areas but was observed over a larger brain area. That is, they observed effects in line with a centralized rather than purely modality specific entrainment mechanism.

The one spurious result in the present study was the lack of an temporal expectancy effect in the unimodal auditory task. Auditory entrainment of rhythms has shown to affect target discrimination of tones (e.g., Jones et al., 2002), as well as response times (Sanabria et al., 2011). In contrast, finding an effect of spatial attention in an auditory detection task, as was observed here, has proven more difficult with many studies reporting a null result (e.g., Posner, 1978; Scharf et al., 1987; Buchtel and Butter, 1988; Hugdahl and Nordby, 1994; although see Spence and Driver, 1994 for a positive effect of auditory spatial attention). Whether the introduction of an auditory spatial task diminished any auditory temporal effects remains unclear, however it seems unlikely as the auditory rhythm in the crossmodal task led to temporal expectancy effects of visual stimuli.

The study aimed to investigate how two processes which have independently shown to influence perception and modulate behavior interacted. The results showed that in no task was there

an interaction between spatial attention and rhythmic cueing ($p = 0.37$, $\eta_p^2 = 0.02$). In other words, any effects of rhythmic cueing were similar regardless if the target appeared at the spatially attended or unattended location. Whilst the results here show both unimodal and crossmodal effects of spatial attention, and unimodal and crossmodal effects of rhythmic cueing, spatial attention and temporal expectancy themselves did not interact, neither at a unimodal nor crossmodal level. This suggests temporal and spatial processes can operate independently in driving behavior, at least as measured with response times. Doherty et al. (2005) found similar independent effects of spatial attention and rhythmic cueing on response times, even though their participants were instructed to use the rhythmic structure to predict an upcoming target and thus introducing a voluntary aspect of rhythmic cueing. In contrast, Rohenkohl et al. (2014) recently showed temporal expectation improved perception when the target appeared at a spatially attended location. However, at unattended locations, temporal expectancy did not affect target processing. Rohenkohl et al. did not use rhythmic cueing but the temporal expectation was top-down. Moreover, they investigated perceptual sensitivity rather than response times which may also account for differences. Cravo et al. (2013) measured response times and perceptual accuracy and found both measured to be improved in a rhythmic compared to arrhythmic condition, but, the two measures showed independent effects. There is thus evidence to suggest perceptual modulation following rhythmic stimuli may be different to response time effects. Future research may wish to explore whether perceptual sensitivity effects of automatic entrainment are also independent from spatial attention effects in unimodal and crossmodal conditions. Moreover, the current study used a target detection task which was relatively easy. Within spatial attention research, endogenous and exogenous effects are typically independent when task demands are low. However, when the attentional and cognitive load increases, the two processes have shown to interact when competing for shared resources (Berger et al., 2005). It is conceivable that top-down spatial attention and bottom-up temporal expectancy effects show a similar pattern. In other words, future research could increase the difficulty of the task to investigate whether endogenous spatial attention and stimulus-driven temporal expectancies are independent even when demands on attentional resources are high.

The automatic effect of presenting rhythmic stimuli demonstrated in the present study is partly in line with research on neural oscillations which have seen a recent increase in popularity in the last decade. Evidence is mounting that, not only does our brain self-generate rhythmic oscillations which drives perception and action (e.g., Buzsaki, 2006; Thut and Miniussi, 2009), but these neural oscillations can also be re-set and driven by rhythms and events in our environment (Lakatos et al., 2008; Arnal and Giraud, 2012). Investigating the function and underlying mechanisms of entrainment will not only further our understanding of what drives our behavior and influences our perception, but recent findings have suggested that certain psychiatric and developmental disorders show abnormal neural oscillation patterns (see Calderone et al., 2014 for a recent review).

ACKNOWLEDGMENTS

The author would like to thank AlHanoof AlBastaki, Nilofar Arman, Daisy Mae Uysin, and Fuyumi Tsukiyama for assistance with data collection.

REFERENCES

Arnal, L. H., and Giraud, A. L. (2012). Cortical oscillations and sensory predictions. *Trends Cogn. Sci.* 16, 390–398. doi: 10.1016/j.tics.2012.05.003

Berger, A., Henik, A., and Rafal, R. (2005). Competition between endogenous and exogenous orienting of visual attention. *J. Exp. Psychol. Gen.* 134, 207–221. doi: 10.1037/0096-3445.134.2.207

Besle, J., Schevon, C. A., Mehta, A. D., Lakatos, P., Goodman, R. R., McKhann, G. M., et al. (2011). Tuning of the human neocortex to the temporal dynamics of attended events. *J. Neurosci.* 31, 3176–3185. doi: 10.1523/JNEUROSCI.4518-10.2011

Bolger, D., Trost, W., and Schön, D. (2013). Rhythm implicitly affects temporal orienting of attention across modalities. *Acta Psychol. (Amst)* 142, 238–244. doi: 10.1016/j.actpsy.2012.11.012

Breska, A., and Deouell, L. Y. (2014). Automatic bias of temporal expectations following temporally regular input independently of high-level temporal expectation. *J. Cogn. Neurosci.* 26, 1555–1571. doi: 10.1162/jocn_a_00564

Buchtel, H. A., and Butter, C. M. (1988). Spatial attention shifts: implications for the role of polysensory mechanisms. *Neuropsychologia* 26, 499–509. doi: 10.1016/0028-3932(88)90107-8

Busch, N. A., Dubois, J., and VanRullen, R. (2009). The phase of ongoing EEG oscillations predicts visual perception. *J. Neurosci.* 29, 7869–7876. doi: 10.1523/JNEUROSCI.0113-09.2009

Buzsaki, G. (2006). *Rhythms of the Brain.* Oxford: Oxford University Press.

Calderone, D. J., Lakatos, P., Butler, P. D., and Castellanos, F. X. (2014). Entrainment of neural oscillations as a modifiable substrate of attention. *Trends Cogn. Sci.* 18, 300–309. doi: 10.1016/j.tics.2014.02.005

Carrasco, M. (2014). "Spatial covert attention: perceptual modulation," in *The Oxford Handbook of Attention*, eds K. Nobre and S. Kastner (Oxford: Oxford University Press), 183–230.

Correa, A., Lupianez, J., and Tudela, P. (2005). Attentional preparation based on temporal expectancy modulates processing at the perceptual level. *Psychon. Bull. Rev.* 12, 328–334. doi: 10.3758/bf03196380

Coull, J. T., and Nobre, A. C. (1998). Where and when to pay attention: the neural systems for directing attention to spatial locations and to time intervals as revealed by both PET and fMRI. *J. Neurosci.* 18, 7426–7435.

Cravo, A. M., Rohenkohl, G., Wyart, V., and Nobre, A. C. (2013). Temporal expectation enhances contrast sensitivity by phase entrainment of low-frequency oscillations in visual cortex. *J. Neurosci.* 33, 4002–4010. doi: 10.1523/jneurosci.4675-12.2013

Davranche, K., Nazarian, B., Vidal, F., and Coull, J. (2011). Orienting attention in time activates left intraparietal sulcus for both perceptual and motor task goals. *J. Cogn. Neurosci.* 23, 3318–3330. doi: 10.1162/jocn_a_00030

Doherty, J. R., Rao, A., Mesulam, M. M., and Nobre, A. C. (2005). Synergistic effect of combined temporal and spatial expectations on visual attention. *J. Neurosci.* 25, 8259–8266. doi: 10.1523/jneurosci.1821-05.2005

Hugdahl, K., and Nordby, H. (1994). Electrophysiological correlates to cued attentional shifts in the visual and auditory modalities. *Behav. Neural Biol.* 62, 21–32. doi: 10.1016/s0163-1047(05)80055-x

Jaramillo, S., and Zador, A. M. (2011). The auditory cortex mediates the perceptual effects of acoustic temporal expectation. *Nat. Neurosci.* 14, 246–251. doi: 10.1038/nn.2688

Jones, M. R. (2010). "Attending to sound patterns and the role of entrainment," in *Attention and Time*, eds A. C. Nobre and J. T. Coull (Oxford: Oxford University Press), 137–330.

Jones, M. R., Johnston, H. M., and Puente, J. (2006). Effects of auditory structure on anticipatory and reactive attending. *Cogn. Psychol.* 53, 59–96. doi: 10.1016/j.cogpsych.2006.01.003

Jones, M. R., Moynihan, H., MacKenzie, N., and Puente, J. (2002). Temporal aspects of stimulusdriven attending in dynamic arrays. *Psychol. Sci.* 13, 313–319. doi: 10.1111/1467-9280.00458

Kayser, C., Petkov, C. I., and Logothetis, N. K. (2008). Visual modulation of neurons in auditory cortex. *Cereb. Cortex* 18, 1560–1574. doi: 10.1093/cercor/bhm187

Lakatos, P., Chen, C. M., O'Connell, M. N., Mills, A., and Schroeder, C. E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron* 53, 279–292. doi: 10.1016/j.neuron.2006.12.011

Lakatos, P., Karmos, G., Mehta, A., Ulbert, I., and Schroeder, C. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science* 320, 110–113. doi: 10.1126/science.1154735

Lakatos, P., Musacchia, G., O'Connel, M. N., Falchier, A. Y., Javitt, D. C., and Schroeder, C. E. (2013). The spectrotemporal filter mechanism of auditory selective attention. *Neuron* 77, 750–761. doi: 10.1016/j.neuron.2012.11.034

Lampar, A., and Lange, K. (2011). Effects of temporal trial-by-trial cuing on early and late stages of auditory processing: evidence from event-related potentials. *Atten. Percept. Psychophys.* 73, 1916–1933. doi: 10.3758/s13414-011-0149-z

Lange, K., Krämer, U. M., and Röder, B. (2006). Attending points in time and space. *Exp. Brain Res.* 173, 130–140. doi: 10.1007/s00221-006-0372-3

Lange, K., and Röder, B. (2006). Orienting attention to points in time improves stimulus processing both within and across modalities. *J. Cogn. Neurosci.* 18, 715–729. doi: 10.1162/jocn.2006.18.5.715

Lawrence, M. A., and Klein, R. M. (2013). Isolating exogenous and endogenous modes of temporal attention. *J. Exp. Psychol. Gen.* 142, 560–572. doi: 10.1037/a0029023

Lima, B., Singer, W., and Neuenschwander, S. (2011). Gamma responses correlate with temporal expectation in monkey primary visual cortex. *J. Neurosci.* 31, 15919–15931. doi: 10.1523/JNEUROSCI.0957-11.2011

Luce, R. D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization.* Oxford: Oxford University Press.

Mangun, G. R., and Hillyard, S. A. (1990). Allocation of visual attention to spatial locations: tradeoff functions for event-related brain potentials and detection performance. *Percept. Psychophys.* 47, 532–550. doi: 10.3758/bf03203106

Mathewson, K. E., Fabiani, M., Gratton, G., Beck, D. M., and Lleras, A. (2010). Rescuing stimuli from invisibility: inducing a momentary release from visual masking with pre-target entrainment. *Cognition* 115, 186–191. doi: 10.1016/j.cognition.2009.11.010

Miller, J. E., Carlson, L. A., and McAuley, J. D. (2013). When what you hear influences when you see: listening to an auditory rhythm influences the temporal allocation of visual attention. *Psychol. Sci.* 24, 11–18. doi: 10.1177/0956797612446707

Naccache, L., Blandin, E., and Dehaene, S. (2002). Unconscious masked priming depends on temporal attention. *Psychol. Sci.* 13, 416–424. doi: 10.1111/1467-9280.00474

Nobre, A. C. (2010). "How can temporal expectations bias perception and action," in *Attention and Time*, eds A. C. Nobre and J. T. Coull (Oxford, UK: Oxford University Press), 137–330.

Nobre, A. C., and Rohenkohl, G. (2014). "Time for the fourth dimension in attention," in *The Oxford Handbook of Attention*, eds K. Nobre and S. Kastner (Oxford: Oxford University Press), 676–724.

Nobre, A. C., Rohenkohl, G., and Stokes, M. G. (2011). "Nervous anticipation -Top down biasing across space and time," in *Cognitive Neuroscience of Attention*. 2nd Edn. ed M. I. Posner (New York: Guilford Publications), 159–186.

Posner, M. I. (1978). *Chronometric Explorations of Mind.* Hillsdale, NJ: Erlbaum.

Posner, M. I. (1980). Orienting of attention. *Q. J. Exp. Psychol.* 32, 3–25. doi: 10.1080/00335558008248231

Posner, M. I., Snyder, C. R., and Davidson, B. J. (1980). Attention and the detection of signals. *J. Exp. Psychol.* 109, 160–174. doi: 10.1037/0096-3445.109.2.160

Praamstra, P., Kourtis, D., Kwok, H. F., and Oostenveld, R. (2006). Neurophysiology of implicit timing in serial choice reaction time performance. *J. Neurosci.* 26, 5448–5455. doi: 10.1523/jneurosci.0440-06.2006

Raichle, M. E. (2010). Two views of brain function. *Trends Cogn. Sci.* 14, 180–190. doi: 10.1016/j.tics.2010.01.008

Rohenkohl, G., Coull, J. T., and Nobre, A. C. (2011). Behavioural dissociation between exogenous and endogenous temporal orienting of attention. *PLoS One* 6:e14620. doi: 10.1371/journal.pone.0014620

Rohenkohl, G., Cravo, A. M., Wyart, V., and Nobre, A. C. (2012). Temporal expectation improves the quality of sensory information. *J. Neurosci.* 32, 8424–8428. doi: 10.1523/jneurosci.0804-12.2012

Rohenkohl, G., Gould, I. C., Pessoa, J., and Nobre, A. C. (2014). Combining spatial and temporal expectations to improve visual perception. *J. Vis.* 14:8. doi: 10.1167/14.4.8

Sanabria, D., Capizzi, M., and Correa, A. (2011). Rhythms that speed you up. *J. Exp. Psychol. Hum. Percept. Perform.* 37, 236–244. doi: 10.1037/a0019956

Santangelo, V., and Spence, C. (2008). Is the exogenous orienting of spatial attention truly automatic? Evidence from unimodal and multisensory studies. *Conscious. Cogn.* 17, 989–1015. doi: 10.1016/j.concog.2008.02.006

Scharf, B., Quigley, S., Aoki, C., Peachey, N., and Reeves, A. (1987). Focused auditory attention and frequency selectivity. *Percept. Psychophys.* 42, 215–223. doi: 10.3758/bf03203073

Spence, C. (2010). Crossmodal spatial attention. *Ann. N Y Acad. Sci.* 1191, 182–200. doi: 10.1111/j.1749-6632.2010.05440.x

Spence, C. J., and Driver, J. (1994). Covert spatial orienting in audition: exogenous and endogenous mechanisms. *J. Exp. Psychol. Hum. Percept. Perform.* 20, 555–574. doi: 10.1037//0096-1523.20.3.555

Spence, C., and Driver, J. (1996). Audiovisual links in endogenous covert spatial attention. *J. Exp. Psychol. Hum. Percept. Perform.* 22, 1005–1030. doi: 10.1037//0096-1523.22.4.1005

Spence, C., and Driver, J. (2004). *Crossmodal Space and Crossmodal Attention.* England: Oxford University Press.

Tang, X., Li, C., Li, Q., Gao, Y., Yang, W., Yang, J., et al. (2013). Modulation of auditory stimulus processing by visual spatial or temporal cue: an event-related potentials study. *Neurosci. Lett.* 553, 40–45. doi: 10.1016/j.neulet.2013.07.022

Thut, G., and Miniussi, C. (2009). New insights into rhythmic brain activity from TMS-EEG studies. *Trends Cogn. Sci.* 13, 182–189. doi: 10.1016/j.tics.2009.01.004

Vroomen, J., and de Gelder, B. (2000). Sound enhances visual perception: cross-modal effects of auditory organization of vision. *J. Exp. Psychol. Hum. Percept. Perform.* 26, 1583–1590. doi: 10.1037//0096-1523.26.5.1583

Wright, R. D., and Ward, L. M. (1994). Shifts of visual attention: an historical and methodological overview. *Can. J. Exp. Psychol.* 48, 151–166. doi: 10.1037/1196-1961.48.2.151

Yeshurun, Y., and Carrasco, M. (1998). Attention improves or impairs visual performance by enhancing spatial resolution. *Nature* 396, 72–75. doi: 10.1038/23936

Zanto, T. P., Pan, P., Liu, H., Bollinger, J., Nobre, A. C., and Gazzaley, A. (2011). Age-related changes in orienting attention in time. *J. Neurosci.* 31, 12461–12470. doi: 10.1523/jneurosci.1149-11.2011

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 28 August 2014; accepted: 05 December 2014; published online: 06 January 2015.

Citation: Jones A (2015) Independent effects of bottom-up temporal expectancy and top-down spatial attention. An audiovisual study using rhythmic cueing. Front. Integr. Neurosci. ***8****:96. doi: 10.3389/fnint.2014.00096*

This article was submitted to the journal Frontiers in Integrative Neuroscience.

frontiers in
INTEGRATIVE NEUROSCIENCE

OPINION ARTICLE
published: 03 March 2015
doi: 10.3389/fnint.2015.00017

Top-down control and early multisensory processes: chicken vs. egg

Rosanna De Meo[1], Micah M. Murray[1,2], Stephanie Clarke[1] and Pawel J. Matusz[3,4,5]*

[1] Neuropsychology and Neurorehabilitation Service, Centre Hospitalier Universitaire Vaudois and University of Lausanne, Lausanne, Switzerland
[2] Electroencephalography Brain Mapping Core, Center for Biomedical Imaging (CIBM), Lausanne and Geneva, Switzerland
[3] The Laboratory for Investigative Neurophysiology, Neuropsychology and Neurorehabilitation Service and Department of Radiology, Centre Hospitalier Universitaire Vaudois and University of Lausanne, Lausanne, Switzerland
[4] Faculty in Wroclaw, University of Social Sciences and Humanities, Wroclaw, Poland
[5] Attention, Brain and Cognitive Development Group, Department of Experimental Psychology, University of Oxford, Oxford, UK
*Correspondence: pawel.matusz@chuv.ch

Edited by:
Salvador Soto-Faraco, Universitat Pompeu Fabra, Spain

Reviewed by:
Mark T. Wallace, Vanderbilt University, USA
Durk Talsma, UGent, Belgium

Keywords: attention, control processes, top-down control, bottom-up, multisensory, EEG/ERP, crossmodal

Traditional views contend that behaviorally-relevant multisensory interactions occur relatively late during stimulus processing and subsequently to influences of (top-down) attentional control. In contrast, work from the last 15 years shows that information from different senses is integrated in the brain also during the initial 100 ms after stimulus onset and within low-level cortices. Critically, many of these early-latency multisensory interactions (hereafter *eMSI*) directly impact behavior. The prevalence of eMSI substantially advances our understanding of how unified perception and goal-related behavior emerge. However, it also raises important questions about the dependency of the eMSI on top-down, goal-based attentional control mechanisms that bias information processing toward task-relevant objects (hereafter *top-down control*). To date, this dependency remains controversial, because eMSI can occur independently of top-down control, making it plausible for (some) multisensory processes to directly shape perception and behavior. In other words, the former is not necessary for these early effects to occur and to link them with perception (see **Figure 1A**). This issue epitomizes the fundamental question regarding direct links between sensation, perception, and behavior (*direct perception*), and also extends it in a crucial way to incorporate the multisensory nature of everyday experience. At the same time, the emerging framework must strive to also incorporate the variety of higher-order control mechanisms that likely influence multisensory stimulus responses but which are not based on task-relevance. This article presents a critical perspective about the importance of top-down control for eMSI: In other words, who is controlling whom?

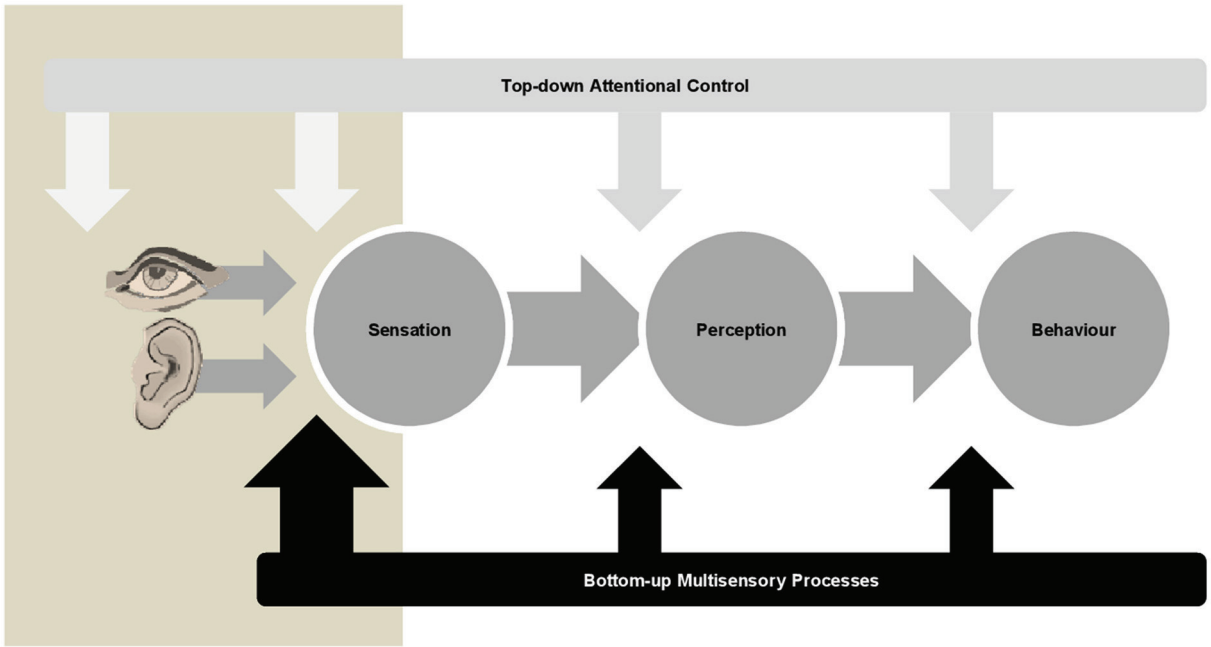
THE UBIQUITY OF eMSI

For the purposes of this article we focus exclusively on auditory-visual interactions and define *eMSI* as those multisensory processes that occur within the first 100 ms post-stimulus onset (but see (Giard and Peronnet, 1999); Giard and Peronnet, who qualified effects <200 ms as early-latency). This definition is in keeping with influential models of visual perception and attentional selection, positing that top-down and recursive inputs manifest after the initial 100 ms of stimulus-driven brain activity, which is believed to be sensory-perceptual and bottom-up in nature (e.g., Luck et al., 1997; Lamme and Roelfsema, 2000). It is likewise important to distinguish between *integration* effects, which are responses elicited by a combination of inputs to different senses, and *cross-modal* effects, which refer to influences of inputs to one sense on activity associated with another sense (e.g., Stein et al., 2010).

The typical perceptual outcome of multisensory integration is that stimulus processing is facilitated (as shown by faster and/or more accurate responses) in contexts where inputs to different senses are carrying similar (*redundant*) information and are presented close in time. This behavioral facilitation is typically accompanied by brain responses to multisensory stimuli that diverge from the summed brain responses to the constituent unisensory signals (*nonlinear responses*; **Figure 1B**). Given the growing evidence for links between the brain and the behavioral responses (reviewed in Murray et al., 2012), one mechanism may be that the temporal co-occurrence of multisensory information lowers the threshold for neural activity that in turn drives perception and action (e.g., Rowland and Stein, 2007).

Based on the extant literature, we argue that these particular multisensory processes, which are reflected by eMSI, are stimulus-driven, *bottom-up* in nature and affect perception and behavior in a direct manner and largely independently of top-down control (**Figure 1A**). The idea of a variety, or even a range of multisensory processes, where some are "automatic" while others dependent on one's current behavioral goals, has until now been systematically investigated mainly in the context of attentional selection of objects in space, rather than their perception *per se* (e.g., Matusz and Eimer, 2011, 2013; Matusz et al., 2015; Talsma and Woldorff, 2005; but see Murray et al., 2004; Soto-Faraco et al., 2004; Tiippana et al., 2004;

A



B

	Technique	Experimental paradigm	Direction	Latency
Human studies				
Giard & Peronnet (1999)	EEG	discrimination (2-forced-choice)	↑	40-90ms
Molholm et al. (2002)	EEG	detection	↓	45-80ms
Fort et al. (2002a)	EEG	detection	↓	45-85ms
Fort et al. (2002b)	EEG	non-redundant target discrimination	↑	125-140ms
Teder-Sälejärvi et al. (2002)	EEG	discrimination (go-nogo)	↓*	98-112ms
Talsma & Woldorff (2005)	EEG	spatial attention	↓**	60-80ms
Gondan & Röder (2006)	EEG	discrimination (go-nogo)	↓	~80-100ms
Talsma et al. (2007)	EEG	detection tasks	↓**/↑	~50ms
Gondan et al. (2007)	EEG	discrimination (go-nogo)	↓*	90-100ms
Vidal et al. (2008)	EEG	passive presentation	↓	55-95ms
Cappe et al. (2010)	EEG	discrimination (go-nogo)	↓*	60-95ms
Raij et al. (2010)	MEG	discrimination (go-nogo)	n.a.	53/75ms
van der Burg et al. (2011)	EEG	visual search	↓	~50-60ms
Cappe et al. (2012)	EEG	discrimination (go-nogo)	↓	73-113ms
Stekelenburg & Vroomen (2012)	EEG	catch-trial detection	↓*°	40-60ms
Stekelenburg et al. (2013)	EEG	catch-trial detection	↓*	85-145ms
Animal studies				
Barth et al. (1995)	LFP	anaesthetised	n.a.	22ms
Miller et al. (2001)	SUA	detection	†	~100-150ms
Ghazanfar et al. (2005)	LFP	fixation	↓↑	~30-150ms
Rowland & Stein (2007)	SUA	anaesthetised	↓↑	<40ms
Kayser et al. (2008)	LFP; S/MUA	fixation	↓↑	50-250ms

FIGURE 1 | (A) Depiction of manners in which top-down attentional control and bottom-up multisensory processes may influence direct perception in multisensory contexts. In this model, the bottom-up multisensory processes that occur early in time (eMSI; beige box) have direct effects on perception and behavior (large black arrow). In turn, top-down attentional control mechanisms, which are typically posited to exert effects at multiple pre-stimulus and post-stimulus stages, do not seem to do so in some multisensory contexts (white arrows). **(B)** Table summarizing principal findings on eMSI from human EEG/MEG studies and animal electrophysiological studies. Note: EEG, electroencephalography; MEG, magnetoencephalography; LFP, local field potentials; SUA, single-unit activity; MUA, multi-unit activity; ↑, sub-additive responses; ↓, super-additive responses; ∗, responses elicited by irrelevant-but-attended multisensory stimuli; ∗∗, responses elicited by unattended multisensory stimuli; °, eMSI found only for spatially congruent audiovisual stimuli; †, eMSI found on the response latency, but not on the response amplitude; n.a., data not available.

Alsius et al., 2005, 2007; Thelen et al., 2012, 2014; Matusz et al., in press). However, control processes are likely to be important for both cognitive functions (e.g., Gunseli et al., 2014); this should hold for both unisensory and multisensory processes, and bottom-up and top-down processes alike (i.e., multisensory processes are not mechanistically "special"; van Atteveldt et al., 2014).

It is difficult to argue with the idea that early responses are a hallmark of bottom-up multisensory processes in the service of perception, if one considers how ubiquitous and context-independent they are in both humans and in the animal models (see **Figure 1B**; reviewed in Murray et al., 2012; Kajikawa et al., 2012). The eMSI in local field potentials as well as spiking activity have been measured in the primary and secondary auditory fields of fixating monkeys (Ghazanfar et al., 2005; Kayser et al., 2008; see also Lakatos et al., 2008; Wang et al., 2008 for cross-modal effects). Importantly, these eMSI occurred for both etiological objects (conspecific communication signals) and simple audiovisual stimuli, though modulated according to bottom-up stimulus salience and neural efficacy. Moreover, non-linear interactions mirroring the behavioral gains in stimulus detection have been recorded in single neurons in the area 4 of the monkey motor cortex within 100–150 ms post-stimulus (Miller et al., 2001).

Electroencephalography and magnetoencephalography (EEG/MEG) studies in humans have likewise demonstrated eMSI across a variety of tasks, ranging from simple detection (Fort et al., 2002a; Molholm et al., 2002) and discrimination (Giard and Peronnet, 1999; Fort et al., 2002b; Teder-Sälejärvi et al., 2002; Gondan and Röder, 2006; Gondan et al., 2007; Raij et al., 2010; Cappe et al., 2010, 2012; Stekelenburg and Vroomen, 2012; Stekelenburg et al., 2013) tasks to multi-stimulus/ multi-stream paradigms necessitating selection (Talsma and Woldorff, 2005; Talsma et al., 2007; van der Burg et al., 2008, 2011). Importantly, eMSI were observed irrespective of whether the multisensory stimuli were targets (e.g., Giard and Peronnet, 1999; Pérez-Bellido et al., 2013), attended but task-irrelevant stimuli (e.g., Cappe et al., 2010) or were presented passively (Vidal et al., 2008). As will be detailed below, data from brain stimulation studies allow causal inference regarding behavioral consequences of eMSI (see below).

The interpretability of the eMSI in terms of bottom-up vs. top-down mechanisms critically depends on their localization. Despite the ubiquity of the eMSI in extant EEG/MEG studies, only few have applied the requisite signal analysis and source reconstruction methods. Localization results support the predominant role of low-level cortices in the eMSI (Cappe et al., 2010; Raij et al., 2010). While the localization of the eMSI to low-level cortices could be taken as evidence for their strictly bottom-up nature, their latency at ∼50–100 ms is sufficiently "late" to provide ample opportunity for recursive processing (Musacchia and Schroeder, 2009; also Moran and Reilly, 2006 for modeling results). This may involve top-down modulation or the extraction and disambiguation of stimulus features (Lamme and Roelfsema, 2000). Thus, care is warranted in regarding all eMSI as indicative of bottom-up multisensory integration. For example, *the pip and pop* effect (van der Burg et al., 2008) triggers eMSI-like responses, but only in the case of targets, not distractors (van der Burg et al., 2011). Thus, dependency of the eMSI on top-down control can be assessed only by analyzing studies where the latter is directly manipulated[1].

THE CHICKEN: TOP-DOWN CONTROL AND ITS LIMITED ROLE IN eMSI

The strongest evidence for the dependence of eMSI on top-down control comes from studies where attended and unattended multisensory stimuli were directly compared (e.g., Alsius et al., 2005, 2007; Talsma and Woldorff, 2005; Talsma et al., 2007). However, the literature seems prone to misconstruing the full breadth of the results. In one study participants detected infrequent targets in one of two central streams of rapidly presented alphanumeric symbols or combinations of beeps and flashes (Talsma et al., 2007). When attended, audiovisual stimuli triggered early enhanced (*super-additive*) nonlinear responses. But, when the competing stream was attended, these nonlinear interactions changed polarity, becoming suppressed (*sub-additive*). One interpretation of these results is that top-down control regulates multisensory integration, from its magnitude and quality to its very presence (Koelewijn et al., 2010). We believe this viewpoint should perhaps be more nuanced. The top-down control manipulations modulated the eMSI, but did not eliminate them. Additionally, the eMSI were observed despite the paradigm manipulating in fact *multiple* top-down mechanisms (inter-modal, but also spatial, feature-based, and object-based). While further research is required to fully characterize the mechanistic underpinnings of super- vs. sub-additive interactions, the results of this study are in line with the importance of top-down control processes revealed by unisensory studies, wherein responses to stimuli are enhanced according to the task-relevance of their location, features or identity (reviewed in Nobre and Kastner, 2013). Talsma et al. (2007) was the first to demonstrate the pivotal role of the task-relevance of multisensory pairings for the *quality* of the eMSI they trigger. However, the *presence* of the eMSI in this study was independent of task-relevance, though some evidence would suggest that the eMSI are preferentially observed in *unattended* contexts (Table 2 in Talsma and Woldorff, 2005). This latter evidence is in line with the eMSI being a hallmark of stimulus-driven processing.

It is difficult to ignore that in these few studies, where top-down control mechanisms were directly manipulated, the eMSI were sub-additive in nature. What is striking is that this is precisely the direction of effects reported in the literature irrespective of whether responses to targets, non-targets or passively presented stimuli are considered (**Figure 1B**). Historically, sub-additive effects were dismissed as confounds related to common activity across both unisensory and multisensory conditions. More recently, they have been increasingly recognized as a canonical mechanism that can convey information particularly efficiently (Kayser et al., 2009; Altieri et al., in press; reviewed in Stevenson et al., 2014). The issue of the quantification of the eMSI is further complicated by the fact that the overwhelming

[1]To our knowledge semantic congruence does not modulate eMSI (Fort et al., 2002b; Molholm et al., 2004; Yuval-Greenberg and Deouell, 2007).

majority of the human EEG studies have used relative, reference-dependent measures of amplitude (cf., Murray et al., 2008).

THE EGG: eMSI AS A BOTTOM-UP PHENOMENON

Several independent lines of research across various species provide converging evidence for the bottom-up nature of the eMSI. On the one hand, there are reports of eMSI in anesthetized animals (e.g., rats, Barth et al., 1995; cats, Rowland and Stein, 2007; see also reviews in Sarko et al., 2012; Rowland and Stein, 2014), where top-down modulations are blocked[2]. On the other hand, sounds have been shown to enhance the excitability of low-level visual cortices, as measured via phosphene perception. Several aspects of this effect demonstrated by TMS studies in humans support the bottom-up nature of the eMSI and the causal links between eMSI and behavior (Romei et al., 2007, 2009, 2013; Spierer et al., 2013).

First, it is modulated by low-level sound features, with greater excitability increases observed for narrowband and higher pitch sounds. Visual cortex excitability is furthermore enhanced selectively by structured approaching (looming) sounds versus stationary or receding sounds as well as non-structured white-noise versions of these sounds. Second, the effect is delimited in time, occurring when sounds precede the TMS by 30–150 ms, in correspondence with the eMSI identified using EEG/MEG. Third, the sound-induced enhancements of visual cortex excitability transpire before subjects can explicitly differentiate between the sounds, i.e., at pre-perceptual processing stages. Relatedly, increases in the occipital excitability occur with sounds that themselves fail to elicit startle responses, arguing against an alerting explanation. Fourth, evidence against a top-down account of these effects comes from studies demonstrating that individuals' attentional preference (as independently measured in an auditory-visual divided attention task) affect late, but not early, stages of the excitability changes.

Finally, the TMS-driven visual cortex activity is behaviorally relevant. Occipital TMS delivered 60–90 ms post-stimulus has opposing effects of roughly equal magnitude (∼15 ms) on reaction times to unisensory auditory and visual stimuli (speeding and slowing, respectively) and has no measurable effect on reaction times to simultaneous auditory-visual multisensory stimuli. Critically, the response speed facilitation obtained from the combination of occipital TMS and an external auditory stimulus was as great as and correlated with that obtained from presenting participants with genuine multisensory stimuli. The TMS-induced cross-modal effects seem to emulate those observed with multisensory stimuli.

CONCLUSIONS AND FUTURE DIRECTIONS

We demonstrated that the eMSI are robust phenomena, observable across species, experimental paradigms and measures of neural activity (**Figure 1B**). To refer more explicitly to the Research Topic of this issue, we subscribe to a view of multiple multisensory processes: The eMSI are a hallmark of bottom-up multisensory processes that facilitate perception and behavior directly, independently of top-down control (**Figure 1A**).

We focused here exclusively on stimulus-locked brain activity. Thus, temporal dynamics complement the understanding of the interplay between bottom-up and top-down mental processes as hitherto provided from the vantage-point of brain oscillations, which assay both intra-population excitability as well as inter-population communication (Thut et al., 2012; van Atteveldt et al., 2014).

A critical next step will be the detailed mechanistic characterization of the eMSI. The sub-additive archetype of the eMSI goes together with the evidence from unisensory research linking reduced responses with more efficient and information-rich processing akin to the repetition suppression phenomena and the predictive coding accounts (e.g., Grill-Spector et al., 2006; Summerfield and Egner, 2009). When and why do top-down control processes flip the sub-additive eMSI to become super-additive? If top-down control affects the nature, rather than the presence, of multisensory processes, then what are the consequences for our understanding of perception? Paradoxically, while the eMSI are on the one hand upturning somewhat dogmatic views on the brain functional organization, they simultaneously are entrenching a classic model of perceptual processing positing direct links between sensation, perception, and behavior. An accurate picture of the nature of perceptual processes is thus provided by studying them in naturalistic, multisensory contexts and where the task demands dynamically vary.

[2]We would hasten to remind the reader that convergent anatomical input is necessary but in and of itself insufficient for eMSI as defined it in this opinion piece. It is true that the anatomic pathways/connectivities as well as their shaping by experiences are prerequisites for multisensory processes. However, the activation of these physical substrates in relation to the cascade of sensory-evoked responses must be sufficiently early so as to influence perception and behavior directly and thus be qualified as eMSI.

AUTHOR CONTRIBUTIONS

All authors have contributed to all aspects of this work. All authors have approved the final version of the manuscript and agreed to be held accountable for all aspects of the work.

ACKNOWLEDGMENTS

Financial support was provided by the Swiss National Science Foundation (grant 320030-149982 to MMM, grant 320030B-141177 to SC), National Centre of Competence in Research project "SYNAPSY, The Synaptic Bases of Mental Disease" [project 51AU40-125759]) and the Swiss Brain League (2014 Research Prize to MMM).

REFERENCES

Alsius, A., Navarra, J., Campbell, R., and Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Curr. Biol.* 15, 839–843. doi: 10.1016/j.cub.2005.03.046

Alsius, A., Navarra, J., and Soto-Faraco, S. (2007). Attention to touch weakens audiovisual speech integration. *Exp. Brain Res.* 183, 399–404. doi: 10.1007/s00221-007-1110-1

Altieri, N., Stevenson, R. A., Wallace, M. T., and Wenger, M. J. (in press). Learning to associate auditory and visual stimuli: behavioral and neural mechanisms. *Brain Topogr.* 28. doi: 10.1007/s10548-013-0333-7

Barth, D. S., Goldberg, N., Brett, B., and Di, S. (1995). The spatiotemporal organization of auditory, visual, and auditory-visual evoked potentials in rat cortex. *Brain Res.* 678, 177–190. doi: 10.1016/0006-8993(95)00182-P

Cappe, C., Thelen, A., Romei, V., Thut, G., and Murray, M. M. (2012). Looming signals reveal synergistic principles of multisensory integration. *J. Neurosci.* 32, 1171–1182. doi: 10.1523/JNEUROSCI.5517-11.2012

Cappe, C., Thut, G., Romei, V., and Murray, M. M. (2010). Auditory–visual multisensory interactions in humans: timing, topography, directionality, and sources. *J. Neurosci.* 30, 12572–12580. doi: 10.1523/JNEUROSCI.1099-10.2010

Fort, A., Delpuech, C., Pernier, J., and Giard, M. H. (2002a). Dynamics of cortico-subcortical cross-modal operations involved in audio-visual object detection in humans. *Cerb. Cortex* 12, 1031–1039. doi: 10.1093/cercor/12.10.1031

Fort, A., Delpuech, C., Pernier, J., and Giard, M. H. (2002b). Early auditory–visual interactions in human cortex during nonredundant target identification. *Cogn. Brain Res.* 14, 20–30. doi: 10.1016/S0926-6410(02)00058-7

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., and Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *J. Neurosci*. 25, 5004–5012. doi: 10.1523/JNEUROSCI.0799-05.2005

Giard, M. H., and Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. *J. Cogn. Neurosci.* 11, 473–490. doi: 10.1162/089892999563544

Gondan, M., and Röder, B. (2006). A new method for detecting interactions between the senses in event-related potentials. *Brain Res.* 1073, 389–397. doi: 10.1016/j.brainres.2005.12.050

Gondan, M., Vorberg, D., and Greenlee, M. W. (2007). Modality shift effects mimic multisensory interactions: an event-related potential study. *Exp. Brain Res.* 182, 199–214. doi: 10.1007/s00221-007-0982-4

Grill-Spector, K., Henson, R., and Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends Cogn. Sci.* 10, 14–23. doi: 10.1016/j.tics.2005.11.006

Gunseli, E., Meeter, M., and Olivers, C. N. (2014). Is a search template an ordinary working memory? Comparing electrophysiological markers of working memory maintenance for visual search and recognition. *Neuropsychologia* 60, 29–38. doi: 10.1016/j.neuropsychologia.2014.05.012

Kajikawa, Y., Falchier, A., Musacchia, G., Lakatos, P., and Schroeder, C. E. (2012). "Audiovisual integration in nonhuman primates: a window into the anatomy and physiology of cognition," in *The Neural Bases of Multisensory Processes*, ed M. M. Murray and M. T. Wallace (Boca Raton, FL: CRC Press), 65–98.

Kayser, C., Montemurro, M. A., Logothetis, N. K., and Panzeri, S. (2009). Spike-phase coding boosts and stabilizes information carried by spatial and temporal spike patterns. *Neuron* 61, 597–608. doi: 10.1016/j.neuron.2009.01.008

Kayser, C., Petkov, C. I., and Logothetis, N. K. (2008). Visual modulation of neurons in auditory cortex. *Cerb. Cortex* 18, 1560–1574. doi: 10.1093/cercor/bhm187

Koelewijn, T., Bronkhorst, A., and Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: a review of audiovisual studies. *Acta Psychol.* 134, 372–384. doi: 10.1016/j.actpsy.2010.03.010

Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I., and Schroeder, C. E. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science* 320, 110–113. doi: 10.1126/science.1154735

Lamme, V. A., and Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* 23, 571–579. doi: 10.1016/S0166-2236(00)01657-X

Luck, S. J., Chelazzi, L., Hillyard, S. A., and Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *J. Neurophysiol.* 77, 24–42.

Matusz, P. J., Broadbent, H., Ferrari, J., Forrest, B., Merkley, R., and Scerif, G. (2015). Multi-modal distraction: insights from children's limited attention. *Cognition* 136, 156–165. doi: 10.1016/j.cognition.2014.11.031

Matusz, P. J., and Eimer, M. (2011). Multisensory enhancement of attentional capture in visual search. *Psychon. B. Rev.* 18, 904–909. doi: 10.3758/s13423-011-0131-8

Matusz, P. J., and Eimer, M. (2013). Top−down control of audiovisual search by bimodal search templates. *Psychophysiology* 50, 996–1009. doi: 10.1111/psyp.12086

Matusz, P. J., Thelen, A., Geiser, E., Anken, J., and Murray, M. M. (in press). The role of auditory cortices in the retrieval of single-trial auditory-visual memories. *Eur. J. Neurosci.*

Miller, J., Ulrich, R., and Lamarre, Y. (2001). Locus of the redundant-signals effect in bimodal divided attention: a neurophysiological analysis. *Percept. Psychophys.* 63, 555–562. doi: 10.3758/BF03194420

Molholm, S., Ritter, W., Javitt, D. C., and Foxe, J. J. (2004). Multisensory visual–auditory object recognition in humans: a high-density electrical mapping study. *Cerb. Cortex* 14, 452–465. doi: 10.1093/cercor/bhh007

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., and Foxe, J. J. (2002). Multisensory auditory–visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Cogn. Brain Res.* 14, 115–128. doi: 10.1016/S0926-6410(02)00066-6

Moran, R. J., and Reilly, R. B. (2006). "Neural mass model of human multisensory integration," in *28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (New York, NY: IEEE), 5559–5562.

Murray, M. M., Brunet, D., and Michel, C. M. (2008). Topographic ERP analyses: a step-by-step tutorial review. *Brain Topogr.* 20, 249–264. doi: 10.1007/s10548-008-0054-5

Murray, M. M., Cappe, C., Romei, V., Martuzzi, R., and Thut, G. (2012). "Auditory-visual multisensory interactions in humans: a synthesis of findings from behavior, ERPs, fMRI, and TMS," in *The New Handbook of Multisensory Processes*, ed B. E. Stein (Cambridge, MA: MIT Press), 223–238.

Murray, M. M., Michel, C. M., de Peralta, R. G., Ortigue, S., Brunet, D., Andino, S. G., et al. (2004). Rapid discrimination of visual and multisensory memories revealed by electrical neuroimaging. *Neuroimage* 21, 125–135.

Musacchia, G., and Schroeder, C. E. (2009). Neuronal mechanisms, response dynamics and perceptual functions of multisensory interactions in auditory cortex. *Hear. Res.* 258, 72–79. doi: 10.1016/j.heares.2009.06.018

Nobre, K., and Kastner, S. (eds.). (2013). *The Oxford Handbook of Attention*. Oxford: Oxford University Press.

Pérez-Bellido, A., Soto-Faraco, S., and López-Moliner, J. (2013). Sound-driven enhancement of vision: disentangling detection-level from decision-level contributions. *J. Neurophys.* 109, 1065–1077. doi: 10.1152/jn.00226.2012

Raij, T., Ahveninen, J., Lin, F. H., Witzel, T., Jääskeläinen, I. P., Letham, B., et al. (2010). Onset timing of cross−sensory activations and multisensory interactions in auditory and visual sensory cortices. *Eur. J. Neurosci.* 31, 1772–1782. doi: 10.1111/j.1460-9568.2010.07213.x

Romei, V., Murray, M. M., Cappe, C., and Thut, G. (2009). Preperceptual and stimulus-selective enhancement of low-level human visual cortex excitability by sounds. *Curr. Biol.* 19, 1799–1805. doi: 10.1016/j.cub.2009.09.027

Romei, V., Murray, M. M., Cappe, C., and Thut, G. (2013). The contributions of sensory dominance and attentional bias to cross-modal enhancement of visual cortex excitability. *J. Cogn. Neurosci.* 25, 1122–1135. doi: 10.1162/jocn_a_00367

Romei, V., Murray, M. M., Merabet, L. B., and Thut, G. (2007). Occipital transcranial magnetic stimulation has opposing effects on visual and auditory stimulus detection: implications for multisensory interactions. *J. Neurosci.* 27, 11465–11472. doi: 10.1523/JNEUROSCI.2827-07.2007

Rowland, B. A., and Stein, B. E. (2007). Multisensory integration produces an initial response enhancement. *Front. Integr. Neurosci.* 1:4. doi: 10.3389/neuro.07.004.2007

Rowland, B. A., and Stein, B. E. (2014). A model of the temporal dynamics of multisensory enhancement. *Neurosci. Biobeh. Rev.* 41, 78–84. doi: 10.1016/j.neubiorev.2013.12.003

Sarko, D. K., Nidiffer, A. R., Powers, I. I. I., A. R., Ghose, D. and Wallace, M. T. (2012). "Spatial and temporal features of multisensory processes," in *The Neural Basis of Multisensory Processes*, ed M. M. Murray and M. T. Wallace (Boca Raton, FL: CRC Press), 191–215.

Soto-Faraco, S., Navarra, J., and Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. *Cognition* 92, B13–B23. doi: 10.1016/j.cognition.2003.10.005

Spierer, L., Manuel, A. L., Bueti, D., and Murray, M. M. (2013). Contributions of pitch and bandwidth to sound-induced enhancement of visual cortex excitability in humans. *Cortex* 49, 2728–2734. doi: 10.1016/j.cortex.2013.01.001

Stein, B. E., Burr, D., Constantinidis, C., Laurienti, P. J., Alex Meredith, M., Perrault, T. J., et al. (2010). Semantic confusion regarding the development of multisensory integration: a practical solution. *Eur. J. Neurosci.* 31, 1713–1720. doi: 10.1111/j.1460-9568.2010.07206.x

Stekelenburg, J. J., Maes, J. P., Van Gool, A. R., Sitskoorn, M., and Vroomen, J. (2013). Deficient multisensory integration in schizophrenia: an event-related potential study. *Schizophr. Res.* 147, 253–261. doi: 10.1016/j.schres.2013.04.038

Stekelenburg, J. J., and Vroomen, J. (2012). Electrophysiological correlates of predictive coding of auditory location in the perception of

natural audiovisual events. *Front. Integr. Neurosci.* 6:26. doi: 10.3389/fnint.2012.00026

Stevenson, R. A., Ghose, D., Fister, J. K., Sarko, D. K., Altieri, N. A., Nidiffer, A. R., et al. (2014). Identifying and quantifying multisensory integration: a tutorial review. *Brain Topogr.* 27, 707–730. doi: 10.1007/s10548-014-0365-7

Summerfield, C., and Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends Cogn. Sci.* 13, 403–409. doi: 10.1016/j.tics.2009.06.003

Talsma, D., Doty, T. J., and Woldorff, M. G. (2007). Selective attention and audiovisual integration: is attending to both modalities a prerequisite for early integration? *Cerb. Cortex* 17, 679–690. doi: 10.1093/cercor/bhk016

Talsma, D., and Woldorff, M. G. (2005). Selective attention and multisensory integration: multiple phases of effects on the evoked brain activity. *J. Cogn. Neurosci.* 17, 1098–1114. doi: 10.1162/0898929054475172

Teder-Sälejärvi, W. A., McDonald, J. J., Di Russo, F., and Hillyard, S. A. (2002). An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings. *Cogn. Brain Res.* 14, 106–114. doi: 10.1016/S0926-6410(02)00065-4

Thelen, A., Cappe, C., and Murray, M. M. (2012). Electrical neuroimaging of memory discrimination based on single-trial multisensory learning. *NeuroImage* 62, 1478–1488. doi: 10.1016/j.neuroimage.2012.05.027

Thelen, A., Matusz, P. J., and Murray, M. M. (2014). Multisensory context portends object memory. *Curr. Biol.* 24, R734–R735. doi: 10.1016/j.cub.2014.06.040

Thut, G., Miniussi, C., and Gross, J. (2012). The functional importance of rhythmic activity in the brain. *Curr. Biol.* 22, R658–R663. doi: 10.1016/j.cub.2012.06.061

Tiippana, K., Andersen, T. S., and Sams, M. (2004). Visual attention modulates audiovisual speech perception. *Eur. J. Cogn. Psychol.* 16, 457–472. doi: 10.1080/09541440340000268

van Atteveldt, N., Murray, M. M., Thut, G., and Schroeder, C. E. (2014). Multisensory integration: flexible use of general operations. *Neuron* 81, 1240–1253. doi: 10.1016/j.neuron.2014.02.044

van der Burg, E., Olivers, C. N., Bronkhorst, A. W., and Theeuwes, J. (2008). Pip and pop: nonspatial auditory signals improve spatial visual search. *J. Exp. Psychol. Hum. Percept. Perform.* 34, 1053–1065. doi: 10.1037/0096-1523.34.5.1053

van der Burg, E., Talsma, D., Olivers, C. N., Hickey, C., and Theeuwes, J. (2011). Early multisensory interactions affect the competition among multiple visual objects. *Neuroimage* 55, 1208–1218. doi: 10.1016/j.neuroimage.2010.12.068

Vidal, J., Giard, M. H., Roux, S., Barthelemy, C., and Bruneau, N. (2008). Cross-modal processing of auditory–visual stimuli in a no-task paradigm: a topographic event-related potential study. *Clin. Neurophysiol.* 119, 763–771 doi: 10.1016/j.clinph.2007.11.178

Wang, Y., Celebrini, S., Trotter, Y., and Barone, P. (2008). Visuo-auditory interactions in the primary visual cortex of the behaving monkey: electrophysiological evidence. *BMC Neurosci.* 9:79. doi: 10.1186/1471-2202-9-79

Yuval-Greenberg, S., and Deouell, L. Y. (2007). What you see is not (always) what you hear: Induced gamma band responses reflect cross-modal interactions in familiar object recognition. *J. Neurosci.* 27, 1090–1096. doi: 10.1523/JNEUROSCI.4828-06.2007

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 12 October 2014; accepted: 13 February 2015; published online: 03 March 2015.

Citation: De Meo R, Murray MM, Clarke S and Matusz PJ (2015) Top-down control and early multisensory processes: chicken vs. egg. Front. Integr. Neurosci. ***9****:17. doi: 10.3389/fnint.2015.00017*

This article was submitted to the journal Frontiers in Integrative Neuroscience.

frontiers in Integrative Neuroscience

REVIEW
published: 26 March 2015
doi: 10.3389/fnint.2015.00019

Predictive coding and multisensory integration: an attentional account of the multisensory mind

Durk Talsma *

Department of Experimental Psychology, Ghent University, Ghent, Belgium

Multisensory integration involves a host of different cognitive processes, occurring at different stages of sensory processing. Here I argue that, despite recent insights suggesting that multisensory interactions can occur at very early latencies, the actual integration of individual sensory traces into an internally consistent mental representation is dependent on both top–down and bottom–up processes. Moreover, I argue that this integration is not limited to just sensory inputs, but that internal cognitive processes also shape the resulting mental representation. Studies showing that memory recall is affected by the initial multisensory context in which the stimuli were presented will be discussed, as well as several studies showing that mental imagery can affect multisensory illusions. This empirical evidence will be discussed from a predictive coding perspective, in which a central top–down attentional process is proposed to play a central role in coordinating the integration of all these inputs into a coherent mental representation.

Keywords: multisensory integration, mental model, predictive coding, attention, top–down, bottom–up

OPEN ACCESS

Edited by:
Salvador Soto-Faraco, Universitat Pompeu Fabra, Spain

Reviewed by:
John S. Butler, Trinity College Dublin, Ireland
Manuel R. Mercier, Albert Einstein College of Medicine of Yeshiva University, USA

***Correspondence:**
Durk Talsma, Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, B-9000 Ghent, Belgium durk.talsma@UGent.be

Received: *24 November 2014*
Paper pending published: *27 January 2015*
Accepted: *03 March 2015*
Published: *26 March 2015*

Citation:
Talsma D (2015) Predictive coding and multisensory integration: an attentional account of the multisensory mind. Front. Integr. Neurosci. 9:19. doi: 10.3389/fnint.2015.00019

An Attentional Account of the Multisensory Mind

Imagine watching a rerun of the famous TV-series the Muppet Show. One popular character, the Swedish chef, is known for its gibberish fake Swedish, which at first appears not to make sense at all, other than its comical effect. Yet by carefully watching the Muppet's mouth movements and the various additional cues given by bodily motions, it becomes suddenly clear that the fake Swedish is actually garbled English. As you watch the end of the clip, you can clearly understand the phrase "The chicken is in the basket" as the chef throws poor Camilla the hen through a basketball ring. Imagine now continuing with a different episode, and you may instantly recognize the chef's response "The dog is in the pot," in response to Miss Piggy's query "what happened to my dog Foo Foo."

This example clearly illustrates a major problem that we humans regularly have to overcome in interpreting sensory information, namely resolving ambiguities. Our understandability of the chef's garbled fake Swedish is greatly enhanced through several non-verbal cues. These cues involve direct visual cues, such as the mouth movements accompanying his speech and several other visual cues that provide the appropriate context for understanding the scene. In addition, memory cues that are based on previous experience with similar scenes may also help us in our interpretation. But exactly how do we manage to integrate all these cues?

To answer this, the discipline of *Multisensory Processing* investigates the mechanisms contributing to the combining of information from our various senses. According to Stein et al. (2010), *Multisensory Integration*, refers to the neural process by which unisensory signals are combined

to form a new product or representation. While multisensory processing studies have greatly increased our understanding of the processes directly involved in combining information from multiple senses, it is still not quite clear how our interpretation of sensory information can be enhanced by other sources of information, such as our existing background knowledge based on prior experience. In this review, I aim to discuss how these cues might be integrated with ongoing sensory input to generate a consistent mental representation.

Multisensory Integration: Top–Down and Bottom–Up Processing

Our understanding of brain function has increased sharply in the last 20 years or so. In multisensory processing research we have equally witnessed a rather dramatic shift in our understanding of the processes that combine information across the individual senses. Before the seminal single cell recording studies in animals that demonstrated the existence of multisensory neurons in the superior colliculus (Wallace et al., 1998; Wallace and Stein, 2001), a predominant view in the late 1980s and early 1990s was that multisensory integration takes place relatively late in the processing stream, in cortical areas known as secondary association areas. For instance, in their influential late 1980s textbook "Brain, Mind, and Behavior," neuroscientists Floyd E. Bloom and Arlyne Lazeron, write:

> *"Association areas in the parietal lobe, for example, are thought to synthesize information from the somatosensory cortex–messages from the skin, muscles, tendons, and joints about the body's position and movement–with information about sight and sound transmitted from the visual and auditory cortices in the occipital and temporal lobes. This integrated information helps us to form an accurate sense of our physical selves as we move through our environment." (Bloom and Lazeron, 1988, pp. 274–275).*

Bloom and Lazeron's (1988) description clearly indicates that the merging of information across the senses was supposed to take place *after* the initial sensory processing had come to completion (see **Figure 1**). Since that time, however, many discoveries have suggested that multisensory integration is more complex than this. For example, in addition to the aforementioned single cell recordings, electrophysiological studies showed that multisensory interactions can already take place as early as 40 ms after stimulus presentation, which is considerably earlier than initially thought possible (Giard and Péronnet, 1999; Molholm et al., 2002).

A Predictive Coding Account

One influential framework to explain the intricacies of multisensory processing is that of predictive coding. The predictive coding framework states that the brain produces a Bayesian estimate of the environment (Friston, 2010). According to this view, stochastic models of the environment exist somewhere in the brain[1], which are updated on the basis of processed sensory information. These stochastic models (see Klemen and Chambers, 2011 for a review) thus provide the brain areas lower in the sensory processing hierarchy with predictions (or in Bayesian terms "priors") that can be used to adjust the processing of ongoing sensory input. A strong mismatch between the prediction and the actual sensory input will then result in a major update of the internal model. For example, when unexpected sensory input is present, our internal model may require updating to deal with this change in representations.

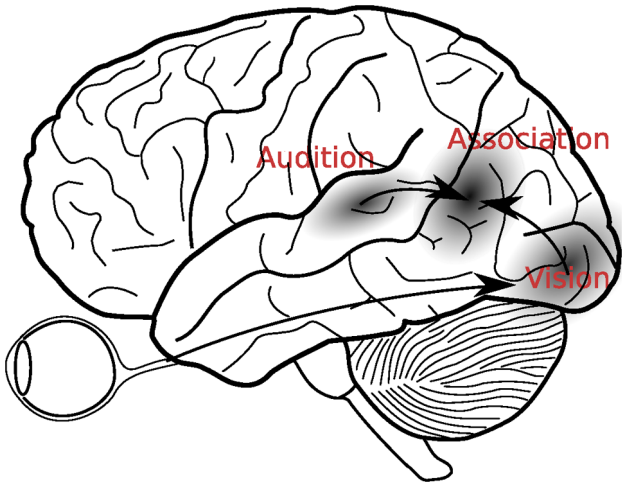


FIGURE 1 | A classical view of multisensory integration. According to this view, visual and auditory signals were first analyzed in the respective sensory cortices, before they were integrated in the secondary association areas, located in the temporo-parietal areas between the auditory and visual cortices.

Predictive Coding: Top–Down vs. Bottom–Up Processing

The aforementioned mismatch is a typical example of bottom–up processing, in which sensory input adjusts the internal model. Conversely, our internal model may also affect the processing of our sensory information. This type of processing is known as top–down and is closely related to selective attention. For example, when we are in a complex environment with many stimuli

[1]Given that the predictive coding framework essentially describes a hierarchy of processing levels, it does not identify specifically which brain areas are involved in creating these internal models. In general, the higher the level of processing, the more distributed the network of brain areas involved can become. For instance, Rao and Ballard (1999) presented a predictive coding model of receptive field effects in striate cortex and stated that, in their simulations, the "internal model is encoded in a distributed manner within the synapses of model neurons at each level." Although it is not yet understood how stochastic models of the environment are coded in the brain, seminal work by Tolman (1948) has shown that cognitive maps can be formed on the basis of relations among salient cues and that these representations support navigation and inferential expression. Subsequently, these maps have been linked to hippocampal functions (see e.g., Morris et al., 1982). The hippocampal and prefrontal cortex interact with each other in decision making (Yu and Frank, 2015). The prefrontal cortex is thought to contain a hierarchical structure of mental representations (Badre, 2008), which in turn are connected to the parietal cortex. This network bears some resemblance to the network of brain areas involved in selective attention (Corbetta and Shulman, 2002) and could jointly support the formation of higher-order stochastic models. This conjecture is speculative, however, and further research in this area is needed.

competing for processing capacity, the most relevant ones need to be prioritized. This is possibly accomplished because the higher-order brain areas that are part of a fronto-parietal network can selectively bias the processing in the lower-order perceptual ones (Corbetta and Shulman, 2002). In other words, attending to task-relevant stimuli might be necessary for them to gain a stronger representation in our neural system.

Viewed within context of the predictive coding framework, the internal representation of our external environment is constantly updated on the basis of sensory input (i.e., forward connections). Sensory processing is, in turn, modulated on the basis of predictions provided by the active representations (i.e., backward connections). It can thus be argued that backward connections from the higher-order to the lower-order brain areas might embody the causal structure of the external world while forward connections only provide feedback about prediction errors to higher areas. In other words, anatomical forward connections are functional feedback connections, and vice versa (Friston, 2005). Mismatches, or –more formally, *prediction errors*– will thus result in strong adjustments in the internal representation and in strong top–down functional feedforward (or anatomical feedback) signals. One possible consequence of such a major prediction error is that the focus of attention shifts to a different aspect of the environment. Seen this way, attention could be considered as a form of predictive coding; a process that establishes an expectation of the moments in time when the relevant, to be integrated stimulus inputs are to arrive (Klemen and Chambers, 2011). It should be noted, however, that while attention may boost the precision of the predicted sensory input (and thus contributes to determining which aspects of our mental representation needs to be updated) the manifestations of attention and expectation can be radically opposite. Whereas expectancy reduces the sensory neural responses, attention enhances them, presumably due to heightening the weighting of the prediction error (Kok et al., 2011, 2012).

The closely related model of optimal Bayesian integration (Anastasio et al., 2000; Deneve and Pouget, 2004), has already been applied to a host of problems related to visual attention, in which attention is considered to provide the lower level visual cortices priors that serve to reduce stimulus ambiguity and therefore enhance the visual search process (Chikkerur et al., 2010). Additionally, it has been successfully applied to explain a number of basic multisensory processing phenomena (Ernst and Banks, 2002; Alais and Burr, 2004; Helbig and Ernst, 2007; Holmes, 2009). Despite these applications, the success with which Bayesian interference models can describe the interaction between attention and multisensory integration remains yet to be answered.

Bottom–Up Principles of Multisensory Integration

Ongoing research that has investigated parts of this interaction may give us at least a partial answer, however. The information contained in the flow of input from the individual senses to the higher-order brain areas can, at least under certain circumstances, determine whether the stimuli contained in these streams are integrated or not. A large number of principles have been uncovered that explain under which conditions inputs to different modalities interact or not (Stein and Meredith, 1993; Noesselt et al., 2007; Stein and Stanford, 2008; Stein and Rowland, 2011). These principles were originally strongly related to the stimulus input characteristics, as well as the individual stimulus processing capabilities of each sensory modality (Stein and Meredith, 1993). For instance, input into the visual modality may influence spatial processing in the auditory modality, while input to the auditory modality may affect temporal processing in the visual modality; two characteristics that have been detailed in the modality appropriate hypothesis (Welch and Warren, 1980; Vatakis and Spence, 2007). Moreover, spatial and temporal proximity (Lewald and Guski, 2003), or the relative intensity (known as the "law of inverse effectiveness"; see Holmes, 2009) of multisensory inputs may be critical factors in determining whether two inputs integrate.

The near-simultaneous stimulation of two or more senses has also been shown to result in increased fMRI BOLD responses (Calvert et al., 2000; Fairhall and Macaluso, 2009) in several brain areas, including the superio-temporal sulcus, superior colliculus, and primary visual cortices. Moreover, increased early latency (∼40 ms after stimulus) event-related potential (ERP) responses to these stimuli (Giard and Péronnet, 1999; Molholm et al., 2002), better performance on stimulus identification tasks (Stein et al., 1996), and visual search benefits (Van der Burg et al., 2008, 2011; Staufenbiel et al., 2011; van den Brink et al., 2014) have been observed. Butler et al. (2012), used EEG recordings and a frequency mismatch paradigm to show that auditory and somatosensory cues elicit a multisensory mismatch, which indeed suggests that these cues can be combined pre-attentatively. Similarly Yau et al. (2009) showed that vibrating somatosensory stimuli for could affect frequency discrimination of auditory stimuli and vice versa (see also; Yau et al., 2010; Butler et al., 2011). Interestingly, however, amplitude judgments were not affected in this fashion. These results show that several stimulus features can automatically influence the processing of stimuli presented in another modality.

Bottom–Up Integration Can Drive Attention

Several behavioral and ERP studies have shown that an object that is simultaneously detected by several sensory systems has a greater potential for capturing one's attention (Van der Burg et al., 2008, 2011; Ngo and Spence, 2010; van den Brink et al., 2014). For instance, Van der Burg et al. (2008), showed that equally task irrelevant auditory stimuli could have strong beneficial impact on participants' ability to detect visual target stimuli. In this study, a visual search task was used to show that a visual target stimulus that was not very salient by itself could become instantly noticeable when it was accompanied by a short tone. This result, labeled the "Pip and Pop" effect, suggests that multisensory stimuli are indeed able to capture attention and therefore that multisensory integration processes themselves operate pre-attentively. In a subsequent ERP study (Van der Burg et al., 2011) we showed that an early latency ERP effect, occurring around 40 ms over posterior scalp areas, correlated significantly with sound-induced performance benefits in this task. Moreover, we also found evidence that the sound

integrated with the visual stimulus in a strongly automatic fashion: whenever a sound was presented, we observed an N2pc component in the ERP waveform, regardless of whether the corresponding visual stimulus was task-relevant or not. Since the N2pc is generally considered to be a neural correlate of automatic bottom–up attentional deployment (Luck, 1994), this can be taken as evidence for automatic integration of the sound with a corresponding visual stimulus. This further suggests that when a sensory modality is processing a stimulus simultaneously with one presented to another modality, these concurrently presented stimuli have a natural tendency to be processed in greater depth than stimuli that are either non-concurrent in time. Thus, these results not only suggest that bottom–up processes have a major influence on multisensory processing, they also show the involvement of early latency unisensory processes in multisensory integration.

Top–Down Influences on Multisensory Processing

Interestingly, multisensory integration is not strictly guided by these principles. For instance Wallace et al. (2004) and Körding et al. (2007) have shown that even when low-level features of multisensory stimuli are perfectly matched, behavioral performance is impaired when these features are perceived to originate from two separate sources. Moreover, using vision and touch, Ernst (2007) trained participants to integrate arbitrary combinations of inputs and found that after training discrimination thresholds had increased for incongruent haptic/visual stimulus combinations. Likewise, Fiebelkorn et al. (2011) modulated the probability of co-occurrence of visual and auditory inputs. They found that hit rates for near-threshold visual inputs depended not only on the mere presence of the auditory input, but also on the participants' expectation: hit-rates for simultaneously presented visual stimuli increased specifically when participants expected the auditory and visual inputs to be simultaneous. Additionally, it has been shown that a stimulus presented in one modality can affect the processing of an accessory stimulus presented in another modality, either due to its task relevance (Busse et al., 2005; Donohue et al., 2014), or because of learned associations between the individual inputs (Molholm et al., 2007; Fiebelkorn et al., 2010). For instance, Busse et al. (2005) paired task-irrelevant auditory stimuli with either attended or unattended visual stimuli and found that processing of the tones that were paired with attended visual stimuli started to differ from that of the tones paired with unattended visual stimuli, around 200 ms after stimulus onset, as expressed by a difference in the ERP waveforms, suggesting that attentional processes in the visual modality strongly affected the processing of irrelevant stimuli in the auditory modality. This difference, which subsequently has been interpreted as a multisensory late processing negativity (Talsma et al., 2007), was found to originate in the primary auditory cortex, as shown using fMRI (Busse et al., 2005; Experiment 2). Thus, these results show that top–down factors can strongly influence multisensory processing.

Following the notion that top–down processing influences multisensory perception, a profound number of recent studies has shifted focus from uncovering the aforementioned basic principles of multisensory integration, to investigating how these principles interact with other cognitive processes. For example, the principle that stimuli are more likely to be integrated when they overlap in space has been found to be more task-dependent than originally assumed (Cappe et al., 2012) and the necessity for temporal correspondence has also been found to depend on tasks and stimulus type (van Atteveldt et al., 2007; Stevenson and Wallace, 2013). These recent findings are somewhat reminiscent of earlier work by Lewald and Guski (2003) who have shown that ones' belief that two stimuli have a common cause might affect whether we perceive cross modal inputs as being one integrated stimulus, or as multiple ones (see also Welch, 1999 for a similar suggestion). Additionally, processes such as attention (Tiippana et al., 2004; Alsius et al., 2005, 2007, 2014; Senkowski et al., 2005, 2007; Talsma and Woldorff, 2005; Talsma et al., 2007; Mozolic et al., 2008; Hugenschmidt et al., 2009) or memory (Thelen et al., 2012) have been shown to affect multisensory processing.

The Multifaceted Interplay between Attention and Multisensory Processing

Thus, it appears that the automaticity of multisensory integration depends on a variety of factors: If the individual stimuli in this bottom–up stream are in themselves salient enough, they can be integrated; specifically when they are approximately matched in time and location with a stimulus processed in another modality. If they are not salient enough, additional prioritizing by an attentional mechanism may be needed (Talsma et al., 2010), suggesting that multisensory integration involves both top–down and bottom–up processes.

If multisensory integration is the result of a complex interaction between top–down and bottom–up processes, then it should take place at multiple stages of processing. So, are we able to identify these stages? Several human electrophysiology studies have shown that *multisensory interactions* can occur at latencies that would exclude the possibility that multisensory processing only occurs after initial sensory analysis has come to completion (Giard and Péronnet, 1999; Molholm et al., 2002; Van der Burg et al., 2011). These interactions indicate that information from one sensory modality can influence the information processing in another one, without necessarily forming a new mental representation. The aforementioned studies thus indicated that although the primary and secondary (uni-) sensory brain areas are possibly involved in multisensory processing, these multisensory processes do not necessarily result in a newly integrated representation. These studies do suggests, however, that multisensory processing is intertwined with basic sensory analysis in a much more intimate fashion than previously thought possible.

To summarize, the findings discussed above show that multisensory integration depends to a much smaller degree on rigid bottom–up principles than originally believed to be the case. By contrast, they show that multisensory integration is by a very large factor determined by top–down processes. The next question now is, how these top–down and bottom–up processes interact, and at which processing stages this occurs.

Early and Late Accounts of Multisensory Processing

As has become clear by now, since the beginning of the 1980s, our understanding of multisensory processing has shifted from relatively rigid and principle-based mechanisms, located late in the processing stream, to a highly flexible process consisting of multiple stages. At least a two sub-processes, one of which can occur very early on in the processing stream, have been identified (Calvert and Thesen, 2004; Talsma et al., 2007; Koelewijn et al., 2010; Baart et al., 2014). In spite of this change in interpretation, there are still a number of arguments to not completely discard the original idea that multisensory integration (partially) takes place after the initial sensory processing has come to completion. It has recently been proposed that the integration of neural representations is an intrinsic property of the brain (Ghazanfar and Schroeder, 2006; van Atteveldt et al., 2014). From this idea it follows that different levels of neural interactions may take place at progressive levels of processing of the sensory inputs.

Multisensory Integration: An Intrinsic Property of the Brain?

van Atteveldt et al. (2014), have suggested that multisensory integration is a process that operates on the basis of the flexible recruitment of several general purpose brain functions that are thought to synchronize activation within several neural pathways. These pathways are thought to connect the sensory cortices, either directly to each other (Falchier et al., 2002), or through cortico-thalamic-cortical pathways (Hackett et al., 2007; Lakatos et al., 2007; van den Brink et al., 2014), suggesting that information can be transferred relatively directly between these brain areas. Another set of these pathways involves recurrent feedback projections from the frontal cortex (notably the frontal eye fields and ventral prefrontal cortex). It is assumed that these feedback mechanisms coordinate activation in the sensory cortices through attention. The general idea is that these recurrent feedback projections can send biasing signals to the perceptual brain areas. The feedback signals can then induce an increase in sensitivity in neurons responsive to the attended feature, while simultaneously causing a decrease in sensitivity of neurons not responsive to the attended feature (Motter, 1994; LaBerge, 1995; Corbetta and Shulman, 2002). This attentional bias can either be expressed overtly, that is, by actively scanning the environment with the oculomotor system, or covertly by scanning the environment using selective attention only. Given the importance of these latter recurrent feedback connections, attentive scanning of the environment is an essential prerequisite for multisensory integration to take place.

A possible function of the direct and cortico-thalamic connections between visual and auditory cortex is that they enable cross-referencing between these cortices. In other words, auditory cortex receives advance information regarding visual processing and vice versa. Viewed from a predictive coding framework, prediction errors in the auditory representation are minimized by additional information presented by the visual system, and vice versa. In our own framework (Talsma et al., 2010), these low level interactions can for instance result in spatio-temporal realignment of the auditory and visual input signal. Thus, the early latency processes appear to cross-feed low-level information between the individual sensory cortices. This cross-feeding may modify the original input signal and can therefore be described as a multisensory interaction, but not necessarily as multisensory integration. Additional research is still required, however, to determine the exact functional role of these direct connections.

Task and Stimulus Type Dependencies

Task Relevance

To further differentiate between early and late multisensory processes, we need to distinguish between two rather strongly differing sets of research. Studies using relatively simple stimuli, such as beeps, and flashes, have predominantly focused on determining the bottom–up driven effects of multisensory processing that have been discussed in detail above (van Wassenhove et al., 2007; Stein and Stanford, 2008; Holmes, 2009; Noesselt et al., 2010; Rach et al., 2011). Studies using more naturalistic, meaningful stimuli, on the other hand, have more strongly emphasized the influence of top–down processing in multisensory integration. For instance, studies using speech fragments and movie clips have indicated that semantic congruence between visual and auditory stimuli also strongly influences multisensory processing (McGurk and MacDonald, 1976; Calvert et al., 2000; Tuomainen et al., 2005; Cappe et al., 2012). Most notably, the McGurk effect, that is, the illusion that speech sounds are being perceived differently when they are combined with non-matching lip-movements is one of the hallmarks of the effectiveness of multisensory integration. It has long been thought that this illusion is highly automatic, although that notion has been challenged by showing that one's susceptibility to the McGurk illusion falters under high attentional demands (Alsius et al., 2005).

Top–Down Effect in Multisensory Speech Perception

The involvement of semantic congruence in multisensory integration in speech processing presumably indicates that access to semantic information constrains the possible interpretation of the bottom–up auditory and visual input streams in a top–down fashion. In other words, access to prior knowledge may restrain the possible interpretations of both the visual and auditory input streams, which may in turn improve the segmentation of the auditory speech signal. Speech signal segmentation is generally problematic (specifically under noisy conditions; see Ross et al., 2007; Foxe et al., 2015), because there is only a very loose connection between speech sounds and the underlying phoneme structure (Liberman et al., 1967). Thus, constraining possible interpretations of the speech signal through top–down processes may further benefit from limitations imposed by information arriving from other modalities. Because of this, speech perception has been considered to be an intrinsic multisensory phenomenon (Stevenson et al., 2014) and it has even been argued

that audio-visual speech perception is a special form of multisensory processing (Tuomainen et al., 2005; but see Vroomen and Stekelenburg, 2011).

Speech Perception and Prior Experience

Interestingly, Vroomen and Baart (2009) showed that speech processing can be affected by lip-reading, but only when their participants could interpret the auditory stimuli as speech signals. This was done by dubbing sine-wave speech (Remez et al., 1981) onto video recordings of lip-moments. One interesting characteristic of sine-wave speech is that to most naïve listeners it sounds just like random sounds from a science fiction movie. Once participants get into speech mode, that is, once they start recognizing the sounds as speech, they usually never fail to ignore the speech component, much in the way that the realization that the Swedish Chef from our example speaks garbled English greatly enhances our comprehension of him. This point thus illustrates that prior experience and background knowledge may influence multisensory processing; a topic that will be discussed in more detail below.

Multiple Stages of Multisensory Integration

Following up on this study, Baart et al. (2014), used ERPs to identify two distinct stages of multisensory integration in the processing of sine-wave speech. The auditory N1 component, a negative component about 100 ms after stimulus onset, peaked earlier for audiovisual stimuli than for auditory stimuli alone, regardless of whether participants were in speech mode or not. By contrast, the P2 component, a positive component peaking at roughly 200 ms after stimulus onset, was also modulated the presence of visual information, but only when participants were in speech mode. It should be noted that the latency of these latter ERP findings, while representative for speech stimuli (e.g., van Wassenhove et al., 2005), occurred somewhat later compared to those typically found in studies using simple beeps and flashes (Giard and Péronnet, 1999; Molholm et al., 2002; Senkowski et al., 2005, 2007; Talsma and Woldorff, 2005; Talsma et al., 2007; Van der Burg et al., 2011). It appears that the N1 component reflects a relatively automatic bottom–up process, while the P2 component reflects a process that is also affected by top–down processing. Despite this difference in latency, the notion of a two-stage approach in multisensory processing is compatible with earlier notions showing separate stages of multisensory processing for simple stimuli (Talsma et al., 2007).

Further evidence for the hypothesis that both top–down and bottom–up processing contribute to multisensory speech processing is provided by an fMRI study from Miller and D'Esposito (2005). These authors presented audiovisual speech fragments in which the relative onsets of the auditory speech stimulus were shifted with respect to the onset of the visual stimulus. Synchronous presentation of the auditory and visual speech signals resulted in a significantly larger activity in several brain areas that are involved in multisensory processing. These areas include Heschl's gyrus, the superior temporal sulcus, the middle intraparietal sulcus, and the inferior frontal gyrus. The involvement of these brain areas provides more evidence that multisensory interactions occur at various stages of processing.

Top Down Processing: Exclusively for Speech Stimuli?

The processing of naturalistic audiovisual stimuli involves both top–down and bottom–up processing. This could lead one to conclude that whereas simple stimuli involve mostly bottom–up processes, complex (speech) stimuli involve both top–down and bottom–up processes. Upon closer inspection, however, this is probably overly simplistic. For example, by using a binocular rivalry paradigm that consisted of a visual stimulus containing looming motion presented to one eye, and radial motion to the other, van Ee et al. (2009) demonstrated that participants were able to hold on to one of the two percepts longer by means of attention. Interestingly, this attentional gain for one of the percepts was prolonged when the attended visual stimulus was accompanied by a sound that matched the temporal characteristics of the attended visual stimulus (**Figure 2**). This pattern of results also suggests a complex interaction between attention and multisensory integration. Although the exact neural mechanisms involved in this process are not yet fully understood, it appears that attention boosts the neural response to one of the competing visual signals, and that this boost, in turn, facilitates integration with the matching auditory signal. This finding suggests that rhythmic congruence between visual and auditory stimuli is another critical principle for multisensory processing. Interestingly, van Ee et al. (2009) also demonstrated that the mere presence of such a matched sound was insufficient. Instead, attention to both the visual and auditory modalities was needed to facilitate attentional facilitation of one of the two percepts. This result shows that multisensory interactions can influence visual awareness, but only in interaction with attention, underscoring that attention plays a pivotal role in multisensory processing.

A Multisensory Representation

The evidence discussed so far shows that information from our various senses fuses at several stages of processing. Additionally, several studies show that multisensory speech processing can be affected by prior experience (Vroomen and Baart, 2009; Navarra et al., 2010; Vroomen and Stekelenburg, 2011; Nahorna et al., 2012). The next question is whether the influence of prior experience is limited to specific forms of speech processing or whether it can be generalized across multiple domains of multisensory processing.

Prior Experience and Multisensory Memories

Thelen et al. (2012) have provided evidence for the role of memory in the formation of a multisensory representation. In this study, participants were required to memorize visual line drawings. These drawings were either presented simultaneously with a meaningless sound (multisensory context), or in isolation (unisensory context). One of the key findings of this study was that recognition accuracy was significantly impaired when the pictures had initially been presented in a multisensory context. This result suggests that the multisensory context provided

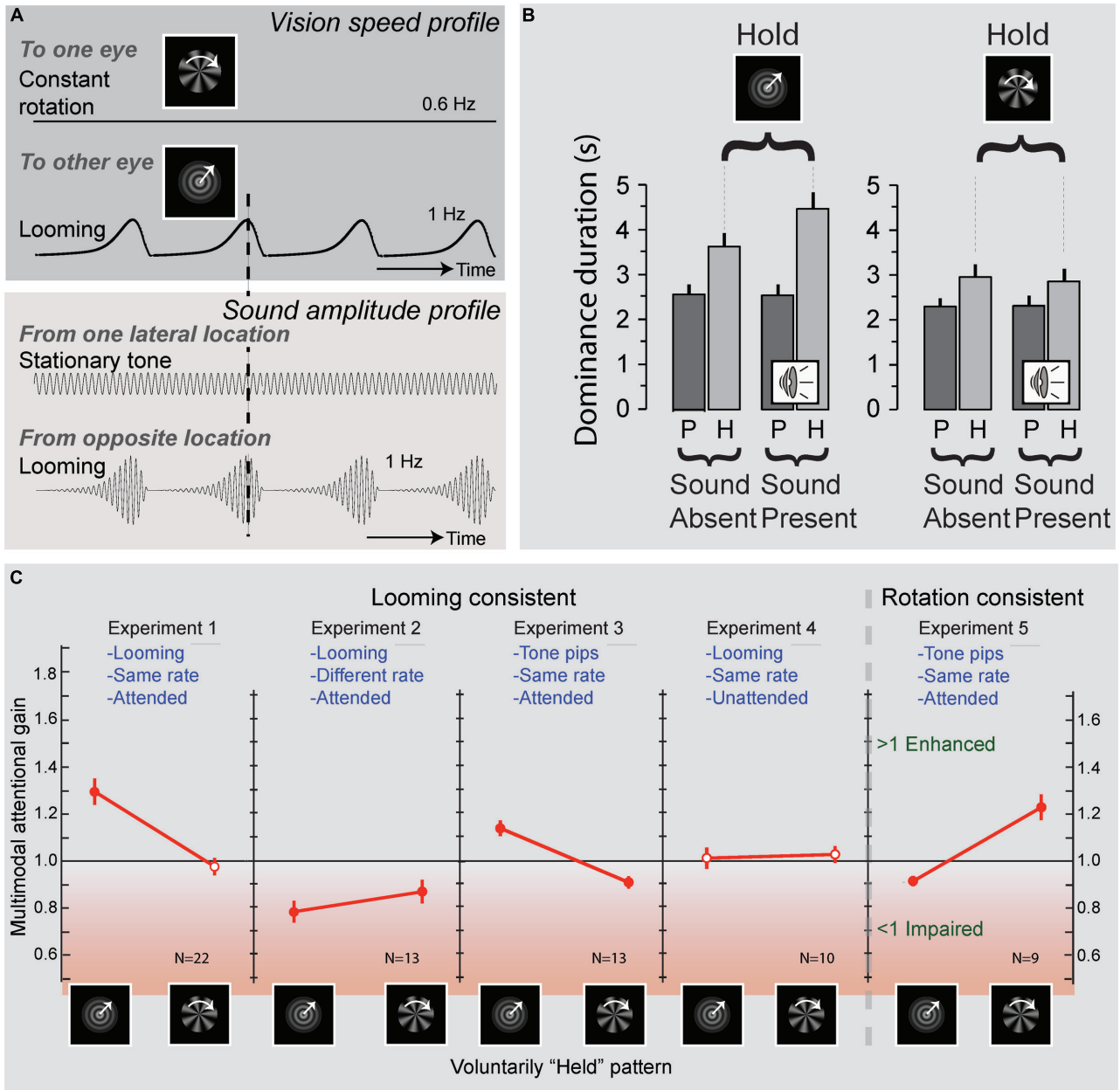


FIGURE 2 | Effects of attention and multisensory integration on conflict resolving in binocular rivalry. (A) Experimental design: an object rotating at a frequency of 0.6 Hz was presented to one eye, while a looming object, expanding at a rate of 1 Hz, was presented to the other eye. Concurrent with the presentation of these visual objects sounds could be presented, consisting of, a stationary "e-chord" sound that was presented to one channel of a headphone, while a looming sound that matched the temporal characteristics of the looming object was presented to the other channel. Participants were required to attend to the looming sound pattern and report when the dominant visual pattern switched from the looming to the rotating image and vice-versa. **(B)** Average durations of the looming (left) and rotating (right) visual patterns being dominant. Duration times were significantly increased when participants were requested to attend and hold on to one of the patterns. Importantly, when the sound pattern was present this effect was enhanced for the (rhythmically congruent) looming visual pattern, but not for the (rhythmically incongruent) rotating visual pattern. These results suggest that attention can affect visual dominance by way of interacting with congruent sound patterns (P, passive viewing; H, hold on to instructed pattern). **(C)** Effects of rhythmic congruency and attention. Experiments 1–4 tested the influence of sounds that were consisted with the looming patterns. Experiments 1 and 3 show an increase in attentional gain (i.e., a prolonging in duration of the held pattern) when a sound was present that was rhythmically congruent with the held pattern. When the sound was rhythmically incongruent (Experiment 2) a decrease in attentional gain was observed, and when the sounds were unattended (Experiment 4) no significant change in attentional gain could be observed. Experiment 5 generalizes the results to rotating visual patterns. Filled red circles indicate attentional gains that significantly deviated from one. Adapted from van Ee et al. (2009) by permission of the Society for Neuroscience.

in the initial presentation has become part of the mental representation. Thelen et al. (2015), subsequently manipulated the semantic relation between the visual and auditory stimuli, and also investigated the effects of multisensory memory formation on auditory processing. We found that auditory object discrimination was enhanced when initial presentations entailed semantically congruent multisensory pairs, and impaired when they entailed incongruent pairs, compared to sounds that had been encountered only in a unisensory manner. This result shows that the subsequent processing of a sensory trace is greatly affected by the initial context in which it was encoded. More specifically, a congruent pair of audiovisual stimuli may facilitate subsequent recall, whereas incongruent, or unrelated auditory and visual stimulus pairings may actually impair such recall. Thus, an internally consistent multisensory stimulus may be remembered more effectively than one that is internally inconsistent (see also Molholm et al., 2007; Fiebelkorn et al., 2010 for evidence of a similar role of learned multisensory associations in attention orienting and; Quak et al., in revision for a literature review on the relation between working memory and multsensory processing).

From a predictive coding perspective, semantically congruent audiovisual stimuli will result in higher-order brain areas receiving consistent information, which will result in a strong and consistent internal model, and a low prediction error. If the information presented is incongruent across modalities, this may result in an inconsistency in the internal model, which in turn may result in a greater error signal being sent back to the sensory cortex, which in turn may result in more effort being invested in encoding the representation, combined with a weaker internal representation.

Following the logic laid out by the predictive coding framework, one would also expect that stimulating only one sensory modality at the time will result in activation of another sensory modality. If, for example, we only present a picture of the Swedish Chef, our background knowledge may strongly affect the internal representation that we derive from this picture. Because his gibberish Swedish is such a characteristic feature, merely presenting an image of the muppet may not only activate our visual representation, it may also activate all concepts related to the Swedish Chef that we have gained through prior experience, including his characteristic manner of speaking. Depending on how strong these associations are, the image may not only trigger one's knowledge of the Chef's manner of speaking, it may even actively trigger recall of specific fragments, such as those given in the examples at the beginning of this article. Likewise, the presentation of a speech fragment may trigger similarly vivid mental images of the Chef's characteristic manners of behavior. In terms of predictive coding, the internal representation would be activated because the sensory input signal matches with an existing representation stored in long-term memory. The resulting internal modal then not only projects feedback information to the original sensory cortex that activated the representation, but also to the other sensory cortices. If this assumption is correct, then we might expect that stimulating one sensory cortex, such as the visual one, might also result in activation of other sensory cortices, such as the auditory one. Next we turn to a number of studies that have provided evidence for this.

Top–Down Induced Sensory Cortex Activation via Mental Imagery

Evidence for the idea that visual cortex might be activated indirectly by auditory stimuli comes from at least two different recent studies. Mercier et al. (2013) used intracranial recordings to show that auditory phase reset, and auditory evoked potentials can be recorded in the visual cortex. This study thus illustrates cross-sensory phase reset by a non-primary stimulus in "unisensory" cortex. Vetter et al. (2014) used fMRI to show that visual cortex was activated either by auditory stimuli, or by imagined stimuli. According to the Vetter et al. (2014). study, sound is initially processed through the classical auditory pathways. The resulting representation causes object-specific neural activation patterns that are subsequently projected back to visual cortex. Interestingly, sounds belonging to two different categories were correctly classified on the basis of the activation pattern observed in visual areas, regardless of whether this pattern was induced by a physical sound or by a mental imagery instruction, suggesting that higher-order cortical networks mediated the visual cortex activation. Vetter et al. (2014) further compared the sound-induced activation patterns with the imagery-induced activation in auditory cortex, but no similarities were found here. Moreover, they compared activation across different exemplars and different classes of stimuli and found that the information projected back appears to convey higher-level semantic information, as was shown using a multivariate pattern analysis. Finally, they found similar activation in multisensory convergence areas, including the precuneus and superior temporal sulcus, suggesting that the visual cortex activation was mainly induced by way of feedback from those multisensory areas.

These results are compatible with recent studies by Berger and Ehrsson (2013). These authors replicated two classical multisensory integration effects, but instead of presenting actual auditory stimuli, they instructed their participants to imagine these stimuli. The cross-bounce illusion consists of two circles moving on a collision course (Sekuler et al., 1997). Phenomenally, the two circles can either be perceived as crossing each other, after making contact, or as bouncing off of each other. When a sound is presented at the moment of contact, participants tend to report more often that two stimuli bounce instead of continuing on their original course. Interestingly, Berger and Ehrsson (2013) showed that these effects could occur not only using actual auditory stimuli but also using imagined ones, a result that is consistent with the Vetter et al. (2014) conclusion that auditory imagery can affect visual processing. Similar results were also found for an imagined version of the McGurk illusion.

In a second experiment, Berger and Ehrsson (2013) also showed that visual imagery can affect auditory processing. This was done using an imagery version of the ventriloquist effect. The ventriloquist effect describes how a sound can be mislocalized because it coincides with a visual stimulus that is presented at a different location. In the imagery version of the ventriloquist effect, participants imagined the presence of a circle at a specific location. Participants' estimates of the location of sounds presented at nearby locations were systematically biased toward the location of the imagined stimulus. In a subsequent fMRI study

(Berger and Ehrsson, 2014), it was found that simultaneous visual imagery and auditory stimulation resulted in an illusory translocation of auditory stimuli that was associated with activity in the left superior temporal sulcus, a key site for the integration of real audiovisual stimuli (Driver and Noesselt, 2008; Ghazanfar et al., 2008; Dahl et al., 2009; Beauchamp et al., 2010; Nath and Beauchamp, 2012). These findings show that processing in brain areas that we considered until recently to be unisensory can be influenced by a variety of sources. This malleability of the sensory cortices can possibly also explain why enhanced peripheral visual processing can be observed in congenitally deaf participants (Scott et al., 2014). Scott et al. (2014) used fMRI to find that congenitally deaf patients show better peripheral vision, a change that is presumably supported by a reorganized auditory cortex. More specifically they found that this increase in peripheral vision related to changes in sensitivity in Herchl's Gyrus, as well as several other visual and multisensory areas, including the posterior parietal cortex, frontal eye fields, anterior cingulated cortex, and the supplementary eye fields. In addition to the already established direct links between the sensory cortices (Falchier et al., 2002; van den Brink et al., 2014), the results discussed in the previous section show that one sensory cortex can also activate another one via a slower route involving higher cortical areas that provide feedback at a more abstract level.

An Attentional Account of the Multisensory Mind

We have seen that the cortical areas that until recently were characterized as "unisensory" are far more interconnected than previously thought possible. The final question is how the interplay between all these connections can result in multisensory integration.

According to the predictive coding framework, mental representations of our external environment are actively constructed by our higher-order brain processes (Friston, 2010), on the basis of sensory input and our existing background knowledge (**Figure 3**). Moreover, these mental representations serve to form predictions about future changes in the external environment so that sensory processing is optimized to predominantly deal with unexpected changes (Baess et al., 2011). Given that backward connections might embody the causal structure of the external world while forward connections only provide feedback about prediction errors to higher areas, it can be argued that both types of connections are needed for integration. The higher-order brain areas containing the conceptual representation provide functional feedforward information to the sensory cortices. Viewed this way, multisensory integration actually takes place because an attentional mechanism combines the information contained in the existing mental representation with general background knowledge and uses the resulting model to update sensory processing, much in the way that attention has been proposed to bind together several stimulus feature within the visual modality (Treisman and Gelade, 1980). Viewed this way, it can be tentatively stated that multisensory integration is largely accounted for by attentional mechanisms.

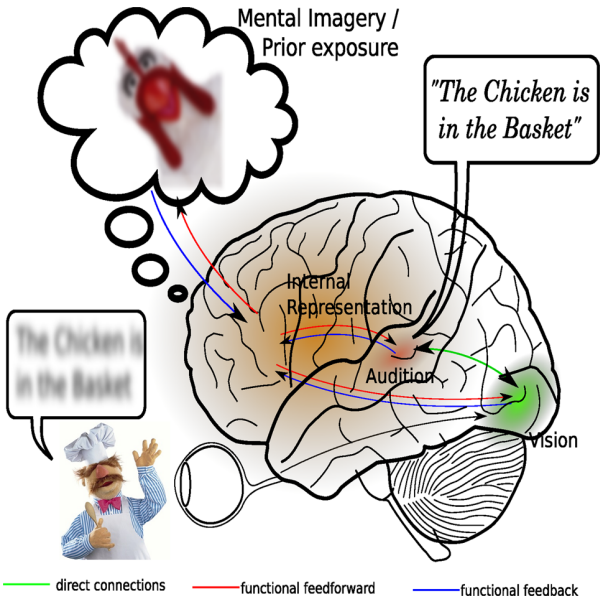


FIGURE 3 | An attentional account of multisensory integration. Central in this revised view of multisensory integration is the presence of a dynamic mental representation, which is updated on the basis of sensory inputs as well as on the basis of representations stored in memory. Shown here is an example of how inaudible speech may benefit from both direct visual stimulation, as well as from the context provided by prior exposure to a similar situation. Processing in the visual and auditory sensory cortices is depending on expectancies generated by the internal models. A mismatch between expected input and actual input, formally known as a prediction error, may result in enhanced activation in the sensory cortices. Multisensory integration here is considered to consist of synchronization of activity in the auditory and visual cortices. This integration is facilitated by direct and thalamo-cortical connections between the auditory and visual cortical areas.

Although this view actually places multisensory integration again at the end of the processing stream from an anatomical perspective, it does not exclude the possibility of multisensory interactions occurring early on in the processing stream. A mismatch between the senses can, depending on the complexity of the input stream, be resolved in the relatively early stages of processing, presumably involving the aforementioned direct or cortico-thalamo-cortical pathways. Presumably, the processes involved here are mainly bottom–up. Although these early sensory interactions may take place at early stages, it should be noted that, following the logic of the predictive coding framework, the individual sensory representations serve nothing more than to update the internal model represented in the higher-order brain areas. It is plausible that the individual senses interact before they update our mental representation, because reducing ambiguities in the individual input channels will inevitably result in a reduction of prediction errors that need to be fed-back to the sensory cortices. Although multisensory interactions can thus take place early on in the processing stream, their presence does not necessarily indicate multisensory integration.

Summary and Conclusion

Despite the initial idea that sensory information integrates in higher-order association areas of the neocortex, substantial amounts of evidence now point toward a much more diffuse process, in which multisensory operations can take place at various stages of processing. Moreover, multisensory processes can be affected by a host of other cognitive processes, including attention, memory, and prior experience.

More generally, this literature review has shown that multisensory processing and attention are strongly related to each other. This brings us to the question whether the role of attention in multisensory integration is a matter of bottom–up or top–down processing. Though speculative, I would argue that while multisensory processing in general involves both bottom–up and top–down processes, the more specific case of multisensory integration is largely subserved by top–down processes: From a predictive coding point of view, it can be argued that integration takes place because higher-order networks actively maintain a mental model of the environment, which generates predictions about the expected sensory input. Sensory processing itself is adjusted on the basis of the (dis)agreement, between the actual sensory activity and the activity predicted by the model. Moreover, this prediction error may also require an update of the model itself. According to this view, it is in this mental model where sight, sound, smell, taste, and touch is integrated with our existing cognitive schemata. Several lines of evidence support this idea; early bottom–up driven processing in one modality can subsequently modify the internal representation of a stimulus in another sensory modality (Busse et al., 2005; Van der Burg et al., 2008, 2011), suggesting that functional feedback from the sensory system results in a change in prediction of another sensory modality. Additional influences from prior experience (Vroomen and Baart, 2009; Thelen et al., 2012, 2015, or mental imagery also actively affect multisensory processing (Berger and Ehrsson, 2013, 2014). Moreover, evidence exists to show that such imagery can, just like actual sensory input, activate processes in another modality (Vetter et al., 2014). Because the processes that are involved in integrating the inputs from such a wide variety of sources are essentially top–down and bearing a strong resemblance to attentional control mechanisms (van Atteveldt et al., 2014), it can be argued that attention plays an essential role in integrating information. Seen this way, attention counts as an essential cognitive faculty in integrating information in the multisensory mind.

Acknowledgments

I wish to thank Elger Abrahamse, Raquel London, Michel Quak, and two reviewers for helpful discussions and comments on earlier drafts of this paper.

References

Alais, D., and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Curr. Biol.* 14, 257–262. doi: 10.1016/j.cub.2004.01.029

Alsius, A., Mottonen, R., Sams, M. E., Soto-Faraco, S., and Tiippana, K. (2014). Effect of attentional load on audiovisual speech perception: evidence from ERPs. *Front. Psychol.* 5:727. doi: 10.3389/fpsyg.2014.00727

Alsius, A., Navarra, J., Campbell, R., and Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Curr. Biol.* 15, 839–843. doi: 10.1016/j.cub.2005.03.046

Alsius, A., Navarra, J., and Soto-Faraco, S. (2007). Attention to touch weakens audiovisual speech integration. *Exp. Brain Res.* 183, 399–404. doi: 10.1007/s00221-007-1110-1

Anastasio, T. J., Patton, P. E., and Belkacem-Boussaid, K. (2000). Using Bayes' Rule to model multisensory enhancement in the superior colliculus. *Neural Comput.* 12, 1165–1187. doi: 10.1162/089976600300015547

Baart, M., Stekelenburg, J., and Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia* 65, 115–121. doi: 10.1016/j.neuropsychologia.2013.11.011i

Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends Cogn. Sci.* 12, 193–200. doi: 10.1016/j.tics.2008.02.004

Baess, P., Horváth, J., Jacobsen, T., and Schröger, E. (2011). Selective suppression of self-initiated sounds in an auditory stream: an ERP study. *Psychophysiology* 48, 1276–1283. doi: 10.1111/j.1469-8986.2011.01196.x

Beauchamp, M. S., Nath, A. R., and Pasalar, S. (2010). fMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the mcGurk effect. *J. Neurosci.* 30, 2414–2417. doi: 10.1523/jneurosci.4865-09.2010

Berger, C. C., and Ehrsson, H. H. (2013). Mental imagery changes multisensory perception. *Curr. Biol.* 23, 1367–1372. doi: 10.1016/j.cub.2013.06.012

Berger, C. C., and Ehrsson, H. H. (2014). The fusion of mental imagery and sensation in the temporal association cortex. *J. Neurosci.* 34, 13684–13692. doi: 10.1523/jneurosci.0943-14.2014

Bloom, F. E., and Lazeron, A. (1988). *Brain, Mind, and Behavior*. New York, NY: W. H. Freeman and Company.

Busse, L., Roberts, K. C., Crist, R. E., Weissman, D. H., and Woldorff, M. G. (2005). The spread of attention across modalities and space in a multisensory object. *Proc. Natl. Acad. Sci. U.S.A.* 102, 18751–18756. doi: 10.1073/pnas.0507704102

Butler, J. S., Foxe, J. J., Fiebelkorn, I. C., Mercier, M. R., and Molholm, S. (2012). Multisensory representation of frequency across audition and touch: high density electrical mapping reveals early sensory-perceptual coupling. *J. Neurosci.* 32, 15338–15344. doi: 10.1523/JNEUROSCI.1796-12.2012

Butler, J. S., Molholm, S., Fiebelkorn, I. C., Mercier, M. R., Schwartz, T. H., and Foxe, J. J. (2011). Common or redundant neural circuits for duration processing across audition and touch. *J. Neurosci.* 31, 3400–3406. doi: 10.1523/JNEUROSCI.3296-10.2011

Calvert, G. A., Campbell, R., and Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Curr. Biol.* 10, 649–657. doi: 10.1016/S0960-9822(00)00513-3

Calvert, G. A., and Thesen, T. (2004). Multisensory integration: methodological approaches and emerging principles in the human brain. *J. Physiol. Paris* 98, 191–205. doi: 10.1016/j.jphysparis.2004.03.018

Cappe, C., Thelen, A., Romei, V., Thut, G., and Murray, M. M. (2012). Looming signals reveal synergistic principles of multisensory integration. *J. Neurosci.* 32, 1171–1182. doi: 10.1523/jneurosci.5517-11.2012

Chikkerur, S., Serre, T., Tan, C., and Poggio, T. (2010). What and where: a Bayesian inference theory of attention. *Vision Res.* 50, 2233–2247. doi: 10.1016/j.visres.2010.05.013

Corbetta, M., and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* 3, 201–215. doi: 10.1038/nrn755

Dahl, C. D., Logothetis, N. K., and Kayser, C. (2009). Spatial organization of multisensory responses in temporal association cortex. *J. Neurosci.* 29, 11924–11932. doi: 10.1523/JNEUROSCI.3437-09.2009

Deneve, S., and Pouget, A. (2004). Bayesian multisensory integration and cross-modal spatial links. *J. Physiol. Paris* 98, 249–258. doi: 10.1016/j.jphysparis.2004.03.011

Donohue, S. E., Todisco, A. E., and Woldorff, M. G. (2014). The rapid distraction of attentional resources toward the Source of incongruent stimulus input during multisensory conflict. *J. Cogn. Neurosci.* 25, 623–635. doi: 10.1162/jocn_a_00336

Driver, J., and Noesselt, T. (2008). Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments. *Neuron* 57, 11–23. doi: 10.1016/j.neuron.2007.12.013

Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *J. Vis.* 7, 1–14. doi: 10.1167/7.5.7

Ernst, M. O., and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415, 429–433. doi: 10.1038/415429a

Fairhall, S. L., and Macaluso, E. (2009). Spatial attention can modulate audiovisual integration at multiple cortical and subcortical sites. *Eur. J. Neurosci.* 29, 1247–1257. doi: 10.1111/j.1460-9568.2009.06688.x

Falchier, A., Clavagnier, S., Barone, P., and Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. *J. Neurosci.* 22, 5749–5759.

Fiebelkorn, I. C., Foxe, J. J., Butler, J. S., Mercier, M. R., Snyder, A. C., and Molholm, S. (2011). Ready, set, reset: stimulus-locked periodicity in behavioral performance demonstrates the consequences of cross-sensory phase reset. *J. Neurosci.* 31, 9971–9981. doi: 10.1523/JNEUROSCI.1338-11.2011

Fiebelkorn, I. C., Foxe, J. J., and Molholm, S. (2010). Dual mechanisms for the cross-sensory spread of attention: how much do learned associations matter? *Cereb. Cortex* 20, 109–120. doi: 10.1093/cercor/bhp083

Foxe, J. J., Molholm, S., Del Bene, V. A., Frey, H. P., Russo, N. N., Blanco, D., et al. (2015). Severe multisensory speech integration deficits in high-functioning school-aged children with autism spectrum disorder (ASD) and their resolution during early adolescence. *Cereb. Cortex* 25, 298–312. doi: 10.1093/cercor/bht213

Friston, K. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. B Biol. Sci.* 360, 815–836. doi: 10.1098/rstb.2005.1622

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Ghazanfar, A. A., Chandrasekaran, C., and Logothetis, N. K. (2008). Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integration in rhesus monkeys. *J. Neurosci.* 28, 4457–4469. doi: 10.1523/JNEUROSCI.0541-08.2008

Ghazanfar, A. A., and Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends Cogn. Sci.* 10, 278–285. doi: 10.1016/j.tics.2006.04.008

Giard, M. H., and Péronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. *J. Cogn. Neurosci.* 11, 473–490. doi: 10.1162/089892999563544

Hackett, T. A., De La Mothe, L. A., Ulbert, I., Karmos, G., Smiley, J., and Schroeder, C. E. (2007). Multisensory convergence in auditory cortex, II. Thalamocortical connections of the caudal superior temporal plane. *J. Comp. Neurol.* 502, 924–952. doi: 10.1002/cne.21326

Helbig, H. B., and Ernst, M. O. (2007). Optimal integration of shape information from vision and touch. *Exp. Brain Res.* 179, 595–606. doi: 10.1007/s00221-006-0814-y

Holmes, N. P. (2009). The principle of inverse effectiveness in multisensory integration: some statistical considerations. *Brain Topogr.* 21, 168–176. doi: 10.1007/s10548-009-0097-2

Hugenschmidt, C. E., Mozolic, J. L., and Laurienti, P. J. (2009). Suppression of multisensory integration by modality-specific attention in aging. *Neuroreport* 20, 349–353. doi: 10.1097/WNR.0b013e328323ab07

Klemen, J., and Chambers, C. D. (2011). Current perspectives and methods in studying neural mechanisms of multisensory interactions. *Neurosci. Biobehav. Rev.* 36, 111–133. doi: 10.1016/j.neubiorev.2011.04.015

Koelewijn, T., Bronkhorst, A., and Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: a review of audiovisual studies. *Acta Psychol.* 134, 372–384. doi: 10.1016/j.actpsy.2010.03.010

Kok, P., Jehee, J. F. M., and de Lange, F. P. (2012). Less is more: expectation sharpens representations in the primary visual cortex. *Neuron* 75, 265–270. doi: 10.1016/j.neuron.2012.04.034

Kok, P., Rahnev, D., Jehee, J. F. M., Lau, H. C., and de Lange, F. P. (2011). Attention Reverses the effect of prediction in silencing sensory signals. *Cereb. Cortex* 22, 2197–2206. doi: 10.1093/cercor/bhr310

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., and Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE* 2:e943. doi: 10.1371/journal.pone.0000943

LaBerge, D. (1995). *Attentionl Processing: The Brain's Art of Mindfulness*. Cambridge, MA: Harvard University Press. doi: 10.4159/harvard.9780674183940

Lakatos, P., Chen, C., O'Connell, M. N., Mills, A., and Schroeder, C. E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron* 53, 279–292. doi: 10.1016/j.neuron.2006.12.011

Lewald, J., and Guski, R. (2003). Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli. *Cogn. Brain Res.* 16, 468–478. doi: 10.1016/s0926-6410(03)00074-0

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychol. Rev.* 74, 431–461. doi: 10.1037/h0020279

Luck, S. J. (1994). Cognitive and neural mechanisms of visual search. *Curr. Opin. Neurobiol.* 4, 183–188. doi: 10.1016/0959-4388(94)90070-1

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Mercier, M. R., Foxe, J. J., Fiebelkorn, I. C., Butler, J. S., Schwartz, T. H., and Molholm, S. (2013). Auditory-driven phase reset in visual cortex: human electrocorticography reveals mechanisms of early multisensory integration. *Neuroimage* 79, 19–29. doi: 10.1016/j.neuroimage.2013.04.060

Miller, L. M., and D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *J. Neurosci.* 25, 5884–5893. doi: 10.1523/jneurosci.0896-05.2005

Molholm, S., Martinez, A., Shpaner, M., and Foxe, J. J. (2007). Object-based attention is multisensory: co-activation of an object's representations in ignored sensory modalities. *Eur. J. Neurosci.* 26, 499–509. doi: 10.1111/j.1460-9568.2007.05668.x

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., and Foxe, J. J. (2002). Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Cogn. Brain Res.* 14, 115–128. doi: 10.1016/S0926-6410(02)00066-6

Morris, R. G. M., Garrud, P., Rawlins, J. N. P., and O'Keefe, J. (1982). Place navigation impaired in rats with hippocampal lesions. *Nature* 297, 681–683. doi: 10.1038/297681a0

Motter, B. C. (1994). Neural correlates of attentive selection for color or luminance in extrastriate area V4. *J. Neurosci.* 14, 2178–2189.

Mozolic, J. L., Hugenschmidt, C. E., Peiffer, A. M., and Laurienti, P. J. (2008). Modality-specific selective attention attenuates multisensory integration. *Exp. Brain Res.* 184, 39–52. doi: 10.1007/s00221-007-1080-3

Nahorna, O., Bethommier, F., and Schwartz, J.-L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *J. Acoust. Soc. Am.* 132, 1061–1077. doi: 10.1121/1.4728187

Nath, A. R., and Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *Neuroimage* 59, 781–787. doi: 10.1016/j.neuroimage.2011.07.024

Navarra, J., Alsius, A., Velasco, I., Soto-Faraco, S., and Spence, C. (2010). Perception of audiovisual speech synchrony for native and non-native language. *Brain Res.* 1323, 84–93. doi: 10.1016/j.brainres.2010.01.059

Ngo, N. K., and Spence, C. (2010). Auditory, tactile, and multisensory cues can facilitate search for dynamic visual stimuli. *Atten. Percept. Psychophys.* 72, 1654–1665. doi: 10.3758/APP.72.6.1654

Noesselt, T., Rieger, J. W., Schoenfeld, M. A., Kanowski, M., Hinrichs, H., Heinze, H. J., et al. (2007). Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *J. Neurosci.* 27, 11431–11441. doi: 10.1523/JNEUROSCI.2252-07.2007

Noesselt, T., Tyll, S., Boehler, C. N., Budinger, E., Heinze, H. J., and Driver, J. (2010). Sound-induced enhancement of low-intensity vision: multisensory influences on human sensory-specific cortices and thalamic bodies relate to perceptual enhancement of visual detection sensitivity. *J. Neurosci.* 30, 13609–13623. doi: 10.1523/jneurosci.4524-09.2010

Rach, S., Diederich, A., and Colonius, H. (2011). On quantifying multisensory interaction effects in reaction time and detection rate. *Psychol. Res.* 75, 77–94. doi: 10.1007/s00426-010-0289-0

Rao, R. P. N., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580

Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science* 212, 947–950. doi: 10.1126/science.7233191

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Molholm, S., Javitt, D. C., and Foxe, J. J. (2007). Impaired multisensory processing in schizophrenia: deficits in the visual enhancement of speech comprehension under noisy environmental conditions. *Schizophr. Res.* 97, 173–183. doi: 10.1016/j.schres.2007.08.008

Scott, G. D., Karns, C. M., Dow, M. W., Stevens, C., and Neville, H. J. (2014). Enhanced peripheral visual processing in congenitally deaf humans is supported by multiple brain regions, including primary auditory cortex. *Front. Hum. Neurosci.* 8:177. doi: 10.3389/fnhum.2014.00177

Sekuler, R., Sekuler, A. B., and Lau, R. (1997). Sound alters visual motion perception. *Nature* 385:308. doi: 10.1038/385308a0

Senkowski, D., Talsma, D., Grigutsch, M., Herrmann, C. S., and Woldorff, M. G. (2007). Good times for multisensory integration: effects of the precision of temporal synchrony as revealed by gamma-band oscillations. *Neuropsychologia* 45, 561–571. doi: 10.1016/j.neuropsychologia.2006.01.013

Senkowski, D., Talsma, D., Herrmann, C. S., and Woldorff, M. G. (2005). Multisensory processing and oscillatory gamma responses: effects of spatial selective attention. *Exp. Brain Res.* 166, 411–426. doi: 10.1007/s00221-005-2381-z

Staufenbiel, M., Van der Lubbe, R. H. J., and Talsma, D. (2011). Spatially uninformative sounds increase sensitivity for visual motion change. *Exp. Brain Res.* 213, 457–464. doi: 10.1007/s00221-011-2797-6

Stein, B. E., Burr, D., Constantinidis, C., Laurienti, P. J., Meredith, A. M., Perrault, T. J., et al. (2010). Semantic confusion regarding the development of multisensory integration: a practical solution. *Eur. J. Neurosci.* 31, 1713–1720. doi: 10.1111/j.1460-9568.2010.07206.x

Stein, B. E., London, N., Wilkinson, L. K., and Price, D. D. (1996). Enhancement of perceived visual intensity by auditory stimuli: a psychophysical analysis. *J. Cogn. Neurosci.* 8, 497–506. doi: 10.1162/jocn.1996.8.6.497

Stein, B. E., and Meredith, M. A. (1993). *The Merging of the Senses*. Cambridge, MA: MIT Press.

Stein, B. E., and Rowland, B. A. (2011). Organization and plasticity in multisensory integration. *Prog. Brain Res.* 191, 145–163. doi: 10.1016/b978-0-444-53752-2.00007-2

Stein, B. E., and Stanford, T. R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nat. Rev. Neurosci.* 9, 255–266. doi: 10.1038/nrn2331

Stevenson, R. A., Segers, M., Ferber, S., Barense, M. D., and Wallace, M. T. (2014). The impact of multisensory integration deficits on speech perception in children with autism spectrum disorders. *Front. Psychol.* 5:379. doi: 10.3389/fpsyg.2014.00379

Stevenson, R. A., and Wallace, M. T. (2013). Multisensory temporal integration: task and stimulus dependencies. *Exp. Brain Res.* 227, 249–261. doi: 10.1007/s00221-013-3507-3

Talsma, D., Doty, T. J., and Woldorff, M. G. (2007). Selective attention and audiovisual integration: is attending to both modalities a prerequisite for early integration? *Cereb. Cortex* 17, 679–690. doi: 10.1093/cercor/bhk016

Talsma, D., Senkowski, D., Soto-Faraco, S., and Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends Cogn. Sci. (Regul. Ed.)* 14, 400–410. doi: 10.1016/j.tics.2010.06.008

Talsma, D., and Woldorff, M. G. (2005). Selective attention and multisensory integration: multiple phases of effects on the evoked brain activity. *J. Cogn. Neurosci.* 17, 1098–1114. doi: 10.1162/0898929054475172

Thelen, A., Cappe, C., and Murray, M. M. (2012). Electrical neuroimaging of memory discrimination based on single-trial multisensory learning. *Neuroimage* 62, 1478–1488. doi: 10.1016/j.neuroimage.2012.05.027

Thelen, A., Talsma, D., and Murray, M. M. (2015). Single-trial multisensory memories affect later auditory and visual object discrimination. *Cognition* 138, 148–160. doi: 10.1016/j.cognition.2015.02.003

Tiippana, K., Andersen, T. S., and Sams, M. (2004). Visual attention modulates audiovisual speech perception. *Eur. J. Cogn. Psychol.* 16, 457–472. doi: 10.1080/09541440340000268

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychol. Rev.* 55, 189–208. doi: 10.1037/h0061626

Treisman, A. M., and Gelade, G. (1980). A feature-integration theory of attention. *Cogn. Psychol.* 12, 97–136. doi: 10.1016/0010-0285(80)90005-5

Tuomainen, J., Andersen, T. S., Tiippana, K., and Sams, M. (2005). Audio-visual speech perception is special. *Cognition* 96, B13–B22. doi: 10.1016/j.cognition.2004.10.004

van Atteveldt, N. M., Formisano, E., Goebel, R., and Blomert, L. (2007). Top–down task effects overrule automatic multisensory responses to letter–sound pairs in auditory association cortex. *Neuroimage* 36, 1345–1360. doi: 10.1016/j.neuroimage.2007.03.065

van Atteveldt, N. M., Murray, M. M., Thut, G., and Schroeder, C. E. (2014). Multisensory integration: flexible use of general operations. *Neuron* 81, 1240–1253. doi: 10.1016/j.neuron.2014.02.044

van den Brink, R. L., Cohen, M. X., van der Burg, E., Talsma, D., Vissers, M. E., and Slagter, H. A. (2014). Subcortical, modality-specific pathways contribute to multisensory processing in humans. *Cereb. Cortex* 24, 2169–2177. doi: 10.1093/cercor/bht069

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., and Theeuwes, J. (2008). Pip and pop: nonspatial auditory signals improve spatial visual search. *J. Exp. Psychol. Hum. Percept. Perform.* 34, 1053–1065. doi: 10.1037/0096-1523.34.5.1053

Van der Burg, E., Talsma, D., Olivers, C. N. L., Hickey, C., and Theeuwes, J. (2011). Early multisensory interactions affect the competition among multiple visual objects. *Neuroimage* 55, 1208–1218. doi: 10.1016/j.neuroimage.2010.12.068

van Ee, R., Van Boxtel, J. J. A., Parker, A. L., and Alais, D. (2009). Multisensory congruency as a mechanism for attentional control over perceptual selection. *J. Neurosci.* 29, 11641–11649. doi: 10.1523/JNEUROSCI.0873-09.2009

van Wassenhove, V., Grant, K. W., and Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proc. Natl. Acad. Sci. U.S.A.* 102, 1181–1186. doi: 10.1073/pnas.0408949102

van Wassenhove, V., Grant, K. W., and Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia* 45, 598–607. doi: 10.1016/j.neuropsychologia.2006.01.001

Vatakis, A., and Spence, C. (2007). Crossmodal binding: evaluating the "unity assumption" using audiovisual speech stimuli. *Percept. Psychophys.* 69, 744–756. doi: 10.3758/BF03193776

Vetter, P., Smith, F. W., and Muckli, L. (2014). Decoding sound and imagery content in early visual cortex. *Curr. Biol.* 24, 1256–1262. doi: 10.1016/j.cub.2014.04.020

Vroomen, J., and Baart, M. (2009). Phonetic recalibration only occurs in speech mode. *Cognition* 110, 254–259. doi: 10.1016/j.cognition.2008.10.015

Vroomen, J., and Stekelenburg, J. J. (2011). Perception of intersensory synchrony in audiovisual speech: not that special. *Cognition* 118, 75–83. doi: 10.1016/j.cognition.2010.10.002

Wallace, M. T., Meredith, M. A., and Stein, B. E. (1998). Multisensory integration in the superior colliculus of the alert cat. *J. Neurophysiol.* 80, 1006–1010.

Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., and Schirillo, J. A. (2004). Unifying multisensory signals across time and space. *Exp. Brain Res.* 158, 252–258. doi: 10.1007/s00221-004-1899-9

Wallace, M. T., and Stein, B. E. (2001). Sensory and multisensory responses in the newborn monkey superior colliculus. *J. Neurosci.* 21, 8886–8894.

Welch, R. B. (1999). "Meaning, attention, and the unity assumption in the intersensory bias of spatial and temporal perceptions," in *Cognitive Contributions to the Perception of Spatial and Temporal Events*, eds G. Aschersleben, T. Bachmann, and J. Müsseler (Amsterdam: Elsevier), 371–387.

Welch, R. B., and Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychol. Bull.* 88, 638–667. doi: 10.1037/0033-2909.88.3.638

Yau, J. M., Olenczak, J. B., Dammann, J. F., and Bensmaia, S. J. (2009). Temporal frequency channels are linked across audition and touch. *Curr. Biol.* 19, 561–566. doi: 10.1016/j.cub.2009.02.013

Yau, J. M., Weber, A. I., and Bensmaia, S. J. (2010). Separate mechanisms for audio-tactile pitch and loudness interactions. *Front. Psychol.* 1:160. doi: 10.3389/fpsyg.2010.00160

Yu, J. Y., and Frank, L. M. (2015). Hippocampal-cortical interaction in decision making. *Neurobiol. Learn. Mem.* 117, 34–41. doi: 10.1016/j.nlm.2014.02.002

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Integrative Neuroscience

ORIGINAL RESEARCH
published: 23 April 2015
doi: 10.3389/fnint.2015.00032

The effects of attention on the temporal integration of multisensory stimuli

Sarah E. Donohue[1,2,3]*, *Jessica J. Green*[1,4] *and Marty G. Woldorff*[1,2,3,5]

[1] Center for Cognitive Neuroscience, Duke University, Durham, NC, USA, [2] Department of Neurology, Otto-von-Guericke University Magdeburg, Magdeburg, Germany, [3] Leibniz Institute for Neurobiology, Magdeburg, Germany, [4] Department of Psychology, University of South Carolina, Columbia, SC, USA, [5] Department of Psychiatry, Duke University, Durham, NC, USA

In unisensory contexts, spatially-focused attention tends to enhance perceptual processing. How attention influences the processing of multisensory stimuli, however, has been of much debate. In some cases, attention has been shown to be important for processes related to the integration of audio-visual stimuli, but in other cases such processes have been reported to occur independently of attention. To address these conflicting results, we performed three experiments to examine how attention interacts with a key facet of multisensory processing: the temporal window of integration (TWI). The first two experiments used a novel cued-spatial-attention version of the bounce/stream illusion, wherein two moving visual stimuli with intersecting paths tend to be perceived as bouncing off rather than streaming through each other when a brief sound occurs near in time. When the task was to report whether the visual stimuli appeared to bounce or stream, attention served to narrow this measure of the TWI and bias perception toward "streaming". When the participants' task was to explicitly judge the simultaneity of the sound with the intersection of the moving visual stimuli, however, the results were quite different. Specifically, attention served to mainly widen the TWI, increasing the likelihood of simultaneity perception, while also substantially increasing the simultaneity judgment accuracy when the stimuli were actually physically simultaneous. Finally, in Experiment 3, where the task was to judge the simultaneity of a simple, temporally discrete, flashed visual stimulus and the same brief tone pip, attention had no effect on the measured TWI. These results highlight the flexibility of attention in enhancing multisensory perception and show that the effects of attention on multisensory processing are highly dependent on the task demands and observer goals.

Keywords: attention, multisensory, audiovisual, cueing, bounce-stream, temporal

OPEN ACCESS

Edited by:
Salvador Soto-Faraco, Universitat Pompeu Fabra, Spain

Reviewed by:
Roberto Martuzzi, Ecole Polytechnique Fédérale de Lausanne, Switzerland
Jessica Hartcher-O'Brien, l'Ecole Normale Superieuer, France

***Correspondence:**
Sarah E. Donohue, Department of Neurology, Otto-von-Guericke University Magdeburg and Leibniz Institute for Neurobiology, Leipziger Strasse 44, 39120 Magdeburg, Germany
donohue.sarah.e@gmail.com

Received: *01 September 2014*
Accepted: *07 April 2015*
Published: *23 April 2015*

Citation:
Donohue SE, Green JJ and Woldorff MG (2015) The effects of attention on the temporal integration of multisensory stimuli. Front. Integr. Neurosci. 9:32. doi: 10.3389/fnint.2015.00032

Introduction

The selective focusing of attention on a particular region in space provides a more accurate representation of the objects that lie within that region than those that lie within unattended regions. With accurate perception being crucial to optimal behavioral responses, the topic of the role that attention plays in enhancing perception has been studied for decades (see Carrasco, 2011 for review). One method that has been used to characterize how attention is focused and

the ramifications of that focus is attentional cueing (Posner, 1980). In spatial cueing studies of visual attention, participants are cued to shift their attention to a particular location, while ignoring other locations, in preparation for an upcoming stimulus that is likely to occur in the cued location (e.g., Posner, 1980; Posner and Cohen, 1984; Weichselgartner and Sperling, 1987; Müller and Rabbitt, 1989; Berger et al., 2005; Giordano et al., 2009). When targets appear at the cued (i.e., attended) location, participants are faster and more accurate to respond to them as compared to targets that appear at an uncued location (e.g., Bashinski and Bacharach, 1980; Hawkins et al., 1990; Coull and Nobre, 1998; Yeshurun and Carrasco, 1999). Data from neural studies of the allocation of spatial attention suggest that this improvement in behavioral performance is the result of a gain in the event-related response of the sensory cortices responsible for processing the targets, as well as surrounding inhibition of the processing of the distractors (Motter, 1993; Luck et al., 1994, 1997; Mangun, 1995; Hopf et al., 2006; Silver et al., 2007).

One key feature of attention that has emerged from recent work is the flexibility with which it operates. When the task is to discriminate the orientation of a gabor patch among noise (i.e., to perform a contrast discrimination), the allocation of spatial attention will enhance the signal from that stimulus, enabling enhanced discrimination (Carrasco et al., 2000). When the task is to make fine color discriminations, attention will serve to enhance the processing of the color information (Wegener et al., 2008). In other tasks, attention can serve to enhance relevant information in the face of conflict (e.g., MacDonald et al., 2000; Botvinick et al., 2001), to spread so as to encompass an entire object (e.g., Egly et al., 1994; Donohue et al., 2011), or to aid in the coding of the direction of motion (Stoppel et al., 2011). Although the majority of data on spatial attention has come from studies of the visual modality, auditory and tactile cues can also serve to direct attention to a particular region of space, producing enhanced processing of visual, auditory, or tactile targets that fall within that region (Eimer and Schröger, 1998; Spence et al., 2000; Wu et al., 2007; Green et al., 2011), demonstrating that attention can be flexibly deployed within and across all the spatial modalities.

Perception is not limited to one modality, however, as we can receive spatially-relevant information from visual, auditory, and tactile modalities concurrently. Input from multiple modalities can arise from a multisensory event or object, and this input is often grouped (or integrated) together. The binding of multisensory input occurs when stimuli are temporally and spatially proximal, with the likelihood of such binding falling off as the spatial and/or temporal separation increase (Meredith and Stein, 1986; Meredith et al., 1987; Slutsky and Recanzone, 2001; Zampini et al., 2005; Donohue et al., 2010; reviewed in Chen and Vroomen, 2013). In speech, for example, this binding of multisensory stimuli allows us to associate the auditory (speech sounds) and visual (mouth movements) input as coming from a single individual and not from multiple sources, which facilitates both the perceptual integration of the separate inputs and the accurate processing of the speech information (Besle et al., 2004). That is, when these redundant inputs (i.e., from the same event or object) are grouped, this can facilitate the perceptual processing of that event or object relative to other stimuli in the environment (see Alais et al., 2010 for review).

With both selective attention and multisensory integration generally enhancing the processing of stimuli (e.g., Miller, 1982; Mangun and Hillyard, 1991; Quinlan and Bailey, 1995; Diederich and Colonius, 2004; Pestilli et al., 2007; Abrams et al., 2010; Gondan et al., 2011), it would seem likely that when these two functional processes occur together, the optimal representation of the environment would be obtained. The interaction between selective attention and multisensory integration is not necessarily additive in nature, however, and the degree to which attention and integration are independent processes, and how they interact during perception has been of much debate recently (see Koelewijn et al., 2010; Talsma et al., 2010 for reviews). One example of such a discrepancy comes from studies of perceptual recalibration wherein the perception of audio-visual simultaneity can be altered by exposure to temporal offsets between the auditory and visual stimuli (e.g., Fujisaki et al., 2004). In such cases, the focus of attention appears to be able to influence the audio-visual pairing to which the perception of simultaneity is recalibrated (Heron et al., 2010; Ikumi and Soto-Faraco, 2014). Yet neural evidence from recordings in animals suggests that auditory and visual stimuli can be temporally integrated without attention being necessary (Meredith et al., 1987).

One possible reason for discrepant findings on the degree to which attention and multisensory integration interact is the nature of the specific tasks that have been used. Some studies have employed tasks that require perceptual discriminations in one modality, with the other modality being task-irrelevant (Keitel et al., 2011; Sarmiento et al., 2012; Marchant and Driver, 2013), whereas others have required attention to both modalities, when a target could be present in one or both modalities (e.g., Talsma et al., 2007). Other studies have used tasks that are orthogonal to the question at hand (Busse et al., 2005; Fairhall and Macaluso, 2009), using measurements of neural activity to infer that multisensory integration has taken place or that enhanced processing results from multisensory stimulation. Still others have not required a task at all, assessing "passive" multisensory integration processes (van Atteveldt et al., 2004, 2007). If attention is as flexible a system as research suggests, then it may be the case that under some of these circumstances attention is necessary for effective multisensory integration, whereas in other tasks it may be less essential or have no influence at all (Bertelson et al., 2000).

Here, we focused on one specific facet of multisensory integration—temporal binding—to determine the circumstances under which attention interacts with audiovisual integration processes. As mentioned above, multisensory stimuli tend to be grouped together when they occur close together in time (see Stein and Stanford, 2008 for review). This temporal binding is not absolute, however, and encompasses a window of approximately ±150 ms, known as the temporal window of integration (TWI;

Spence et al., 2001; Zampini et al., 2005; van Wassenhove et al., 2007). In general, when an auditory stimulus and a visual stimulus occur within this temporal range they are more likely to be linked together and integrated, whereas with larger temporal separations the stimuli are more likely to be segregated. The breadth of this temporal window, therefore, reflects the temporal precision of the integration process, with a narrow window indicating integration that is in line with physical simultaneity and a broad window indicating integration that occurs relatively far beyond physical simultaneity. Such a temporal spread, therefore, can serve as a useful tool in characterizing the way attention and multisensory integration processes interact.

Utilizing what is known about attention within and across modalities, several possible hypotheses can be generated about the ways in which attention could interact with the TWI. If attention serves to sharpen perception, giving more precision to judgments of what is physically present in the environment, then attention should act to narrow the TWI (**Figure 1A**). Conversely, because attention is a flexible process, it may be the case that in tasks that are facilitated by multisensory processing, attention will tend to serve to broaden the TWI, making integration more likely over a broader temporal range, thus enhancing the multisensory integration process itself (**Figure 1B**). Lastly, attention could have no effect on the TWI, with the same amount of temporal integration observed whether stimuli are attended or are unattended.

Here we performed three experiments that used three different tasks and that manipulated the allocation of spatial attention and temporal onsets of the auditory and visual stimuli to determine if attention would sharpen or broaden the TWI. Our results suggest that attention interacts with audiovisual integration processes in a flexible and adaptive manner—broadening the TWI, sharpening the TWI, or not modulating the TWI at all—depending on the requirements of the task and the amount of perceptual uncertainty due to the complexity of the stimuli.

Experiment 1

In this experiment, we created a cued-attention version of the classic "bounce/stream paradigm" to measure audio-visual integration as a function of attention. In the bounce-stream paradigm, two visual objects (e.g., circular disks) move toward each other, overlap, and then move away from each other (Watanabe and Shimojo, 1998). This pattern of motion can be perceived either as two objects streaming through each other or as two objects bouncing off each other. When the visual objects are presented in this configuration, participants generally tend to perceive them (correctly) as streaming through each other. However, when a sound is presented near the time of overlap, participants are more likely to perceive the objects as bouncing off of each other (Sekuler et al., 1997; Watanabe and Shimojo, 2001; Bushara et al., 2003). That is, although the motion of the visual stimuli is always physically identical, the mere presence of an irrelevant sound can alter the perception of the visual stimuli.

The above-described phenomenon, known as the auditory bounce effect (ABE), has been proposed to be the result of audio-visual integrative processes based on several pieces of evidence. First, the ABE is dependent on the type of sound used, with sounds that are more collision-like in nature producing higher percepts of bouncing (Grassi and Casco, 2009, 2010). Second, when the objects are perceived as bouncing vs. as streaming (under the same physical conditions), multisensory brain regions

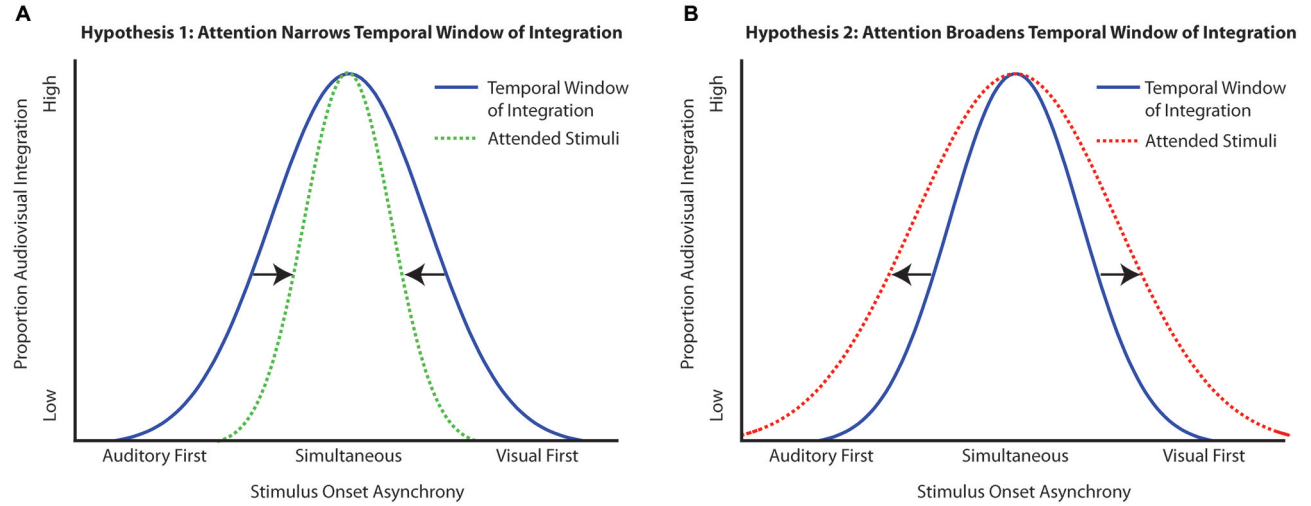


FIGURE 1 | Possible outcomes of the interaction of attention with the temporal window of integration (TWI). **(A)** Hypothesis 1 depicts the narrowing of the TWI under the influence of attention, which would make the integration more precise (i.e., closer in line with the physical offsets of the stimuli). **(B)** Hypothesis 2 depicts the broadening of the TWI when multisensory stimuli are presented in the presence of attention. In this circumstance, attention serves to enhance the integration process overall, thereby producing a greater time range over which the multisensory stimulus components are likely to be integrated.

are activated (Bushara et al., 2003). Third, transcranial magnetic stimulation (TMS) to the right posterior parietal cortex (a region implicated in multisensory processing) decreases the perception of bouncing responses (Maniglia et al., 2012). Finally, the ABE has been shown to decrease as the auditory stimulus is presented farther away in time from being coincident with the visual stimulus (Watanabe and Shimojo, 2001; Remijn et al., 2004), indicating the temporal dependency of this multisensory effect.

In the classic bounce-stream paradigm, the visual stimuli are presented centrally, are the only stimuli presented, and thus occur within the focus of attention. In the current experiment, participants were given an attention-directing cue toward the left or right visual field that indicated where the bouncing/streaming objects were most likely to appear, allowing us to examine responses when the stimuli were occurring within the cued (attended) location vs. when they were occurring within the uncued (unattended) location. In addition, we manipulated the temporal delay between the auditory stimulus and intersection of the visual stimulus pair, as the perception of bouncing should decrease as the temporal discrepancy increases. If attention has no influence on multisensory integration, then we would expect the same pattern of integration-reflecting behavior regardless of whether the auditory and visual events occurred within or outside the focus of attention. If, however, attention interacts with multisensory integration, we would expect a different pattern of perception as a function of attentional allocation. More specifically, we hypothesized that attention would interact with multisensory integration by specifically serving to narrow the TWI, highlighting its ability to provide more accurate representations of objects that fall within its focus (**Figure 1A**).

Methods

Participants

Twenty healthy adults with normal vision and hearing participated in this study (6 male; Mean age = 24.6 years, SD = 4.1 years). One additional participant was excluded due to a failure to understand the task instructions. All procedures were approved by the Institutional Review Board of the Duke University Health System, and all participants gave written informed consent prior to the start of the experiment.

Stimuli and Task

Each trial began with a central cue at fixation (the letter "L" or "R") that instructed participants to direct their attention to the left or right hemifield, respectively (**Figure 2**). At the same time as the cue, four white disks onset bilaterally, two in the upper left hemifield and two in the upper right hemifield. Each disk was 1.5° in diameter and presented 4° above fixation, with the innermost disks 4.9° and the outermost disks 10° to the left and right of fixation. All stimuli were presented on a black background. The cue lasted for 250 ms, after which there was a cue-target interval of 650 ms wherein the disks (and a fixation cross) remained stationary on the screen. After the delay period, each pair of disks began to move toward one another (i.e., the left disks moved toward each other and the right disks moved toward each other). On each trial, one pair of disks continued to move (at a rate of 11 degrees per second), intersected with 100% overlap, and then continued their trajectory until they were 4° below fixation. The pattern of the motion, therefore, was an "X" in shape, and the disks physically always streamed through each other. The second pair of disks stopped moving prior to the intersection such that they never overlapped (See **Figure 2** for trial sequence). On 75% of trials the full motion stimulus appeared at the validly cued location and the stopped motion appeared at the uncued location. On the remaining 25% of trials the full motion stimulus appeared at the uncued location, and the stopped motion at the cued location. The next trial began 750 ms after the full motion stimulus ended. Participants were instructed to make a bounce/stream judgment for the full motion stimulus, regardless of the location at which it appeared, and to respond via button press as quickly as possible.

On 25% of trials the visual stimuli were presented alone (Visual Only; VO), allowing us to examine the effects of spatial attention on the perception of the motion stimuli in the absence of any multisensory interactions. On 75% of the trials an auditory stimulus was presented (500 Hz tone, 16 ms duration with 5 ms rise/fall, 50 dB SPL) via speakers positioned adjacent

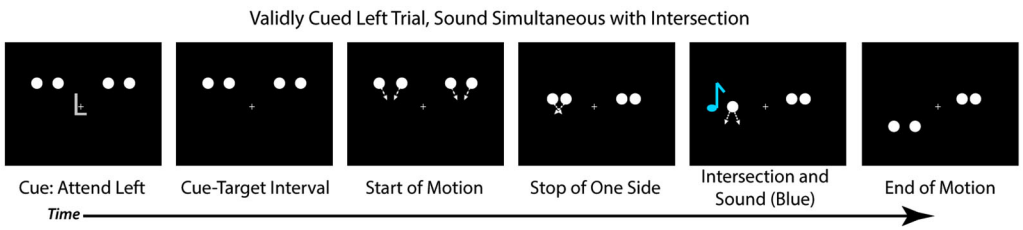


FIGURE 2 | Trial Structure. Each trial started with a centrally presented cue ("L" or "R" for left or right), followed by a cue-target delay interval. After the cue-target delay, both sets of disks started moving toward each other. Before they intersected, one set of disks (here depicted on the right) would stop moving while the other would continue to move, intersect and overlap, and move apart, continuing upon the trajectory. In 25% of the trials a sound would occur simultaneously with the disks' intersection (D0, depicted here), and other trials the sound could be delayed by 150 ms after the intersection (25% of trials; D150), be delayed by 300 ms after the intersection (25% of trials; D300), or not be present at all (25% of trials; VO). Participants were asked to judge if they perceived the disks that continuously moved as bouncing off of each other or streaming through each other.

to the computer monitor. On these multisensory trials, the sound could be presented simultaneously with the intersection of the disks (25% of all trials; 0 ms audio-visual delay; [D0]), presented 150 ms after the intersection of the disks (25% of all trials; 150 ms Delay [D150]), or presented 300 ms after the intersection of the disks (25% of all trials; 300 ms Delay [D300]). The sounds were always presented from the speaker on the same side as the full motion visual stimulus to avoid effects of spatial incongruency between the auditory and visual stimuli and increasing the likelihood of multisensory integration (Meredith and Stein, 1986; Slutsky and Recanzone, 2001). Participants were told that some trials would contain a non-informative sound which was not relevant for their responses.

Each participant completed one practice block followed by six experimental blocks. Participants' eye movements were monitored online via video feed to ensure they were maintaining central fixation. Each block contained 72 valid trials and 24 invalid trials, equally distributed across locations (left/right) and SOAs (D0/D150/D300), for a total of 144 valid and 48 invalid trials for each SOA for each participant. In the VO condition there were also a total of 144 valid and 48 invalid trials presented during the experiment.

Behavioral Data Analysis

The proportion of "bounce" responses was compared with a repeated-measures ANOVA with factors for validity (2 levels: validly cued targets, invalidly cued targets) and audio-visual delay (3 levels: D0, D150, D300). The VO trials were separately compared for valid vs. invalid cuing with a paired-samples *t*-test. For the response-time data, a 2 × 4 ANOVA was run with the factors of validity (2 levels: validly cued targets, invalidly cued targets) and of condition (4 levels: VO, DO, D150, D300). Greenhouse-Geisser adjusted *p*-values are reported where applicable.

Results

Response Times

Prior to performing any analyses of the bounce judgments, response times (RTs) for the valid trials were compared to those for the invalid trials to ensure that the attentional manipulation had been effective. Participants were significantly faster to respond when the target stimuli occurred at the validly cued location as compared to when they occurred at the invalidly cued location (Mean Valid RT = 590 ms, SD = 140; Mean Invalid RT = 666 ms, SD = 140, $F_{(1,19)} = 28.45$, $p < 0.001$, $\eta_p^2 = 0.60$), indicating that the participants were, indeed, attending to the cued side of the display. There was an additional main effect of condition ($F_{(3,57)} = 15.86$, $p < 0.001$, $\eta_p^2 = 0.46$), with the responses to the D0 stimuli being faster than those to the visual alone (Mean D0 = 582 ms, Mean VO = 637 ms, $t_{(19)} = 4.52$, $p < 0.001$), the responses to the D0 condition being faster than the D150 condition (Mean D150 = 647 ms, $t_{(19)} = 5.15$, $p < 0.001$), and the responses to the D0 condition being faster than the responses to the D300 condition (Mean D300 = 646 ms, $t_{(19)} = 4.46$, $p < 0.001$). Of note, all the aforementioned pair-wise comparisons remained significant at the Bonferroni-corrected alpha level of 0.008). For the RTs, there was also a significant interaction between validity and condition ($F_{(3,57)} = 3.95$, $p = 0.01$, $\eta_p^2 = 0.17$), which was driven by the validity effect for the D300 condition being significantly larger than the validity effect for the VO condition ($t_{(19)} = 3.44$, $p = 0.003$).

Bounce/Stream Judgments

The proportion "bounce" responses as a function of cue validity is shown in **Figure 3**. The ANOVA revealed a main effect of validity ($F_{(1,19)} = 7.98$, $p = 0.01$, $\eta_p^2 = 0.30$), a main effect of SOA ($F_{(2,38)} = 35.46$, $p < 0.001$, $\eta_p^2 = 0.65$), and a significant interaction of validity and SOA ($F_{(2,38)} = 23.99$, $p < 0.001$, $\eta_p^2 = 0.56$). *Post hoc t*-tests showed a significant difference (at the Bonferroni-corrected alpha of 0.02) between the proportion "bounce" responses for validly and invalidly cued targets at both the 150 ms ($t_{(19)} = 3.01$, $p = 0.007$) and 300 ms delays ($t_{(19)} = 4.03$, $p = 0.001$), with both SOAs showing a higher proportion of "bounce" responses when the stimuli were presented on the invalidly-cued side. In contrast, the simultaneous condition did not reveal any significant differences as a function of cue validity ($t_{(19)} = 0.87$, $p = 0.40$). An analysis of the visual-alone condition also revealed a lower proportion of "bounce" responses for valid compared to invalid trials ($t_{(19)} = 3.80$, $p = 0.001$). Thus, in this experiment the effect of spatial attention was to narrow the TWI, by steepening the roll-off of the SOA function over which the audio-visual information was integrated into a "bouncing" percept (see **Figure 3**).

Discussion of Experiment 1

The results of Experiment 1 show that attention does indeed alter the temporal binding of multisensory stimuli. In line with

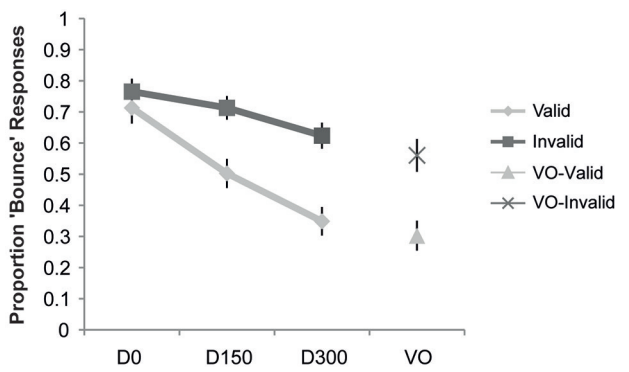


FIGURE 3 | Results of Experiment 1. The proportion "bounce" responses are plotted as a function of condition. The respective points represent when there was no auditory stimulus (Visual Only, VO), and when the auditory stimulus occurred simultaneously with the visual intersection (D0), delayed by 150 ms from the visual intersection (D150), and delayed by 300 ms from the visual intersection (D300), separately for the validly and invalidly cued trial types. Compared to the validly cued trials, a significant increase in the bouncing percept was observed in the invalid trials for the VO, 150, and 300 conditions, thus indicating a narrowing of the TWI with attention. Error bars represent the standard error of the mean (SEM).

previous studies, the perception of bouncing decreased as the auditory stimulus occurred farther away in time from the visual intersection (Watanabe and Shimojo, 2001; Remijn et al., 2004). This was true for both the validly and invalidly cued trials, and highlights the importance of temporal coincidence in this multisensory percept. However, when the visual and auditory events fell outside of the focus of attention (i.e., occurred at the invalidly cued location), participants were more likely to perceive bouncing when the auditory stimuli were delayed in time. This modulation of perception by attention, even at a delay that is typically considered outside the TWI (300 ms), suggests that the TWI was broadened when attention was not present, or, conversely, that the presence of attention narrowed/steepened the TWI).

One question that arises, however, is whether attention was altering multisensory integration *per se*, or if the effects seen here are driven primarily by attentional modulation of visual perception, as the visual-alone condition showed a similar effect of cue validity. Previous findings have suggested the importance of local motion cues in determining the accurate representation of streaming (Kawabe and Miura, 2006) and that when attention is drawn away from the local motion of the two objects they are more likely to be perceived as bouncing (Watanabe and Shimojo, 1998). Although this visual modulation may be contributing to the effects observed here, it cannot fully account for our results, as the increase in "bounce" perception was not uniform across temporal intervals. The combination of an absence of a validity effect on the bounce perception when the audio-visual stimuli occurred simultaneously and the increasing cue validity effects with increasing temporal disparity suggests that attention interacted with the TWI. Moreover, several recent studies have linked the bounce/stream illusion to multisensory integration by demonstrating that the perception of bouncing is highly dependent on the type of auditory stimuli (with more collision-like stimuli giving a higher proportion of bouncing percepts (Grassi and Casco, 2009, 2010), and that the perception of bouncing both activates multisensory areas in neuroimaging studies and is dependent on the functional integrity of those areas (Bushara et al., 2003; Maniglia et al., 2012).

The pattern observed here is thus consistent with the idea that attention serves to provide the most accurate representation of the information within its focus, whether it be visual alone or visual combined with auditory information. Indeed, participants were more likely to perceive the visual stimulus by itself as streaming (i.e., its veridical physical movement), rather than bouncing, when it was presented inside the focus of attention. Importantly, however, the visual information in this experiment was always attended and the auditory information was always task-irrelevant.

Experiment 2

Although the results of Experiment 1 provide one way in which the focus of attention can alter the temporal pattern of multisensory integration, the actual judgment of the temporal binding of the visual and auditory stimuli was inferred through a somewhat indirect measure (i.e., the proportion "bounce" responses of the visual stimuli, with only the visual stimuli being relevant). In the second experiment we wanted to more directly assess the temporal binding processes by having both the audio and visual stimuli be task relevant and by making the task entail an explicit judgment of the relative timing of these two stimuli. Accordingly, we asked an independent group of participants to perform a different task using the same stimuli, namely to judge the temporal coincidence of the auditory stimulus and the intersection of the moving circles. We hypothesized that the pattern of temporal integration would be altered in a similar manner as in Experiment 1 such that there would be increased integration at the 150 and 300 ms SOAs when the audio-visual events occurred in an unattended compared to attended visual location (**Figure 1A**).

Participants

Twenty participants (9 male) participated in this experiment (Mean age = 22.3 years, SD = 3.2 years). None of the participants in this study had participated in Experiment 1. All participants gave written informed consent and all procedures were approved by the Institutional Review Board of the Duke University Health System.

Stimuli and Task

The stimuli and experimental conditions were identical to those used in Experiment 1 with the exception of the visual-only condition, which was eliminated due to the task requiring the presence of both the audio and visual stimuli on every trial. In particular, rather than participants judging if the visual disks appeared to bounce or stream through each other, they now performed a simultaneity judgment task. Specifically, participants were asked to determine if the sound occurred at the same time as the intersection of the visual stimuli or if it was offset in time. All responses were made via button press, and participants were instructed to respond as quickly and as accurately as possible. Participants were monitored via a live video feed to ensure they were maintaining fixation.

Behavioral Data Analysis

Analogous to Experiment 1, the proportion of "simultaneous" responses in the various conditions was compared with a repeated-measures ANOVA, with factors for validity (2 levels: validly cued targets, invalidly cued targets) and audio-visual delay (3 levels: D0, D150, D300). A separate ANOVA with identical factors was conducted for the response time data. Greenhouse-Geisser adjusted p-values are reported where applicable.

Results

Response Times

As above, in order to assess the efficacy of the attentional manipulation, RTs to discrimination task for validly cued stimuli were compared to those to invalidly cued stimuli. As in Experiment 1, participants were significantly faster when the multisensory stimuli appeared on the cued (valid) side than on the uncued (invalid) side (Mean valid RT = 649 ms, SD = 162 ms; Mean invalid RT = 704 ms, SD = 188 ms; $F_{(1,19)}$ = 11.05,

$p = 0.003$, $\eta_p^2 = 0.36$), There was also a main effect of audio-visual delay $F_{(2,40)} = 77.43$, $p < 0.001$, $\eta_p^2 = 0.80$), with participants responding significantly faster (at the Bonferroni-corrected alpha level of 0.017) in the D0 condition (Mean = 589 ms) than in the D150 condition (Mean = 703 ms; $t_{(19)} = 9.26$, $p < 0.001$), as well as faster in the D0 condition than in the D300 condition (Mean = 737 ms; $t_{(19)} = 912.07$, $p < 0.001$). There was also a significant validity by audio-visual delay interaction ($F_{(2,40)} = 7.36$, $p = 0.003$, $\eta_p^2 = 0.30$), which was driven by the validity effect for the D0 condition being greater (at the Bonferroni-corrected alpha level of 0.017) than the validity effect for the D150 condition ($t_{(19)} = 4.14$, $p = 0.001$) and the validity effect for the D0 condition being greater than the validity effect for the D300 condition ($t_{(19)} = 2.87$, $p = 0.009$).

Simultaneity Judgments

Results of the simultaneity judgment task can be seen in **Figure 4**. Similar to Experiment 1, we observed a main effect of validity ($F_{(1,19)} = 10.46$, $p = 0.004$, $\eta_p^2 = 0.36$), a main effect of SOA ($F_{(2,38)} = 68.23$, $p < 0.001$, $\eta_p^2 = 0.78$), and an interaction of validity and SOA ($F_{(2,38)} = 11.40$, $p < 0.001$, $\eta_p^2 = 0.38$). However, *post hoc t*-tests (significant at the Bonferroni-corrected alpha level of 0.02) revealed that there were significantly fewer "simultaneous" responses for invalidly cued trials at both the D0 (Mean Valid = 93.1%, Mean Invalid = 67.2%; $t_{(19)} = 4.02$, $p = 0.001$) and 150 ms delay (Mean Valid = 66.8%, Mean Invalid = 46.2%; $t_{(19)} = 3.15$, $p = 0.005$). No cue validity effect was observed at the longest SOA.

Discussion of Experiment 2

In this second experiment, now using a simultaneity judgment task, attention also interacted robustly with the temporal pattern of multisensory integration; however, this interaction manifested in a completely different manner than in Experiment 1. Whereas in Experiment 1 the measure for integration (a "bounce" perception) was *more* likely to occur at the unattended location, at least as the multisensory components were more separated in time, here the measure for integration (a simultaneity judgment) was *less* likely to occur at the unattended location, particularly when the audio and visual events were closer in time even physically simultaneous and would have been expected to be temporally integrated. This discrepancy indicates that the specifics of the task, such as whether only one or both modalities are attended or whether the temporal relationship of the stimuli is task relevant, can dramatically influence the interaction of attention and stimulus binding or integration processes.

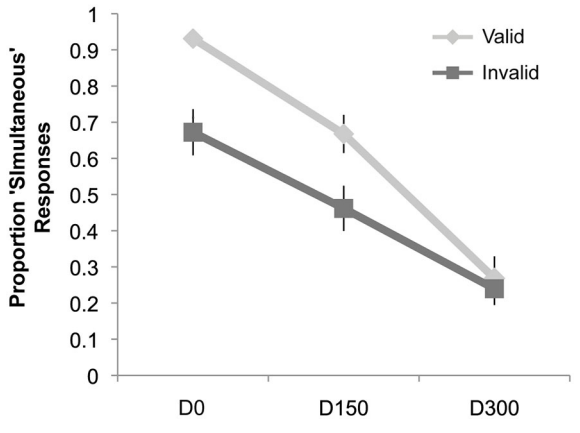


FIGURE 4 | Results of Experiment 2. The proportion "simultaneous" responses are plotted as a function of condition: simultaneous auditory stimulus with visual intersection (D0), auditory stimulus delayed by 150 ms from the visual intersection (D150), and the auditory stimulus delayed by 300 ms (D300), separately for the validly and invalidly cued trial types. There was a significant decrease in the proportion of "simultaneous" responses for the D0 and D150 conditions when these trial types were presented on the invalidly cued side (compared to the validly cued side), but no difference for the D300 condition.

One important difference between the two above experiments was the task-relevance of the auditory stimuli. The audiovisual stimuli were physically identical in both experiments, but in Experiment 1 only the visual stimuli were task-relevant whereas in Experiment 2 information about both the auditory and visual events was necessary for responding correctly. Thus, the simultaneity judgment task may have imparted an increased level of uncertainty about the stimuli, or at least an increased level of complexity of the task: Not only did participants need to determine when the visual stimuli intersected, but they also needed to decide if the auditory stimulus coincided with this intersection. Having attention directed away from the uncued side may have increased the uncertainty of the timing of both the visual intersection and that of the auditory stimulus, while also increasing the complexity of the task, which may have resulted in the shift in criterion that we see here. In other words, if both the intersection of the disks and the auditory tone were happening outside the focus of attention, the greater temporal ambiguity of the two relevant events may have caused participants to be less likely to link them in time (as measured by their judgment of the simultaneity).

A second interpretation of the pattern of results observed here is that when, as in Experiment 2, the task specifically required the discrimination of the temporal relationship of the stimuli and thus the explicit discrimination of the temporal binding, attention served to increase the precision of such binding by integrating events that were temporally concurrent (i.e., in the simultaneous condition), while still segregating appropriately those events that were temporally separate (i.e., the D300 condition). However, when the measure of multisensory integration is more direct as in the bouncing/streaming task of Experiment 1 (although it is perhaps a more indirect measure of temporal linking), attention served to provide the most precise representation of the visual motion, with more veridical streaming percepts when the stimuli occurred within its focus.

A slightly different interpretation of these results could be one of a shift in criterion. When the auditory and visual stimuli were presented at the attended location, it could be the case that participants shifted their criteria toward judging the stimuli as being simultaneous. In other words, with the exception of the widest SOA (D300), participants were more prone to responding "simultaneous" when attention was present. Again, given that

this effect did not occur for every SOA, it was not independent of the timing, but rather served to increase the "simultaneous" responses particularly for the D0 and D150 conditions. From such an interpretation, it would follow that attention serves to enhance the temporal binding of auditory-visual stimuli, making them more likely to be perceived as being temporally simultaneous, and tending to create more certainty for making a "simultaneous" response.

Overall these first two experiments both clearly show that attention can strongly modulate the processes related to the temporal integration of multisensory stimuli. Although the measures of temporal integration and segregation differed for the two tasks, in both Experiments 1 and 2 there was a clear modulation of widely used measures of integration by attention. The strikingly different patterns of this attentional modulation depended strongly on the measures being used and on what task was being performed. More fundamentally, the results suggest that the role of attention in multisensory integration processes can change depending on the task being performed with the sensory stimuli, in a way that is in line with the idea that attention serves to resolve ambiguity in our perception, depending on what type of information is important for veridical perception. It also suggests that the temporal binding of multisensory stimuli (as measured with the simultaneity judgment task) differs from the perceptual integration of multisensory stimuli (measured in the bounce/stream task).

Both the bounce/stream task and the simultaneity judgment task used here required precise tracking of moving visual stimuli over one second, with successful performance depending on determining what was occurring during one specific moment of this motion (i.e., the stimulus intersection). The disruption of this trajectory when participants had to switch to the uncued side likely increased the uncertainty of the disk intersection timing, as well as the nature of this perceptual event. Not only do these complex stimuli increase perceptual uncertainty overall when they occur on the unattended side, the long-duration motion stimuli are quite different than the static, highly-controlled, flashed visual stimuli often used in multisensory integration tasks (Spence et al., 2001; Zampini et al., 2005; Donohue et al., 2010) Thus, it remained unclear what role attention would play in audiovisual integration under situations with little-to-no perceptual uncertainty, such as with discrete temporal events. Accordingly, we further wanted to ascertain how specific this modulation was to the stimuli that we had been using, and therefore conducted a cued-variant of a simultaneity judgment task using temporally discrete stimuli in both the visual and auditory modalities.

Experiment 3

Experiment 3 utilized the same task as in Experiment 2 (a cued simultaneity-judgment task) but replaced the complex visual motion stimuli with a flash of a static checkerboard image. We hypothesized that this simple task, involving a temporally discrete visual stimulus and a temporally discrete auditory one, would involve less perceptual uncertainty and less task complexity, and thus that the influence of attention on the TWI would be reduced, as attentional resources would be less essential for accurate task performance.

Participants

Twenty participants took part in this experiment (one left-handed, nine male, Mean age = 23.6, SD = 4.5 years). None of these participants had taken part in either of the previous experiments. Three additional participants were excluded for failure to do the task properly (i.e., having less than 50% accuracy for reporting the simultaneous trials as simultaneous on the validly cued side). All participants gave written informed consent, and all procedures were approved by the Institutional Review Board of the Duke University Health System.

Stimuli and Task

The auditory stimuli used were the same as in Experiment 2 (brief tone pips), but the visual motion stimuli were replaced by a briefly flashed (16 ms duration) static black and white checkerboard (1.8° × 1.8°) that could be presented 12.6° to the left or right of fixation and presented on a gray background (**Figure 5**). In addition, on each trial, the checkerboards were only presented on one side of fixation (as opposed to the bilateral displays used in the previous experiments). The checkerboard image was presented on the validy cued side 75% of the time and on the invalidly cued side 25% of the time). There were 18 trials in each block for each of the valid conditions and 6 trials in each block for each of the invalidly-cued conditions, with six blocks in total completed by each participant.

As in Experiment 2, participants performed a simultaneity judgment of the audiovisual events. Participants were asked to determine if the checkerboard image and the tone occurred at the same time or if they occurred at slightly different times, and to respond as quickly and as accurately as possible via button press. As with all experiments, participants were monitored over a closed-circuit video camera to ensure they were maintaining fixation and remaining attentive to the task.

Data Analysis

All data analyses of the simultaneity judgment's task measures were performed in an identical manner to Experiment 2.

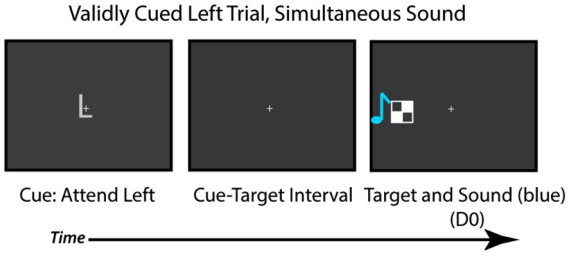


FIGURE 5 | Task used in Experiment 3. Participants were cued to attend to the left or right visual field. After a cue-target interval, the checkerboard target appeared on the validly cued (75% of time) or invalidly-cued side (25% of the time), and participants had to judge if the sound and visual target occurred simultaneously or at separate times.

Results

Response Time

Participants were significantly faster to respond to the validly cued trials compared to the invalidly cued trials (Mean valid RT = 630 ms, SD = 60 ms; Mean invalid RT = 651 ms, SD = 69 ms; ($F_{(1,19)} = 12.15$, $p = 0.002$, $\eta_p^2 = 0.39$), demonstrating participants were focusing their attention toward the cued location. There was also a main effect of audio-visual SOA (SOA ($F_{(2,38)} = 109.46$, $p < 0.001$, $\eta_p^2 = 0.85$) with all conditions differing significantly from each other at the Bonferroni-corrected alpha level of 0.0167 (Mean RT D0 = 560 ms, Mean RT D150 = 640 ms, Mean RT D300 = 720 ms, all p's < 0.001). The interaction between validity and audio-visual significance did not reach significance ($F < 1$).

Simultaneity Judgments

The repeated-measures ANOVA on the proportion of "simultaneous" responses (see **Figure 6**) revealed a main effect of SOA ($F_{(2,38)} = 36.74$, $p < 0.001$, $\eta_p^2 = 0.66$), but no significant effect of validity nor a significant interaction of validity and SOA (all p's > 0.05).

Discussion of Experiment 3

In Experiment 3, although the proportion of "simultaneous" responses showed the classic decrease with the temporal separation of the visual and auditory stimuli, there were no differences observed at any of the SOAs as a function of attention. Thus, with the use of simple, briefly presented stimuli in both modalities, attention did not influence the TWI. Importantly, there was still a validity effect on the RTs, suggesting that participants were appropriately focusing their attention to the cued location (as in Yeshurun and Carrasco, 1999).

These data suggest that attention is not necessary for processes related to multisensory integration in all cases. Indeed, much of the pioneering work on the neural basis of multisensory integration was done with recordings in the superior colliculus of anesthetized cats (Meredith and Stein, 1986; Meredith et al., 1987). In this case, the animals were not paying attention to the simple audiovisual stimuli (flashes and tones) and yet the temporal and spatial properties of multisensory integration were still observed (Meredith and Stein, 1986; Meredith et al., 1987).

More likely, because the stimuli in Experiment 3 were simple, discrete perceptual events, without the added ambiguity or complexity of predicting motion trajectories, attention was not needed to help resolve the perceptual uncertainty of their timing. The brief flash of a checkerboard image on the screen provides a temporally isolated event, whereas in the case of the bounce/stream stimuli, the intersection of the two disks is part of a continuously moving image. Moreover, the checkerboard was only presented on one side of the screen, reducing any location uncertainty and thus also perhaps increasing binding between the single auditory and single visual stimulus. In contrast, in Experiments 1 and 2 there was always an incomplete motion stimulus on the opposite side of the screen of the full motion target. This may have reduced the ability or speed for participants to shift their attention to the invalidly cued location in response to the onset of motion, as desired for those paradigms, while also adding to the perceptual complexity of the task.

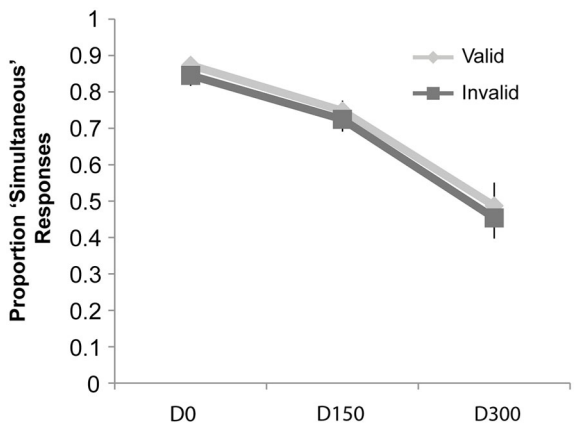


FIGURE 6 | Proportion "Simultaneous" responses for the cued simultaneity judgment task of Experiment 3. While there was a main effect of SOA, no effects of attention or an interaction of attention and validity were observed.

General Discussion

In a series of experiments, we observed three different behavioral patterns that reflect ways in which attention can interact with the TWI for audiovisual stimuli. In the first experiment, when the relative stimulus timing was not relevant for the participants' task, attention served to narrow the TWI. In the second experiment, when the task involved explicitly judging the audiovisual synchrony, attention served mainly to broaden the TWI, making the SOA function both wider and steeper. In the third experiment, also requiring explicit judgment of temporal synchrony but when simple, unambiguous, briefly presented stimuli were used in both the visual and auditory modalities, attention had no effect on the measured TWI. Although at first glance it would appear that these findings are contradictory, each result provides a piece of evidence that together demonstrate the adaptable nature of attention. Specifically, when attention acts in the context of multisensory stimuli it is able to do so in a flexible manner, enhancing that which is most relevant for a given task.

In Experiment 1, the visual stimuli were configured such that they could be perceived to bounce off of each other or stream through each other. Thus, perceptual ambiguity was created by both the local motion of the visual stimuli over a period of time, with the intersection occurring very briefly within that period, and by the presence of task-irrelevant yet perceptually influential auditory stimuli. In order to be successful at this task, attention must be implemented in such a way as to enhance accurate perception of the visual motion. When the intersection occurred on the unattended side, the participants likely did not have as clear a representation of the motion trajectory because they had not been attending to

that side in the first place, and some information about the motion trajectory may have been lost in the switch of attention to the unattended side. Additionally, attention could help by segregating or suppressing the irrelevant auditory stimulus so that it was not perceived as part of the multisensory event. In line with this interpretation, we observed fewer reports of "bouncing" when the auditory stimulus was offset in time from the visual event and attention was present, thereby narrowing the TWI, as proposed in the potential outcomes in **Figure 1A**. Further, in the unimodal (visual only) condition, having attention present also helped participants achieve the veridical perception of streaming, confirming previous findings of the importance of attention to local motion cues in the bounce/stream paradigm (Watanabe and Shimojo, 1998).

Despite the use of identical stimuli in Experiment 2, the effects of attention were quite different from those in Experiment 1. Simply switching the task from a bounce/stream judgment of the visual stimuli, where the sounds were an irrelevant distraction, to a simultaneity judgment, where the sounds were necessary for task performance, resulted in a very different pattern of effects of attention on the TWI function. With the simultaneity judgment task, attention served to cause some broadening of the temporal window by increasing the amount of temporal disparity that the multisensory inputs could have and still result in a perception of simultaneity, particularly in the D150 condition. However, the results were not completely in line with the simple "broadening" process shown in **Figure 1B**. The greatest effect here was at the SOA of 0 (i.e., where the stimulus events were actually simultaneous), in that attention served to increase the number of "simultaneous" responses reported (from ∼68% for unattended to ∼93% for attended—see **Figure 4**). Notably, this D0 effect also meant that the TWI was actually both broadened *and* steepened by attention in this experiment.

Importantly, the perceptual challenge of the task in Experiment 2 was to determine if the auditory stimulus occurred at the precise moment where the moving visual stimuli intersected. For this to be successfully done, the participant had to focus on the motion trajectory of stimuli while preparing for the auditory stimulus. Thus, the large D0 effect indicates that attention served to enhance perception in the simultaneous condition by appropriately grouping the auditory and visual events together. It is somewhat less clear why attention would serve to increase the proportion of "Simultaneous" responses in the D150 condition; one possibility, however, is that anticipatory attention builds up as the visual motion nears the point of intersection, enhancing audio-visual integration, and then tapers off after the intersection has occurred. The continued enhancement for the 150 ms delay would therefore be merely a by-product of the enhanced integration that was temporally aligned with the visual intersection that had not yet had time to dissipate.

The results of the first two experiments, together, suggest that when auditory and visual stimuli are presented such that multiple interpretations of what physically occurred are possible, attention aids in resolving this ambiguity in a flexible manner depending on task demands. These findings demonstrate that attention can act in multisensory contexts much as it does in unimodal contexts: by helping the processing of that which is relevant for the most successful, appropriate behavior. Additionally, just as in unimodal contexts, attention here did not appear to have a strict gating mechanism; that is, there was still some processing of stimuli that was the same in the presence or absence of attention (the D0 condition in Experiment 1 and the D300 condition in Experiment 2). Such low-level interactions support previous research that has suggested that multisensory integration can occur without the presence of attention (Vroomen et al., 2001), and these interactions can still fall off as the temporal separation between the stimuli increases (Meredith et al., 1987).

In Experiment 3, however, when simple, temporally discrete, brief stimuli were used in both modalities, attention appeared to have no effect on the temporal integration and segregation of multisensory stimuli. As most of the previous work on multisensory temporal processing in both humans (Spence et al., 2001; Zampini et al., 2005; Donohue et al., 2010, 2011) and animals (Meredith et al., 1987) has used simple stimuli, it is not surprising that robust multisensory interactions have still been found neurally in the absence of attention (Meredith et al., 1987). Critically, here we show that, under circumstances with very simple brief stimuli, with no motion trajectories needed to calculate, attention did not seem to be needed for, or to show any influence on, temporal coincidence judgments. It is also the case that the discrete onsets/offsets in Experiment 3 of both the auditory and the visual stimuli were more likely able to exogenously capture attention such that the judgments of the relative timing could be accomplished even when attention was not initially in place at the spatial location of the stimuli. Indeed, the difference in reaction times between the attended and unattended sides was smallest in Experiment 3, suggesting that attention was indeed more easily and rapidly shifted to the unattended side in this case, and this likely influenced the pattern of behavior we observed here. Moreover, the overall RTs were faster in the simultaneity judgment task in Experiment 3 as compared to that in Experiment 2, further suggesting that the task itself was overall easier, consistent with the view that attention would not need to play as important a role there.

Another important issue to consider here is whether the multisensory stimuli were actually being perceptually integrated in all of these experiments, or if other aspects of multisensory processing were being measured which do not necessarily involve perceptual integration. In the bounce/stream task used here in Experiment 1, the task involves reporting the perception of visual stimuli, namely whether the perception of two crossing visual stimuli is perceived as crossing or bouncing, and this perception varies as a function of the presentation of an irrelevant auditory stimulus. This influence by the auditory stimulus on the perception of the visual stimuli would thus appear to accurately reflect the perceptual integration of these multisensory inputs, and these effects have been previously interpreted in this way. Moreover, previous research has provided converging evidence that the integration of audio-visual information underlies the enhanced perception of bouncing in the bounce/stream task by demonstrating the involvement of multisensory regions in the "bounce" percept (Bushara et al., 2003; Maniglia et al., 2012) and the role that the attributes of the sound stimuli play in

these judgments (Grassi and Casco, 2009, 2010). Accordingly, the bounce/stream task would seem to be is a fitting direct measure of multisensory integration (although perhaps a somewhat indirect measure of temporal linking).

Simultaneity judgment paradigms, on the other hand, where the explicit task is to judge the temporal relationship between two events in different modalities, are a less direct measure of multisensory integration *per se*. Although there is strong evidence to suggest that when stimuli are temporally coincident, or are perceived as temporally coincident, they will tend to be integrated into a single event (see Stein and Stanford, 2008 for review), it is possible that auditory and visual stimuli could be judged as occurring at the same time without actually being perceptually integrated, just as two visual stimuli could be judged as occurring simultaneously even though they are perceived as separate and discrete perceptual events. Without other evidence to support the actual occurrence of multisensory integration in the simultaneity judgment experiments, another explanation for the discrepant findings here is that attention may play an entirely different role when two multisensory stimuli are being perceptually integrated vs. when they just temporally interact or their temporal relationship is being discriminated.

Within the multisensory integration literature, there has been some evidence to suggest that temporal binding and the perceptual integration of multisensory stimuli are different processes. Such an example can be found in the McGurk illusion, wherein a subjects is presented with a face mouthing the sound "ga" and while hearing the sound "ba", resulting in the perception of "da" (McGurk and MacDonald, 1976). In a series of studies using this illusion, (Soto-Faraco and Alsius, 2007, 2009) it was found that when participants were presented with these stimuli at various SOAs and asked to report both their percept and if the auditory and visual stimuli occurred simultaneously (or the temporal order of the stimuli), they observed that these two types of judgments did not necessarily line up. At some SOAs participants correctly identified the temporal order of the stimuli (or perceived them as temporally asynchronous) despite still perceiving the McGurk illusion, suggesting the stimuli were fused into one multisensory percept. Of course, these processes are not completely independent, given that temporal binding may often be necessary for auditory and visual events to be integrated (a moving mouth and voice would not be associated as coming from the same source if they were substantially temporally offset). Nevertheless, these previous results provide evidence that simultaneity judgment tasks (or temporal-order judgment tasks) do not always produce judgments of the temporal aspects of stimuli that directly correspond to perceptual integration of the exact same stimuli.

Moreover, the present results indicate more generally that the notion of multisensory integration vs. temporal binding is a key point that requires further research. Although it would seem to be the case that temporal binding is necessary for multisensory integration to occur, it may not be sufficient for such a process, in that it would seem that events could be temporally bound or linked without being perceptually integrated into a multisensory object. If these two processes were always identical, then the way in which attention should interact with them should also be identical; our current results, however, clearly indicate otherwise. Thus, although tasks such as simultaneity judgment tasks may be useful to assess the temporal linking of stimuli, they are likely not tapping into the multisensory integration processes in the same way that a bounce/stream task does.

Although attention modulated the functional patterns of our measures of the TWI, the absence of attention did not eliminate integration or binding. Across all experiments, integration still followed a temporal gradient, with more temporally coincident stimuli being more likely to be integrated and more temporally disparate stimuli tending to be segregated. It is not surprising, however, that some perceptual information is still getting through even on the unattended side. Classic ERP studies of visual attention show that although attention serves to enhance a sensory-evoked response to a visual stimulus, it is not that the visual stimuli in the unattended location do not evoke any neural response at all; in fact unattended stimuli clearly evoke sensory responses, albeit smaller (e.g., Voorhis and Hillyard, 1977). Thus, multisensory integration processes can occur in the absence of attention, and may tend to be fairly accurate, especially for simple, more unambiguous stimuli. With increasingly complex and ambiguous stimuli, however, attention would appear to enhance our ability to successfully integrate or segregate auditory and visual inputs as required by the task at hand. As such, our findings are not consistent with there being only one mechanism of the influence of attention on multisensory processing (**Figure 1**), but rather would appear to indicate a more complex, dynamic process.

Our pattern of results is consistent with prior work on the interactions of attention and multisensory integration, with some important expansion of previous findings. First, the simultaneity judgment task of Experiment 3, which showed that attention did not modulate the TWI is consistent with some studies in humans suggesting that attention does not modulate the ventriloquism effect (Bertelson et al., 2000) and with studies in animals showing integration in anesthetized cats (e.g., Meredith et al., 1987). Second, the stimuli which contained a motion aspect in the present study (Experiments 1 and 2) showed that attention can modulate the binding of multisensory stimuli, consistent with other studies showing that attention can modulate the BOLD response to congruent audio-visual speech (Fairhall and Macaluso, 2009), can modulate multisensory integration at early perceptual stages (~80 ms; Talsma and Woldorff, 2005), and can enhance EEG components associated with the perception of a multisensory "extra-flash" illusion (Mishra et al., 2010). Together, many of these studies have found enhancements of a neural response to multisensory stimuli in the presence of attention, and our behavioral data extend these findings suggesting that what is "enhanced" can be highly dependent on the goals and needs of the task.

Based on these results, the hypotheses put forth in the introduction (**Figure 1**) can be reformulated to reflect broader principles concerning multisensory processing and integration. One overarching principle is that the influence of attention on the multisensory TWI is flexible and can adapt to the quality and complexity of the incoming information and to the perceiver's task needs. Relatedly, the influence of attention

on the window of integration also depends on whether the measures of that window reflect true multisensory perceptual integration or reflect simply a temporal linking process. Thus, when only one modality is relevant for a task, as in Experiment 1 for example, attention will tend to narrow the temporal window over which input from an irrelevant modality will be perceptually integrated. In contrast (Experiment 2), when the task involves judging if two stimuli from the different modalities are simultaneous (and thus both are relevant), attention appears to broaden the window of integration, while also making physically simultaneous stimuli more likely to be viewed as such, thereby also resulting in a taller center of the integration window. Another major principle of the influence of attention on multisensory processing is that it tends to decrease as the quality of the sensory input increases, such that in simpler, less ambiguous situations the interactions between attention and multisensory integration processes will be smaller. Thus, with very simple, discrete stimuli, as in Experiment 3, attention will tend to not alter the perception of temporal relationships between them, likely because perception is already more than sufficient and doesn't require attentional enhancement. More generally, the influence of attention on multisensory processing will vary depending on the task-relevance of the stimulus information from the different modalities, the nature and complexity of those stimuli, and the specific task goals.

Our results suggest that multisensory integration, at least temporal judgments of integration, can sometimes be a bottom-up process operating largely independently from attention, but it can also be substantially modulated by top-down attentional goals, particularly in situations with more complex sensory input or task needs. We propose that the interactions between attention and multisensory stimuli should not be thought of as a single process operating the same way in all cases, but rather as being context- and task-dependent, providing perceptual enhancements of multisensory stimuli as needed to maximize perception and performance.

Acknowledgments

The authors thank Cameron McKay, Brittany Zullkiewicz, Jennifer Hong, and Josh Stivers for assistance with data collection. This work was supported by a grant from the U.S. National Institute for Neurological Disorders and Stroke (R01-NS051048) to MGW and by funding from Otto-von-Guericke University in Magdeburg.

References

Abrams, J., Barbot, A., and Carrasco, A. (2010). Voluntary attention increases perceived spatial frequency. *Atten. Percept. Psychophys.* 72, 1510–1521. doi: 10.3758/APP.72.6.1510

Alais, D., Newell, F. N., and Mamassian, P. (2010). Multisensory processing in review: from physiology to behaviour. *Seeing Perceiving* 23, 3–38. doi: 10.1163/187847510x488603

Bashinski, H. S., and Bacharach, V. R. (1980). Enhancement of perceptual sensitivity as the result of selectively attending to spatial locations. *Percept. Psychophys.* 28, 241–248. doi: 10.3758/bf03204380

Berger, A., Henik, A., and Rafal, R. (2005). Competition between endogenous and exogenous orienting of visual attention. *J. Exp. Psychol. Gen.* 134, 207–221. doi: 10.1037/0096-3445.134.2.207

Bertelson, P., Vroomen, J., de Gelder, B., and Driver, J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. *Percept. Psychophys.* 62, 321–332. doi: 10.3758/bf03205552

Besle, J., Fort, A., Delpuech, C., and Giard, M.-H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. *Eur. J. Neurosci.* 20, 2225–2234. doi: 10.1111/j.1460-9568.2004.03670.x

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., and Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychol. Rev.* 108, 624–652. doi: 10.1037/0033-295X.108.3.624

Bushara, K. O., Hanakawa, T., Immisch, I., Toma, K., Kansaku, K., and Hallett, M. (2003). Neural correlates of cross-modal binding. *Nat. Neurosci.* 6, 190–195. doi: 10.1038/nn993

Busse, L., Roberts, K. C., Crist, R. E., Weissman, D. H., and Woldorff, M. G. (2005). The spread of attention across modalities and space in a multisensory object. *Proc. Natl. Acad. Sci. U S A* 102, 18751–18756. doi: 10.1073/pnas.0507704102

Carrasco, M. (2011). Visual attention: the past 25 years. *Vision Res.* 51, 1484–1525. doi: 10.1016/j.visres.2011.04.012

Carrasco, M., Penpeci-Talgar, C., and Eckstein, M. (2000). Spatial covert attention increases contrast sensitivity across the CSF: support for signal enhancement. *Vision Res.* 40, 1203–1215. doi: 10.1016/s0042-6989(00)00024-9

Chen, L., and Vroomen, J. (2013). Intersensory binding across space and time: a tutorial review. *Atten. Percept. Psychophys.* 75, 790–811. doi: 10.3758/s13414-013-0475-4

Coull, J. T., and Nobre, A. C. (1998). Where and when to pay attention: the neural systems for directing attention to spatial locations and to time intervals as revealed by both PET and fMRI. *J. Neurosci.* 18, 7426–7435.

Diederich, A., and Colonius, H. (2004). Bimodal and trimodal multisensory enhancement: effects of stimulus onset and intensity on reaction time. *Percept. Psychophys.* 66, 1388–1404. doi: 10.3758/bf03195006

Donohue, S. E., Roberts, K. C., Grent-'t-Jong, T., and Woldorff, M. G. (2011). The cross-modal spread of attention reveals differential constraints for the temporal and spatial linking of visual and auditory stimulus events. *J. Neurosci.* 31, 7982–7990. doi: 10.1523/JNEUROSCI.5298-10.2011

Donohue, S. E., Woldorff, M. G., and Mitroff, S. R. (2010). Video game players show more precise multisensory temporal processing abilities. *Atten. Percept. Psychophys.* 72, 1120–1129. doi: 10.3758/APP.72.4.1120

Egly, R., Driver, J., and Rafal, R. D. (1994). Shifting visual attention between objects and locations: evidence from normal and parietal lesion subjects. *J. Exp. Psychol. Gen.* 123, 161–177. doi: 10.1037/0096-3445.123.2.161

Eimer, M., and Schröger, E. (1998). ERP effects of intermodal attention and cross-modal links in spatial attention. *Psychophysiology* 35, 313–327. doi: 10.1017/s004857729897086x

Fairhall, S. L., and Macaluso, E. (2009). Spatial attention can modulate audiovisual integration at multiple cortical and subcortical sites. *Eur. J. Neurosci.* 29, 1247–1257. doi: 10.1111/j.1460-9568.2009.06688.x

Fujisaki, W., Shimojo, S., Kashino, M., and Nishida, S. (2004). Recalibration of audiovisual simultaneity. *Nat. Neurosci.* 7, 773–778. doi: 10.1038/nn1268

Giordano, A. M., McElree, B., and Carrasco, M. (2009). On the automaticity and flexibility of covert attention: a speed-accuracy trade-off analysis. *J. Vis.* 9:30. doi: 10.1167/9.3.30

Gondan, M., Blurton, S. P., Hughes, F., and Greenlee, M. W. (2011). Effects of spatial and selective attention on basic multisensory integration. *J. Exp. Psychol. Hum. Percept. Perform.* 37, 1887–1897. doi: 10.1037/a0025635

Grassi, M., and Casco, C. (2009). Audiovisual bounce-inducing effect: attention alone does not explain why the discs are bouncing. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 235–243. doi: 10.1037/a0013031

Grassi, M., and Casco, C. (2010). Audiovisual bounce-inducing effect: when sound congruence affects grouping in vision. *Atten. Percept. Psychophys.* 72, 378–386. doi: 10.3758/APP.72.2.378

Green, J. J., Doesburg, S. M., Ward, L. M., and McDonald, J. J. (2011). Electrical neuroimaging of voluntary audiospatial attention: evidence for a

supramodal attention control network. *J. Neurosci.* 31, 3560–3564. doi: 10.1523/JNEUROSCI.5758-10.2011

Hawkins, H. L., Hillyard, S. A., Luck, S. J., Mouloua, M., Downing, C. J., and Woodward, D. P. (1990). Visual attention modulates signal detectability. *J. Exp. Psychol. Hum. Percept. Perform.* 16, 802–811. doi: 10.1037//0096-1523.16.4.802

Heron, J., Roach, N. W., Whitaker, D., and Hanson, J. V. M. (2010). Attention regulates the plasticity of multisensory timing. *Eur. J. Neurosci.* 31, 1755–1762. doi: 10.1111/j.1460-9568.2010.07194.x

Hopf, J.-M., Boehler, C. N., Luck, S. J., Tsotsos, J. K., Heinze, H. J., and Schoenfeld, M. A. (2006). Direct neurophysiological evidence for spatial suppression surrounding the focus of attention in vision. *Proc. Natl. Acad. Sci. U S A* 103, 1053–1058. doi: 10.1073/pnas.0507746103

Ikumi, N., and Soto-Faraco, S. (2014). Selective attention modulates the direction of audio-visual temporal recalibration. *PLoS One* 9:e99311. doi: 10.1371/journal.pone.0099311

Kawabe, T., and Miura, K. (2006). Effects of the orientation of moving objects on the perception of streaming/bouncing motion displays. *Percept. Psychophys.* 68, 750–758. doi: 10.3758/bf03193698

Keitel, C., Schröger, E., Saupe, K., and Müller, M. M. (2011). Sustained selective intermodal attention modulates processing of language-like stimuli. *Exp. Brain Res.* 213, 321–327. doi: 10.1007/s00221-011-2667-2

Koelewijn, T., Bronkhorst, A., and Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: a review of audiovisual studies. *Acta Psychol. (Amst)* 134, 372–384. doi: 10.1016/j.actpsy.2010.03.010

Luck, S. J., Chelazzi, L., Hillyard, S. A., and Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2 and V4 of macaque visual cortex. *J. Neurophysiol.* 77, 24–42.

Luck, S. J., Hillyard, S. A., Mouloua, M., Woldorff, M. G., Clark, V. P., and Hawkins, H. L. (1994). Effects of spatial cuing on luminance detectability: psychophysical and electrophysiological evidence for early selection. *J. Exp. Psychol. Hum. Percept. Perform.* 20, 887–904. doi: 10.1037//0096-1523.20.4.887

MacDonald, A. W., Cohen, J. D., Stenger, A., and Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science* 288, 1835–1838. doi: 10.1126/science.288.5472.1835

Mangun, G. R. (1995). Neural mechanisms of visual selective attention. *Psychophysiology* 32, 4–18. doi: 10.1111/j.1469-8986.1995.tb03400.x

Mangun, G. R., and Hillyard, S. A. (1991). Modulations of sensory-evoked brain potentials indicate changes in perceptual processing during visual-spatial priming. *J. Exp. Psychol. Hum. Percept. Perform.* 17, 1057–1074. doi: 10.1037//0096-1523.17.4.1057

Maniglia, M., Grassi, M., Casco, C., and Campana, G. (2012). The origin of the audiovisual bounce inducing effect: a TMS study. *Neuropsychologia* 50, 1478–1482. doi: 10.1016/j.neuropsychologia.2012.02.033

Marchant, J. L., and Driver, J. (2013). Visual and audiovisual effects of isochronous timing on visual perception and brain activity. *Cereb. Cortex* 23, 1290–1298. doi: 10.1093/cercor/bhs095

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Meredith, M. A., Nemitz, J. W., and Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *J. Neurosci.* 7, 3215–3229.

Meredith, M. A., and Stein, B. E. (1986). Spatial factors determine the activity of multisensory neurons in cat superior colliculus. *Brain Res.* 365, 350–354. doi: 10.1016/0006-8993(86)91648-3

Miller, J. (1982). Divided attention: evidence for coactivation with redundant signals. *Cogn. Psychol.* 14, 247–279. doi: 10.1016/0010-0285(82)90010-x

Mishra, J., Martínez, A., and Hillyard, S. A. (2010). Effect of attention on early cortical processes associated with the sound-induced extra flash illusion. *J. Cogn. Neurosci.* 22, 1714–1729. doi: 10.1162/jocn.2009.21295

Motter, B. C. (1993). Focal attention produces spatially selective processing in visual cortical areas V1, V2 and V4 in the presence of competing stimuli. *J. Neurophysiol.* 70, 909–919.

Müller, H. J., and Rabbitt, P. M. (1989). Reflexive and voluntary orienting of visual attention: time course of activation and resistance to interruption. *J. Exp. Psychol. Hum. Percept. Perform.* 15, 315–330. doi: 10.1037//0096-1523.15.2.315

Pestilli, F., Viera, G., and Carrasco, M. (2007). How do attention and adaptation affect contrast sensitivity? *J. Vis.* 7:9. doi: 10.1167/7.7.9

Posner, M. I. (1980). Orienting of attention. *Q. J. Exp. Psychol.* 32, 3–25. doi: 10.1080/00335558008248231

Posner, M. I., and Cohen, Y. (1984). "Components of visual orienting," in *Attention and Performance*, eds X. H. Bouma and D. G. Bowhuis (Hilldale, N.J.: Erlbaum), 531–556.

Quinlan, P. T., and Bailey, P. J. (1995). An examination of attentional control in the auditory modality: further evidence for auditory orienting. *Percept. Psychophys.* 57, 614–628. doi: 10.3758/bf03213267

Remijn, G. B., Ito, H., and Nakajima, Y. (2004). Audiovisual integration: an investigation of the "streaming-bouncing" phenomenon. *J. Physiol. Anthropol. Appl. Human Sci.* 23, 243–247. doi: 10.2114/jpa.23.243

Sarmiento, B. R., Shore, D. I., Milliken, B., and Sanabria, D. (2012). Audiovisual interactions depend on context of congruency. *Atten. Percept. Psychophys.* 74, 563–574. doi: 10.3758/s13414-011-0249-9

Sekuler, R., Sekuler, A. B., and Lau, R. (1997). Sound alters visual motion perception. *Nature* 385:308. doi: 10.1038/385308a0

Silver, M. A., Ress, D., and Heeger, D. J. (2007). Neural correlates of sustained spatial attention in human early visual cortex. *J. Neurophysiol.* 97, 229–237. doi: 10.1152/jn.00677.2006

Slutsky, D. A., and Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect. *Neuroreport* 12, 7–10. doi: 10.1097/00001756-200101220-00009

Soto-Faraco, S., and Alsius, A. (2007). Conscious access to the unisensory components of a cross-modal illusion. *Neuroreport* 18, 347–350. doi: 10.1097/wnr.0b013e32801776f9

Soto-Faraco, S., and Alsius, A. (2009). Deconstructing the McGurk-MacDonald illusion. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 580–587. doi: 10.1037/a0013483

Spence, C., Pavani, F., and Driver, J. (2000). Crossmodal links between vision and touch in covert endogenous spatial attention. *J. Exp. Psychol. Hum. Percept. Perform.* 26, 1298–1319. doi: 10.1037//0096-1523.26.4.1298

Spence, C., Shore, D. I., and Klein, R. M. (2001). Multisensory prior entry. *J. Exp. Psychol. Gen.* 130, 799–832. doi: 10.1037/0096-3445.130.4.799

Stein, B. E., and Stanford, T. R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nat. Rev. Neurosci.* 9, 255–266. doi: 10.1038/nrn2331

Stoppel, C. M., Boehler, C. N., Strumpf, H., Heinze, H.-J., Noesselt, T., Hopf, J.-M., et al. (2011). Feature-based attention modulates direction-selective hemodynamic activity within human MT. *Hum. Brain Mapp.* 32, 2183–2192. doi: 10.1002/hbm.21180

Talsma, D., Doty, T. J., and Woldorff, M. G. (2007). Selective attention and audiovisual integration: is attending to both modalities a prerequisite for early integration? *Cereb. Cortex* 17, 679–690. doi: 10.1093/cercor/bhk016

Talsma, D., Senkowski, D., Soto-Faraco, S., and Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends Cogn. Sci.* 14, 400–410. doi: 10.1016/j.tics.2010.06.008

Talsma, D., and Woldorff, M. G. (2005). Selective attention and multisensory integration: multiple phases of effects on the evoked brain activity. *J. Cogn. Neurosci.* 17, 1098–1114. doi: 10.1162/0898929054475172

van Atteveldt, N. M., Formisano, E., Blomert, L., and Goebel, R. (2007). The effect of temporal asynchrony on the multisensory integration of letters and speech sounds. *Cereb. Cortex* 17, 962–974. doi: 10.1093/cercor/bhl007

van Atteveldt, N., Formisano, E., Goebel, R., and Blomert, L. (2004). Integration of letters and speech sounds in the human brain. *Neuron* 43, 271–282. doi: 10.1016/j.neuron.2004.06.025

van Wassenhove, V., Grant, K., and Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia* 45, 598–607. doi: 10.1016/j.neuropsychologia.2006.01.001

Voorhis, S. V., and Hillyard, S. A. (1977). Visual evoked potentials and selective attention to points in space. *Percept. Psychophys.* 22, 54–62. doi: 10.3758/bf03206080

Vroomen, J., Bertelson, P., and de Gelder, B. (2001). The ventriloquist effect does not depend on the direction of automatic visual attention. *Atten. Percept. Psychophys.* 63, 651–659. doi: 10.3758/bf03194427

Watanabe, K., and Shimojo, S. (1998). Attentional modulation in perception of visual motion events. *Perception* 27, 1041–1054. doi: 10.1068/p271041

Watanabe, K., and Shimojo, S. (2001). When sound affects vision: effects of auditory grouping on visual motion perception. *Psychol. Sci.* 12, 109–116. doi: 10.1111/1467-9280.00319

Wegener, D., Ehn, F., Aurich, M. K., Galashan, F. O., and Kreiter, A. K. (2008). Feature-based attention and the suppression of non-relevant object features. *Vision Res.* 48, 2696–2707. doi: 10.1016/j.visres.2008.08.021

Weichselgartner, E., and Sperling, G. (1987). Dynamics of automatic and controlled visual-attention. *Science* 238, 778–780. doi: 10.1126/science.3672124

Wu, C. T., Weissman, D. H., Roberts, K. C., and Woldorff, M. G. (2007). The neural circuitry underlying the executive control of auditory spatial attention. *Brain Res.* 1134, 187–198. doi: 10.1016/j.brainres.2006.11.088

Yeshurun, Y., and Carrasco, M. (1999). Spatial attention improves performance in spatial resolution tasks. *Vision Res.* 39, 293–306. doi: 10.101/S0042-6989(98)00114-X

Zampini, M., Guest, S., Shore, D. I., and Spence, C. (2005). Audio-visual simultaneity judgments. *Percept. Psychophys.* 67, 531–544. doi: 10.3758/bf03193329

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in
INTEGRATIVE NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 11 September 2014
doi: 10.3389/fnint.2014.00070

A phonologically congruent sound boosts a visual target into perceptual awareness

Ruth Adam [1,2,3]* ***and Uta Noppeney*** [1,4]

[1] *Cognitive Neuroimaging Group, Max Planck Institute for Biological Cybernetics, Tuebingen, Germany*
[2] *Department of General Psychiatry, Center of Psychosocial Medicine, University of Heidelberg, Heidelberg, Germany*
[3] *Institute for Stroke and Dementia Research, Ludwig-Maximilian-University, Munich, Germany*
[4] *Department of Psychology, Centre for Computational Neuroscience and Cognitive Robotics, University of Birmingham, Birmingham, UK*

Edited by:
John J. Foxe, Albert Einstein College of Medicine, USA

Reviewed by:
Walter Ritter, Albert Einstein College of Medicine, USA
Kyle Elliott Mathewson, University of Illinois, USA
Joy Geng, University of California Davis, USA

***Correspondence:**
Ruth Adam, Institute for Stroke and Dementia Research, Ludwig-Maximilian-University, Max-Lebsche-Platz 30, 81377 Munich, Germany
e-mail: ruth.adam@med.uni-muenchen.de

Capacity limitations of attentional resources allow only a fraction of sensory inputs to enter our awareness. Most prominently, in the attentional blink the observer often fails to detect the second of two rapidly successive targets that are presented in a sequence of distractor items. To investigate how auditory inputs enable a visual target to escape the attentional blink, this study presented the visual letter targets T1 and T2 together with phonologically congruent or incongruent spoken letter names. First, a congruent relative to an incongruent sound at T2 rendered visual T2 more visible. Second, this T2 congruency effect was amplified when the sound was congruent at T1 as indicated by a T1 congruency × T2 congruency interaction. Critically, these effects were observed both when the sounds were presented in synchrony with and prior to the visual target letters suggesting that the sounds may increase visual target identification via multiple mechanisms such as audiovisual priming or decisional interactions. Our results demonstrate that a sound around the time of T2 increases subjects' awareness of the visual target as a function of T1 and T2 congruency. Consistent with Bayesian causal inference, the brain may thus combine (1) prior congruency expectations based on T1 congruency and (2) phonological congruency cues provided by the audiovisual inputs at T2 to infer whether auditory and visual signals emanate from a common source and should hence be integrated for perceptual decisions.

Keywords: attentional blink, audiovisual synchrony, awareness, Bayesian causal inference, crossmodal integration, multisensory integration

INTRODUCTION

In our natural multisensory environment, our sensory systems are exposed to a constant inflow of sensory signals. Yet, only a small subset of those signals reaches our perceptual awareness. Attentional selection has been proposed as a critical processing bottleneck that determines whether sensory signals enter our awareness (Pashler, 1984; Tombu et al., 2011). Since attentional resources are limited, allocation of attention to one stimulus may impair perception of other competing stimuli co-occurring close in time. In the laboratory, the attentional blink paradigm (Broadbent and Broadbent, 1987; Raymond et al., 1992) is a prime example illustrating limitations in attentional capacity for two rapidly successive stimuli (Chun and Potter, 1995; Marois et al., 2004; Shapiro et al., 2006; Adam et al., 2014). In an attentional blink paradigm, participants are impaired when reporting the second (T2) of two targets (T1 and T2) that are presented within a 500 ms interval amongst a rapid visual sequence of distractor items (Shapiro et al., 1997b; Dux and Marois, 2009 see Olson et al., 2001 for phonological material).

Several mechanisms have been suggested to account for the attentional blink (see Dux and Marois, 2009; Martens and Wyble, 2010 for review). Classical "bottleneck models" attribute the attentional blink to capacity limitations that prevent the second target from consolidation into working memory (Chun and Potter, 1995; Jolicoeur, 1998; Dux and Harris, 2007; Dell'acqua et al., 2009). However, explanations based on capacity limitations have recently been challenged by studies demonstrating that the attentional blink can be reduced by various factors such as (i) changing the allocation of attentional resources to T1, distracters or T2 (Nieuwenstein, 2006), or (ii) adding a distractor task to the attentional blink paradigm. In the latter case, participants showed less attentional blinks, when they were concurrently engaged in a distractor task such as free associating. The authors attributed this paradoxical pattern to a widening of participants' attention that allowed them to process T2 in addition to T1 (Olivers and Nieuwenhuis, 2005). Collectively, these studies suggest that the attentional blink may be a product of active attentional control that selectively allocates attention to target 1 and 2 and reduces attention to the distractor items (Di Lollo et al., 2005; Olivers and Nieuwenhuis, 2005; Nieuwenstein, 2006; Olivers et al., 2007).

While most previous research has focused on the visual modality, an attentional blink has also been demonstrated for auditory or tactile processing pointing toward fundamental processing limitations of the human cognitive system (Duncan et al., 1997; Arnell and Jolicoeur, 1999; Hillstrom et al., 2002; Dell'acqua et al., 2006; Shen and Mondor, 2006; Vachon and Tremblay, 2008;

Horvath and Burgyan, 2011). Moreover, a so-called crossmodal attentional blink has also been observed when target 1 and target 2 were presented in different modalities suggesting that at least some processing limitations or attentional control emerge at later potentially crossmodal processing stages (Arnell and Jolicoeur, 1999; Soto-Faraco et al., 2002; Arnell and Jenkins, 2004; Ptito et al., 2008; though see Duncan et al., 1997; Potter et al., 1998; Soto-Faraco and Spence, 2002; Martens et al., 2010). Likewise, a recent EEG study showed that the auditory mismatch negativity is enhanced for trials with visual attentional blink indicating that attentional resources are shared and commonly controlled across sensory modalities (Haroush et al., 2011).

Visual attention is thought to be guided by top-down biases as well as by bottom-up stimulus salience (Desimone and Duncan, 1995; Egeth and Yantis, 1997; Buschman and Miller, 2007). It is therefore not surprising that the probability of an attentional blink depends on the salience or behavioral relevance of the second stimulus. Previous studies have shown that T2 identification rate is enhanced for physically dissimilar items (Chun and Potter, 1995; Raymond et al., 1995; Maki et al., 1997; Nieuwenstein et al., 2005), the participant's own name (Shapiro et al., 1997a) and emotional stimuli (Anderson and Phelps, 2001). A more recent study has also demonstrated that an otherwise uninformative sound presented together with T2 enables T2 to escape the attentional blink (Olivers and Van Der Burg, 2008). Importantly, an increase in T2 identification rate was observed only if the brief sound was emitted simultaneously with the second target, but not when presented 100–300 ms prior to the target. This temporal profile argues against alerting as the underlying mechanism. It suggests that the salience of the visual T2 target is amplified by a concurrent sound via genuine multisensory mechanisms that depend on audiovisual co-occurrence.

Indeed, in our multisensory world the salience of stimuli should be determined by integrating inputs from all senses. Yet, when bombarded with many different signals the brain faces the challenge to integrate only signals that are generated by a common event or object, but segregate those from different events (Roach et al., 2006). Thus, multisensory integration inherently involves solving the so-called "causal inference" problem (Welch and Warren, 1980; Shams and Beierholm, 2010). In other words, the brain needs to infer whether two sensory signals are caused by common or two different events. From a Bayesian perspective, the brain may solve this causal inference problem by combining two sorts of knowledge: (i) top-down prior knowledge and (ii) bottom-up congruency cues. First, participants have prior knowledge or expectations about whether or not two sensory signals emanate from a common source. For instance, having encountered a series of congruent audiovisual signals that were caused by a common cause participants have high expectations that future auditory and visual signals are also generated by a common event. Conversely, after incongruent audiovisual signals participants will decrease (resp. increase) their congruency (resp. incongruency) expectations. Formally, these (in)congruency expectations are referred to as common source prior. Second, participants can infer whether signals are caused by common cause from "multisensory" congruency cues that are derived from the new incoming sensory signals (i.e., the likelihood of the two signals given a common source) (Ernst and Bulthoff, 2004; Kording et al., 2007; Beierholm et al., 2009; Yu et al., 2009). The brain may use multiple cues that are abstracted from the sensory inputs at multiples levels to infer whether two signals in different modalities are generated by the same event. Most prominently, sensory signals from a common source should coincide in time and space (Wallace et al., 1996, 2004; Macaluso and Driver, 2005; Van Atteveldt et al., 2007; Lewis and Noppeney, 2010; Vroomen and Keetels, 2010; Donohue et al., 2011). Likewise, higher order congruency cues that are defined in terms of semantics or phonology (e.g., syllables) can impose important constraints on multisensory integration (Laurienti et al., 2004; Van Atteveldt et al., 2004; Noppeney et al., 2008; Adam and Noppeney, 2010).

This study used a visual attentional blink paradigm to investigate how a task-irrelevant and unattended auditory signal boosts a visual signal into subjects' awareness depending on the congruency of the audiovisual (AV) signals and participants' prior congruency expectations. Specifically, in two experiments we investigated how phonologically congruent and incongruent sounds that are presented concurrently with (i.e., in synchrony) or prior to (i.e., auditory leading asynchrony) visual T1 and T2 influence subjects' T2 identification accuracy. The first experimental design factorially manipulated (1) the phonological congruency of sound 1 with T1, (2) the phonological congruency of sound 2 with T2, and (3) the lag between T1 and T2 (**Figure 1A**). After each trial, subjects reported the identity of T1, the identity of T2 and rated the visibility of T2 (invisible, unsure, visible). By contrast, the second experiment manipulated (1) the phonological congruency of sound 1 with T1, (2) the phonological congruency of sound 2 with T2, and (3) the synchrony between the sounds and the visual targets (**Figure 1C**). After each trial, subjects reported the identity of T1 and the identity of T2.

From the perspective of Bayesian causal inference, we expected an increase in T2 visibility as well as in T2 identification accuracy (i.e., a decrease in the number of attentional blinks) for phonologically congruent relative to incongruent audiovisual T2 pairs. Further, this "T2 congruency effect" should be amplified when T2 is preceded by a phonologically congruent as compared to incongruent AV T1 pair, because phonological congruency at T1 induces prior congruency expectations (i.e., a common source prior). In other words, a congruent (resp. incongruent) T1 pair will increase (resp. decrease) participant's expectations that the audiovisual signals at T2 are congruent. These prior congruency expectations will increase participants' tendency to attend to and integrate auditory and visual inputs at T2 into a unified percept resulting in an increase in accuracy for congruent trials, yet a decrease in accuracy for incongruent trials where the sound is incompatible with the visual T2 letter.

Critically, auditory, and visual signals might interact at multiple processing stages possibly implemented at different levels of the cortical hierarchy (Werner and Noppeney, 2010a,b). It is assumed that predominantly lower integration processes depend on the synchrony of the audiovisual signals, while higher order integration processes, for instance at the decisional level, are less sensitive to the precise temporal co-occurrence of the stimuli.

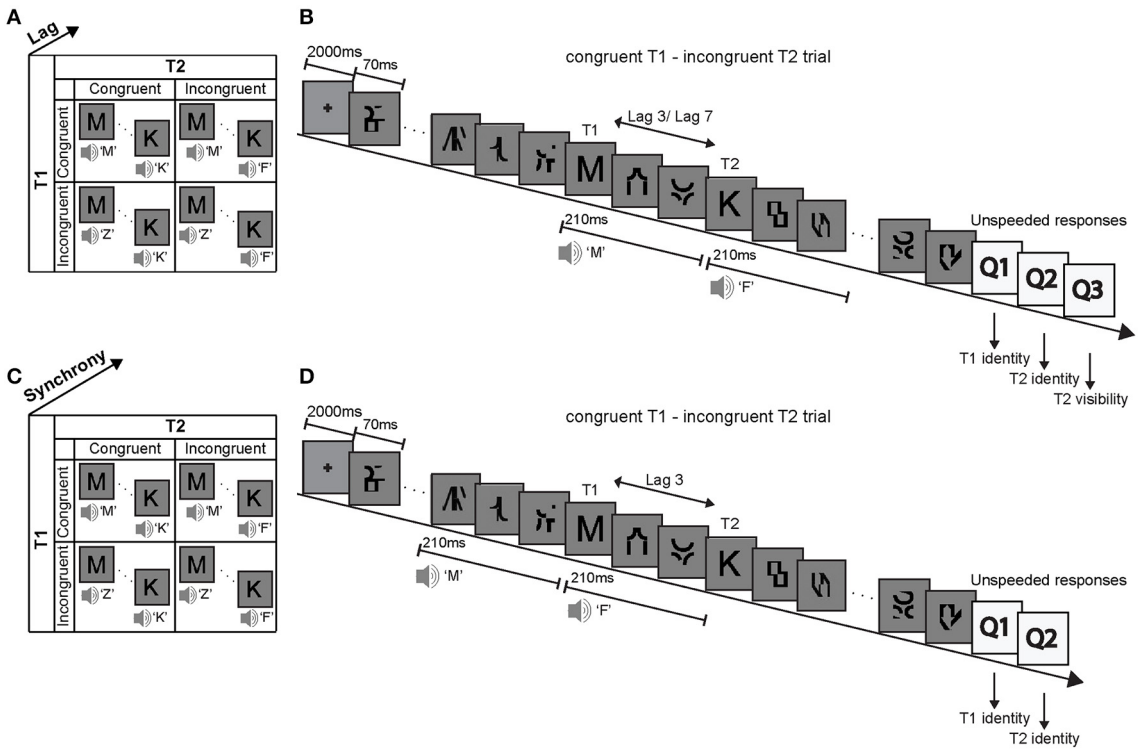


FIGURE 1 | Experimental design, example trial and stimuli. Experiment 1: **(A)** The 2 × 2 × 2 factorial design with the factors (i) T1 AV-congruency (congruent vs. incongruent), (ii) T2 AV-congruency (congruent vs. incongruent), and (iii) lag (lag 3 vs. lag 7). **(B)** Example trial and stimuli. In an audiovisual attentional blink paradigm, participants were presented with two distinct visual target letters T1 and T2 that were accompanied by congruent or incongruent spoken letter names in a series of distractor items. Participants identified visual letter targets T1 and T2 and rated the visibility of T2. Experiment 2: **(C)** The 2 × 2 × 2 factorial design with the factors (i) T1 AV-congruency (congruent vs. incongruent), (ii) T2 AV-congruency (congruent vs. incongruent), and (iii) AV synchrony (synchrony vs. auditory-leading). **(D)** Example trial and stimuli of an auditory-leading trial. The congruent or incongruent spoken letter names were presented 210 ms before the target letters onset. T1: first target, T2: second target.

Likewise, a prior sound may facilitate visual letter identification via crossmodal priming mechanisms that do not rely on audiovisual temporal co-occurrence (e.g., if a congruent spoken syllable precedes the visual target letter T2 identification may be facilitated).

To dissociate between mechanisms of multisensory interactions that differ in their temporal sensitivity, a follow-up experiment 2 manipulated the synchrony of the sound with respect to visual T1 and T2. If the sound and T1 or T2 are integrated into a unified percept via low level temporally sensitive mechanisms, the increase in letter identification due to congruent AV signals should depend on the synchrony of the audiovisual signals. The T2 identification accuracy should be reduced when the sound precedes T2. By contrast, we would expect a similar reduction in identification accuracy for both synchronous and asynchronous presentations when audiovisual interactions are mediated via priming or higher order decisional mechanisms.

Finally, as previously shown we expect an audiovisually incongruent T1 to reduce T2 identification accuracy (Van Der Burg et al., 2010), since audiovisual incongruent T1 pairs require greater processing demands and thereby decrease the attentional resources to be allocated to T2.

EXPERIMENT 1

MATERIAL AND METHODS

Subjects

Thirty seven healthy subjects (20 females, mean age 26.9 years, range 18–45 years) participated in experiment 1. All subjects had normal or corrected to normal vision and reported normal hearing. Thirty five were German native speakers.

Five subjects were excluded from the analysis because they either reported themselves to be Bulgarian native speakers and were thus less familiar with German phonology (two subjects), did not complete the experiment (one subject) or they misunderstood the task and responded almost exclusively to the sound, leading to missing values in several conditions (two subjects).

Subjects gave written informed consent prior to the study as approved by the joint human research review committee of the local ethics committee of the University of Tübingen.

Stimuli

Visual stimuli consisted of 12 targets and 12 distractors centered on a gray background (15.4 cd/m^2). Targets were capital Latin letters that were selected from two sets that were distinct for T1 (i.e., C, H, M, S, T, or Z) and T2 (i.e., F, J, K, N, P, or U). The letters were selected and grouped carefully according to the distinctiveness

of their written letters and their spoken letter names. In addition, salient and meaningful letter combinations (e.g., T1 = P and T2 = C ⇒ PC) were avoided. Distractors were meaningless symbols created by spatially shuffling the image segments of the target letters to match the mean luminance of distractors and targets.

To decrease training effects, six stimulus sets were created, each containing the same target letters presented in a different font.

Auditory stimuli (sampling rate: 44,100 Hz, stereo, 16 bits, presented at 66 dB SPL) were the spoken German letter names corresponding to the visual target letters. Each auditory waveform was contracted to 210 ms, which left the spoken letter names fully recognizable, yet shortened their presentation time to the duration of three targets in the rapid serial visual presentation (RSVP). To avoid auditory clicks, a linear ramp of 18 ms was added to the beginning and end of the sound.

Design and procedure

In a visual attentional blink paradigm, subjects were presented with two visual targets (i.e., target 1: T1, target 2: T2) in a sequence of 13 rapidly presented distractor items. The visual targets were written letters selected from two non-overlapping sets of Latin letters for T1 and T2 to avoid response interference between T1 and T2 (see Stimuli section).

On each trial targets and distractors were presented at approximately 14.3 Hz (i.e., presentation duration: 70 ms, presented at visual angle 1°) in a RSVP after an initial 2000 ms fixation period (**Figure 1B**). T1 was presented equally often at positions 3, 4, 5, and 6. T2 was presented either 3 positions after T1 (i.e., lag 3 session) or 7 positions after T1 (i.e., lag 7 session), in separate sessions.

Concurrently with T1 and T2 onsets, a spoken letter name was presented that was phonologically congruent or incongruent to the visual target letter with an equal probability of 0.5. For instance, for congruent T1, the visual target letter "C" was presented together with the spoken letter name "Ce." Conversely, for incongruent T1, the visual target letter "C" was presented for instance with the spoken letter name "Ha." The auditory sound in this paradigm is exactly 50% of the time congruent and 50% of the time incongruent. Hence, if subjects responded consistently according to the sound, they would obtain 50% accuracy when averaging across all conditions. Hence, the 2 × 2 × 2 factorial design manipulated (i) T1 AV-congruency (congruent, incongruent), (ii) T2 AV-congruency (congruent, incongruent) and (3) lag between T1 and T2 (lag 3, lag 7) (**Figure 1A**).

In a visual selective attention paradigm, participants were instructed to attend to the visual stimuli and ignore the sounds. After each trial, subjects responded to three questions as accurately as possible in an unspeeded fashion: (1) What is the identity of T1 (C, H, M, S, T, or Z)? (2) What is the identity of T2 (F, J, K, N, P, or U)?, and (3) Rate the visibility of T2 (invisible, unsure, visible). For the identification questions, subjects were instructed to make a forced choice guess, even if they could not identify the targets. They indicated their responses on a customized keyboard. The keypress for the visibility response then triggered the next trial. Thus, our experimental paradigm combined an objective (= identification accuracy) and subjective (= visibility) criterion of observer's awareness.

Each session included 30 trials per condition amounting to 120 trials in total. Please note that all trials were of the same lag in one session, so that each session included only 4 conditions, either at lag 3 or the control condition lag 7 (Maclean and Arnell, 2012). We performed lag 3 and 7 in different sessions to make our results comparable to other studies that included only one lag, as otherwise the temporal expectancies would introduce additional variance. The order of conditions was pseudo-randomized and the letter identity was randomized with each letter appearing equally often in each condition. The assignment of lag 3 and 7 trials to separate sessions was counterbalanced. During the post-experiment inquiry, only one subject reported noticing time-differences between the two lags. In total, subjects performed nine sessions, six with lag 3 resulting in 180 trials per lag 3 condition, and three sessions with lag 7 resulting in 90 trials per lag 7 condition. This substantial number of trials was required to ensure sufficient trials per condition and visibility rating. As our study focused in particular on the lag 3 trials, we included more trials for the short T1-T2 time window (lag 3) which was our main focus. In each session, the target letters were presented in a different font to minimize learning effects that reduce the number of attentional blinks. Prior to each session, subjects were familiarized with the stimuli in the particular font setting. The familiarization procedure included four repetitions of the 12 target letters accompanied by their congruent sounds while subjects pressed the keyboard-key corresponding to the visual letter. Prior to the experiment, participants performed one practice session which included two trials per condition.

Apparatus

The experiment was conducted in a dimly lit experimental room. Visual stimuli were displayed on a CRT monitor (1600 × 1200 resolution, 100 Hz refresh rate, 21″ Sony CPD-G520, Japan), approximately 56 cm from the subjects' eyes. Auditory stimuli were presented at approximately 66 dB SPL, using headphones (Sennheiser HD 555MR, Germany). Experimental sessions were presented using the Cogent 2000 v1.25 (developed by the Cogent 2000 team at the FIL and the ICN and Cogent Graphics developed by John Romaya at the LON at the Wellcome Department of Imaging Neuroscience, UCL, London, UK; http://www.vislab.ucl.ac.uk/cogent.php) running under MATLAB (Mathworks Inc., Natick, MA, USA) on a Windows PC.

Data analysis

Operationally, awareness was defined based on subjects' report at the end of the trial. In experiment 1, we employed two different reports: visual letter identification and visibility judgment. Data analysis was limited to trials where subjects correctly identified the T1 letter. In other words, all measures were contingent on T1 correctness.

We assessed observer's awareness of the T2 using two criteria (following recommendation by Dehaene and Changeux, 2011). First, in accordance with most attentional blink studies, we employed subjects' visual letter identification accuracy at T2 as an objective index of visual awareness. Critically, visual letter identification at T2 was limited to only those trials where participants correctly identified T1 (i.e., % correct T2 identification contingent on correct T1 identification: %T2|T1). Second,

we used subjects' visibility judgment (i.e., the percentage judged visible) as a subjective criterion again limited to only those trials where T1 was correctly identified (Sergent and Dehaene, 2004; Nieuwenhuis and De Kleijn, 2011). The objective index is thought to be independent of subjects' response criterion, yet may overestimate visual awareness, because subjects can perform better than chance even for stimuli they are not aware of (e.g., correct responses in blindsight; Weiskrantz et al., 1974; Persaud and Lau, 2008). Conversely, the subjective index depends on where subjects set their internal visibility criterion, yet may be more inclusive.

RESULTS AND DISCUSSION

The overall mean T1 identification accuracy (±s.e.m.) was 82.7 ± 2.3%. A 2 × 2 repeated measures ANOVA of % T1 identification accuracy with the within subject factors lag (3 vs. 7) and T1 AV-congruency (congruent vs. incongruent) revealed a T1 congruency main effect on T1 performance [$F_{(1, 31)} = 25.42$, $p < 0.001$, partial $\eta^2 = 0.451$], with reduced accuracy for incongruent (77.0 ± 3.0%) relative to congruent (88.4 ± 2.0%) AV pairs. No other effects were significant.

Objective awareness criterion: T2 identification accuracy (given T1 is correct)

The 2 (lag: 3 vs. 7) × 2 (T1 congruency: congruent vs. incongruent) × 2 (T2 congruency: congruent vs. incongruent) repeated measures ANOVA of % T2 identification accuracy (given correct identification of T1) revealed main effects of lag, T1 congruency and T2 congruency. Consistent with the well-established time-course of the attentional blink, T2 accuracy was increased for lag 7 relative to lag 3 validating our attentional blink paradigm (Raymond et al., 1992). Nevertheless, identification accuracy was still reduced even for lag 7 trials, potentially because the audiovisual T1 pairs (especially the incongruent target-sound pairs, Van Der Burg et al., 2010) are more difficult to process than the standard purely visual T1 thereby protracting the attentional blink. Further, T2 identification accuracy decreased both for incongruent T1 and incongruent T2 pairs as indicated by the two congruency main effects. In other words, fewer attentional blinks were observed when the auditory sound matched T2 (79.8 ± 2.5% for congruent vs. 67.2 ± 3.1% for incongruent T2 pair) (see **Table 1**). Yet, these main effects need to be interpreted with caution as we also observed a 3 way interaction (see below).

We also observed a significant 2-way interaction between lag x T2 congruency with greater T2 congruency effects for lag 3 vs. lag 7 [*post-hoc t*-test for lag 3: $t_{(31)} = 6.01$, $p < 0.001$, mean difference = 14.3%; *post-hoc t*-test for lag 7: $t_{(31)} = 5.35$, $p < 0.001$, mean difference = 10.9%]. Critically, there was a trend for T1 congruency × T2 congruency interaction and in particular a significant 3-way interaction. To further evaluate this 3-way interaction, we tested for the T1 congruency × T2 congruency effects separately for the two lags. These additional ANOVAs revealed a significant T1 × T2 interaction only for lag 3 [$F_{(1, 31)} = 6.84$, $p = 0.014$, partial $\eta^2 = 0.181$], but not for lag 7 [$F_{(1, 31)} = 0.1$, $p = 0.755$, partial $\eta^2 = 0.003$]. Follow up *post-hoc t*-tests on the interaction at lag 3 showed significant but stronger T2 congruency effects when T1 is congruent [$t_{(31)} = 5.13$, $p < 0.001$, mean difference = 17.3%] relative to when it is incongruent [$t_{(31)} = 6.98$, $p < 0.001$, mean difference = 11.2%]. These results demonstrate that the audiovisual T2 congruency effect is amplified for audiovisually congruent T1 pairs at lag 3 (**Figure 2**). This T1 × T2 interaction at lag 3 was hypothesized based on models of Bayesian causal inference. Basically, as participants have some tendency to integrate audiovisual signals that are close in time and space, we observe higher identification accuracy when the auditory signal provide congruent (i.e., facilitatory) relative to incongruent (i.e., interfering) information. Importantly, if T1 is congruent and participants expect T2 audiovisual signals to be congruent, audiovisual integration will be amplified at T2 leading to enhanced audiovisual T2 congruency effects.

Critically, the interpretation of this interaction remains to some extent ambiguous, as our experimental paradigm did not include any "neutral" audiovisual condition that is neither congruent nor incongruent. In fact, we would argue that a truly neutral condition does not exist. One may suggest a unisensory condition without any auditory T2 may be included as a neutral

Table 1 | Statistical results of experiment 1.

Factor	Objective reports	Subjective reports
Statistical results from the three-way ANOVAs (*df*: 1,31)		
Lag	$F = 28.24$, $p < 0.001$* partial $\eta^2 = 0.477$	$F = 15.38$, $p < 0.001$* partial $\eta^2 = 0.332$
T1 congruency	$F = 34.85$, $p < 0.001$* partial $\eta^2 = 0.529$	$F = 38.57$, $p < 0.001$* partial $\eta^2 = 0.554$
T2 congruency	$F = 35.61$, $p < 0.001$* partial $\eta^2 = 0.535$	$F = 18.15$, $p < 0.001$* partial $\eta^2 = 0.369$
T1 congruency × lag	$F = 1.41$, $p = 0.244$ partial $\eta^2 = 0.044$	$F = 0.001$, $p = 0.977$ partial $\eta^2 < 0.001$
T2 congruency × lag	$F = 6.37$, $p = 0.017$* partial $\eta^2 = 0.171$	$F = 0.48$, $p = 0.493$ partial $\eta^2 = 0.015$
T1 congruency × T2 congruency	$F = 2.92$, $p = 0.097$ partial $\eta^2 = 0.086$	$F = 6.14$, $p = 0.019$* partial $\eta^2 = 0.165$
T1 congruency × T2 congruency × lag	$F = 6.42$, $p = 0.017$* Partial $\eta^2 = 0.172$	$F = 0.64$, $p = 0.430$ partial $\eta^2 = 0.020$
Mean ± s.e.m. identification accuracy and visibility judgment (given T1 correct) in the 8 conditions		
T1 congruent & T2 congruent & lag 3	0.80 ± 0.03	0.49 ± 0.05
T1 congruent & T2 incongruent & lag 3	0.62 ± 0.04	0.42 ± 0.04
T1 incongruent & T2 congruent & lag 3	0.73 ± 0.03	0.43 ± 0.05
T1 incongruent & T2 incongruent & lag 3	0.62 ± 0.03	0.38 ± 0.05
T1 congruent & T2 congruent & lag 7	0.86 ± 0.02	0.57 ± 0.05
T1 congruent & T2 incongruent & lag 7	0.75 ± 0.03	0.51 ± 0.04
T1 incongruent & T2 congruent & lag 7	0.81 ± 0.03	0.50 ± 0.05
T1 incongruent & T2 incongruent & lag 7	0.70 ± 0.03	0.47 ± 0.04

*$p < 0.05$.

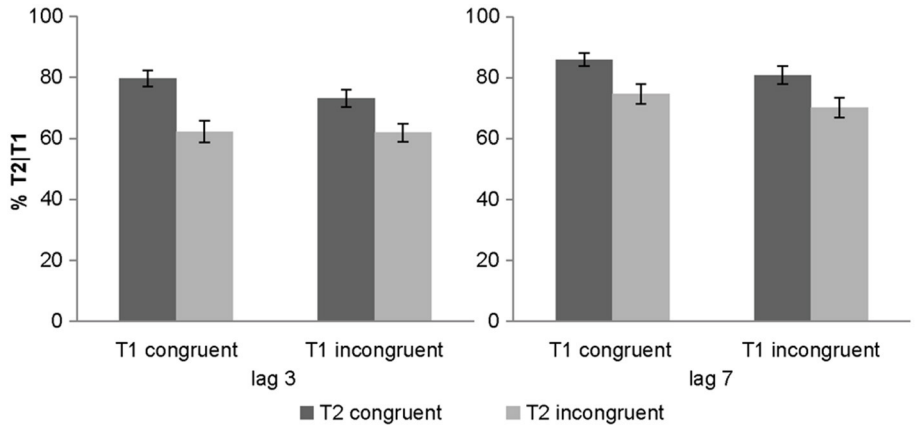


FIGURE 2 | Objective awareness criterion in experiment 1. T2 identification accuracy (% T2 correct conditional on T1 correct) (across subjects' mean ± s.e.m.) for the 8 different conditions.

condition. However, a previous study demonstrated that even a simple beep changes the attentional processing at T2 (Olivers and Van Der Burg, 2008). Likewise, a "beep" is not an ideal "neutral" control condition, as it differs in sound complexity and cognitive processing demands from the spoken syllables. Hence, it seems difficult or even impossible to generate a neutral condition that is neither congruent nor incongruent and yet tightly matched to the spoken syllables in terms of processing demands (e.g., phonemic recognition etc.). The absence of a neutral condition makes the interpretation of participant's response profile ambiguous.

At first sight, the accuracy profile for lag 3 conditions in **Figure 2** may suggest that T1 congruency increases the accuracy on T2 congruent trials without reducing the accuracy on T2 incongruent trials. In other words, T1 congruency only facilitates identification of congruent T2 without inducing interference for incongruent T2 trials. This would be a surprising finding because from the perspective of Bayesian causal inference, we would expect T1 congruency to increase participants' congruency expectations and hence their tendency to integrate audiovisual signals at T2 irrespective of T2 congruency. Enhanced audiovisual integration at T2 should then lead to both an increase in accuracy for congruent T2 pairs (= AV facilitation) and a decrease in accuracy for incongruent T2 pairs (= AV interference).

Yet, we may also explain this response profile by assuming that incongruent T1 pairs exert two distinct effects. First, as previously suggested, incongruent T1 should place more demands on processing and therefore generally decrease T2 accuracy for both congruent and incongruent T2 signals (Van Der Burg et al., 2010). Second, as described above incongruent T1 signals should also make subjects less likely to integrate AV signals at T2 again regardless of their congruency. This second mechanisms should then lead to a decrease in accuracy for congruent T2 signals and an increase in accuracy for incongruent T2 signals (by reducing the interference from the incongruent auditory signal at T2). Thus, T1 (in)congruency would have opposite effects on processing incongruent T2 signals via those to mechanisms; yet, T1 (in)congruency would have the same effect on congruent T2 signals. Assuming that T1 (in)congruency influences T2 processing concurrently via both mechanisms, the T1 (in)congruency effect on incongruent T2 signals may be canceled out.

In conclusion, a combination of a general main effect of T1 (in)congruency (i.e., incongruent relative to congruent T1 signals decrease accuracy for both T2 congruent and incongruent trials) and an interaction between T1 × T2 congruency (i.e., incongruent relative to congruent T1 signals decrease accuracy for congruent T2 and increase accuracy for incongruent T2 trials) may then induce an accuracy profile where T1 congruency apparently leads only to a facilitation for congruent T2, but no interference for incongruent T2 trials (i.e., no decrease in accuracy for incongruent relative to congruent T1 on incongruent T2 trials).

To further investigate whether T1 congruency influences the audiovisual binding of incongruent T2 pairs, we therefore analyzed subjects' error responses on T2 incongruent trials. The basic hypothesis was that if audiovisual T1 congruency induces a congruency prior that generally increases the binding of audiovisual signals at T2, subjects should more frequently misidentify T2 according to the spoken letter name, when T1 is congruent relative to incongruent.

Hence, we computed the fraction of T2 incongruent trials where subjects reported the identity of the spoken letter name rather than an unrelated letter name. A 2 (lag: 3 vs. 7) × 2 (T1 congruency: congruent vs. incongruent) repeated measures ANOVA on the fraction of trials in which the spoken letter name was reported out of all incorrect trials revealed a significant main effect of T1 congruency (**Table 2**). More specifically, the identity of the spoken letter name was more frequently reported when the trial started with a congruent T1 (42.6 ± 3.6%) relative to an incongruent T1 (36.3 ± 2.4%). This is in line with the prediction of Bayesian causal inference where prior congruency expectations will increase audiovisual interference if the two signals are incongruent.

Table 2 | Reports according to sound in experiment 1: statistical results from the Two-Way ANOVA.

Factor (*df*: 1, 31)	
Lag	$F = 0.04$, $p = 0.841$ partial $\eta^2 = 0.001$
T1 congruency	$F = 5.64$, $p = 0.024$* partial $\eta^2 = 0.154$
T1 congruency × Lag	$F = 0.63$, $p = 0.433$ partial $\eta^2 = 0.020$

*$p < 0.05$.

Subjective awareness criterion: visibility judgment (given T1 correct)

Percentage of T2 targets judged visibly was used as a complementary subjective measure of awareness. The 2 (lag: 3 vs. 7) × 2 (T1 congruency: congruent vs. incongruent) × 2 (T2 congruency: congruent vs. incongruent) repeated measures ANOVA of % judged visible revealed a significant main effect of T1 congruency, T2 congruency and lag. T2 visibility was increased for congruent T1, congruent T2 and lag 7 (see **Table 1**). Furthermore, there was a significant interaction between T1 and T2 congruency. Follow up *post-hoc t*-tests on the T2 congruency effects for visibility judgments showed significant but stronger T2 congruency effects when T1 is congruent [$t_{(31)} = 4.01$, $p < 0.001$, mean difference = 6.5%] relative to when it is incongruent [$t_{(31)} = 3.88$, $p = 0.001$, mean difference = 3.6%]. In other words, T2 target visibility was enhanced for congruent relative to incongruent T2 pairs, and this T2 congruency effect was enhanced by congruent T1 pairs (**Figure 3**). Importantly, even though the objective and subjective awareness indices showed some small differences in results pattern (e.g., 3-way interaction only for objective index), they both converged in showing an interaction between T1 and T2 congruency at least for short lag as expected under Bayesian causal inference.

EXPERIMENT 2

MATERIAL AND METHODS

The second experiment investigated whether the congruency effects that we observed in the first experiment for lag 3 were dependent on audiovisual synchrony. Thus, the experimental paradigm was basically identical to the first experiment apart from the following modifications:

Subjects

16 healthy subjects participated in the second experiment (11 females, mean age 25.1 years, range 19–30 years). As experiment 2 was partly a replication of experiment 1 and we could therefore use directed tests based on strong a priori hypotheses, we included fewer subjects in this experiment. One subject was excluded due to problems with the setup, resulting 15 subjects in the final analysis. All subjects were German native speakers, had normal or corrected to normal vision and reported normal hearing.

Design and procedure

The 2 × 2 × 2 factorial design manipulated (i) T1 AV-congruency (congruent, incongruent), (ii) T2 AV-congruency (congruent, incongruent), and (iii) AV synchrony (synchronous, auditory-leading) (**Figure 1C**).

In a visual attentional blink paradigm, subjects were presented with T1 and T2 embedded in a sequence of 13 rapidly presented distractor items. T1 was presented equally often at positions 5, 6, 7, and 8. In this way we avoided presenting the sounds in synchrony with distractor one in the asynchronous auditory-leading case. T2 was always presented at lag 3 where most attentional blinks occur. As in experiment 1, a spoken letter name was played together with T1 and T2 onset in synchronous trials. In the auditory-leading condition, the sound onset was 210 ms prior to the target presentation. Thus, in auditory-leading trials, the T1 sound onset was synchronous with a distractor and the T2 sound onset was synchronous with the presentation of visual T1 (**Figure 1D**). If the effect of the sounds on visual identification is strictly dependent on audiovisual synchrony, the presentation of the 2nd sound in synchrony with T1 should induce an incongruency effect irrespective of T2 congruency. Hence, the observation of a T1 × T2 congruency interaction despite this design choice would point toward neural mechanisms that do not strictly depend on audiovisual synchrony. However, the effect of the spoken T2 syllables on T1 identification may be minimal, because T1 and T2 were selected two distinct stimulus sets.

As the subjective and objective indices of awareness provided basically equivalent results in experiment 1, experiment 2 focused only on the objective awareness index that is traditionally used in attentional blink paradigms. Thus, after each trial, participants were asked only to report: (1) What is the identity of T1? (2) What is the identity of T2?

Subjects performed four sessions for synchronous and four sessions for asynchronous audiovisual presentations amounting to 120 trials per condition. Audiovisual synchrony was manipulated across sessions in order to control for temporal expectancies and make the results comparable across our two experiments. The order of the audiovisual synchrony sessions was pseudo-randomized. Prior to the experiment, participants performed one practice session which included two trials per condition.

Apparatus

The experiment was conducted in a dimly lit cubicle. Visual stimuli were displayed on a LCD monitor (1600 × 12000 resolution, 60 Hz refresh rate, 20.1″, DELL 2007FP, US), placed approximately 56 cm from the subjects' eyes.

RESULTS AND DISCUSSION

The overall mean T1 identification accuracy was 82.04 ± 3.7%.A 2 × 2 repeated measures ANOVA of % T1 identification accuracy with the factors AV synchrony (synchronous vs. auditory-leading) and T1 AV-congruency (congruent vs. incongruent) revealed a main effect of T1 congruency [$F_{(1, 14)} = 8.03$, $p = 0.013$, partial $\eta^2 = 0.365$], with deceased accuracy for incongruent relative to congruent stimuli (88.5 ± 4.0% for congruent and, 75.6 ± 4.7% accuracy for incongruent T1). No other effects were significant.

Objective awareness criterion: T2 identification accuracy (given T1 is correct)

The 2 (AV synchrony: synchronous vs. auditory-leading) × 2 (T1 congruency: congruent vs. incongruent) × 2 (T2 congruency:

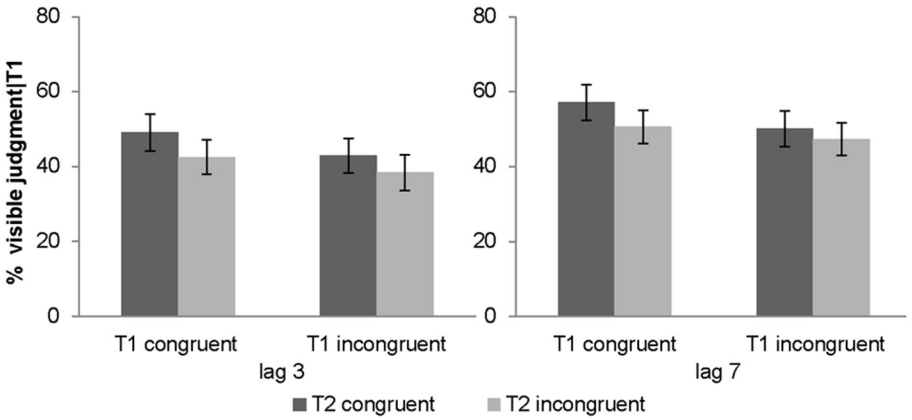


FIGURE 3 | Subjective awareness criterion in experiment 1 (visibility judgment). Percentage of visible targets given T1 correct (across subjects' mean ± s.e.m.) for the 8 different conditions.

Table 3 | Statistical results of experiment 2.

Factor	Objective reports
Statistical results from the three-way ANOVA (*df*: 1,14)	
Synchrony	$F = 0.19$, $p = 0.669$ partial $\eta^2 = 0.013$
T1 congruency	$F = 4.00$, $p = 0.065$^ partial $\eta^2 = 0.222$
T2 congruency	$F = 16.16$, $p = 0.001$* partial $\eta^2 = 0.536$
T1 congruency × synchrony	$F < 0.1$, $p = 1.000$ partial $\eta^2 = 0.00$
T2 congruency × synchrony	$F = 0.31$, $p = 0.587$ partial $\eta^2 = 0.022$
T1 congruency × T2 congruency	$F = 4.23$, $p = 0.059$^ partial $\eta^2 = 0.232$
T1 congruency × T2 congruency × synchrony	$F = 0.55$, $p = 0.819$ partial $\eta^2 = 0.004$
Mean ± s.e.m. identification accuracy (given T1 correct) in the 8 conditions	
T1 congruent & T2 congruent & synchronous	0.82 ± 0.04
T1 congruent & T2 incongruent & synchronous	0.57 ± 0.06
T1 incongruent & T2 congruent & synchronous	0.75 ± 0.05
T1 incongruent & T2 incongruent & synchronous	0.59 ± 0.05
T1 congruent & T2 congruent & auditory-leading	0.82 ± 0.04
T1 congruent & T2 incongruent & auditory-leading	0.55 ± 0.07
T1 incongruent & T2 congruent & auditory-leading	0.75 ± 0.05
T1 incongruent & T2 incongruent & auditory-leading	0.57 ± 0.06

*$p < 0.05$, ^$p < 0.10$.

congruent vs. incongruent) repeated measures ANOVA of % T2 accuracy indicated a significant main effect of T2 congruency and a trend for main effect of T1 congruency ($p = 0.065$). In line with experiment 1, T2 identification accuracy decreased for incongruent T2 pairs (78.6 ± 4.2%, 57.1 ± 5.8% accuracy for congruent and incongruent T2, respectively) (see **Table 3**).

Importantly, there was a trend for a two way T1 congruency × T2 congruency interaction ($p = 0.059$). Experiment 1 demonstrated an interaction between T1 × T2 congruency which serves as a directed a priori hypothesis for experiment 2. Hence, based on this a priori hypothesis, we could test for a directed interaction resulting in a p-value = 0.03. As in experiment 1, T1 congruency amplified the congruency effect of T2 for both synchronous and asynchronous conditions (**Figure 4**). *Post-hoc* t-tests on the T2 congruency effects showed significant but stronger T2 congruency effects when T1 is congruent [$t_{(14)} = 4.05$, $p < 0.001$, mean difference = 26.3%] relative to when it is incongruent [$t_{(14)} = 3.31$, $p = 0.005$, mean difference = 16.8%].

In summary, experiment 2 replicated the effects we observed in experiment 1 for both synchronous and asynchronous (i.e., auditory-leading) conditions. The slightly less significant effects are most likely due to smaller number of subjects included in experiment 2. Note, however, that the magnitude of the difference between the congruent and the incongruent conditions was larger compared to the one observed in experiment 1. Importantly, we did not observe any interactions between synchrony and T1 or T2 congruency indicating that the congruency effects do not always rely critically on the synchrony of the audiovisual signals. Collectively, these results suggest that a sound can boost the visual target into awareness also via mechanisms that do not critically depend on audiovisual timing (e.g., audiovisual priming in the asynchronous condition or interactions at the decisional level).

GENERAL DISCUSSION

In our natural environment our senses are constantly bombarded by many different signals with only a small fraction of them entering our awareness (Raymond et al., 1992; Simons and Chabris, 1999; Sergent et al., 2005; Pourtois et al., 2006). This study investigated how the brain selects visual signals for conscious perception. Specifically, we examined whether the awareness of visual signals is influenced by auditory signals. Using the attentional blink paradigm, we demonstrate that spoken syllables boost visual letters into subjects' awareness depending on audiovisual congruency and subjects' prior congruency expectations. As the

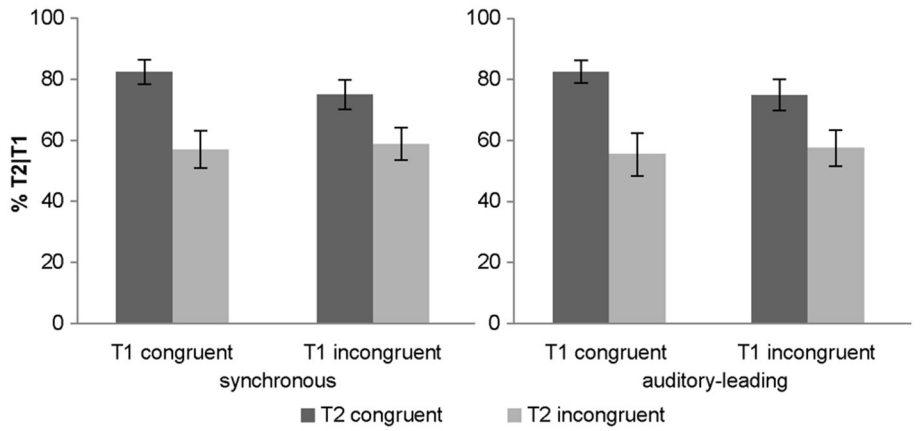


FIGURE 4 | Objective awareness criterion in experiment 2. T2 identification accuracy (% T2 correct conditional on T1 correct) (across subjects' mean ± s.e.m.) for the 8 different conditions.

audiovisual congruency effects did not always rely critically on audiovisual synchrony, they may be mediated potentially via multiple mechanisms such as audiovisual binding, crossmodal priming or even interference/facilitation at the decisional level.

Our results suggest that audiovisual interactions play a critical role in shaping visual awareness as measured by participants' accuracy in the letter identification task and subjective visibility judgments. Previous research into perceptual awareness has focused primarily on signals from one sensory domain. Most prominently, visual, auditory and tactile signals were shown to evade conscious perception when presented in a rapid stream of distractor items (Sergent and Dehaene, 2004; Dell'acqua et al., 2006; Horvath and Burgyan, 2011). Yet, the question whether sensory signals are selected for awareness independently for each sensory modality or interactively across the senses remains open (see related research on multistability and rivalry in a multisensory context: van Ee et al., 2009; Conrad et al., 2010, 2012, 2013; Lunghi et al., 2010, 2014). In the latter case, auditory signals may influence subjects' visual awareness via several multisensory mechanisms.

To investigate whether and how auditory signals modulate subjects' visual awareness, we presented the written T1 and T2 letters together with spoken letter names in an attentional blink paradigm (Raymond et al., 1992). The spoken letter names were either congruent or incongruent with respect to the written T1 and T2 letters. As congruent and incongruent spoken letter names were presented with equal probability, subjects that relied solely on the spoken letter names for making their decision should obtain 50% accuracy averaged across all conditions. In the following, we will first discuss the main effects of T1 and T2 congruency on identification accuracy and then the critical interaction between T1 and T2 congruency within the framework of Bayesian Causal Inference.

First, we demonstrate that incongruent T1 pairs decreased both T1 identification accuracy and T2 identification accuracy in particular for congruent audiovisual T2 signals (for related findings see Van Der Burg et al., 2010). Thus, audiovisually incongruent T1 pairs place greater processing demands at T1 and thereby reduce the attentional resources available for T2 processing resulting in decreased performance (Visser, 2007; Giesbrecht et al., 2009; Burt et al., 2011).

Second and more importantly, we investigated the effect of audiovisual congruency at T2 on visual awareness. From the perspective of Bayesian causal inference, audiovisual congruency is an important cue informing the brain whether visual and auditory signals are generated by a common source and should hence be combined for a perceptual decision or even integrated into a unified percept (Roach et al., 2006; Shams and Seitz, 2008). Hence, we expected audiovisual congruency at T2 to facilitate audiovisual processing, which in turn should enable recognition of visual signals. Indeed, subjects were more likely to report the correct written T2 letter, when it was presented together with a congruent spoken letter name. Convergent results were provided by the subjective criterion of awareness, i.e. the visibility judgment of T2 letter. Critically, this subjective criterion of awareness showed the same profile across conditions with an increase in visibility for audiovisually congruent relative to incongruent T2. This increase in stimulus perceptibility for congruent relative to incongruent T2 targets suggests that auditory signals influence visual awareness. Next, we investigated whether audiovisual facilitation relies strictly on audiovisual synchrony as would be expected for low level automatic integration processes. Yet in contrast to this conjecture, experiment 2 demonstrated that a prior sound that preceded the visual target by 210 ms induced a similar increase in letter identification. These results suggest that the facilitation of T2 identification in the attentional blink paradigm does not necessitate time-sensitive audiovisual integration mechanisms. Instead, several mechanisms may be involved in mediating the facilitation induced by a prior congruent relative to an incongruent sound. Most prominently, a prior congruent sound (e.g., in the context of asynchronous presentation) may facilitate T2 identification via mechanisms of audiovisual (i.e., crossmodal) priming. Alternatively, auditory and visual signals may interact at higher processing levels that are less constrained by temporal co-occurrence.

In the next step, we examined whether audiovisual congruencies at T1 and T2 interact as predicted by Bayesian causal inference where a top-down congruency prior is combined with bottom-up congruency cues derived from new sensory signals to infer whether two sensory signals should be integrated. Indeed, a congruent T1 pair amplified the increase in visibility and performance accuracy for congruent relative to incongruent T2 pairs both for synchronous and auditory-leading presentation.

Conversely, subjects responded more frequently according to the spoken letter name, when incongruent T2 pairs were preceded by a congruent T1 pair. In other words, subjects' response was more strongly influenced by the incongruent auditory letter name in trials that started with a congruent T1. Thus, in line with Bayesian causal inference, a congruent T1 pair induces observers to form a congruency prior, i.e., the prior expectation that subsequent auditory and visual signals pertain to the same event and should hence be integrated. The congruency expectations then in turn enhance audiovisual interactions at T2 leading to greater benefits for congruent T2 pairs (facilitation) and/or audiovisual interference for incongruent T2 pairs. As our study did not include any neutral condition, these two aspects (i.e., interference for incongruent or facilitation for congruent audiovisual signals) cannot be distinguished. Collectively, our results suggest that participants combine prior congruency expectations (formed on the basis of T1) with incoming phonological congruency cues (provided by T2) to determine whether auditory and visual signals should be combined for perceptual decisions. In the congruent case, audiovisual interactions boost visual signals into awareness leading to higher identification accuracy and visibility. Conversely, in the incongruent case, they lead to audiovisual interference. Importantly, these audiovisual congruency effects were observed for both audiovisual synchronous and auditory-leading presentations suggesting that the audiovisual interactions emerge potentially via several mechanisms at least some of which do not critically rely on temporal synchrony such as crossmodal priming in the asynchronous conditions.

Yet, as a cautionary note we should add that awareness in this and many other paradigms is operationally defined based on whether or not participants are able to correctly report T2 letter identity at the end of the trial. Hence, as an alternative explanatory mechanism audiovisual integration may not facilitate awareness *per se*, but stabilize memory representations such that they are more reportable at the end of the trial. This alternative mechanism may be further investigated in paradigms that also manipulate the delay between audiovisual stimulation and report of target identity.

Collectively, our results demonstrate that audiovisual interactions may affect perceptual awareness in attentional blink paradigms at multiple levels. First, audiovisual integration or priming (in the asynchronous case) mechanisms (Soto-Faraco et al., 2004; Lewis and Noppeney, 2010; Talsma et al., 2010; Werner and Noppeney, 2010a) may boost the bottom-up salience of the visual stimulus thereby facilitating perceptual awareness. As awareness in the attentional blink paradigm is closely related to attentional selection, some of these mechanisms may act preattentively. Second, audiovisual interactions may influence perceptual decision mechanisms as previously described in audiovisual congruency manipulations (Adam and Noppeney, 2010; Conrad et al., 2010; Noppeney et al., 2010; Werner and Noppeney, 2010a; Hsiao et al., 2012), Stroop (Banich et al., 2000; MacDonald et al., 2000; Kane and Engle, 2003; Egner and Hirsch, 2005; Egner, 2007) and flanker (Gratton et al., 1992; Botvinick et al., 1999; Lavie et al., 2003; Egner, 2007; Yu et al., 2009) tasks. Audiovisual interactions at all stages ranging from audiovisual integration or priming in the absence of awareness to decisional processes may be governed by Bayesian causal inference (Kording et al., 2007; Yu et al., 2009) as normative computational principles that enable optimal perception of the environment. Bayesian causal inference normatively describes the computational principles that the brain should use to determine whether or not to combine information from multiple sources in processes that range from low level automatic audiovisual interactions to higher order perceptual decisions. The brain may determine whether sensory signals should interact or be segregated by combining prior congruency information (based on T1) and incoming sensory evidence (T2).

Future neuroimaging studies (e.g., fMRI, EEG, MEG) are needed to track and dissociate the neural processes underlying multisensory interactions at multiple levels of the processing hierarchy throughout unaware and aware processing stages. For instance, prior congruency expectations may affect multisensory integration through modulatory activity in the left prefrontal cortex that has previously been implicated in cognitive control (Kerns et al., 2004; Rushworth et al., 2004; Brown and Braver, 2005; Carter and Van Veen, 2007; Orr and Weissman, 2009). Thus, in the Stroop color-naming task (naming the ink-color of a color word), prior incongruent trials increased inferior frontal sulcus (IFS) activation and top-down modulation which in turn reduced interference from irrelevant and incongruent information on subsequent trials (Kerns et al., 2004). Conversely, different types of incongruency relationships may be processed at distinct levels of the cortical hierarchy including temporal congruency at the primary cortical level (e.g., Noesselt et al., 2007; Lewis and Noppeney, 2010; Lee and Noppeney, 2014) and phonological or semantic congruency at higher order association areas (Ojanen et al., 2005; Pekkola et al., 2006; Von Kriegstein and Giraud, 2006; Hein et al., 2007; Van Atteveldt et al., 2007; Adam and Noppeney, 2010; Yoncheva et al., 2010).

ACKNOWLEDGMENTS

We would like to thank Karin Bierig and Isabelle Bülthoff for their help with the second experiment. This research was supported by the Max-Planck-Gesellschaft and an ERC starter grant.

REFERENCES

Adam, R., and Noppeney, U. (2010). Prior auditory information shapes visual category-selectivity in ventral occipito-temporal cortex. *Neuroimage* 52, 1592–1602. doi: 10.1016/j.neuroimage.2010.05.002

Adam, R., Schonfelder, S., Forneck, J., and Wessa, M. (2014). Regulating the blink: cognitive reappraisal modulates attention. *Front. Psychol.* 5:143. doi: 10.3389/fpsyg.2014.00143

Anderson, A. K., and Phelps, E. A. (2001). Lesions of the human amygdala impair enhanced perception of emotionally salient events. *Nature* 411, 305–309. doi: 10.1038/35077083

Arnell, K. M., and Jenkins, R. (2004). Revisiting within-modality and cross-modality attentional blinks: effects of target-distractor similarity. *Percept. Psychophys.* 66, 1147–1161. doi: 10.3758/BF03196842

Arnell, K. M., and Jolicoeur, P. (1999). The attentional blink across stimulus modalities: evidence for central processing limitations. *J. Exp. Psychol. Hum. Percept. Perform.* 25, 630–648. doi: 10.1037/0096-1523.25.3.630

Banich, M. T., Milham, M. P., Atchley, R., Cohen, N. J., Webb, A., Wszalek, T., et al. (2000). fMRI studies of Stroop tasks reveal unique roles of anterior and posterior brain systems in attentional selection. *J. Cogn. Neurosci.* 12, 988–1000. doi: 10.1162/08989290051137521

Beierholm, U. R., Quartz, S. R., and Shams, L. (2009). Bayesian priors are encoded independently from likelihoods in human multisensory perception. *J. Vis.* 9, 23.1–23.9. doi: 10.1167/9.5.23

Botvinick, M., Nystrom, L. E., Fissell, K., Carter, C. S., and Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature* 402, 179–181. doi: 10.1038/46035

Broadbent, D. E., and Broadbent, M. H. (1987). From detection to identification: response to multiple targets in rapid serial visual presentation. *Percept. Psychophys.* 42, 105–113. doi: 10.3758/BF03210498

Brown, J. W., and Braver, T. S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science* 307, 1118–1121. doi: 10.1126/science.1105783

Burt, J. S., Howard, S., and Falconer, E. K. (2011). T1 difficulty affects the AB: manipulating T1 word frequency and T1 orthographic neighbor frequency. *Atten. Percept. Psychophys.* 73, 751–765. doi: 10.3758/s13414-010-0054-x

Buschman, T. J., and Miller, E. K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science* 315, 1860–1862. doi: 10.1126/science.1138071

Carter, C. S., and Van Veen, V. (2007). Anterior cingulate cortex and conflict detection: an update of theory and data. *Cogn. Affect. Behav. Neurosci.* 7, 367–379. doi: 10.3758/CABN.7.4.367

Chun, M. M., and Potter, M. C. (1995). A two-stage model for multiple target detection in rapid serial visual presentation. *J. Exp. Psychol. Hum. Percept. Perform.* 21, 109–127. doi: 10.1037/0096-1523.21.1.109

Conrad, V., Bartels, A., Kleiner, M., and Noppeney, U. (2010). Audiovisual interactions in binocular rivalry. *J. Vis.* 10, 27. doi: 10.1167/10.10.27

Conrad, V., Kleiner, M., Bartels, A., Hartcher O'brien, J., Bulthoff, H. H., and Noppeney, U. (2013). Naturalistic stimulus structure determines the integration of audiovisual looming signals in binocular rivalry. *PLoS ONE* 8:e70710. doi: 10.1371/journal.pone.0070710

Conrad, V., Vitello, M. P., and Noppeney, U. (2012). Interactions between apparent motion rivalry in vision and touch. *Psychol. Sci.* 23, 940–948. doi: 10.1177/0956797612438735

Dehaene, S., and Changeux, J. P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron* 70, 200–227. doi: 10.1016/j.neuron.2011.03.018

Dell'acqua, R., Jolicoeur, P., Luria, R., and Pluchino, P. (2009). Reevaluating encoding-capacity limitations as a cause of the attentional blink. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 338–351. doi: 10.1037/a0013555

Dell'acqua, R., Jolicoeur, P., Sessa, P., and Turatto, M. (2006). Attentional blink and selection in the tactile domain. *Eur. J. Cogn. Psychol.* 18, 537–559. doi: 10.1080/09541440500423186

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* 18, 193–222. doi: 10.1146/annurev.ne.18.030195.001205

Di Lollo, V., Kawahara, J., Shahab Ghorashi, S. M., and Enns, J. T. (2005). The attentional blink: resource depletion or temporary loss of control? *Psychol. Res.* 69, 191–200. doi: 10.1007/s00426-004-0173-x

Donohue, S. E., Roberts, K. C., Grent-'T-Jong, T., and Woldorff, M. G. (2011). The cross-modal spread of attention reveals differential constraints for the temporal and spatial linking of visual and auditory stimulus events. *J. Neurosci.* 31, 7982–7990. doi: 10.1523/JNEUROSCI.5298-10.2011

Duncan, J., Martens, S., and Ward, R. (1997). Restricted attentional capacity within but not between sensory modalities. *Nature* 387, 808–810. doi: 10.1038/42947

Dux, P. E., and Harris, I. M. (2007). On the failure of distractor inhibition in the attentional blink. *Psychon. Bull. Rev.* 14, 723–728. doi: 10.3758/BF03196828

Dux, P. E., and Marois, R. (2009). The attentional blink: a review of data and theory. *Atten. Percept. Psychophys.* 71, 1683–1700. doi: 10.3758/APP.71.8.1683

Egeth, H. E., and Yantis, S. (1997). Visual attention: control, representation, and time course. *Annu. Rev. Psychol.* 48, 269–297. doi: 10.1146/annurev.psych.48.1.269

Egner, T. (2007). Congruency sequence effects and cognitive control. *Cogn. Affect. Behav. Neurosci.* 7, 380–390. doi: 10.3758/CABN.7.4.380

Egner, T., and Hirsch, J. (2005). Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. *Nat. Neurosci.* 8, 1784–1790. doi: 10.1038/nn1594

Ernst, M. O., and Bulthoff, H. H. (2004). Merging the senses into a robust percept. *Trends Cogn. Sci.* 8, 162–169. doi: 10.1016/j.tics.2004.02.002

Giesbrecht, B., Sy, J. L., and Lewis, M. K. (2009). Personal names do not always survive the attentional blink: behavioral evidence for a flexible locus of selection. *Vision Res.* 49, 1378–1388. doi: 10.1016/j.visres.2008.02.013

Gratton, G., Coles, M. G., and Donchin, E. (1992). Optimizing the use of information: strategic control of activation of responses. *J. Exp. Psychol. Gen.* 121, 480–506. doi: 10.1037/0096-3445.121.4.480

Haroush, K., Deouell, L. Y., and Hochstein, S. (2011). Hearing while blinking: multisensory attentional blink revisited. *J. Neurosci.* 31, 922–927. doi: 10.1523/JNEUROSCI.0420-10.2011

Hein, G., Doehrmann, O., Muller, N. G., Kaiser, J., Muckli, L., and Naumer, M. J. (2007). Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. *J. Neurosci.* 27, 7881–7887. doi: 10.1523/JNEUROSCI.1740-07.2007

Hillstrom, A. P., Shapiro, K. L., and Spence, C. (2002). Attentional limitations in processing sequentially presented vibrotactile targets. *Percept. Psychophys.* 64, 1068–1082. doi: 10.3758/BF03194757

Horvath, J., and Burgyan, A. (2011). Distraction and the auditory attentional blink. *Atten. Percept. Psychophys.* 73, 695–701. doi: 10.3758/s13414-010-0077-3

Hsiao, J. Y., Chen, Y. C., Spence, C., and Yeh, S. L. (2012). Assessing the effects of audiovisual semantic congruency on the perception of a bistable figure. *Conscious. Cogn.* 21, 775–787. doi: 10.1016/j.concog.2012.02.001

Jolicoeur, P. (1998). Modulation of the attentional blink by on-line response selection: evidence from speeded and unspeeded task1 decisions. *Mem. Cogn.* 26, 1014–1032. doi: 10.3758/BF03201180

Kane, M. J., and Engle, R. W. (2003). Working-memory capacity and the control of attention: the contributions of goal neglect, response competition, and task set to Stroop interference. *J. Exp. Psychol. Gen.* 132, 47–70. doi: 10.1037/0096-3445.132.1.47

Kerns, J. G., Cohen, J. D., Macdonald, A. W. 3rd., Cho, R. Y., Stenger, V. A., and Carter, C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science* 303, 1023–1026. doi: 10.1126/science.1089910

Kording, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., and Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE* 2:e943. doi: 10.1371/journal.pone.0000943

Laurienti, P. J., Kraft, R. A., Maldjian, J. A., Burdette, J. H., and Wallace, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Exp. Brain Res.* 158, 405–414. doi: 10.1007/s00221-004-1913-2

Lavie, N., Ro, T., and Russell, C. (2003). The role of perceptual load in processing distractor faces. *Psychol. Sci.* 14, 510–515. doi: 10.1111/1467-9280.03453

Lee, H., and Noppeney, U. (2014). Temporal prediction errors in visual and auditory cortices. *Curr. Biol.* 24, R309–R310. doi: 10.1016/j.cub.2014.02.007

Lewis, R., and Noppeney, U. (2010). Audiovisual synchrony improves motion discrimination via enhanced connectivity between early visual and auditory areas. *J. Neurosci.* 30, 12329–12339. doi: 10.1523/JNEUROSCI.5745-09.2010

Lunghi, C., Binda, P., and Morrone, M. C. (2010). Touch disambiguates rivalrous perception at early stages of visual analysis. *Curr. Biol.* 20, R143–R144. doi: 10.1016/j.cub.2009.12.015

Lunghi, C., Morrone, M. C., and Alais, D. (2014). Auditory and tactile signals combine to influence vision during binocular rivalry. *J. Neurosci.* 34, 784–792. doi: 10.1523/JNEUROSCI.2732-13.2014

Macaluso, E., and Driver, J. (2005). Multisensory spatial interactions: a window onto functional integration in the human brain. *Trends Neurosci.* 28, 264–271. doi: 10.1016/j.tins.2005.03.008

MacDonald, A. W. 3rd., Cohen, J. D., Stenger, V. A., and Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science* 288, 1835–1838. doi: 10.1126/science.288.5472.1835

Maclean, M. H., and Arnell, K. M. (2012). A conceptual and methodological framework for measuring and modulating the attentional blink. *Atten. Percept. Psychophys.* 74, 1080–1097. doi: 10.3758/s13414-012-0338-4

Maki, W. S., Couture, T., Frigen, K., and Lien, D. (1997). Sources of the attentional blink during rapid serial visual presentation: perceptual interference and retrieval competition. *J. Exp. Psychol. Hum. Percept. Perform.* 23, 1393–1411. doi: 10.1037/0096-1523.23.5.1393

Marois, R., Yi, D. J., and Chun, M. M. (2004). The neural fate of consciously perceived and missed events in the attentional blink. *Neuron* 41, 465–472. doi: 10.1016/S0896-6273(04)00012-1

Martens, S., Kandula, M., and Duncan, J. (2010). Restricted attentional capacity within but not between sensory modalities: an individual differences approach. *PLoS ONE* 5:e15280. doi: 10.1371/journal.pone.0015280

Martens, S., and Wyble, B. (2010). The attentional blink: past, present, and future of a blind spot in perceptual awareness. *Neurosci. Biobehav. Rev.* 34, 947–957. doi: 10.1016/j.neubiorev.2009.12.005

Nieuwenhuis, S., and De Kleijn, R. (2011). Consciousness of targets during the attentional blink: a gradual or all-or-none dimension? *Atten. Percept. Psychophys.* 73, 364–373. doi: 10.3758/s13414-010-0026-1

Nieuwenstein, M. R. (2006). Top-down controlled, delayed selection in the attentional blink. *J. Exp. Psychol. Hum. Percept. Perform.* 32, 973–985. doi: 10.1037/0096-1523.32.4.973

Nieuwenstein, M. R., Chun, M. M., Van Der Lubbe, R. H. J., and Hooge, I. T. C. (2005). Delayed attentional engagement in the attentional blink. *J. Exp. Psychol. Hum. Percept. Perform.* 31, 1463–1475. doi: 10.1037/0096-1523.31.6.1463

Noesselt, T., Rieger, J. W., Schoenfeld, M. A., Kanowski, M., Hinrichs, H., Heinze, H. J., et al. (2007). Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *J. Neurosci.* 27, 11431–11441. doi: 10.1523/JNEUROSCI.2252-07.2007

Noppeney, U., Josephs, O., Hocking, J., Price, C. J., and Friston, K. J. (2008). The effect of prior visual information on recognition of speech and sounds. *Cereb. Cortex* 18, 598–609. doi: 10.1093/cercor/bhm091

Noppeney, U., Ostwald, D., and Werner, S. (2010). Perceptual decisions formed by accumulation of audiovisual evidence in prefrontal cortex. *J. Neurosci.* 30, 7434–7446. doi: 10.1523/JNEUROSCI.0455-10.2010

Olivers, C. N., and Nieuwenhuis, S. (2005). The beneficial effect of concurrent task-irrelevant mental activity on temporal attention. *Psychol. Sci.* 16, 265–269. doi: 10.1111/j.0956-7976.2005.01526.x

Olivers, C. N., and Van Der Burg, E. (2008). Bleeping you out of the blink: sound saves vision from oblivion. *Brain Res.* 1242, 191–199. doi: 10.1016/j.brainres.2008.01.070

Olivers, C. N., Van Der Stigchel, S., and Hulleman, J. (2007). Spreading the sparing: against a limited-capacity account of the attentional blink. *Psychol. Res.* 71, 126–139. doi: 10.1007/s00426-005-0029-z

Olson, I. R., Chun, M. M., and Anderson, A. K. (2001). Effects of phonological length on the attentional blink for words. *J. Exp. Psychol. Hum. Percept. Perform.* 27, 1116–1123. doi: 10.1037/0096-1523.27.5.1116

Ojanen, V., Mottonen, R., Pekkola, J., Jaaskelainen, I. P., Joensuu, R., Autti, T., et al. (2005). Processing of audiovisual speech in Broca's area. *Neuroimage* 25, 333–338. doi: 10.1016/j.neuroimage.2004.12.001

Orr, J. M., and Weissman, D. H. (2009). Anterior cingulate cortex makes 2 contributions to minimizing distraction. *Cereb. Cortex* 19, 703–711. doi: 10.1093/cercor/bhn119

Pashler, H. (1984). Processing stages in overlapping tasks: evidence for a central bottleneck. *J. Exp. Psychol. Hum. Percept. Perform.* 10, 358–377. doi: 10.1037/0096-1523.10.3.358

Pekkola, J., Laasonen, M., Ojanen, V., Autti, T., Jaaskelainen, I. P., Kujala, T., et al. (2006). Perception of matching and conflicting audiovisual speech in dyslexic and fluent readers: an fMRI study at 3 T. *Neuroimage* 29, 797–807. doi: 10.1016/j.neuroimage.2005.09.069

Persaud, N., and Lau, H. (2008). Direct assessment of qualia in a blindsight participant. *Conscious. Cogn.* 17, 1046–1049. doi: 10.1016/j.concog.2007.10.001

Potter, M. C., Chun, M. M., Banks, B. S., and 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, 979–992. doi: 10.1037/0278-7393.24.4.979

Pourtois, G., De Pretto, M., Hauert, C. A., and Vuilleumier, P. (2006). Time course of brain activity during change blindness and change awareness: performance is predicted by neural events before change onset. *J. Cogn. Neurosci.* 18, 2108–2129. doi: 10.1162/jocn.2006.18.12.2108

Ptito, A., Arnell, K., Jolicoeur, P., and Macleod, J. (2008). Intramodal and crossmodal processing delays in the attentional blink paradigm revealed by event-related potentials. *Psychophysiology* 45, 794–803. doi: 10.1111/j.1469-8986.2008.00677.x

Raymond, J. E., Shapiro, K. L., and Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: an attentional blink? *J. Exp. Psychol. Hum. Percept. Perform.* 18, 849–860. doi: 10.1037/0096-1523.18.3.849

Raymond, J. E., Shapiro, K. L., and Arnell, K. M. (1995). Similarity determines the attentional blink. *J. Exp. Psychol. Hum. Percept. Perform.* 21, 653–662. doi: 10.1037/0096-1523.21.3.653

Roach, N. W., Heron, J., and McGraw, P. V. (2006). Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration. *Proc. Biol. Sci.* 273, 2159–2168. doi: 10.1098/rspb.2006.3578

Rushworth, M. F., Walton, M. E., Kennerley, S. W., and Bannerman, D. M. (2004). Action sets and decisions in the medial frontal cortex. *Trends Cogn. Sci.* 8, 410–417. doi: 10.1016/j.tics.2004.07.009

Sergent, C., Baillet, S., and Dehaene, S. (2005). Timing of the brain events underlying access to consciousness during the attentional blink. *Nat. Neurosci.* 8, 1391–1400. doi: 10.1038/nn1549

Sergent, C., and Dehaene, S. (2004). Is consciousness a gradual phenomenon? Evidence for an all-or-none bifurcation during the attentional blink. *Psychol. Sci.* 15, 720–728. doi: 10.1111/j.0956-7976.2004.00748.x

Shams, L., and Beierholm, U. R. (2010). Causal inference in perception. *Trends Cogn. Sci.* 14, 425–432. doi: 10.1016/j.tics.2010.07.001

Shams, L., and Seitz, A. R. (2008). Benefits of multisensory learning. *Trends Cogn. Sci.* 12, 411–417. doi: 10.1016/j.tics.2008.07.006

Shapiro, K., Schmitz, F., Martens, S., Hommel, B., and Schnitzler, A. (2006). Resource sharing in the attentional blink. *Neuroreport* 17, 163–166. doi: 10.1097/01.wnr.0000195670.37892.1a

Shapiro, K. L., Caldwell, J., and Sorensen, R. E. (1997a). Personal names and the attentional blink: a visual "cocktail party" effect. *J. Exp. Psychol. Hum. Percept. Perform.* 23, 504–514. doi: 10.1037/0096-1523.23.2.504

Shapiro, K. L., Raymond, J. E., and Arnell, K. M. (1997b). The attentional blink. *Trends Cogn. Sci.* 1, 291–296. doi: 10.1016/S1364-6613(97)01094-2

Shen, D., and Mondor, T. A. (2006). Effect of distractor sounds on the auditory attentional blink. *Percept. Psychophys.* 68, 228–243. doi: 10.3758/BF03193672

Simons, D. J., and Chabris, C. F. (1999). Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception* 28, 1059–1074. doi: 10.1068/p2952

Soto-Faraco, S., Navarra, J., and Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. *Cognition* 92, B13–B23. doi: 10.1016/j.cognition.2003.10.005

Soto-Faraco, S., and Spence, C. (2002). Modality-specific auditory and visual temporal processing deficits. *Q. J. Exp. Psychol. A* 55, 23–40. doi: 10.1080/02724980143000136

Soto-Faraco, S., Spence, C., Fairbank, K., Kingstone, A., Hillstrom, A. P., and Shapiro, K. (2002). A crossmodal attentional blink between vision and touch. *Psychon. Bull. Rev.* 9, 731–738. doi: 10.3758/BF03196328

Talsma, D., Senkowski, D., Soto-Faraco, S., and Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends Cogn. Sci.* 14, 400–410. doi: 10.1016/j.tics.2010.06.008

Tombu, M. N., Asplund, C. L., Dux, P. E., Godwin, D., Martin, J. W., and Marois, R. (2011). A Unified attentional bottleneck in the human brain. *Proc. Natl. Acad. Sci. U.S.A.* 108, 13426–13431. doi: 10.1073/pnas.1103583108

Vachon, F., and Tremblay, S. (2008). Modality-specific and amodal sources of interference in the attentional blink. *Percept. Psychophys.* 70, 1000–1015. doi: 10.3758/PP.70.6.1000

Van Atteveldt, N., Formisano, E., Goebel, R., and Blomert, L. (2004). Integration of letters and speech sounds in the human brain. *Neuron* 43, 271–282. doi: 10.1016/j.neuron.2004.06.025

Van Atteveldt, N. M., Formisano, E., Blomert, L., and Goebel, R. (2007). The effect of temporal asynchrony on the multisensory integration of letters and speech sounds. *Cereb. Cortex* 17, 962–974. doi: 10.1093/cercor/bhl007

Van Der Burg, E., Brederoo, S. G., Nieuwenstein, M. R., Theeuwes, J., and Olivers, C. N. (2010). Audiovisual semantic interference and attention: evidence from the attentional blink paradigm. *Acta Psychol. (Amst.)* 134, 198–205. doi: 10.1016/j.actpsy.2010.01.010

van Ee, R., van Boxtel, J. J. A., Parker, A., and Alais, D. (2009). Multisensory congruency as a mechanism for attentional control over perceptual selection. *J. Neurosci.* 29, 11641–11647 doi: 10.1523/JNEUROSCI.0873-09.2009

Visser, T. A. (2007). Masking T1 difficulty: processing time and the attenional blink. *J. Exp. Psychol. Hum. Percept. Perform.* 33, 285–297. doi: 10.1037/0096-1523.33.2.285

Von Kriegstein, K., and Giraud, A. L. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biol.* 4:e326. doi: 10.1371/journal.pbio.0040326

Vroomen, J., and Keetels, M. (2010). Perception of intersensory synchrony: a tutorial review. *Atten. Percept. Psychophys.* 72, 871–884. doi: 10.3758/APP.72.4.871

Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., and Schirillo, J. A. (2004). Unifying multisensory signals across time and space. *Exp. Brain Res.* 158, 252–258. doi: 10.1007/s00221-004-1899-9

Wallace, M. T., Wilkinson, L. K., and Stein, B. E. (1996). Representation and integration of multiple sensory inputs in primate superior colliculus. *J. Neurophysiol.* 76, 1246–1266.

Weiskrantz, L., Warrington, E. K., Sanders, M. D., and Marshall, J. (1974). Visual capacity in the hemianopic field following a restricted occipital ablation. *Brain* 97, 709–728. doi: 10.1093/brain/97.1.709

Welch, R. B., and Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychol. Bull.* 88, 638–667. doi: 10.1037/0033-2909.88.3.638

Werner, S., and Noppeney, U. (2010a). Distinct functional contributions of primary sensory and association areas to audiovisual integration in object categorization. *J. Neurosci.* 30, 2662–2675. doi: 10.1523/JNEUROSCI.5091-09.2010

Werner, S., and Noppeney, U. (2010b). Superadditive responses in superior temporal sulcus predict audiovisual benefits in object categorization. *Cereb. Cortex* 20, 1829–1842. doi: 10.1093/cercor/bhp248

Yoncheva, Y. N., Zevin, J. D., Maurer, U., and McCandliss, B. D. (2010). Auditory selective attention to speech modulates activity in the visual word form area. *Cereb. Cortex* 20, 622–632. doi: 10.1093/cercor/bhp129

Yu, A. J., Dayan, P., and Cohen, J. D. (2009). Dynamics of attentional selection under conflict: toward a rational Bayesian account. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 700–717. doi: 10.1037/a0013553

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 12 May 2014; accepted: 20 August 2014; published online: 11 September 2014.

Citation: Adam R and Noppeney U (2014) A phonologically congruent sound boosts a visual target into perceptual awareness. Front. Integr. Neurosci. ***8****:70. doi: 10.3389/fnint.2014.00070*

This article was submitted to the journal Frontiers in Integrative Neuroscience.

frontiers in Integrative Neuroscience

ORIGINAL RESEARCH
published: 10 July 2015
doi: 10.3389/fnint.2015.00045

Crossmodal semantic congruence can affect visuo-spatial processing and activity of the fronto-parietal attention networks

Serena Mastroberardino[1], Valerio Santangelo[1,2] and Emiliano Macaluso[1]*

[1] Neuroimaging Laboratory, Santa Lucia Foundation, Rome, Italy, [2] Department of Philosophy, Social Sciences & Education, University of Perugia, Perugia, Italy

OPEN ACCESS

Edited by:
Salvador Soto-Faraco, Universitat Pompeu Fabra, Spain

Reviewed by:
Tobias Andersen, Technical University of Denmark, Denmark
Pawel J. Matusz, Centre Hospitalier Universitaire Vaudois – University of Lausanne, Switzerland

***Correspondence:**
Serena Mastroberardino, Neuroimaging Laboratory, Santa Lucia Foundation, Via Ardeatina, 306, 00100 Rome, Italy
smastr@gmail.com

Received: 06 October 2014
Accepted: 25 June 2015
Published: 10 July 2015

Citation:
Mastroberardino S, Santangelo V and Macaluso E (2015) Crossmodal semantic congruence can affect visuo-spatial processing and activity of the fronto-parietal attention networks. Front. Integr. Neurosci. 9:45. doi: 10.3389/fnint.2015.00045

Previous studies have shown that multisensory stimuli can contribute to attention control. Here we investigate whether irrelevant audio–visual stimuli can affect the processing of subsequent visual targets, in the absence of any direct bottom–up signals generated by low-level sensory changes and any goal-related associations between the multisensory stimuli and the visual targets. Each trial included two pictures (cat/dog), one in each visual hemifield, and a central sound that was semantically congruent with one of the two pictures (i.e., either "meow" or "woof" sound). These irrelevant audio–visual stimuli were followed by a visual target that appeared either where the congruent or the incongruent picture had been presented (valid/invalid trials). The visual target was a Gabor patch requiring an orientation discrimination judgment, allowing us to uncouple the visual task from the audio–visual stimuli. Behaviourally we found lower performance for invalid than valid trials, but only when the task demands were high (Gabor target presented together with a Gabor distractor vs. Gabor target alone). The fMRI analyses revealed greater activity for invalid than for valid trials in the dorsal and the ventral fronto-parietal attention networks. The dorsal network was recruited irrespective of task demands, while the ventral network was recruited only when task demands were high and target discrimination required additional top–down control. We propose that crossmodal semantic congruence generates a processing bias associated with the location of congruent picture, and that the presentation of the visual target on the opposite side required updating these processing priorities. We relate the activation of the attention networks to these updating operations. We conclude that the fronto-parietal networks mediate the influence of crossmodal semantic congruence on visuo-spatial processing, even in the absence of any low-level sensory cue and any goal-driven task associations.

Keywords: attention, multisensory, semantics, space, fronto-parietal networks

Introduction

Over the last 30 years, multisensory processing and the integration of signals across sensory modalities has gained much interest (for a review see Calvert, 2001; see also Stevenson et al., 2014). An outstanding issue in this field concerns to what extent crossmodal interactions occur in a fully automatic manner or whether there are significant couplings between multisensory processing and attention control. While traditional views emphasized pre-attentive mechanisms of multisensory integration, recent studies highlighted that attention and multisensory processing can influence each other in many different ways (McDonald et al., 2001; Koelewijn et al., 2010; Talsma et al., 2010; Santangelo and Macaluso, 2012, for reviews). Here we sought to contribute to this debate by asking whether crossmodal semantic congruence between visual and auditory signals presented at different locations can generate spatial attention biases and affect the processing of subsequent visual stimuli. Specifically, we made use of a paradigm where the audio–visual signals were fully task-irrelevant and did not provide any low-level spatial cues that might affect the processing of the visual targets. Therefore, any crossmodal spatial influence on visual processing can be attributed to crossmodal semantic processing rather than other low-level/bottom–up or goal-related factors directly linking the multisensory input to visual-spatial attention control processes.

Previous studies have shown that auditory stimuli can affect visual spatial processing, consistent with supramodal mechanisms of attention control (e.g., Vecera and Farah, 1994). Crossmodal spatial cueing studies have shown that a lateralised auditory stimulus (non-predictive cue) can influence the response to a subsequent visual target, with better performance when the target is presented at the same location as the cue (valid trials) than on the opposite side (invalid trials; Driver, 1996; Spence and Driver, 1997, 1998; McDonald et al., 2000). These crossmodal cueing effects suggest that a sudden auditory onset at one location can attract visual attention toward that location, and that processing targets on the opposite side requires additional processes (e.g., disengaging from the cued location, shifting/re-orienting, and re-engaging at the position of the visual target, see Posner and Cohen, 1984; Brown and Denney, 2007; Chen, 2012). Other studies have highlighted the influence of spatially non-informative auditory cues on visual search tasks. One example of this is the "pip-and-pop" effect, where the binaural presentation of a sound synchronized with a color change of the visual target can boost search performance (Van der Burg et al., 2008; van den Brink et al., 2014). While cueing and search paradigms differ in many ways (e.g., role of temporal vs. spatial correspondences between the two modalities), they both rely on spatially localized low-level changes in the sensory input. Indeed, one possible mechanism generating these crossmodal interactions is that the between-modalities (spatial and/or temporal) correspondence of the physical change makes the target location more salient via bottom–up, stimulus-driven attention control (e.g., Van der Burg et al., 2008; see also Talsma et al., 2010, for review).

However, high-level factors can also contribute to crossmodal influences on visuo-spatial processing. For example, semantic congruence plays an important role during the processing of complex audio–visual stimuli and has been found to influence visual attention. Using a search task, Iordanescu et al. (2008) presented pictures of natural objects/animals together with a centrally presented non-informative sound. They found faster target localization when the target object (e.g., a picture of a cat) was presented together with a semantically congruent sound (i.e., a meow), compared with an unrelated sound or a sound associated with a distractor picture (see also Iordanescu et al., 2010). These findings suggest that audio–visual semantic congruence can bias visuo-spatial processing, e.g., via enhanced representation of the visual target (Iordanescu et al., 2008, 2010), even in the absence of any spatially localized sensory change linking the central sound and the visual target. However, in these visual search studies and the pip-and-pop effect studies, the visual component of the "interacting" audio–visual stimuli was always task-relevant (i.e., a visual target). An exception to this is Experiment 5 in Van der Burg et al. (2008) that revealed a marginal effect/cost for sounds coupled with distractors. However, it should be considered that during serial search, participants will voluntarily shift attention between the various elements of the visual display, including the distractors. Therefore, audio–visual interactions for sounds synchronized with a distractor-change will sometimes involve visual stimuli (i.e., the synchronized distractor) that might be attended to in a goal-driven manner.

In the studies discussed above, goal-driven attention was directed toward the multisensory stimuli (or – at least – the visual component of these), which is likely to have a significant impact on how/whether the two modalities interacted with each other (see below; and Koelewijn et al., 2010, for a review). In the context of cueing studies, one approach to assess whether crossmodal spatial interactions also occur between task-irrelevant stimuli consists in using bimodal non-predictive cues. For example, Santangelo et al. (2006) presented audio–visual cues followed by unimodal visual targets. They found that spatially congruent bimodal cues on the same side of the visual target lead to faster discriminations, but this effect was not larger than the cueing effect elicited by unisensory auditory or visual cues. While this null finding suggests that audio–visual stimuli do not interact with each other when fully task-irrelevant, later studies showed that bimodal cues can affect ERPs over and above any effect of unimodal cues (Santangelo et al., 2008b) and that, unlike unimodal cues, they influence visual target discrimination also under high-load, dual-task conditions (Santangelo and Spence, 2007; Santangelo et al., 2008a; see, for a review, Santangelo and Spence, 2008). Additional evidence for the influence of irrelevant audio–visual stimuli on visuo-spatial processing comes from a study by Matusz and Eimer (2011). In this study, each trial included a first array of irrelevant visual stimuli coupled with a centrally presented sound, followed by a visual search display. The results showed improved search performance when the sound was coupled with a color change in the first display, at the same location of the subsequent visual target. This effect did not depend on the relationship between the color of the cue and the currently relevant target color (cf. contingent attentional capture, Folk et al., 1992), consistent with pure bottom–up mechanisms

of attentional capture. In a subsequent study, Matusz et al. (2015) further investigated the possible influence of top–down signals during the processing of irrelevant audio–visual stimuli, now using semantic matching. Each trial required the discrimination of a visual target that was either presented in isolation (low top–down task demand) or embedded in visual distractors (high demand). In the critical audio–visual distractor trials, a task-irrelevant colored visual stimulus was presented together with a voice saying the color (e.g., a red square shape coupled with a spoken "red"). In adult participants, the results showed that task-irrelevant audio–visual stimuli that included goal-relevant information (e.g., the color "red" was also a relevant feature of the target) interfered with target discrimination irrespective of task demands. In summary, several studies have demonstrated the influence of irrelevant audio–visual stimuli on visual attention, but they have always involved either spatially localized "bottom–up" physical changes in the sensory input (e.g., Santangelo and Spence, 2007; Santangelo et al., 2008a; Matusz and Eimer, 2011); or shared features between the audio–visual stimuli and the task-relevant visual target (interfering distractors, in Matusz et al., 2015).

Accordingly, the main aim of the current study was to investigate crossmodal spatial influences of task-irrelevant audio–visual stimuli on visual processing, in the absence of any low-level spatial cue related to the onset of the stimuli, or any goal-related signal linking the audio–visual stimuli with the subject's current task. For this, we presented sounds together with semantically related/unrelated pictures (cf., Iordanescu et al., 2008), which were completely irrelevant to participants' task. The task of the participants was to perform an orientation discrimination of a Gabor patch presented after the audio–visual stimuli. The target Gabor patch was presented either on the same side (valid trails) or on the opposite side (invalid trials) of the picture that was semantically congruent with the centrally presented sound. We hypothesized that the semantic relationship between the central sound and one picture would influence visuo-spatial attention, which in turn would affect the processing of the subsequent visual targets.

From the neuroimaging perspective, several previous studies have investigated the neural substrate of crossmodal semantic congruence by presenting in-/congruent pictures and sounds (e.g., Taylor et al., 2006; Noppeney et al., 2007; see Doehrmann and Naumer, 2008, for a review). These studies highlighted that audio–visual semantic interactions can affect activity in polysensory regions of the superior temporal sulcus, as well as higher-order areas in the medial temporal cortex and the left prefrontal cortex. In the current study, we might expect the involvement of these brain areas, but note that our main "valid/invalid" comparisons entailed trials with identical audio–visual input, with one picture that is always congruent with the sound and one that is incongruent. Because of this, areas involved in audio–visual semantic matching are unlikely to show any differential condition-specific effect. In contrast, because we expected crossmodal interactions to influence visuo-spatial processing, we would predict condition-specific effects in fronto-parietal networks associated with visuo-spatial attention control (Corbetta and Shulman, 2002). These networks have also been found to activate in studies of attention control in modalities other than vision (e.g., see Yantis et al., 2002; Macaluso et al., 2003; Krumbholz et al., 2009; Hill and Miller, 2010, for the dorsal network; and Downar et al., 2000; Macaluso et al., 2002, for the ventral network), which makes them the ideal candidates for mediating the influence of non-visual signals on visuo-spatial attention control.

In the current paradigm, we aimed to uncouple the audio–visual stimuli from any goal-related signals associated with the subject's task, and we eliminated any stimulus-driven spatial cue by avoiding spatially localized sensory changes when presenting the audio–visual stimuli (see above). Arguably, the semantic matching of the sound with the semantically related picture still entails endogenous processes such as internal object representations required to combine visual and auditory signals (Iordanescu et al., 2008; see also Fiebelkorn et al., 2010). These endogenous effects should be distinguished from more traditional goal-directed processes associated, for example, with predictive cues that provide participants with task-relevant information (i.e., signaling the most likely location of the upcoming target) and can be used to strategically control spatial attention in a goal-driven manner. Nonetheless, the involvement of endogenous processes for crossmodal semantic matching and our main expectation that these will affect processing of the task-relevant visual targets lead us to hypothesize the involvement of dorsal fronto-parietal regions associated with top–down control (Corbetta and Shulman, 2002), as well as ventral regions where top–down and stimulus-driven signals jointly contribute to visuo-spatial orienting (Corbetta et al., 2008; Geng and Vossel, 2013; Macaluso and Doricchi, 2013; see also Natale et al., 2009, for a fMRI study using non-predictive cues but involving top–down control).

In order to gain further insights into the relative contributions of top–down and stimulus-driven control in the current paradigm, the design included several additional manipulations. First, we varied the time between the offset of the irrelevant audio–visual stimuli and the onset of the visual target (ISIs = 0 or 250 ms). We expected that if audio–visual semantic matching generates spatial signals analogous to those typically associated with non-predictive peripheral cues, maximal effects should occur with the shortest ISI. In contrast, if semantic matching generates a top–down signal analogous to goal-related signals typically associated with predictive cues, the effects should be largest with the longer ISI (e.g., Ruz and Lupiáñez, 2002; and Rauschenberger, 2003, for reviews). Second, we manipulated top–down task demands by varying the general difficulty of target discrimination (i.e., easy vs. difficult tilt judgment, see **Figure 1A**) and by varying the amount of spatial competition during the visual judgment task (target Gabor only, in Experiment 1; target plus one distractor Gabor in Experiment 2; see **Figure 1A**). When two Gabor patches were presented, the participants had to make use of internal information about the current task-set (i.e., a relevant color defining the target Gabor), which implies additional top–down control during the target phase of the trial (see Indovina and Macaluso, 2007, who used an analogous procedure to demonstrate the role of top–down control for the activation of the inferior parietal cortex in a purely visual

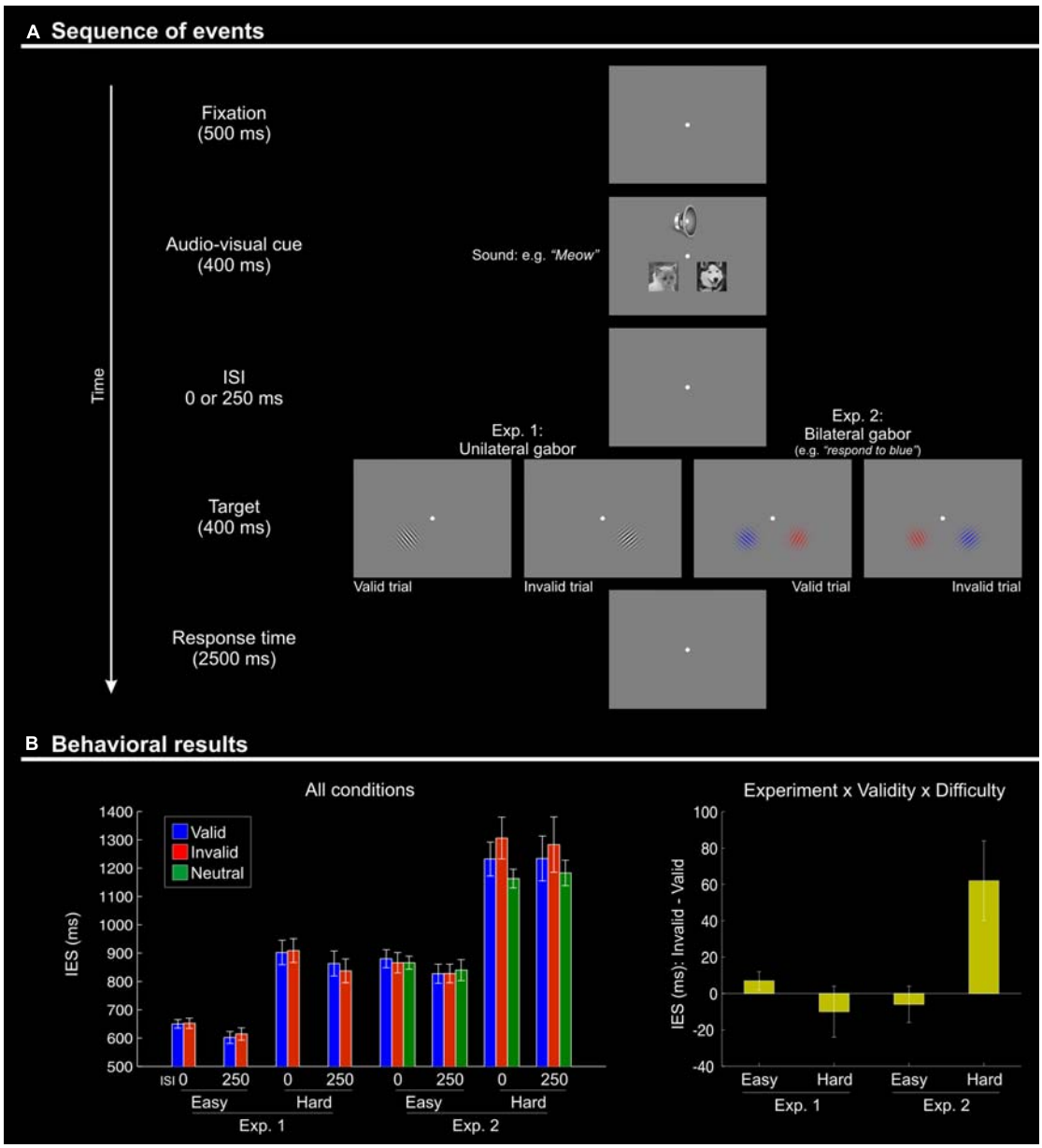


FIGURE 1 | (A) Schematic diagram showing the sequence of events during one trial. Each trial began with the presentation of a fixation point for 500 ms. Two pictures (cat and dog) were then presented simultaneously with a centrally presented sound: a "meow" or a "woof". The stimuli in both modalities lasted for 400 ms. After a variable inter-stimulus interval (ISI) of 0 or 250 ms, a single (Experiment 1) or two (one in each visual hemifield; Experiment 2) Gabor patches were presented for 400 ms. Subjects had 2500 ms to respond to the orientation of the single Gabor in Experiment 1, or to the target Gabor of the relevant color in Experiment 2. Visual targets appeared either in the location of the picture that was semantically congruent with the sound ("valid") or on the other side ("invalid"). After an inter-trial interval (ITI) of 1500 ms with a blank screen, a new trial began. **(B)** Behavioral results (inverse efficiency scores, IES) showing a Validity effect (invalid > valid) only for the difficult orientation discrimination trials ("hard" conditions) of Experiment 2. Left panel: mean IES plotted separately for each experimental condition; Right panel: the Validity effect (invalid – valid) plotted separately for "easy" and "hard" conditions, in the two experiments. The selective effect of Validity in the "hard" condition of Experiment 2 (cf. rightmost bar in this plot) highlights the 3-way interaction between Validity, Difficulty, and Experiment.

task). Based on previous studies that used dual-task procedures to engage processing resources away from multisensory stimuli (e.g., Alsius et al., 2005; Santangelo and Spence, 2007; and Koelewijn et al., 2010, for a review), here one might predict a reduction of any crossmodal effect of semantic congruence in conditions of high task-demands (see also Eimer and Kiss, 2008; for a purely visual study showing that changes of target-related task demands can influence effects associated with preceding

non-predictive cues, albeit in the context of a contingent capture paradigm). However, if the multisensory stimuli impact top–down control mechanisms engaged only under high task demands, one might expect crossmodal influences specifically in conditions engaging these additional control processes.

In summary, we asked whether audio–visual semantic congruence can bias the processing of subsequent visual targets, specifically when the audio–visual stimuli are task-irrelevant and do not generate any spatially localized sensory change. We presented two pictures, one in the left and one in the right hemifield, together with a central sound that was semantically congruent with one of the two pictures. Shortly after, we presented the visual target either on the side of the picture congruent with the central sound (valid trials) or on the opposite side (invalid trials). We hypothesized that the picture-sound semantic congruence would affect visuo-spatial attention and, thus, that the processing of the subsequent visual target would change as a function of the trial in-/validity. We predicted that these crossmodal effects would impact primarily on the activity of the fronto-parietal attention networks, where any spatial signal associated with semantic congruence may interact with other task-related factors that regulate the functioning of these attention control systems.

Materials and Methods

Participants

Nineteen volunteers participated in Experiment 1 and 20 in Experiment 2. None of the subjects participated in both experiments. All participants were neurologically intact, were not on psychotropic or vasoactive medication, and had no history of psychiatric or neurological disease. They had normal or corrected-to-normal vision (i.e., with contact lenses) and reported normal hearing. Before scanning, all participants were tested in a 20-min training session. The training session comprised one block of 192 trials. The stimuli and the task were identical to those presented during the imaging sessions (see Stimuli and Procedure), with the exception that at the end of each trial the participant received visual feedback of his/her performance. Subjects who failed to reach 80% accuracy in the training session did not participate in the imaging experiment ($n = 2$ in Experiment 1, and $n = 4$ in Experiment 2). Four subjects were excluded from data analysis of Experiment 2 for excessive within-fMRI-run head movements (larger than 2 mm or 2°). Therefore, the final analyses included 17 participants for Experiment 1 (7 males; mean age $= 22.3 \pm 3.3$) and 12 participants for Experiment 2 (7 males; mean age $= 24.3 \pm 3.0$). All subjects gave written informed consent to participate in the study, which was approved by the independent Ethics Committee of the Santa Lucia Foundation (Scientific Institute for Research Hospitalization and Health Care).

Stimuli and Procedure

Stimulus presentation was controlled with Matlab 7.1 (The MathWorks Inc., Natick, MA, USA), using the Cogent2000 Toolbox (Wellcome laboratory of Neurobiology, University College London). Visual stimuli were presented on a gray background using a rear projection system. Participants were instructed to maintain fixation on a central dot during the scanning sessions. The auditory stimuli were presented using MRI compatible headphones.

In both experiments, the participants were presented with two pictures (cat or dog), one on each side of the central fixation point, plus an auditory stimulus presented binaurally. The pictures and the sound were presented for 400 ms, which should be sufficient time to identify the audio–visual object before the onset of the subsequent visual target (e.g., see De Lucia et al., 2010, who showed neuro-physiological signatures of auditory categorization of complex sounds at around 200 ms post-stimulus onset, even with a large pool of 160 stimuli rather than just 2–3 sounds used here). The pictures were displayed in black and white (resolution 200 × 200 pixels), centerd 2.7° to the left and to right of the central fixation and 2° below it, subtending a visual angle of 3.8° × 3.8° (see **Figure 1A**). The auditory stimulus consisted of a cat's meow or a dog's bark, presented binaurally and perceived centrally. Experiment 2 included a third sound type (a frog's croak) that was used for the "neutral" trials (see below).

After a variable delay (ISI = 0 or 250 ms), the participant was presented with the visual target. In Experiment 1, the target was a single Gabor patch (visual angle size 3.8° × 3.8°) presented in either the left or right hemifield, in place of the picture of the cat or the dog. In the *valid* conditions, the target was presented on the side where the picture was congruent with the sound (e.g., Gabor on the left when preceded by a picture of a cat on the left coupled with a "meow" sound). In the *invalid* conditions, the target was presented on the opposite side of the congruent picture (e.g., Gabor on the left when preceded by a picture of a dog on the right and a "woof" sound). The position of the congruent picture was not predictive of the target position (i.e., 50% "cue validity"). Despite this, participants might have sought to find a systematic relationship between the position of the congruent picture and the visual target, thus using the irrelevant audio–visual stimuli to direct spatial attention in a goal-driven manner. While this cannot be excluded, it should be emphasized that all participants underwent a 20-min training session, and it is unlikely that they continued looking for this inexistent relationship throughout the whole fMRI experiment.

On each trial, the target Gabor patch had one out of eight possible orientations. The task of the subject was to discriminate whether the Gabor orientation was smaller or larger than 45° and to report this by pressing a button either with the index finger (larger) or with the middle finger (smaller). In the *easy* conditions, the target was oriented at either 30 or ±60°; while in the *hard* conditions the tilt was either ±40° or ±50°. Subjects had 2500 ms to provide a response. This was followed by a variable inter-trial interval (1–3 s, uniformly distributed). Overall, Experiment 1 comprised 384 trials, equally divided in three fMRI runs. Each run of 128 trials comprised 16 repetitions of each of the eight experimental conditions [2 (Validity) × 2 (ISI) × 2 (Difficulty)].

In Experiment 2, the target phase of the trial comprised the presentation of two Gabor patches, flashed simultaneously

in the left and right visual fields. One patch was colored in red while the other was in blue (see **Figure 1A**, panels on the right). Before starting the experiment, the participants were instructed which color of the Gabor patch was task relevant (counterbalanced across participants). This defined the location of the visual target that required discrimination and response and, thus, whether the trial was valid or invalid. Experiment 2 also included a *neutral* condition, where the cat/dog pictures were coupled with the sound of a frog's croak. This neutral condition allowed us to address the additional question of whether any effect of Validity (invalid vs. valid trials) resulted from a cueing "benefit" on valid trails, a "cost" on invalid trials, or both. Valid, invalid and neutral cues were equally likely and not predictive of target location. Experiment 2 comprised a total of 576 trials, equally divided into three fMRI runs. In each run of 192 trials, each of the 12 experimental conditions [3 (Validity) × 2 (ISI) × 2 (Difficulty)] was repeated 16 times.

Eye Movement Recording

To make sure that participants maintained central fixation through the experimental sessions, eye position was monitored using an infrared ASL eye-tracking system, adapted for use in the scanner (Applied Science Laboratories, Bedford, MA, USA; Model 504, sampling rate 60 Hz). Changes in horizontal eye position greater than ±2° of visual angle in a time window of 1550 ms (i.e., from trial onset to target offset, inclusive of the longer ISI of 250 ms) were classified as failure to maintain fixation. Overall, participants made few eye movements away from central fixation (4% Experiment 1; 5% Experiment 2).

Magnetic Resonance Imaging

A Siemens Allegra (Siemens Medical Systems, Erlangen, Germany) operating at 3T and equipped for echo-planar imaging (EPI) was used to acquire the functional magnetic resonance images. A quadrature volume head coil was used for radio frequency transmission and reception. Head movement was minimized by mild restraint with cushions. Thirty-two slices of functional MR images were acquired using blood oxygenation level-dependent imaging (3 mm × 3 mm, 2.5 mm thick, 50% distance factor, repetition time = 2.08 s, time echo = 30 ms), covering the entirety of the cortex.

Image pre-processing and data analysis were performed using SPM8 (Wellcome Department of Cognitive Neurology) implemented in Matlab 7.1 (The MathWorks Inc., Natick, MA, USA). In Experiment 1, we collected a total of 885 fMRI volumes (295 × 3 runs); while in Experiment 2 we collected a total of 1275 fMRI volumes (425 × 3 runs). For each participant, after having discarded the first four volumes of each run, all images were corrected for head movements. All images were normalized using the SPM8 standard EPI template, re-sampled to a 2-mm isotropic voxel size and spatially smoothed using an isotropic Gaussian kernel of 8 mm FWHM. The time series at each voxel for each participant was high-pass filtered at 220 s and pre-whitened by means of the autoregressive model AR (1) (Friston et al., 2002).

For statistical inference, we used a random effects approach (Penny and Holmes, 2004). This comprised two steps. First, for each participant the time series at each voxel was best-fitted by model parameters based on a linear combination of effects of interest. These were delta functions representing: for Experiment 1, the onsets of the eight conditions given by our 2 × 2 × 2 factorial design [Validity (valid, invalid); Difficulty (easy, hard); ISI [0, 250)]; and for Experiment 2, the onsets of the 12 conditions of our 3 × 2 × 3 factorial design [Validity (valid, invalid, neutral); Difficulty (easy, hard); ISI (0, 250)]. All onsets were convolved with the SPM8 hemodynamic response function. The onset of the hemodynamic response function was aligned with the onset of the multisensory cue, with duration = 0. Onsets of trials in which an erroneous response occurred were included in the general linear model as covariates of no interest, and excluded from any further analysis of the imaging data.

For statistical inference at the group level, we considered the data of both experiments together, allowing us to formally assess the effect of Validity (invalid vs. valid trials) as a function of task-demands (high vs. low in Experiment 2 vs. Experiment 1, respectively)]. We tested for the main effect of Validity (invalid vs. valid) and any interaction between this and the factors associated with top–down control (i.e., ISI, Difficulty, and Experiment) in ANOVAs. For each participant, we computed the contrast "invalid minus valid trials" separately for the two ISIs (0, 250 ms) and the two levels of discrimination Difficulty (easy, hard). The resulting four conditions/effects per subject were entered in the AVOVA that included the Validity effects of both groups (Experiment 1 and 2) and enabled us to test our main hypothesis of the effect of crossmodal semantic congruence on visuo-spatial attention: i.e., the overall effect of Validity (invalid > valid) and any modulation by the three top–down factors (e.g., larger re-orienting effects when the task required focused spatial selection: $(\text{invalid} - \text{valid})_{\text{Exp2}} > (\text{invalid} - \text{valid})_{\text{Exp1}}$; i.e., the interaction "Validity × Experiment").

In addition, we carried out two separate ANOVAs that tested for the effects of ISIs and discrimination Difficulty, irrespective of Validity. For each participant, we computed the contrast ISI "0 − 250" (irrespective of Validity and Difficulty); and the contrast "hard minus easy" (irrespective of Validity and ISI). These were entered in two separate ANOVAs, where we tested for the mean effect of ISI/Difficulty across Experiments, and the interactions between ISI/Difficulty and Experiment.

Finally, for Experiment 2 only, we compared the overall effect of "valid and invalid cues" against the "neutral cues" in a separate group analysis. First, for each subject, we computed the contrast "[(valid + invalid)/2] minus neutral" separately for the two ISIs and the two levels of task Difficulty. We then ran another ANOVA to test for the overall effect of "valid/invalid versus neutral" trials and for any interaction between this and the other two factors included in Experiment 2 (i.e., ISI and task Difficulty).

All ANOVAs were corrected for non-sphericity (Friston et al., 2002) to account for possible differences in error variance across conditions. Statistical thresholds were set to p-FWE-corrected (family wise error) = 0.05 at cluster level (cluster size estimated

at p-uncorrected, p-uncorrected = 0.005), considering the whole brain as the volume of interest.

Results

Behavioral Data

To allow comparisons among conditions accounting for any possible effect of speed-accuracy tradeoffs, we computed and analyzed inverse efficiency scores (IES; Townsend and Ashby, 1983; see also Bruyer and Brysbaert, 2011). For completeness, we also report the analyses of the reaction times and the error rates (RTs and ER, see legend of **Table 1**).The behavioral data were analyzed with SPSS (Statistical Package for Social Science, version 13.0). The Greenhouse-Geisser procedure was used to correct for any violations of sphericity. Analogous to the imaging analyses, the main behavioral analysis considered the two experiments together allowing us to test for condition-by-experiment interactions (see also Magnetic Resonance Imaging, above).

The IES are shown in **Figure 1B** and in **Table 1**. We carried out a four-way mixed ANOVA with "Experiment" as a between-subjects factor (1 vs. 2) and the following three within-subjects factors: "Validity" (valid, invalid), "Difficulty" (easy, hard), and "ISI" (0, 250 ms). The ANOVA revealed significant main effects of "Experiment," "Difficulty," and "ISI." Participants were more accurate and faster in judging the target in Experiment 1 than in Experiment 2 {IES means: 754 vs. 1057 ms; [$F(1,27) = 31.7$, $p < 0.001$]}. Discrimination performance was better for easy compared to hard trials {IES: 740 vs. 1071 ms, [$F(1,27) = 166.3$, $p < 0.001$]}, and a decrease in discrimination performance was found for short compared to long ISIs (IES means: 925 vs. 866 ms; [$F(1,27) = 7.3$, $p = 0.012$]). Analogous results were obtained for the RT data (see **Table 1**).

While the main effect of "Validity" was not significant, the ANOVA revealed a significant three-way interaction between Validity, Experiment, and Difficulty [$F(1,27) = 9.8$, $p < 0.004$]. This complex interaction was driven by a Validity effect (invalid > valid trials) only in the difficult discrimination conditions of Experiment 2 [IES mean: 62 ms; $t(11) = 2.4$, $p = 0.036$; see **Figure 1B**, right graph]. No other effect reached significance.

We further explored the Validity effect in Experiment 2 by evaluating costs vs. benefits for invalid/valid trials compared to the neutral cues in separate t-tests. We considered only the "hard" conditions of Experiment 2, averaging across ISIs. We tested for cueing costs in invalid trials (invalid > neutral) and for cueing benefits in valid trials (neutral > valid), using one tailed t-tests. These showed the expected costs of invalid trials [$IES_{invalid - neutral}$: 122 ms, $t(11) = 1.81$; $p < 0.049$]; while the valid trials did not lead to any benefits and actually showed numerically lower performance than the neutral trials [$IES_{neutral - valid}$: −60 ms, $t(11) = -0.95$; $p > 0.8$].

To summarize, the behavioral results demonstrated an effect of crossmodal semantic congruence on the processing of the subsequent visual targets and showed a role of the current task demands on this behavioral effect. Specifically,

TABLE 1 | Behavioral data.

		Difficulty: Easy						Difficulty: Hard					
		ISI 0			ISI 250			ISI 0			ISI 250		
		Valid	Invalid	Neutral	Valid	Invalid	Neutral	Valid	Invalid	Neutral	Valid	Invalid	Neutral
Experiment 1	IES(ms)	650 ± 15	652 ± 18		602 ± 21	614 ± 22		902 ± 43	909 ± 42		863 ± 44	837 ± 42	
	RT(ms)	633 ± 17	631 ± 17		584 ± 20	589 ± 20		715 ± 26	720 ± 29		67031 ±	671 ± 28	
	ER(%)	3 ± 1	3 ± 2		3 ± 1	4 ± 1		20 ± 2	20 ± 3		22 ± 2	19 ± 3	
Experiment 2	IES(ms)	880 ± 32	866 ± 36	866 ± 23	827 ± 34	828 ± 33	840 ± 37	1232 ± 60	1306 ± 74	1163 ± 33	1234 ± 79	1283 ± 98	1183 ± 45
	RT(ms)	840 ± 29	832 ± 31	838 ± 23	802 ± 36	799 ± 33	805 ± 31	973 ± 44	995 ± 52	929 ± 26	968 ± 57	961 ± 62	906 ± 24
	ER(%)	4 ± 1	4 ± 2	3 ± 1	3 ± 1	3 ± 1	4 ± 1	21 ± 2	23 ± 3	20 ± 3	21 ± 2	24 ± 3	22 ± 3

Mean (±SEs) inverse efficiency scores (IES), reaction times (RTs), and error rates (proportion), for the different conditions in the two Experiments. RTs and ERs were tested in a four-way mixed ANOVA with the factors of: Validity (valid, invalid), Difficulty (easy vs. hard), ISI (0 vs. 250 ms), and Experiment (1 vs. 2). This revealed main effects of Experiment [RT means: 652 vs. 896 ms, $F(1,27) = 32.1$, $p < 0.001$; ER means: 12 vs. 13%, $F(1,27) < 1$, n.s.]; main effects of Difficulty [RT: 714 vs. 834 ms, $F(1,27) = 60.5$, $p < 0.001$; ER: 3 vs. 21%, $F(1,27) = 227.7$, $p < 0.001$]; a main effect of ISI, for RT only [RT: 793 vs. 755 ms, $F(1,27) = 25.6$, $p < 0.001$; ER: 12 vs. 12%, $F(1,27) < 1$, n.s.]; without any significant main effect of Validity. However, we found a three-way interaction between Validity, Experiment, and Difficulty for the ER [$F(1,27) = 7.2$, $p = 0.013$], but not for RT [$F(1,27) < 1$, n.s.], see also the main text for the results of the IES analyses. Val/Inv/Neu, valid/invalid/neutral trials.

we found a significant difference between invalid vs. valid trials ($IES_{invalid} > IES_{valid}$) only when the primary visual task had high demands ("hard" conditions of Experiment 2, see **Figure 1B**).

fMRI Results

The main fMRI analysis compared "valid" and "invalid" trials with the aim of revealing any spatial effect of crossmodal semantic congruence on the processing of the subsequent visual targets; and assessed this under different task constraints: i.e., 0/250 ms ISI; easy/hard target discrimination; Experiment 1/2, with single vs. two Gabor patches in the target phase of the trial. The corresponding main effects and interactions were tested in a ANOVA that for each subject considered the contrast "invalid minus valid trials," modeling the effects of ISI, Difficulty, and Experiment at the group level (see Materials and Methods, Magnetic Resonance Imaging).

Irrespective of task constraints, we found a main effect of "invalid > valid" trials in dorsal fronto-parietal regions, including the frontal eye-fields (FEF) bilaterally and the right superior parietal lobule (SPL, see **Figure 2A**; **Table 2**). As shown in the corresponding signal plots, these regions showed larger activity for invalid than valid trials across trial types (see positive effect sizes, on average, in these plots).

In contrast, in the ventral fronto-parietal cortex we found that the Validity effect was significantly modulated by the current task demands. Specifically, we found a 3-way interaction among Validity, Experiment, and ISI in the right inferior frontal gyrus (IFG) and in the right inferior parietal cortex, with a cluster comprising the temporo-parietal junction and the angular gyrus (TPJ and AngG, see **Figure 2B**; **Table 2**). The signal plots in **Figure 2B** show primarily a Validity effect (invalid > valid) for "ISI 0" trials of Experiment 2 (Bars 5 and 7, in these plots). In Experiment 1 the same Validity effect was larger for "ISI 250" than "ISI 0". The finding of opposite effects in the two experiments may reflect that, in voxel-wise analyses, voxels with opposite effects of one factor under the two levels of the other factor will obtain high interaction-statistics and appear as peaks in the corresponding whole-brain map [e.g., the interaction "(A1 – A2) – (B1 – B2)" will be largest in voxels where: "A1 > A2" and "B2 > B1"].

Aside these main results concerning the effect of Validity and the interaction of this with the other experimental factors, we also tested for the effects of target discrimination Difficulty and cue-to-target ISI, irrespective of the in/-validity of the multisensory cues. For this we used two separate ANOVAs: one pooling Validity and ISI (testing for Difficulty and Difficulty-by-Experiment interactions) and the other

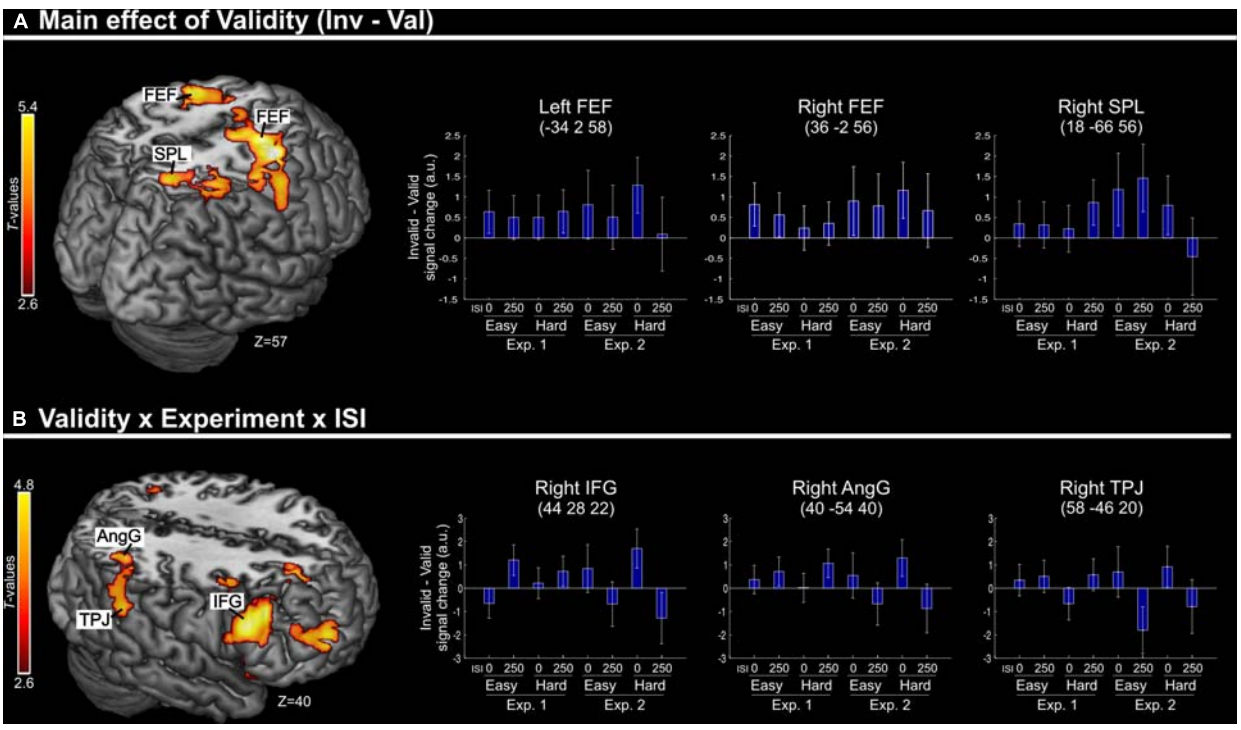


FIGURE 2 | (A) Transversal section through a 3D rendering of the canonical MNI template showing activations for the main effect of Validity (invalid minus valid trials), revealing recruitment of the dorsal fronto-parietal attention network. The corresponding signal plots show greater activity for invalid vs. valid trials for all these regions, irrespective of the other experimental factors related to top–down task demands. **(B)** Transversal section through a 3D rendering of a canonical MNI template showing activations modulated by the interaction among Validity, Experiment, and ISI, revealing that crossmodal semantic congruence interacted with top–down task-related factors in the right ventral fronto-parietal attention network. For display purposes, all activation maps are displayed at a threshold of p-uncorrected < 0.005. Signal plots report "invalid minus valid" trials (error bars represent 90% C.I.) in arbitrary units (a.u.). SPL, superior parietal lobule; FEF, frontal eye-fields; IFG, inferior frontal gyrus; TPJ, temporo-parietal junction; AngG, angular gyrus.

TABLE 2 | Brain activations associated with the effect of Validity.

Contrast	Area	*P*-corrected	Cluster size	*T*-value	*x y z*
Main effect of Validity	R FEF	<0.001	3082	5.53	36 –2 56
	L FEF			4.39	–34 2 58
	R SPL	0.025	613	3.96	18 –60 56
Interaction: Validity ×ISI× Experiment	R IFG	0.003	899	4.80	44 28 22
	R TPJ	0.026	608	3.69	58 –46 20
	R AngG			3.49	40 –54 40

*MNI coordinates of the peak (x, y, z), cluster size (number of supra-threshold voxels, estimated at p-uncorrected = 0.005), T-values, and p-FWE-corrected values are shown for areas showing a significant main effect of Validity (invalid minus valid trials, see **Figure 2A**) and for the interaction among Validity, ISI, and Experiment (see **Figure 2B**). R/L, left/right hemisphere; FEF, frontal eye-fields; SPL, superior parietal lobe; IFG, inferior frontal gyrus; TPJ, temporo-parietal junction; AngG, angular gyrus.*

pooling Validity and Difficulty (testing for ISI and ISI-by-Experiment interactions; see also Materials and Methods, Magnetic Resonance Imaging).

Across experiments, we found larger activation in dorsal fronto-parietal areas in "hard" compared to "easy" trials (see **Figure 3A**; panels on the left, including SPL and FEF), as well as for "long" compared to "short" ISIs (see **Figure 3B**, panel on the left; see also **Tables 3** and **4**). These comparisons also activated medial areas of the pre-motor cortex (SMA), as well as, the insula and the IFG, bilaterally (see **Figures 3A,B**, right panel). The medial pre-motor cortex often co-activates with other dorsal fronto-parietal areas and possibly plays a role in top–down attention control (e.g., Kastner and Ungerleider, 2001).

The IFG cluster found for "hard > easy" overlapped considerably with the IFG cluster for "ISI 250 > ISI 0", but this region of overlap was located more posteriorly than the IFG cluster involved in the significant 3-way interaction among Validity, Experiment, and ISI (see above, cf. **Figure 2A**). The opposite contrasts ("easy > hard" and "ISI 0 > ISI 250") also revealed a common region in the angular gyrus bilaterally (see **Figures 3A,B,** right panels, and **Tables 3** and **4**). Additionally, the contrast comparing "easy > hard" trials showed significant effects in medial frontal and medial parietal areas, traditionally associated with the default-mode network (Raichle et al., 2001; Greicius et al., 2003; see **Table 3**), plus the cerebellum that is seldom reported in studies of attention but possibly plays a role in visuo-spatial orienting (see Striemer et al., 2015).

Finally, for Experiment 2 we compared the two trial types that included a picture congruent with the centrally presented sound (i.e., valid and invalid trials) vs. the neutral trials, when the sound was unrelated to either of the two pictures. The corresponding ANOVA considered the average of "valid and

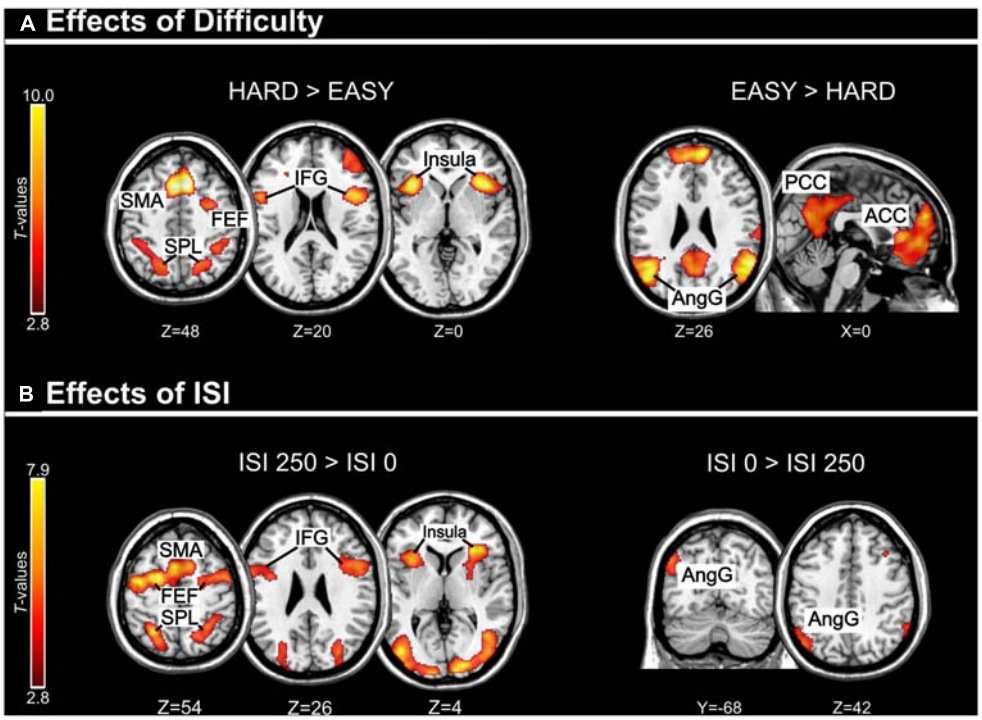


FIGURE 3 | (A) Effects of Difficulty. Left panel: Axial sections showing the effect of Hard minus Easy trials, which recruited the dorsal fronto-parietal attention network, plus the insula and the IFG, bilaterally. Right panel: Axial and sagittal sections showing Easy trials, compared to Hard trials, recruited the default mode network. **(B)** Effects of ISI. Left panel: Axial sections showing the effect of long ISI250 minus short ISI0 trials, which recruited the dorsal fronto-parietal attention network, plus the insula and the IFG, bilaterally. Right panel: Coronal and axial sections showing the effect of ISI0 minus ISI250, which recruited the left angular gyrus. For display purposes, all activation maps were displayed at a threshold of *p*-uncorrected < 0.005. SPL, superior parietal lobule; FEF, frontal eye-fields; SMA, supplementary motor cortex; IFG, inferior frontal gyrus; PCC/ACC, posterior/anterior cingulate cortex; AngG, angular gyrus.

TABLE 3 | Brain activations associated with the main effects of Difficulty.

Area	*P*-corrected	Cluster size	*T*-value	*x y z*
Hard > Easy				
SMA	<0.001	2792	9.97	8 18 52
R FEF	0.034	558	4.99	34 2 54
L FEF	*0.224*	*328*	*5.32*	*−24 2 54*
R SPL	0.007	766	4.25	24 −62 46
R aIPS	0.048	514	4.59	34 −32 36
L SPL	0.001	1066	4.63	−18 −64 46
L aIPS			4.38	−36 −38 36
R MFG	0.024	606	4.75	44 40 28
L Ins	0.002	935	7.45	−36 20 2
R Ins	<0.001	2209	7.10	42 20 −2
R IFG			5.91	48 6 18
L IFG	*0.079*	*454*	*3.48*	*−34 6 28*
Easy > Hard				
R AngG	<0.001	2002	7.60	60 −52 24
L AngG	<0.001	2511	7.19	−44 −54 20
PCC	<0.001	4410	5.61	−12 −40 40
ACC	<0.001	7251	7.02	8 56 24
L PHc	<0.001	1680	5.64	−24 −24 −22
Cereb	< 0.001	2821	5.99	28 −32 −36

MNI coordinates of the peak (x, y, z), cluster size (estimated at p-uncorrected = 0.005), T-values, and p-FWE-corrected values for the areas activated by the comparison between "Hard vs. Easy". Peaks in italics did not survive correction for multiple comparisons but are reported here if contralateral region survived correction. L/R, left/right hemisphere; SMA, supplementary motor area; FEF, frontal eye fields; SPL, superior parietal lobe; aIPS, anterior intra-parietal sulcus; MFG, middle frontal gyrus; Ins, insula; IFG, inferior frontal gyrus; AngG, angular gyrus; PCC, posterior cingulate cortex; ACC, anterior cingulate cortex; PHc, parahippocampal cortex; Cereb, cerebellum.

TABLE 4 | Brain activations associated with the main effects of ISI.

Area	*P*-corrected	Cluster size	*T*-value	*x y z*
ISI 250 > ISI 0				
SMA	<0.001	5426	6.16	−6 10 50
L FEF			6.85	−24 −2 56
R FEF			5.05	28 −4 50
L IFG			6.03	−56 4 40
R IFG			4.80	58 10 34
L OCC	<0.001	4569	7.86	−24 −96 12
L SPL			5.80	−26 −58 54
L aIPS			4.05	−44 −28 42
R OCC	<0.001	4097	6.61	22 −92 4
R SPL			4.82	24 −58 58
R aIPS			3.98	38 −28 40
R Ins	0.010	777	6.36	32 28 6
L Ins	*0.081*	*486*	*5.41*	*−32 18 4*
ISI 0 > ISI 250				
L AngG	0.041	577	4.63	−50 −68 42

MNI coordinates of the peak (x, y, z), cluster size (estimated at p-uncorrected = 0.005), T-values, and p-FWE-corrected values for areas activated by the comparison between the two ISI conditions (0 vs 250 ms). Peaks in italics did not survive correction for multiple comparisons but are reported here if contralateral region survived correction. L/R, left/right hemisphere; SMA, supplementary motor area; FEF, frontal eye fields; IFG, inferior frontal gyrus; OCC, visual occipital cortex; SPL, superior parietal lobe; aIPS, anterior intra-parietal sulcus; Ins, insula; AngG, angular gyrus.

invalid" trails minus the neutral condition and modeled the effects of ISI and Difficulty at the group level (see also Materials and Methods, Magnetic Resonance Imaging). Irrespective of ISI and discrimination Difficulty, this revealed activity in Heschl's gyrus, corresponding to the primary auditory cortex (right hemisphere: $x\ y\ z = 60\ -8\ -2$; cluster size = 935; $t = 7.17$; p-corrected = 0.001; left hemisphere: $x\ y\ z = -54\ -18\ 0$; cluster size = 1214; $t = 7.01$; p-corrected < 0.001). An additional activation was found in the right superior occipital gyrus ($x\ y\ z = 28\ -78\ 42$; cluster size = 1073; $t = 4.05$; p-corrected < 0.001), where the effect of the valid/invalid trials vs. neutral trials was larger for the hard than the easy trials. Since the acoustic characteristics of the sound in the valid and invalid trials were not matched with the neutral condition (i.e., cat/dog's "meow/woof" vs. frog's "croak"), here we will only underline the lack of activation in dorsal and ventral fronto-parietal regions without further discussing these effects in the auditory and occipital cortices.

Discussion

The aim of the present study was to assess whether crossmodal interactions between semantically related, but spatially separated, audio–visual signals can affect the processing of subsequent visual targets. Specifically, we investigated these influences when the audio–visual stimuli were task-irrelevant and did not produce any low-level spatial cue for visuo-spatial orienting (e.g., physical changes at the same/opposite location of the visual target). Behavioral and imaging data showed that audio–visual semantic congruence can influence the processing of visual targets, modulating the activity in dorsal and ventral regions of the parietal and the premotor cortices. The localization of these effects most likely correspond to the dorsal (SPL and FEF, cf. **Figure 2A**) and ventral (rTPJ and rIFG, cf. **Figure 2B**) attention control networks (e.g., see Corbetta and Shulman, 2002).

At the behavioral level, we found that the subjects' performance decreased when the visual targets were presented away from the location of the picture congruent with the sound, compared with targets presented at the same location ("invalid vs. valid" trials). A possible account of this crossmodal effect might be that the semantic relationship between the centrally presented sound and the congruent picture lead to a shift of visuo-spatial attention toward that picture. The presentation of the target on the opposite side would then require the re-orienting of visual attention from the location of the congruent picture to the position of the visual target, with corresponding behavioral costs (e.g., cueing costs for "invalid trials" in standard spatial cueing paradigms, see Posner and Cohen, 1984). However, the results of our study suggest that a more complex sequence of processes is taking place here. First, the behavioral data of Experiment 2 indicated that there was no cueing benefit for the congruent/valid conditions, but rather, that both invalid and valid trials lead to a reduction in performance compared to the

"neutral" trials (central sound unrelated to either pictures). The lack of cueing-benefits appears at odds with previous studies that have demonstrated faster orienting toward the location of a target-picture congruent with a central sound and no costs when the sound was congruent with a distractor-picture (Iordanescu et al., 2008, 2010).

However, in these previous studies, the pictures were task-relevant, whereas in our study the pictures were fully task-irrelevant. Therefore, in previous studies any effect associated with crossmodal semantic congruence (e.g., enhanced representation of the congruent picture, Iordanescu et al., 2008) would match the current task set/target template: that is, stimulus-related semantic congruence and task-related, goal-driven attention work together to boost the processing of the same picture (i.e., the search target; see also Iordanescu et al., 2010). In contrast, in our study the objects displayed in the pictures were irrelevant; the subject's only task was to judge the orientation of the Gabor patches presented after the audio–visual objects. We suggest that in this situation, the brain detected the semantic correspondence between the central sound and one of the two the pictures, registering the position of the congruent picture. However, because the task did not involve any object discrimination or spatial orienting toward that picture (cf. Iordanescu et al., 2008, 2010), goal-related attention control generated inhibitory rather than boosting signals, thus with opposite effects of semantic congruence and goal-directed attention. In this view, even on "valid" trials (congruent-picture on the same side as the Gabor-target), the pictures in our study would be more comparable with the distractor pictures than the target pictures used in previous search tasks that also combined pictures with semantically in-/congruent central sounds. Thus, the interplay between semantic congruency and task goal might explain the overall decrease of performance when the trial included a congruent picture (valid and invalid conditions) compared with trials containing a sound unrelated to both pictures (neutral condition). In the latter case, there would be no need to ignore and suppress the crossmodally enhanced object representation, making the discrimination of the subsequent visual targets faster.

Despite the lack of relationship between the object/picture associated with semantic congruence and the current task-goal may have reduced the processing the irrelevant pictures, our data showed that the presentation of the visual target in the hemifield opposite to the congruent-picture lead to several behavioral and imaging effects. Behaviourally, we found a further reduction of performance for invalid vs. valid trials but only in the "most difficult" conditions of Experiment 2 (finer orientation discrimination and additional distractor Gabor). Similarly, we found an effect of validity and task demand in our imaging results: activity in the ventral fronto-parietal network was modulated by the interaction between validity (invalid > valid) and task demands (see **Figure 2B**) while in the dorsal fronto-parietal network, the effect of validity was observed across all conditions (**Figure 2A**).

The imaging findings in the dorsal system indicate that, in spite of any object-related suppressive mechanism as discussed, the brain did register the location for the task-irrelevant picture coupled with the semantically congruent sound. Several recent studies have reported activation of dorsal fronto-parietal regions in response to salient visual stimuli, even when these were task-irrelevant (e.g., Bogler et al., 2011; Nardo et al., 2011; see also Schall and Hanes, 1993; Constantinidis and Steinmetz, 2001). In the context of multisensory spatial processing, Nardo et al. (2014) showed that the saliency of sounds in complex and naturalistic audio–visual video clips modulated activity in the posterior parietal cortex. This saliency-related modulation of activity in the parietal cortex was found only when the auditory stimuli were spatially congruent (i.e., on the same side) as the main visual event in the scene. Analogous with that study, we propose that here the semantic congruence between one picture and the centrally presented sound generated a processing priority bias in dorsal parietal regions (see Gottlieb et al., 1998; and Awh et al., 2012; Ptak, 2012, for reviews), which required updating when the task-relevant visual target was presented on the opposite side to congruent audio–visual pairs (invalid trials). While inhibitory interactions between object-related crossmodal processing and goal-related attention did not result in any behavioral benefit on valid trials, the spatial updating operations affected activity in the dorsal attention network and were observed using fMRI in all invalid conditions.

In contrast, the activation of the right ventral attention system (rIFG and rTPJ, see **Figure 2B**) was observed only when the visual discrimination task required top–down control to identify the task-relevant Gabor patch in Experiment 2. Many previous imaging studies have reported activation of the ventral attention network comparing invalid versus valid trials following predictive central cues (e.g., Arrington et al., 2000; Corbetta et al., 2000). While these effects might be associated with stimulus-driven shifts of spatial attention triggered by the onset of a stimulus at an unattended location, recent evidence indicates that the activation of the ventral system reflects a more complex interplay between the stimulus-driven signals and other factors associated with current task demands (e.g., Kincade et al., 2005; Indovina and Macaluso, 2007; Natale et al., 2009; see Corbetta et al., 2008, for a review). In the current study, the presence of a bilateral stimulation in the target phase of Experiment 2 lead to high demands on top–down control (see also overall low performance in Experiment 2 vs. Experiment 1, **Figure 1B**). Here, we suggest that the need of selecting one of the two Gabor patches based on *a priori* internal information (i.e., task instructions) produced top–down demands that interacted with any spatial priority bias associated with the congruent picture, leading to the activation of the right ventral network. This process might not strictly involve any shift of spatial attention but possibly entails the update of spatial predictions generated during the processing of the multisensory stimuli (Geng and Vossel, 2013; Macaluso and Doricchi, 2013; see also Shams and Beierholm, 2010, for a related and more formal framework).

It should be noted that such an expectation/prediction framework has been previously put forward in the context of endogenous spatial cueing, while here we suggest that the initial priority bias was generated by the irrelevant and non-predictive (audio–visual) stimuli. This difference could explain

the finding here that ventral network activity was observed only at the short ISI (cf. interaction between Validity, Experiment, and ISI; **Figure 2B**). Unlike endogenous cues that are typically associated with long-lasting spatial effects, here we would expect any bias generated by the irrelevant audio–visual stimuli to be relatively short-lived, and therefore, any process triggered by the interaction between these crossmodal effects and top–down control signals (i.e., identification and selection of the target Gabor) would take place only when the visual targets were presented in close temporal proximity of the audio–visual stimuli, i.e., at short ISIs. It should be noted that at the short ISI, we also observed an overall reduction of behavioral performance (cf. main effect of ISI). Because the visual targets were presented at the same location as the irrelevant pictures, this suggests a possible role of forward masking when ISI = 0. However, the behavioral data also showed orientation discrimination accuracies up to 95% (see **Table 1**, "easy" conditions) demonstrating the targets were well visible, incompatible with a major role of forward masking.

Aside from the specific mechanisms and interpretations that we proposed above, the current findings provide us with novel insights about the interplay between top–down and bottom–up signals in the context of multisensory processing. Extensive research has highlighted the complex interplay between these types of signals in multisensory integration (e.g., see Talsma et al., 2010, for a review). In the current study, we did not measure the crossmodal binding between the sound and the congruent picture nor of perceived sound location shift toward the congruent picture. However, we observed changes in visuo-spatial processing (valid vs. invalid trials) that provide us with an indirect measure that semantic crossmodal interactions affect how attention is allocated in the visual space. Most importantly, our experimental setup allowed us to demonstrate these effects in the absence of any physical low-level change or task-related association between the crossmodal cues the subsequent visual target. In this setup, the results cannot be attributed to any direct effect of purely stimulus-driven or purely goal-driven attention. In contrast, previous studies typically relied on physical changes of the sensory input at peripheral locations (e.g., spatial cueing paradigms, Santangelo et al., 2008a,b; see Spence and Santangelo, 2009; and Spence, 2010 for a review; see also Van der Burg et al., 2008; Matusz and Eimer, 2011); and/or on the existence of some goal-related relationship between the audio–visual stimuli and the to-be-judged visual stimuli (e.g., Matusz et al., 2015). In the former case, stimulus-driven signals are likely to play a direct role, with the spatial and/or temporal correspondence of the sensory changes acting as the main cue triggering associations across modalities (see Santangelo et al., 2008a,b; Van der Burg et al., 2008). Here, the presentation of two pictures and a centrally presented sound in all conditions eliminates any such low-level cues and ensures that crossmodal influences on visual attention could be specifically attributed to the semantic correspondence between the auditory and visual input.

Uncoupling the audio–visual stimuli (animal pictures and sounds) from the sole visual task (Gabor orientation discrimination) also enabled us to demonstrate crossmodal semantic influences in the absence of any goal-related signal linking the audio–visual stimuli with the task-relevant visual target. Such task-based relationships characterized previous studies where the crossmodally enhanced visual object was also the target of the search task (Iordanescu et al., 2008, 2010) or shared some task-relevant feature with the search target (Matusz et al., 2015). We extend these previous results by demonstrating that crossmodal semantic congruence affects visual attention despite opposing top–down task constraints. These results are in line with results from previous studies indicating that crossmodal effects can occur in the absence of goal-directed attention toward the audio–visual stimuli (Alsius et al., 2005; see also Santangelo and Spence, 2007; but note that in these studies physical changes, rather than semantic congruence, might have contributed to associate the auditory and visual signals; see also Van der Burg et al., 2012). Nonetheless, other studies have found that increasing the demands of a primary task, while presenting participants with multisensory stimuli, can reduce the interaction between the multisensory input (e.g., Alsius et al., 2005; see also Talsma et al., 2010, for a review). Contrary to these studies, we found several crossmodal effects only when the demands of the primary visual task were high (cf. behavioral results and the pattern of activation in the ventral attention system in Experiment 2). A possible reason for these differences is that here task demands specifically concerned the target phase of the trial and not the resources available to process the audio–visual stimuli. Thus, task demands can have multifaceted consequences on the processing of multisensory stimuli, including suppression (Alsius et al., 2005; using a dual-task approach), no effect (Matusz et al., 2015; presence of competing distractors in a crossmodal interference paradigm), and selective influences only under high demands (the current study, where crossmodal signals interact with attention control operations under high demands only, as discussed above); see also Koelewijn et al. (2010) and Talsma et al. (2010) for reviews on the impact of task demands on multisensory processing.

Conclusion

The present study demonstrated that crossmodal semantic congruence between spatially separated audio–visual stimuli can affect visual-spatial attention control. We found these crossmodal effects in the absence of any physical change that might capture visual attention in a direct bottom–up manner, and of any goal-related relationship between the audio–visual stimuli and the visual targets; that is, with fully task-irrelevant audio–visual stimuli. We discussed these effects in relation to multiple signals associated with the processing of irrelevant audio–visual stimuli and the top–down task demands of the primary visual task. We propose that the semantic congruence between the task-irrelevant audio–visual stimuli generates processing biases that require updating when a subsequent task-relevant visual target is presented at a different location. We relate these updating operations to saliency representations in the dorsal

attention network and with the interplay between stimulus- and task-related signals in the ventral attention network. We conclude that crossmodal semantic congruence can affect visual-spatial processing in the absence of any direct bottom–up or goal-related influences, and highlight the role of the fronto-parietal attention control networks in mediating the effect of multisensory processing on visual attention.

Acknowledgments

The Neuroimaging Laboratory, Santa Lucia Foundation, is supported by The Italian Ministry of Health. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ ERC grant agreement n. 242809.

References

Alsius, A., Navarra, J., Campbell, R., and Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Curr. Biol.* 15, 839–843. doi: 10.1016/j.cub.2005.03.046

Arrington, C., Carr, T., Mayer, A., and Rao, S. (2000). Neural mechanisms of visual attention: object-based selection of a region in space. *J. Cogn. Neurosci.* 12, 106–117. doi: 10.1162/089892900563975

Awh, E., Belopolsky, A. V., and Theeuwes, J. (2012). Top-down versus bottom-up attentional control: a failed theoretical dichotomy. *Trends Cogn. Sci.* 16, 437–443. doi: 10.1016/j.tics.2012.06.010

Bogler, C., Bode, S., and Haynes, J. D. (2011). Decoding successive computational stages of saliency processing. *Curr. Biol.* 21, 1667–1671. doi: 10.1016/j.cub.2011.08.039

Brown, J. M., and Denney, H. I. (2007). Shifting attention into and out of objects: evaluating the processes underlying the object advantage. *Percept. Psychophys.* 69, 606–618. doi: 10.3758/BF03193918

Bruyer, R., and Brysbaert, M. (2011). Combining speed and accuracy in cognitive psychology: is the Inverse Efficiency Score (IES) a better dependent variable than the mean Reaction Time (RT) and the Percentage of Errors (PE)? *Psychol. Belg.* 51, 5–13. doi: 10.5334/pb-51-1-5

Calvert, G. A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cereb. Cortex* 11, 1110–1123. doi: 10.1093/cercor/11.12.1110

Chen, Z. (2012). Object-based attention: a tutorial review. *Atten. Percept. Psychophys.* 74, 784–802. doi: 10.3758/s13414-012-0322-z

Constantinidis, C., and Steinmetz, M. A. (2001). Neuronal responses in area 7a to multiple-stimulus displays: I. Neurons encode the location of the salient stimulus. *Cereb. Cortex* 11, 581–591. doi: 10.1093/cercor/11.7.581

Corbetta, M., Kincade, J. M., Ollinger, J. M., McAvoy, M. P., and Shulman, G. L. (2000). Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nat. Neurosci.* 3, 292–297. doi: 10.1038/73009

Corbetta, M., Patel, G., and Shulman, G. L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron* 58, 306–324. doi: 10.1016/j.neuron.2008.04.017

Corbetta, M., and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* 3, 201–215. doi: 10.1038/nrn755

De Lucia, M., Clarke, S., and Murray, M. M. (2010). A temporal hierarchy for conspecific vocalization discrimination in humans. *J. Neurosci.* 30, 11210–11221. doi: 10.1523/JNEUROSCI.2239-10.2010

Doehrmann, O., and Naumer, M. J. (2008). Semantics and the multisensory brain: how meaning modulates processes of audio-visual integration. *Brain Res.* 1242, 136–150. doi: 10.1016/j.brainres.2008.03.071

Downar, J., Crawley, A. P., Mikulis, D. J., and Davis, K. D. (2000). A multimodal cortical network for the detection of changes in the sensory environment. *Nat. Neurosci.* 3, 277–283. doi: 10.1038/72991

Driver, J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature* 381, 66–68. doi: 10.1038/381066a0

Eimer, M., and Kiss, M. (2008). Involuntary attentional capture is determined by task set: evidence from event-related brain potentials. *J. Cogn. Neurosci.* 20, 1423–1433. doi: 10.1162/jocn.2008.20099

Fiebelkorn, I. C., Foxe, J. J., Schwartz, T. H., and Molholm, S. (2010). Staying within the lines: the formation of visuospatial boundaries influences multisensory feature integration. *Eur. J. Neurosci.* 31, 1737–1743. doi: 10.1111/j.1460-9568.2010.07196.x

Folk, C. L., Remington, R., and Johnson, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *J. Exp. Psychol. Hum. Percept. Perform.* 19, 676–681. doi: 10.1037//0096-1523.18.4.1030

Friston, K. J., Glaser, D. E., Henson, R. N., Kiebel, S., Phillips, C., and Ashburner, J. (2002). Classical and Bayesian inference in neuroimaging: applications. *Neuroimage* 16, 484–512. doi: 10.1006/nimg.2002.1091

Geng, J. J., and Vossel, S. (2013). Re-evaluating the role of TPJ in attentional control: contextual updating? *Neurosci. Biobehav. Rev.* 37, 2608–2620. doi: 10.1016/j.neubiorev.2013.08.010

Gottlieb, J. P., Kusunoki, M., and Goldberg, M. E. (1998). The representation of visual salience in monkey parietal cortex. *Nature* 391, 481–484. doi: 10.1038/35135

Greicius, M. D., Krasnow, B., Reiss, A. L., and Menon, V. (2003). Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* 100, 253–258. doi: 10.1073/pnas.0135058100

Hill, K. T., and Miller, L. M. (2010). Auditory attentional control and selection during cocktail party listening. *Cereb. Cortex* 20, 583–590. doi: 10.1093/cercor/bhp124

Indovina, I., and Macaluso, E. (2007). Dissociation of stimulus relevance and saliency factors during shifts of visuospatial attention. *Cereb. Cortex* 17, 1701–1711. doi: 10.1093/cercor/bhl081

Iordanescu, L., Grabowecky, M., Franconeri, S., Theeuwes, J., and Suzuki, S. (2010). Characteristic sounds make you look at target objects more quickly. *Atten. Percept. Psychophys.* 72, 1736–1741. doi: 10.3758/APP.72.7.1736

Iordanescu, L., Guzman-Martinez, E., Grabowecky, M., and Suzuki, S. (2008). Characteristic sounds facilitate visual search. *Psychon. Bull. Rev.* 15, 548–554. doi: 10.3758/PBR.15.3.548

Kastner, S., and Ungerleider, L. G. (2001). The neural basis of biased competition in human visual cortex. *Neuropsychologia* 39, 1263–1276. doi: 10.1016/S0028-3932(01)00116-6

Kincade, J. M., Abrams, R. A., Astafiev, S. V., Shulman, G. L., and Corbetta, M. (2005). An event-related functional magnetic resonance imaging study of voluntary and stimulus-driven orienting of attention. *J. Neurosci.* 25, 4593–4604. doi: 10.1523/JNEUROSCI.0236-05.2005

Koelewijn, T., Bronkhorst, A., and Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: a review of audiovisual studies. *Acta Psychol.* 134, 372–384. doi: 10.1016/j.actpsy.2010.03.010

Krumbholz, K., Nobis, E. A., Weatheritt, R. J., and Fink, G. R. (2009). Executive control of spatial attention shifts in the auditory compared to the visual modality. *Hum. Brain Mapp.* 30, 1457–1469. doi: 10.1002/hbm.20615

Macaluso, E., and Doricchi, F. (2013). Attention and predictions: control of spatial attention beyond the endogenous-exogenous dichotomy. *Front. Hum. Neurosci.* 7:685. doi: 10.3389/fnhum.2013.00685

Macaluso, E., Eimer, M., Frith, C. D., and Driver, J. (2003). Preparatory states in crossmodal spatial attention: spatial specificity and possible control mechanisms. *Exp. Brain Res.* 149, 62–74. doi: 10.1007/s00221-002-1335-y

Macaluso, E., Frith, C. D., and Driver, J. (2002). Supramodal effects of covert spatial orienting triggered by visual or tactile events. *J. Cogn. Neurosci.* 14, 389–401. doi: 10.1162/089892902317361912

Matusz, P. J., Broadbent, H., Ferrari, J., Forrest, B., Merkley, R., and Scerif, G. (2015). Multi-modal distraction: insights from children's limited attention. *Cognition* 136, 156–165. doi: 10.1016/j.cognition.2014.11.031

Matusz, P. J., and Eimer, M. (2011). Multisensory enhancement of attentional capture in visual search. *Psychon. Bull. Rev.* 18, 904–909. doi: 10.3758/s13423-011-0131-8

McDonald, J. J., Teder-Sälejärvi, W. A., and Hillyard, S. A. (2000). Involuntary orienting to sound improves visual perception. *Nature* 407, 906–908. doi: 10.1038/35038085

McDonald, J. J.,Teder-Sälejärvi, W. A., and Ward, L. M. (2001). Multisensory integration and crossmodal attention effects in the human brain. *Science* 292, 1791–1791. doi: 10.1126/science.292.5523.1791a

Nardo, D., Santangelo, V., and Macaluso, E. (2011). Stimulus-driven orienting of visuo-spatial attention in complex dynamic environments. *Neuron* 69, 1015–1028. doi: 10.1016/j.neuron.2011.02.020

Nardo, D., Santangelo, V., and Macaluso, E. (2014). Spatial orienting in complex audiovisual environments. *Hum. Brain Mapp.* 35, 1597–1614. doi: 10.1002/hbm.22276

Natale, E., Marzi, C. A., and Macaluso, E. (2009). FMRI correlates of visuo-spatial reorienting investigated with an attention shifting double-cue paradigm. *Hum. Brain Mapp.* 30, 2367–2381. doi: 10.1002/hbm.20675

Noppeney, U., Josephs, O., Hocking, J., Price, C. J., and Friston, K. J. (2007). The effect of prior visual information on recognition of speech and sounds. *Cereb. Cortex* 18, 598–609. doi: 10.1093/cercor/bhm091

Penny, W. D., and Holmes, A. P. (2004). "Random effects analysis," in *Human Brain Function*, 2 Edn, eds R. S. J. Frackowiak, K. J. Friston, R. Frith, K. J. Dolan, C. J. Price, S. Zeki, et al. (San Diego, CA: Academic Press), 843–850.

Posner, M. I., and Cohen, Y. (1984). "Components of visual orienting," in *Attention and Performance*, eds H. Bouma and D. Bowhuis (Hillsdale, NJ: Lawrence Erlbaum), 531–556. doi: 10.1093/cercor/bhm091

Ptak, R. (2012). The frontoparietal attention network of the human brain action, saliency, and a priority map of the environment. *Neuroscientist* 18, 502–515. doi: 10.1177/1073858411409051

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., and Shulman, G. L. (2001). A default mode of brain function. *Proc. Natl. Acad. Sci. U.S.A.* 98, 676–682. doi: 10.1073/pnas.98.2.676

Rauschenberger, R. (2003). Attentional capture by auto-and allo-cues. *Psychon. Bull. Rev.* 10, 814–842. doi: 10.3758/BF03196545

Ruz, M., and Lupiáñez, J. (2002). A review of attentional capture: on its automaticity and sensitivity to endogenous control. *Psicológica* 23, 283–369.

Santangelo, V., Ho, C., and Spence, C. (2008a). Capturing spatial attention with multisensory cues. *Psychon. Bull. Rev.* 15, 398–403. doi: 10.3758/pbr.15.2.398

Santangelo, V., Van der Lubbe, R. H., Olivetti Belardinelli, M., and Postma, A. (2008b). Multisensory integration affects ERP components elicited by exogenous cues. *Exp. Brain Res.* 185, 269–277. doi: 10.1007/s00221-007-1151-5

Santangelo, V., and Macaluso, E. (2012). "Spatial attention and audiovisual processing," in *The New Handbook of Multisensory Processes*, ed. B. E. Stein (Cambridge, MA: The MIT Press), 359–370.

Santangelo, V., and Spence, C. (2007). Multisensory cues capture spatial attention regardless of perceptual load. *J. Exp. Psychol. Hum. Percept. Perform.* 33, 1311–1321. doi: 10.1037/0096-1523.33.6.1311

Santangelo, V., and Spence, C. (2008). Is the exogenous orienting of spatial attention truly automatic? Evidence from unimodal and multisensory studies. *Conscious. Cogn.* 17, 989–1015. doi: 10.1016/j.concog.2008.02.006

Santangelo, V., Van der Lubbe, R. H., Olivetti Belardinelli, M., and Postma, A. (2006). Spatial attention triggered by unimodal, crossmodal, and bimodal exogenous cues: a comparison of reflexive orienting mechanisms. *Exp. Brain Res.* 173, 40–48. doi: 10.1007/s00221-006-0361-6

Schall, J. D., and Hanes, D. P. (1993). Neural basis of saccade target selection in frontal eye field during visual search. *Nature* 366, 467–469. doi: 10.1038/366467a0

Shams, L., and Beierholm, U. R. (2010). Causal inference in perception. *Trends Cogn. Sci.* 14, 425–432. doi: 10.1016/j.tics.2010.07.001

Spence, C. (2010). Crossmodal spatial attention. *Ann. N. Y. Acad. Sci.* 1191, 182–200. doi: 10.1111/j.1749-6632.2010.05440.x

Spence, C., and Driver, J. (1997). Audiovisual links in exogenous covert spatial orienting. *Percept. Psychophys.* 59, 1–22. doi: 10.3758/BF03206843

Spence, C., and Driver, J. (1998). Inhibition of return following an auditory cue: the role of central reorienting events. *Exp. Brain Res.* 118, 352–360. doi: 10.1007/s002210050289

Spence, C., and Santangelo, V. (2009). Capturing spatial attention with multisensory cues: a review. *Hear. Res.* 258, 134–142. doi: 10.1016/j.heares.2009.04.015

Stevenson, R. A., Wallace, M. T., and Altieri, N. (2014). The interaction between stimulus factors and cognitive factors during multisensory integration of audiovisual speech. *Front. Psychol.* 5:352. doi: 10.3389/fpsyg.2014.00352

Striemer, C. L., Chouinard, P. A., Goodale, M. A., and de Ribaupierre, S. (2015). Overlapping neural circuits for visual attention and eye movements in the human cerebellum. *Neuropsychologia* 69, 9–21. doi: 10.1016/j.neuropsychologia.2015.01.024

Talsma, D., Senkowski, D., Soto-Faraco, S., and Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends Cogn. Sci.* 14, 400–410. doi: 10.1016 /j.tics.2010.06.008

Taylor, K. I., Moss, H. E., Stamatakis, E. A., and Tyler, L. K. (2006). Binding crossmodal object features in perirhinal cortex. *Proc. Natl. Acad. Sci. U.S.A.* 103, 8239–8244. doi: 10.1073/pnas.0509704103

Townsend, J. T., and Ashby, F. G. (1983). *Stochastic Modeling of Elementary Psychological Processes*. Cambridge: Cambridge University Press.

van den Brink, R. L., Cohen, M. X., van der Burg, E., Talsma, D., Vissers, M. E., and Slagter, H. A. (2014). Subcortical, modality-specific pathways contribute to multisensory processing in humans. *Cereb. Cortex* 24, 2169–2177. doi: 10.1093/cercor/bht069

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., and Theeuwes, J. (2008). Pip and pop: nonspatial auditory signals improve spatial visual search. *J. Exp. Psychol. Hum. Percept. Perform.* 34, 1053–1065. doi: 10.1037/0096-1523.34.5.1053

Van der Burg, E., Olivers, C., and Theeuwes, J. (2012). The attentional window modulates capture by audiovisual events. *PLoS ONE* 7:e39137. doi: 10.1371/journal.pone.0039137

Vecera, S. P., and Farah, M. J. (1994). Does visual attention select objects or locations? *J. Exp. Psychol. Gen.* 123, 146–160. doi: 10.1037/0096-3445.123.2.146

Yantis, S., Schwarzbach, J., Serences, J. T., Carlson, R. L., Steinmetz, M. A., Pekar, J. J., et al. (2002). Transient neural activity in human parietal cortex during spatial attention shifts. *Nat. Neurosci.* 5, 995–1002. doi: 10.1038/nn921

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in
INTEGRATIVE NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 08 December 2014
doi: 10.3389/fnint.2014.00092

Content congruency and its interplay with temporal synchrony modulate integration between rhythmic audiovisual streams

Yi-Huang Su*

Department of Movement Science, Faculty of Sport and Health Sciences, Technical University of Munich, Munich, Germany

Edited by:
Salvador Soto-Faraco, Universitat Pompeu Fabra, Spain

Reviewed by:
Ian C. Fiebelkorn, Princeton University, USA
Manuel R. Mercier, Albert Einstein College of Medicine of Yeshiva University, USA

***Correspondence:**
Yi-Huang Su, Department of Movement Science, Faculty of Sport and Health Sciences, Technical University of Munich, Georg-Brauchle-Ring 62, 80992 Munich, Germany
e-mail: yihuang.su@tum.de

Both lower-level stimulus factors (e.g., temporal proximity) and higher-level cognitive factors (e.g., content congruency) are known to influence multisensory integration. The former can direct attention in a converging manner, and the latter can indicate whether information from the two modalities belongs together. The present research investigated whether and how these two factors interacted in the perception of rhythmic, audiovisual (AV) streams derived from a human movement scenario. Congruency here was based on sensorimotor correspondence pertaining to rhythm perception. Participants attended to bimodal stimuli consisting of a humanlike figure moving regularly to a sequence of auditory beat, and detected a possible auditory temporal deviant. The figure moved either downwards (congruently) or upwards (incongruently) to the downbeat, while in both situations the movement was either synchronous with the beat, or lagging behind it. Greater cross-modal binding was expected to hinder deviant detection. Results revealed poorer detection for congruent than for incongruent streams, suggesting stronger integration in the former. False alarms increased in asynchronous stimuli only for congruent streams, indicating greater tendency for deviant report due to visual capture of asynchronous auditory events. In addition, a greater increase in perceived synchrony was associated with a greater reduction in false alarms for congruent streams, while the pattern was reversed for incongruent ones. These results demonstrate that content congruency as a top-down factor not only promotes integration, but also modulates bottom-up effects of synchrony. Results are also discussed regarding how theories of integration and attentional entrainment may be combined in the context of rhythmic multisensory stimuli.

Keywords: multisensory integration, rhythm, content congruency, audiovisual synchrony, attention

INTRODUCTION

A key function of the perceptual system is its ability to continuously track and integrate information originating from different sensory modalities. Previous investigations of multisensory integration, employing paradigms with relatively simple bimodal stimuli (Meredith and Stein, 1983; Alvarado et al., 2007; Stein and Stanford, 2008; Stevenson et al., 2014a), have identified several factors related to the stimulus features that mediate the integration process. Amongst the most robust findings is that temporal proximity between the bimodal events promotes cross-modal integration (Chen and Vroomen, 2013). Integration is typically shown as enhanced neuronal response as well as behavioral advantages to concurrent multisensory information, compared to those in the most effective unisensory situation. Findings along this line suggest that temporally convergent information directs (or "captures") attention in a stimulus-driven, bottom-up manner (Van der Burg et al., 2008), which facilitates subsequent binding of the inter-sensory signals (Fiebelkorn et al., 2010; Koelewijn et al., 2010; Talsma et al., 2010).

Integration in more complex multisensory stimuli can also be modulated by aspects of higher-level stimulus content. One such factor that especially concerns the present research is content congruency, i.e., the perceived content match between the bimodal stimuli based on their semantic correspondence or consistency (Doehrmann and Naumer, 2008). Stimuli that are matched in content tend to be treated as originating from the same source, and are thus more likely to be integrated by the perceptual system—also referred to as the *unity assumption* (Welch and Warren, 1980). This has been demonstrated in audiovisual (AV) speech, in which integration is favored when the spoken sound matches the gender of the talking face (Vatakis and Spence, 2007), or when the spoken syllable matches the facial articulatory movement (van Wassenhove et al., 2007; Ten Oever et al., 2013), compared to when they mismatch. In non-speech AV human actions, stronger integration has been found for a drumming movement paired with congruent than with incongruent impact sounds (Arrighi et al., 2006; Petrini et al., 2009a). In a similar vein, effects of AV content congruency have also been shown in biological motion perception. In

those studies, visual detection of a walking humanlike point-light figure, ("PLF", Johansson, 1973; Blake and Shiffrar, 2007) embedded in random dots is enhanced if the accompanying sounds convey natural footstep information compared to artificial tones (Thomas and Shiffrar, 2010, 2013), or when the direction of the moving sounds matches that of the walking PLF (Brooks et al., 2007; Schouten et al., 2011; Wuerger et al., 2012). In the scenarios discussed thus far, perceived content congruency relies on various learned associations between the bimodal stimuli. Such prior knowledge represents a cognitive factor that modulates multisensory integration in a top-down manner, which may also interact with lower-level stimulus factors (e.g., temporal relation) in the perceptual decision (Ten Oever et al., 2013; Stevenson et al., 2014b). Similarly, while temporal alignment drives attention in a bottom-up manner for cross-modal binding (i.e., through attentional spread), highly learned associations between bimodal stimuli can additionally activate a top-down attentional mechanism for integration (Fiebelkorn et al., 2010).

Perhaps not surprisingly, both the speech and non-speech AV stimuli mentioned above involve human movements, in which the sounds are consequent upon the viewed actions. That is, the auditory and the visual information is causally linked. Thus, based on prior experiences, a perceiver will generate certain expectations upon stimulus presentation, which can be used for temporal prediction in the ongoing bimodal streams (Lee and Noppeney, 2014; van Atteveldt et al., 2014). For example, in natural AV speech, the lip movements and the spoken sounds are temporally correlated, and the former typically precedes the latter (Chandrasekaran et al., 2009). This makes it possible for an observer to use the visual cues to predict when the sounds should occur (van Wassenhove et al., 2005; Zion Golumbic et al., 2013), by which attention can be directed to the expected points in time to support auditory processing (Lakatos et al., 2008) and, eventually, multisensory integration (van Atteveldt et al., 2014). Similarly, in non-speech AV actions such as drumming, the trajectory of the arm movement predicts the temporal occurrence of the impact sounds. The availability of visual movement cues for cross-modal prediction is also found to affect the strength of integration in this scenario (Arrighi et al., 2006; Petrini et al., 2009b). Notably, the predictive mechanism can be influenced by cognitive factors such as content congruency. Streams matched in content tend to be attributed to the same source of action, which then increases the likelihood that a perceiver would use cues in one modality to predict event occurrences in the other modality.

Given the role of the stimulus (temporal) and the cognitive factors, as well as the predictive mechanism in multisensory integration, one question may arise from here. In the course of AV action perception, besides the cross-modal prediction that is perpetuated by the stimulus correlation and the content match, there exists a possibility of temporal prediction within each modality. This may be especially true for bimodal stimuli that yield a perceivable periodicity in both sensory streams. The most prominent examples are rhythmic human movements that produce rhythmic sounds, e.g., drumming (Arrighi et al., 2006; Petrini et al., 2009b), hand clapping (Sevdalis and Keller, 2010), or walking (Thomas and Shiffrar, 2010, 2013). Speech, albeit with temporal variations, is also rhythmic along various time scales (Rothermich et al., 2012; Ghazanfar, 2013; Patel, 2014). For each modality, the underlying periodicity in the rhythmic stimulus can entrain attention accordingly, leading the perceiver to generate expectations/predictions of event occurrences at regular points in time (*Dynamic Attending Theory*, "DAT", Large and Jones, 1999). As a result, stimulus processing is enhanced at these expected moments. This has been most frequently reported in the auditory modality (Jones et al., 2002; Large and Snyder, 2009; Repp, 2010); however, recent studies demonstrate that temporal entrainment can occur cross-modally, such that attention entrained by auditory rhythms can facilitate visual processing (Bolger et al., 2013, 2014), and the other way around (Su, 2014a). As such, in the course of multisensory perception of rhythmic human movements, both within-modal and cross-modal predictions may occur, and both mechanisms can deploy attention to convergent points in time that in turn promotes integration. Because integration is often measured by tasks that require judging the relation between both streams, i.e., synchrony judgment (SJ) or temporal order judgment (TOJ; Vroomen and Keetels, 2010), it is difficult to disentangle these two modes of prediction. It thus remains unclear to what extent each prediction mode contributes to the attentional deployment in multisensory perception, and whether either or both interact with other stimulus and cognitive factors.

Motivated by these issues, the present study set out to address several questions in multisensory perception involving continuous, rhythmic human movements. First, as opposed to causally linked AV actions, would the top-down effect of content congruency on integration be obtained in scenarios where the sounds are *not* caused by the movement, but rather that the movement is coordinated with extraneous sounds? The rationale behind was that content congruency can be based on various forms of association, and its effect has also been found for stimuli exhibiting abstract, synesthetic correspondences (Parise and Spence, 2009). In terms of humans moving along with sounds, such as dancing to music, a correspondence may exist as to which kind of movement is typically performed with regard to the rhythm of continuous sounds: For example, humans tend to move their body vertically to a musical beat (Toiviainen et al., 2010), and they most often move downwards rather than upwards to the beat (Miura et al., 2011; Su, 2014b). As no study has examined congruency regarding such action-perception association, this constituted the first question of interest in the present research. The next question asked whether, in this particular scenario, temporal proximity (i.e., synchrony) between the auditory and visual streams would also direct attention in a bottom-up manner to promote integration. More importantly, the focus was whether this stimulus-driven, temporal factor would interact with the cognitive factor of content congruency, which has recently been shown in AV perception of speech syllables (Ten Oever et al., 2013) but has not been investigated in a non-speech action domain. Finally, as both the auditory and visual streams were rhythmic in this case, it was of interest to examine whether within-modal or cross-modal predictive mechanism plays a dominant role when the task probes the perceptual outcome in one modality.

To this end, the present study employed an AV paradigm that resembled the scenario of observing a person moving to music. Here, a humanlike figure performed a whole-body bouncing movement vertically and periodically (as in Su, 2014a,c) to a sequence of regular auditory beat. The movement could be either congruent (moving *down* to the beat) or incongruent (moving *up* to the beat) to the auditory rhythm, and in both cases the movement could be either synchronous with the beat, or lagging behind the beat. Instead of a SJ or TOJ task, the present task required detection of a temporal deviant only in the auditory stream. Because the auditory sequence had a clear periodicity and the task was only auditory, there should be no effect of any of the visual manipulations if auditory prediction alone were adopted to perform the task. However, if the visual information were obligatorily incorporated into the auditory percept, i.e., if integration took place, then the AV streams should become temporally bound as a whole in perception. Consequently, one might become less sensitive to a slight deviation in one stream, resembling the reserved version of "temporal ventriloquism" (Fendrich and Corballis, 2001; Morein-Zamir et al., 2003). That is, the stronger the integration, the more the visual stream would temporally "capture" the auditory deviant, making it less salient than otherwise. As such, factors contributing to AV integration—synchrony, congruency, or both—should lead to *decreased* detection of the auditory deviant. Of interest, then, was whether synchrony and congruency operate independently, or whether they interact with each other in this process.

METHODS

PARTICIPANTS

Fourteen paid volunteers (five male, mean age 27 years, SD = 6) participated in this experiment. All reported normal or corrected-to-normal vision and normal hearing. Participants were not pre-screened for musical training and varied in the length of training. The training duration ranged from 0–20 years (all amateur musicians), with a mean duration of 8 years (SD = 6). Amongst the amateur musicians (13), the learned instruments included piano or keyboard (10), percussion (2), and guitar (1). This study had been approved by the ethic commission of Technical University of Munich, and was conducted in accordance with the ethical standards of the 1964 Declaration of Helsinki. All participants gave written informed consent prior to the experiment.

STIMULI AND MATERIALS

Visual Stimuli. The visual stimuli consisted of a humanlike PLF performing a repetitive whole-body bouncing movement (i.e., repetitive knee flexion and extension), without the feet leaving the ground. The PLF was initially constructed by recording a practiced actor performing this movement continuously using a 3D motion capture system (Qualisys Oqus, 8 cameras), with a sampling rate of 200 Hz. 13 markers in total were attached to the major joints (Johansson, 1973). The recorded motion data were converted into a 2D (without depth information) point-light display in Matlab ®R2012b (Mathworks) using Psychophysics Toolbox extensions version 3 (Brainard, 1997), and the animation was down-sampled to 100 Hz to match the monitor's frame frequency. The PLF was represented by 13 white discs against a black background, each of which subtended 0.4° of visual angle (°). In order to convey the human figure unambiguously, white lines were added to connect the discs[1]. The whole PLF subtended approximately 5° and 12° when viewed at 80 cm, and was centered in the middle of the screen (See also Figure 1 in Su (2014c)).

Each movement cycle consisted of a downward (knee flexion) and an upward (knee extension) phase. The former corresponded to 345 ms and the latter 255 ms on average across all the moving discs, as shown in the recorded motion data. The PLF movement was presented at a tempo corresponding to an inter-bounce interval of 600 ms, i.e., the temporal interval between the lowest positions (the "bounce") of two consecutive cycles was 600 ms. Very similar visual stimuli were employed in three recent studies (Su, 2014a,b,c), in which steps of motion data processing and relevant parameters were described in detail. In Su (2014b), detailed information regarding the motion profile of the PLF movement can also be found. Here, as in Su (2014a,c), the PLF movement was presented as iterations of a single cycle. Slight temporal and spatial interpolations had been applied to the motion data to ensure that there was no temporal or spatial discrepancy when the movement was displayed cyclically.

Auditory stimuli. The auditory stimuli consisted of repetitive cycles of alternating "downbeat" and "upbeat" tones as employed in Su (2014b). The sounds were generated as wave files by the music software Logic 8 Express (Apple Inc. California). The downbeat tones had a synthesized sound of the instrument "bongo" with 50 ms tone duration, and the upbeat tones had a synthesized sound of the instrument "high hat" with 47 ms tone duration. The inter-downbeat interval was 600 ms, corresponding to a cycle of the PLF movement. To match the auditory temporal structure to the uneven movement phases of the PLF, the interval between a downbeat and its following upbeat was 255 ms/345 ms for stimuli in the AV congruent/incongruent conditions (see Section Procedure and Design). The downbeat tones had a lower timbre, and the upbeat tones were attenuated by 10 dB relative to the downbeat tones. As such, regular accents in the auditory sequence were unambiguously perceived at the downbeat positions (Su, 2014b).

PROCEDURE AND DESIGN

The experimental program was controlled by a customized Matlab script using Psychophysics Toolbox version 3 routines running on a Mac OSX environment. The visual stimuli were displayed on a 17-inch CRT monitor (Fujitsu X178 P117A) with a frame frequency of 100 Hz at a spatial resolution of 1024 × 768 pixels. Participants sat with a viewing distance of 80 cm. Sounds were presented at a sampling rate of 44,100 Hz through closed studio headphones (AKG K271 MKII).

[1]As noted in Su (2014a), this constituted a departure from the original nature of a point-light display, where the figure motion is perceived from unconnected moving discs (Blake and Shiffrar, 2007).

Each trial started with a fixation cross in the center of the screen for 1000 ms, followed by a presentation of five cycles of concurrent visual and auditory sequences. The visual sequence was a periodically bouncing PLF, and the auditory sequence consisted of repetitive downbeats and upbeats in alternation. The visual and auditory sequences were presented in four combinations that varied in terms of content congruency and temporal synchrony between the two streams. Content congruency was based on the correspondence between the movement phase and the auditory beat: In half of all the trials, the PLF was bouncing *downwards* to the downbeat (congruent); in the other half, the PLF was bouncing *upwards* to the downbeat (incongruent). In terms of synchrony, in half of all the trials the PLF moved synchronously to the auditory beat; namely, the lowest/highest position of the movement coincided with the auditory downbeat in the congruent/incongruent condition (as in (Su (2014a), Exp 2)). In the other half of the trials, the visual stream was phase shifted with a delay of 150 ms relative to the auditory stream. This lag was chosen on the basis of it slightly exceeding the temporal integration window for the same auditory and visual streams as measured in a previous study (Su, 2014b).

In this task, participants were instructed to attend to both the auditory and the visual sequences while focusing more on the auditory one, as it was task relevant. In half of all the trials, a temporal perturbation could occur in the auditory sequence, such that one of the five auditory downbeats could be delayed or advanced (with equal probability) by 6% of the inter-downbeat interval (i.e., 36 ms). The perturbation could occur either on the second, the third, or the fourth downbeat, with equal probability. Participants were required to respond in each trial whether or not there was any temporal irregularity in the auditory sequence (**Figure 1**). They were informed that the deviant could only occur in one of the downbeat tones (the "heavier" tones), and never in the upbeat tones. Participants gave their response by pressing one of the two predefined keys. Participants were also informed that the PLF could be moving either downwards or upwards to the downbeats on different trials, and that this was irrelevant to the requested task. To ensure visual attention, in each trial following the response of auditory deviation detection, participants were also asked to recall whether the auditory and visual streams were synchronous or not by pressing one of the two predefined keys (different keys from those for the detection task). They were instructed to base their SJ solely on the subjective impression; it was also stressed that auditory deviation detection was the more important task that should be prioritized, whereas SJ was secondary. This instruction was imposed to avoid compromising the performance of the detection task, which was the primary task of interest.

Each participant underwent five practice trials before starting the experiment. The experiment followed a 2 (AV congruency) × 2 (AV synchrony) × 2 (auditory perturbation) within-participant design, each with 36 repetitions. The total trials were assigned to three experimental blocks of 96 trials each. All the experimental conditions, including the position of auditory perturbation and the nature of perturbation, were balanced across blocks. Within each block the conditions were presented in a randomized order. The entire experiment lasted around 1 h, completed in a single session. A break was required after each block of around 15 min.

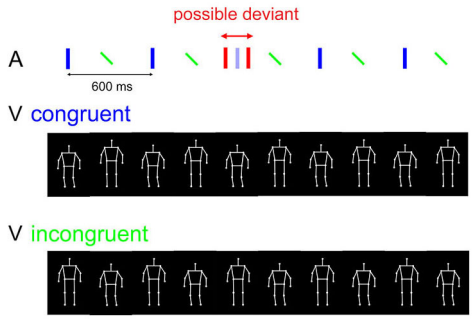


FIGURE 1 | Stimuli and trial procedure of the experiment. In each trial, an auditory beat sequence (upper panel) consisted of five alternating downbeats (blue vertical bars) and upbeats (tilted green bars), where accents were perceived in the former. The beat sequence was combined with either a congruently moving PLF (middle panel; moving *down* to the downbeat) or an incongruently moving PLF (lowest panel; moving *up* to the downbeat). Red bars depict a potential auditory deviant, which could be a delayed or advanced downbeat relative to its original temporal position (shown in opaque blue). The manipulation of synchrony between the auditory and visual streams is not illustrated here.

PILOT EXPERIMENT

It should be noted that the asynchronous AV condition in the present task was implemented by delaying, but *not* advancing, the visual stream relative to the auditory one. This manipulation was based on the result of a pilot experiment, which examined whether the relation between the visual movement phase and the auditory beat was consistently perceived across all the AV combinations. In the pilot experiment, AV synchronous, visual leading (by 150 ms), and visual lagging (by 150 ms) conditions were combined with AV congruent and incongruent presentations as described above, with ten trial repetitions per condition presented in a random order. Ten observers responded in each trial whether they perceived the PLF as moving downwards or moving upwards to the auditory downbeat. It was found that perception of movement phase relative to the downbeat was largely consistent when the auditory and visual streams were synchronous: On average 96% and 99% of the response indicated "downwards" and "upwards" for the congruent and incongruent conditions, respectively. The response was also consistent when the visual stream lagged the auditory one, with 94% and 94% of the response on average indicating "downwards" and "upwards" for the congruent and incongruent conditions. By contrast, when the visual stream led the auditory one, it became less clear to the participants whether the PLF was moving downwards or upwards to the downbeat (on average 51% and 62% of the response for the congruent and incongruent conditions). As the present study intended to manipulate the perceived content congruency with regard to how the PLF moved to the beat, only conditions that yielded consistent perception of such were selected for the main experiment, i.e., synchronous auditory and visual streams, and asynchronous streams in which the visual stream lagged the auditory one.

RESULTS

PERCENTAGE OF DEVIANT DETECTION (HIT RATE)

Of primary interest was the effect of congruency and synchrony on deviant detection. However, as the present task employed streams of continuous rhythmic stimuli, analyses including the auditory perturbation position as an additional factor may reveal effects related to the predictive nature of the stimuli, as well as its possible interplay with the two main factors. For this purpose, the percentage of correctly detecting an auditory deviant (i.e., the hit rate) for each experimental condition was calculated individually as a first index of the task performance. Individual hit rates were submitted to a 2 (AV congruency) × 2 (AV synchrony) × 3 (auditory perturbation position) within-subject ANOVA. A main effect of synchrony was found, $F_{(1,13)} = 17.17$, $p < 0.002$, $\eta^2 = 0.57$, showing a greater hit rate when the AV streams were *a*synchronous than when they were synchronous. A main effect of position was also found, $F_{(2,26)} = 26.53$, $p < 0.001$, $\eta^2 = 0.67$. *Post-hoc* tests (Tukey HSD) revealed better detection when the perturbation occurred in the third or fourth beat than in the second beat of the auditory sequence, both *p*s < 0.001. The three-way interaction was not significant, $p > 0.7$. The two-way interaction between congruency and synchrony was significant, $F_{(1,13)} = 6.12$, $p < 0.03$, $\eta^2 = 0.32$. As perturbation position did not yield an interaction with any of the two other factors, within-subject means were computed across all positions and submitted to the follow-up one-way ANOVAs conducted for each congruency condition separately. Hit rate was found higher for asynchronous than for synchronous AV streams when they were congruent (i.e., PLF bounced downward to the beat), $F_{(1,13)} = 17.0$, $p < 0.002$, $\eta^2 = 0.57$. By contrast, no effect of synchrony was observed when the AV streams were incongruent (i.e., PLF bounced upward to the beat), $p > 0.2$ (**Figure 2**). Thus, the effect of synchrony on hit rate—i.e., more hits for asynchronous than for synchronous streams—appeared mostly driven by the AV congruent condition.

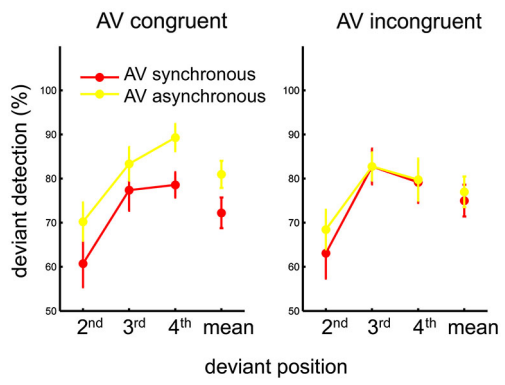


FIGURE 2 | Mean percentage of auditory deviant detection as a function of the deviant position, for each AV congruency and AV synchrony condition. The mean across deviant positions for each condition is also plotted in the respective graph. Error bars represent standard error of the means.

SENSITIVITY (D')

To assess perceptual sensitivity to the auditory deviants, *d'* was calculated following signal detection theory analysis ("SDT", Stanislaw and Todorov, 1999) individually for each of the four experimental conditions based on congruency and synchrony. *d'* was calculated as the z-score transformed hit rate minus the z-score transformed false alarm rate. The within-subject *d'*s were submitted to a 2 (AV congruency) × 2 (AV synchrony) repeated-measures ANOVA. A main effect of congruency was found, $F_{(1,13)} = 5.17$, $p < 0.05$, $\eta^2 = 0.28$, showing a greater *d'* in the incongruent than in the congruent condition. The effect of synchrony was marginally significant, $p = 0.07$, with a trend of greater *d'* in asynchronous than in synchronous conditions. The interaction between congruency and synchrony was not significant, $p > 0.6$ (**Figure 3A**). In short, participants were less sensitive to a deviant auditory beat when the observed PLF moved downwards than when it moved upwards to the beat. To some extent, sensitivity to an auditory deviant also seemed lower when the auditory and visual streams were synchronous than when they were asynchronous.

RESPONSE CRITERION (C)

To examine whether synchrony and congruency also affected processes in the decisional level, the response criterion (*c*) as defined by SDT (averaging the z-score transformed hit rate and the z-score transformed false alarm rate, then multiplied by minus one) was calculated individually for each experimental condition, and submitted to the 2 (AV congruency) × 2 (AV synchrony) within-subject ANOVA. A significant main effect of synchrony was found, $F_{(1,13)} = 8.21$, $p < 0.02$, $\eta^2 = 0.39$, showing that participants were more liberal with their response in the asynchronous than in the synchronous condition. The interaction between congruency and synchrony was also significant, $F_{(1,13)} = 5.59$, $p < 0.04$, $\eta^2 = 0.30$. Follow-up one-way ANOVAs revealed that the difference in response criterion between synchronous and asynchronous conditions was only evident when the AV streams were congruent, $F_{(1,13)} = 9.19$, $p < 0.01$, $\eta^2 = 0.41$, whereas no such difference was found for incongruent AV streams, $p > 0.6$ (**Figure 3B**). On average, the response criterion as indexed by

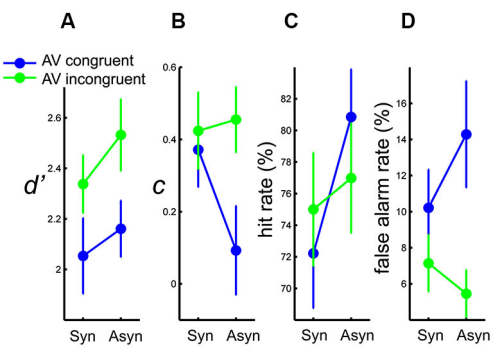


FIGURE 3 | Results of (A) mean *d'*, (B) mean *c*, (C) mean hit rate, and (D) mean false alarm rate, as a function of AV synchrony, for each AV congruency condition. Error bars represent standard error of the means.

c was positive in all the experimental conditions, showing that participants in this task tended overall to be more conservative than neutral. Participants were more liberal in the asynchronous than in the synchronous condition, but only when observing the PLF moving downwards to the auditory beat.

FALSE ALARM RATE

Following the main effect and interaction found in the response criterion, false alarm rates were analyzed to reveal how synchrony and congruency affected the error behavior. (See Section Percentage of Deviant Detection (Hit Rate) for results of hit rate analysis. Results of hit rates were re-plotted here as **Figure 3C** for better visualization.) Individual false alarm rates were submitted to a 2 (AV congruency) × 2 (AV synchrony) within-subject ANOVA. Only a main effect of congruency was found, $F_{(1,13)} = 6.09$, $p < 0.05$, $\eta^2 = 0.32$, showing a higher false alarm rate in the congruent than in the incongruent condition. The interaction between congruency and synchrony was marginally significant, $p = 0.08$ (**Figure 3D**). As shown, there were generally more false alarms when the PLF moved congruently to the auditory beat. From the marginally significant interaction and the trend of the mean data, it would seem as if participants tended to make more false alarms in the asynchronous than in the synchronous condition for congruent streams.

PERCEIVED SYNCHRONY

To explore whether response in the secondary task (SJ) differed across congruency conditions, individual percentages of responding "synchronous" for each of the experimental conditions were also submitted to a 2 (AV congruency) × 2 (AV synchrony) within-subject ANOVA. A main effect of synchrony was found, $F_{(1,13)} = 12.97$, $p < 0.01$, $\eta^2 = 0.50$, with on average 81% and 70% of the response being "synchronous" for the experimental synchronous and asynchronous condition, respectively. The interaction between the two factors was close to significant, $F_{(1,13)} = 4.61$, $p = 0.051$, $\eta^2 = 0.26$. A trend was observed of a greater difference in perceived synchrony in the congruent condition (on average 80% and 63% of the response was "synchronous" for synchronous and asynchronous stimuli, respectively) than in the incongruent condition (on average 81% and 77%).

RELATION BETWEEN EACH INDEX AND THE PERCEIVED SYNCHRONY

Although there were only two objective levels of implemented synchrony (i.e., synchronous or asynchronous), the degree of subjectively perceived synchrony across these two levels may differ amongst individuals (c.f. Su, 2014b). Thus, it was of interest whether and how each dependent variable was related to the extent of perceived synchrony, and whether this relation was varied by AV congruency. To this end, correlational analyses (Pearson's correlation) were carried out on an individual level ($N = 14$), for the AV congruent and AV incongruent conditions separately, between the following two measures: (1) the difference in the percentage of synchrony response (i.e., the response being "synchronous") between AV synchronous and AV asynchronous conditions; and (2) the difference in each of the indexes reported thus far (i.e., *d'*, *c*, hit rate, and false alarm rate) between AV synchronous and AV asynchronous conditions.

Results revealed significant correlations only in *c* and in false alarm rate, but not in *d'* or hit rate (**Figure 4**). Regarding *c*, a positive correlation was found in the AV congruent condition, $r = 0.61$, $p = 0.02$, showing that a greater shift to conservative response was associated with a greater increase in perceived synchrony. In the AV incongruent condition, by contrast, the correlation was negative, $r = -0.53$, $p = 0.05$, showing that a greater shift to liberal response was associated with a greater increase in perceived synchrony (**Figure 4**, 2nd column). As for the false alarm rate, which accounted for the correlations found in *c*, a negative correlation was found in the AV congruent condition, $r = -0.61$, $p = 0.02$, showing that a greater *reduction* in false alarms was associated with a greater increase in perceived synchrony. In the AV incongruent condition, a positive correlation was found, $r = 0.58$, $p = 0.03$, showing that a greater *increase* in false alarms was associated with a greater increase in perceived synchrony (**Figure 4**, 4th column)[2]. In sum, the difference in response criterion and that in false alarm rate were each correlated with the difference in subjectively perceived synchrony of the AV streams. This correlation, critically, exhibited opposite patterns between congruent and incongruent AV conditions.

DISCUSSION

The present study investigated how content congruency and temporal synchrony between concurrent rhythmic auditory and visual streams influenced AV integration, as indicated by auditory deviant perception. Participants attended to AV stimuli consisting of a PLF moving regularly to a sequence of auditory beat, and detected a possible auditory temporal deviant. The PLF could move congruently (downwards) or incongruently (upwards) to the beat, while in both situations the movement could be either synchronous with the beat, or lagging behind it. The main results show that, as evidenced by *d'* (**Figure 3A**), participants were better at detecting an auditory deviant when the PLF moved *in*congruently than congruently to the beat, suggesting stronger integration—or greater visual temporal capture of the auditory beat—in the latter. Similarly, a trend can be noted of stronger visual capture (i.e., lower *d'* for auditory deviant detection) for synchronous than for asynchronous AV streams. Thus, both content congruency and (to some extent) temporal synchrony appeared to promote AV integration in the present scenario.

Specific to the congruent AV stimuli, more hits as well as more false alarms were observed with *a*synchronous than with synchronous AV streams. This synchrony-dependent difference was not seen when the PLF moved incongruently to the beat (**Figures 3C,D**). Although the increased hit rate in the asynchronous and congruent condition could have been associated with better deviant detection due to lower cross-modal

[2]Although the correlations found in the AV congruent conditions might have been driven by one extreme data point, there is no apparent reason (e.g., experimental errors) to consider this data point illegitimate for inclusion. In the presence of a potential outlier, correlation analyses were performed again on square root transformed data (Osborne, 2002). While the correlation found in the AV congruent condition was not significant any more for c ($r = 0.46$, $p = 0.1$), the negative correlation remained significant for the false alarm rate ($r = -0.56$, $p = 0.03$).

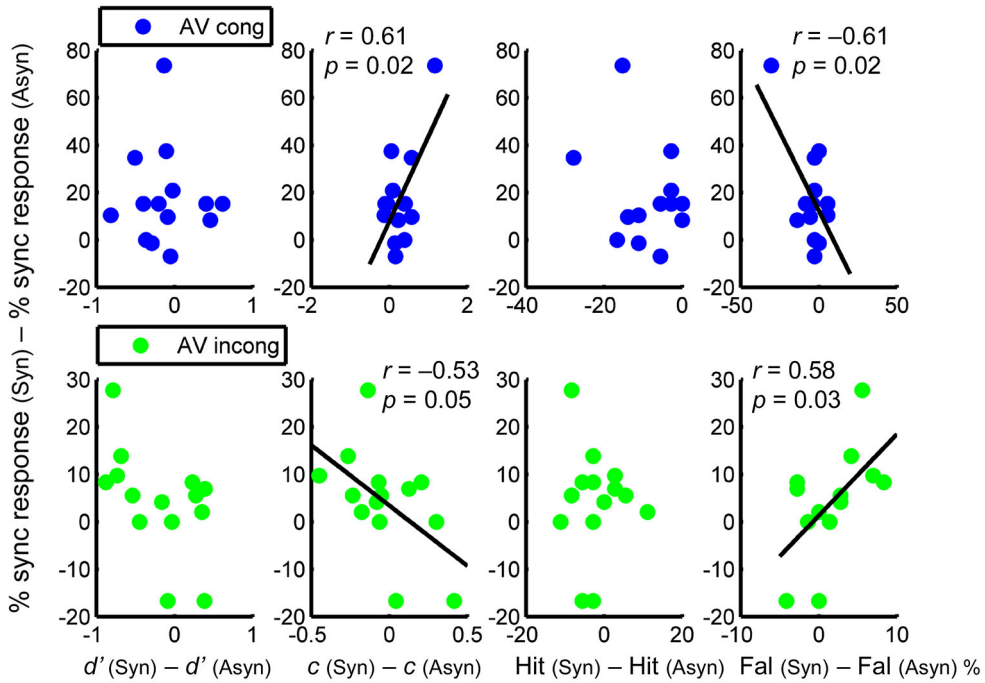


FIGURE 4 | Relationship between the difference in perceived synchrony and the difference in each parameter (calculated as the difference between AV synchronous and asynchronous conditions), for AV congruent (upper panel) and AV incongruent (lower panel) condition separately. Columns from left to right: *d'*, *c*, hit rate, and false alarm rate. Pairwise correlations are only significant for *c* and for false alarm rate.

binding, this explanation is challenged by the shift of response criterion (to be more liberal) in this particular condition (**Figure 3B**), as well as the lack of a corresponding interaction between congruency and synchrony in *d'*. As such, and given the corroborating pattern in false alarms (albeit only marginally significant), this result may rather be explained by the error behavior: namely, more false positives and an increased tendency to report a deviant in asynchronous than in synchronous streams for congruent stimuli. Moreover, congruency also modulated how individual errors were associated with subjectively perceived synchrony: When the PLF moved congruently/incongruently to the beat, a greater increase in perceived synchrony was associated with a greater reduction/increase in false alarms (**Figure 4**). Thus, in the present task, errors of false alarm were modulated by an interaction between congruency and synchrony. Possible mechanisms underlying these errors will be discussed in Section Content Congruency Modulates Synchrony Effect.

CROSS-MODAL ATTENTION IS ASSOCIATED WITH INTEGRATION

Owing to the rhythmic nature of both sensory stimuli, the present paradigm afforded the possibility of auditory temporal prediction for the auditory task, which would have rendered the result largely independent of the visual conditions. However, effects of visual manipulation were evident, suggesting that concurrent visual movement information was readily integrated with the auditory rhythm in perception (Su, 2014c). This supports the idea that when multisensory information is available and associated with each other (Lee and Noppeney, 2014; van Atteveldt et al., 2014), cross-modal rather than within-modal attention dominates temporal prediction in each stream, even if the latter alone would have sufficed for the task. Cross-modal prediction has often been shown to underlie perception of AV stimuli that are causally bound in an action, such as AV speech (Zion Golumbic et al., 2013) or AV drumming movements (Petrini et al., 2009b; see also Vroomen and Stekelenburg, 2010, for similar results of artificial visual motion paired with an impact-like sound). Importantly, here it shows that this prediction mode also applies to AV stimuli that are related to each other by means of action-perception coupling (Prinz, 1997), such as observed movements coordinated with external sounds (as in the example of observing dancers moving to music). In this case, the visual movement information is associated with the auditory stream due to the observer's understanding, or internal representation, of how humans move to rhythmic sounds. For such bimodal rhythmic stimuli, the (possibly obligatory) visual prediction of auditory stream may facilitate coupling between cortical oscillations entrained to each stream, which in turn supports AV integration (Senkowski et al., 2008; Schroeder and Lakatos, 2009).

Depending on the temporal relation and the content match between modalities, effects on AV integration were reflected in how strongly the visual stream attracted an auditory deviant temporally, making it less distinct in some conditions than in others. Such temporal binding of cross-modal stimuli, typically known as "temporal ventriloquism", has mainly been reported as

auditory event(s) shifting the perceived visual onset(s), and not the other way around (Fendrich and Corballis, 2001; Morein-Zamir et al., 2003; Recanzone, 2003). The same modality asymmetry in temporal capture has also been shown in a rhythmic context: Finger taps synchronized to an isochronous visual flashes are considerably attracted to a concurrent but phase-shifted auditory sequence, whereas taps synchronized to tones were rather uninfluenced by concurrent visual distractors (Aschersleben and Bertelson, 2003; Repp and Penel, 2004). The direction of this capture is often taken as evidence of superior temporal processing in the auditory compared to the visual modality (Welch and Warren, 1980). However, recent studies demonstrate that visual rhythm perception and synchronization is much improved when the visual stimulus consists of spatiotemporal periodicity, such as communicated by a moving object (Grahn, 2012a; Hove et al., 2013a,b). Furthermore, previous works have revealed that the same periodic PLF movement as a visual stimulus can modulate auditory rhythm perception (Su, 2014a) as well as improve auditory synchronization (Su, 2014c), and the behavioral gain in the latter study is suggestive of multisensory integration. In this light, the present study presents a new case of visual capture of auditory event in the temporal domain, using visual stimuli derived from biological motion. The integration effect likely originates from perceptual binding of AV information, which occurs when observing a rhythmic human movement while listening to an auditory rhythm (Su, 2014c). Specifically, auditory rhythm perception entails internal motor representation of the rhythm in the listener (Repp and Su, 2013; see also Grahn (2012b), for a review of cortical and sub-cortical motor areas involved in this process). Likewise, observing a human movement elicits internal motor representation (or simulation) of the action in the observer (Jeannerod, 2001). An association between auditory rhythm and rhythmic visual movement that leads to AV binding is proposed to be based on such internal sensorimotor coupling (see Su, 2014c for more relevant discussions).

CONTENT CONGRUENCY MODULATES AUDIOVISUAL INTEGRATION

The main findings of the present study are twofold: Multisensory integration was modulated by AV content congruency, as well as by an interaction between AV congruency and AV synchrony. Congruency affected auditory deviant detection, whereas the interaction between congruency and synchrony modulated false alarms and response criterion. Given the effects on these parameters, AV congruency and synchrony appear to modulate integration in both the perceptual and the decisional processes (Meyer and Wuerger, 2001; Wuerger et al., 2003; Sanabria et al., 2007).

Perceptual effects as indexed by *d'* are most consistently associated with congruency, i.e., lower sensitivity to a deviant (indicating greater AV integration) for congruent than for incongruent stimuli. This result is straightforward, and it confirms that cognitive factors such as perceived content match promote integration, as previously shown in AV speech or drumming actions using a SJ or TOJ task (Petrini et al., 2009a; van Wassenhove et al., 2007; Vatakis and Spence, 2007). Notably, congruency can be derived from various forms of AV correspondence (Parise and Spence, 2009; Spence, 2011), and stimuli of abstract correspondences are shown to be processed cortically in a manner similar to multisensory integration (Bien et al., 2012). In this light, the present result reveals a new congruency effect based on whether an observed movement matches an individual's own motor repertoire coordinated with an auditory beat, i.e., whether it matches how one would naturally move to a beat (see also Su (2014b)). An observed downward movement appears to be favored for integration with an auditory downbeat, compared to an upward movement.

To some extent, synchronous AV streams seem to be associated with stronger visual capture (as indicated by poorer detection) of an auditor temporal deviant, compared to asynchronous ones. This trend is consistent with a large body of literature on intersensory binding (Chen and Vroomen, 2013), showing that temporal alignment between the two streams may direct attention in a converging manner to facilitate integration. This pattern is also consistent with the role of cross-modal prediction in multisensory integration (Zion Golumbic et al., 2013; Lee and Noppeney, 2014; van Atteveldt et al., 2014), as visual information in the present asynchronous condition (i.e., visual stream lagging the auditory one) is of little predictive value for the auditory system, thus leading to less integration than in the synchronous condition. However, while some studies show that AV synchrony is critical for auditory enhancement of visual biological motion detection (Saygin et al., 2008; Arrighi et al., 2009), others fail to find support for its importance (Thomas and Shiffrar, 2013). The currently mixed findings may be associated with differences in visual stimuli (e.g., a whole-body figure or only part of it) or the required task (e.g., detection of a walker, other temporal aspects of the movement, or of auditory patterns as presently probed). It may also be that, in studies where temporal synchrony does not modulate multisensory perception, the measured effect reflects inter-sensory priming (Noppeney et al., 2008; Chen and Spence, 2010) rather than integration, which can occur without strict temporal co-occurrence.

Regarding possible neural correlates, in the present task, the observed higher-level, cognitive influence on sensory (here, auditory) processes seems in line with neural findings of higher multisensory regions feedback-modulating lower sensory areas in the course of AV integration (Driver and Noesselt, 2008; Musacchia and Schroeder, 2009). Specifically, the effect of AV congruency is consistent with evidence that neuronal processing in cortical unisensory areas is enhanced by congruent multisensory stimuli, but much less so by incongruent ones (Kayser et al., 2010). Such top-down modulations may be achieved through cortical oscillations between higher-level and lower-level areas (Senkowski et al., 2008; Klemen and Chambers, 2012). In agreement with that, cortical oscillations underlying multisensory integration is also modulated by congruency between dynamic AV stimuli (Gleiss and Kayser, 2014). Finally, semantically congruent and incongruent AV stimuli are often found to engage different cortical multisensory areas, i.e., temporal and inferior frontal regions, respectively (Hein et al., 2007; Doehrmann and Naumer, 2008; van Atteveldt et al., 2010). This pattern is proposed to reflect well-learned associations, or multisensory objects, represented in the temporal regions (e.g., superior temporal sulcus), and conflict monitoring in the inferior frontal areas. It remains to be tested whether the presently proposed top-down influence on

lower sensory processes may originate from different multisensory regions depending on content congruency.

CONTENT CONGRUENCY MODULATES SYNCHRONY EFFECT

Effects pertaining to post-perceptual, decisional processes were reflected in false alarms and the shift of response criterion. First, there were more false alarms with congruent than with incongruent stimuli. Although it is not entirely clear why, one speculation is that the perceived auditory timing might have been shifted by a mixture of position and velocity cues in the continuous PLF movement trajectory (Su, 2014b). This shift was likely not stable or constant for all the auditory events, causing occasional fluctuation in the perceived auditory onsets and thus erroneous judgment of a deviant. The observed effect was greater for congruent than for incongruent stimuli, arguing for stronger AV binding in the former. Next, in particular, a curious pattern of increased false alarms and more liberal response was seen in *a*synchronous streams, and the effect was mainly evident when the observed movement was congruent with the auditory rhythm. One plausible explanation for this pattern, paradoxically, also rests upon visual temporal capture of auditory beats: In an asynchronous AV situation, the auditory events might be temporally shifted by the visual stream due to AV binding (i.e., temporal ventriloquism). If, as suspected, this shift is occasional and not constant throughout the auditory sequence, the perceived irregularity could be erroneously taken as a deviant, leading to a false positive response. Notably, this effect is specific to the congruent AV stimuli, suggesting that content congruency can promote (potentially erroneous) integration of AV information at greater temporal distance. As such, the effect of temporal proximity as a low-level stimulus factor on integration seems to be modulated by higher-level cognitive factors, such as the perceived content match. One question, then, is whether this result pattern might be associated with the observation that the difference in perceived synchrony between synchronous and asynchronous conditions (as measured in the secondary task) seems greater in congruent than in incongruent stimuli. Put in another word, is subjective AV asynchrony directly linked to the auditory susceptibility to visual temporal capture? There seems to be evidence against this speculation (Stevenson et al., 2012): A narrower AV temporal integration window (i.e., lower tendency to perceive asynchronous stimuli as synchronous) is correlated with a *lower* tendency to integrate asynchronous stimuli, and thus—in the present case—it should have led to fewer, and not more, false alarms.

A similar interaction between stimulus timing and content congruency has been described in a recent study of AV speech (syllable) perception (Ten Oever et al., 2013), in which semantically congruent AV stimuli compared to incongruent ones are integrated at greater temporal disparity. This leads to the proposal that, as opposed to lower-level stimulus features (e.g., timing) and higher-level cognitive factors (e.g., semantic congruency) operating serially and hierarchically, these two factors may in fact work in parallel to reach a perceptual outcome (Stevenson et al., 2014b). In line with this proposal, the present results extend the principle to a non-speech action domain involving continuous AV stimuli, whose congruency is derived from internal motor simulation (Jeannerod, 2001). It may be argued that such top-down cognitive mechanisms, based on sensorimotor coupling, operate in parallel with bottom-up, synchrony-driven attention (Van der Burg et al., 2008; Fiebelkorn et al., 2010) in the course of multisensory integration of rhythmic stimuli.

Also regarding the interaction between the two factors, under congruent conditions, a greater increase in perceived synchrony was associated with a greater *decrease* in false alarms across individuals. Under incongruent conditions, however, a greater increase in perceived synchrony was associated with a greater *increase* in false alarms. These patterns may be explained in terms of individual differences in AV synchrony perception predisposing the strength of AV binding (Stevenson et al., 2012), and this tendency leads to different consequences of error, depending on content congruency. With congruent content, the more an individual is able to discern synchronous from asynchronous streams, the more the streams may be unambiguously integrated in the former and less in the latter, thus reducing the chance of visual capture of asynchronous auditory stimuli and the subsequent false alarms. By contrast, incongruent content may increase uncertainty in synchronous situations, possibly due to conflicting information regarding the unity of stimuli (Welch and Warren, 1980), i.e., the incompatible movement relative to the beat deters the perceptual system from integration, whereas synchrony between the streams promotes it. As a result, individuals who can better tell apart synchronous from asynchronous situations are subject to greater perceptual conflict, leading to more errors.

The interaction between content congruency and temporal synchrony seems to occur later in the decisional stage (as reflected in the response criterion) compared to its perceptual effect (as reflected in sensitivity). From the literature, congruency seems to modulate the time course of multisensory processing, with a larger early response to congruent (compared to incongruent) stimuli (Naci et al., 2012), followed by a later response to incongruent (compared to congruent) ones (Meyer et al., 2013). Although it remains speculative at present, it is possible that an earlier feedback modulation through congruent AV stimuli (Naci et al., 2012) would contribute to temporal capture or integration in the auditory cortices (Musacchia and Schroeder, 2009; Marchant and Driver, 2013), whereas feedback from incongruent stimuli may occur later and, rather than interacting with stimulus timing for integration, it would be more involved in conflict resolution.

Finally, it is worth mentioning that in the pilot experiment, an asymmetry was evident regarding how AV temporal order influenced the perceived movement direction relative to the beat: Judgment of direction (implying congruency) was more ambiguous when the visual stream led—compared to when it lagged—the auditory one. Together with the observation in the main experiment that congruency appeared to influence perceived synchrony, it is possible that AV temporal relation and content congruency in the present scenario interact with each other both-ways in perception. Although a detailed discussion on this point is beyond the scope of the present research, future investigations using different paradigms are warranted to gather further evidence of this interaction, and its implication in multisensory perception.

MULTISENSORY INTEGRATION VS. ATTENTIONAL ENTRAINMENT IN RHYTHMIC STIMULI

In the domain of multisensory integration, attention is often considered to be a mechanism that can facilitate cross-modal binding (Fiebelkorn et al., 2010; Koelewijn et al., 2010; Talsma et al., 2010). There is, however, a different framework pertaining to the role of attention, namely that of the *Dynamic Attending Theory* ("DAT", Jones and Boltz, 1989; Large and Jones, 1999) as briefly mentioned in the Introduction, which is also relevant in the context of bimodal rhythmic stimuli. Discussions are thus warranted as to possible overlaps and discrepancies between DAT and theories of integration when explaining multisensory perception. DAT proposes that attention can be seen as an oscillatory energy, and it is temporally entrained by the periodicity of the external sensory rhythms, leading to enhanced stimulus processing at the expected points in time. Findings in support of this theory typically show that a deviant is better detected when its expected occurrence coincides with the entrained periodicity (Jones et al., 2002, 2006; Repp, 2010; Su, 2014a). This model is further corroborated by possible neural correlates, such as cortical oscillations in the beta band being phase-locked to a regular auditory beat (Large and Snyder, 2009; Iversen et al., 2009; Fujioka et al., 2012). Within this framework, synchronous multisensory rhythms compared to asynchronous ones are expected to facilitate such processing by entraining attention to convergent points in time (Nozaradan et al., 2012; Su, 2014c). As such, in the present case, DAT would predict that synchronous AV streams should yield better auditory deviation detection than asynchronous ones, while content congruency should not play a critical role. These predictions run contrary to those made with regard to AV integration and intersensory capture. At first sight, the present results seem to support the latter.

Can these two accounts—thus far situated in somewhat different research domains and yet both tapping onto the operation of attention—be reconciled in addressing multisensory perception of rhythmic stimuli? Inspection of the present data suggests that these two accounts may be combined to explain the results. First, deviants were better detected in later temporal positions, which appears to reflect the effect of attentional entrainment, as expectation can be more strongly and precisely generated with more repetitions of intervals preceding a possible deviant (Haenschel et al., 2005). This effect was independent of AV congruency and synchrony, i.e., the mechanism exists independently of the concurrent visual information, suggesting that it functions as a perceptual basis for rhythmic stimuli at least in the task-relevant modality. On top of that, auditory deviant detection varied according to AV congruency and to some extent synchrony, and the effect was consistent with predictions of AV integration rather than of bimodal entrainment alone. Based on these results, the present research proposes the following: In the context of multisensory rhythms, attention is temporally entrained by the (especially task-relevant) stimulus rhythmicity, likely in a bottom-up, automatic manner (Bolger et al., 2013). This temporal orienting serves a general perceptual frame for stimulus processing that is less sensitive to specificities of multisensory information. Indeed, literature on attentional entrainment consistently shows that enhanced attention can be flexibly transferred across modalities and tasks (Escoffier et al., 2010; Bolger et al., 2013; Brochard et al., 2013). However, owing to the heightened attention entrained by the stimulus rhythmicity, multisensory binding around these points in time is also enhanced, which is then subject to modulations of variables critical for integration, such as congruency and synchrony. Presently it would seem as if the same attentional capacity is deployed for temporal entrainment and multisensory integration in a hierarchical manner, with the former serving the basis for the latter.

There seems to be a link between the DAT model and multisensory integration: With respect to attentional entrainment, a body of neurophysiological research demonstrates that rhythmic cortical oscillations can be entrained (i.e., the neuronal excitatory phase being aligned) to the rhythmicity of external stimuli, such that neuronal responses to the sensory input are amplified (e.g., Lakatos et al., 2008; Schroeder and Lakatos, 2009). This operation is especially instantiated by deploying attention to the task-relevant stimulus stream amongst other phase-shifted streams in a different modality (Lakatos et al., 2008, 2013). Most critical in the context of multisensory stimuli are proposals that oscillations in one lower sensory area can be phase-reset, in a predictive manner, by concurrent input from another modality (Lakatos et al., 2007; Schroeder et al., 2008), a mechanism that is argued to underlie multisensory integration (van Atteveldt et al., 2014). These findings seem to support the hypothesis proposed above. Namely, the task-relevant rhythmic stream entrains internal processes (i.e., oscillations) in the temporal domain through attentional deployment, while input from another modality—dependent upon multisensory correspondence—modulates the processes on top of this entrainment, thus enhancing or impeding integration. From here on, other relevant hypotheses can be tested, e.g., in rhythmic stimuli comprising several hierarchical levels of periodicity, whether the strength of integration would vary according to the saliency (and thus potential for entrainment) of each periodicity.

In conclusion, the present study highlights the effect of the cognitive factor (content congruency), as well as its interaction with the stimulus factor (temporal synchrony), on integration of continuous, rhythmic AV information related to human movements and extraneous sounds. A new form of congruency is demonstrated here, based on whether the observed movement matches how humans typically move to an auditory beat (i.e., action-perception coupling). This content congruency influences integration, as well as whether attention may be spread despite inter-sensory asynchrony to support integration. Consistent with previous findings in AV speech, perception of complex AV actions may also entail parallel processing of lower-level stimulus parameters and higher-level content correspondence. As a multitude of environmental and biological signals are multisensory and rhythmic (Arnal and Giraud, 2012), possible interplays amongst factors of integration and rhythm perception remain an interesting scenario for further explorations.

ACKNOWLEDGMENTS

This work and the author were supported by a grant from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), SU 782/1-1. The author thanks the reviewers for their constructive comments.

REFERENCES

Alvarado, J. C., Vaughan, J. W., Stanford, T. R., and Stein, B. E. (2007). Multisensory versus unisensory integration: contrasting modes in the superior colliculus. *J. Neurophysiol.* 97, 3193–3205. doi: 10.1152/jn.00018.2007

Arnal, L. H., and Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. *Trends Cogn. Sci.* 16, 390–398. doi: 10.1016/j.tics.2012.05.003

Arrighi, R., Alais, D., and Burr, D. (2006). Perceptual synchrony of audiovisual streams for natural and artificial motion sequences. *J. Vis.* 6, 260–268. doi: 10.1167/6.3.6

Arrighi, R., Marini, F., and Burr, D. (2009). Meaningful auditory information enhances perception of visual biological motion. *J. Vis.* 9, 25.1–25.7. doi: 10.1167/9.4.25

Aschersleben, G., and Bertelson, P. (2003). Temporal ventriloquism: crossmodal interaction on the time dimension. 2. Evidence from sensorimotor synchronization.. *Int. J. Psychophysiol.* 50, 157–163. doi: 10.1016/s0167-8760(03)s00131-4

Bien, N., ten Oever, S., Goebel, R., and Sack, A. T. (2012). The sound of size: crossmodal binding in pitch-size synesthesia: a combined TMS, EEG and psychophysics study. *Neuroimage* 59, 663–672. doi: 10.1016/j.neuroimage.2011.06.095

Blake, R., and Shiffrar, M. (2007). Perception of human motion. *Annu. Rev. Psychol.* 58, 47–73. doi: 10.1146/annurev.psych.57.102904.190152

Bolger, D., Coull, J. T., and Schön, D. (2014). Metrical rhythm implicitly orients attention in time as indexed by improved target detection and left inferior parietal activation. *J. Cogn. Neurosci.* 26, 593–605. doi: 10.1162/jocn_a_00511

Bolger, D., Trost, W., and Schön, D. (2013). Rhythm implicitly affects temporal orienting of attention across modalities. *Acta Psychol. (Amst)* 142, 238–244. doi: 10.1016/j.actpsy.2012.11.012

Brainard, D. H. (1997). The psychophysics toolbox. *Spat. Vis.* 10, 433–436. doi: 10.1163/156856897x00357

Brochard, R., Tassin, M., and Zagar, D. (2013). Got rhythm. . . for better and for worse. Cross-modal effects of auditory rhythm on visual word recognition. *Cognition* 127, 214–219. doi: 10.1016/j.cognition.2013.01.007

Brooks, A., van der Zwan, R., Billard, A., Petreska, B., Clarke, S., and Blanke, O. (2007). Auditory motion affects visual biological motion processing. *Neuropsychologia* 45, 523–530. doi: 10.1016/j.neuropsychologia.2005.12.012

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., and Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Comp. Biol.* 5:e1000436. doi: 10.1371/journal.pcbi.1000436

Chen, Y.-C., and Spence, C. (2010). When hearing the bark helps to identify the dog: semantically-congruent sounds modulate the identification of masked pictures. *Cognition* 114, 389–404. doi: 10.1016/j.cognition.2009.10.012

Chen, L., and Vroomen, J. (2013). Intersensory binding across space and time: a tutorial review. *Atten Percept. Psychophys.* 75, 790–811. doi: 10.3758/s13414-013-0475-4

Doehrmann, O., and Naumer, M. J. (2008). Semantics and the multisensory brain: how meaning modulates processes of audio-visual integration. *Brain Res.* 1242, 136–150. doi: 10.1016/j.brainres.2008.03.071

Driver, J., and Noesselt, T. (2008). Multisensory interplay reveals crossmodal influences on "sensory-specific" brain regions, neural responses and judgments. *Neuron* 57, 11–23. doi: 10.1016/j.neuron.2007.12.013

Escoffier, N., Sheng, D. Y. J., and Schirmer, A. (2010). Unattended musical beats enhance visual processing. *Acta Psychol. (Amst)* 135, 12–16. doi: 10.1016/j.actpsy.2010.04.005

Fendrich, R., and Corballis, P. (2001). The temporal cross-capture of audition and vision. *Percept. Psychophys.* 63, 719–725. doi: 10.3758/bf03194432

Fiebelkorn, I. C., Foxe, J. J., and Molholm, S. (2010). Dual mechanisms for the cross-sensory spread of attention: how much do learned associations matter? *Cereb. Cortex* 20, 109–120. doi: 10.1093/cercor/bhp083

Fujioka, T., Trainor, L. J., Large, E. W., and Ross, B. (2012). Internalized timing of isochronous sounds is represented in neuromagnetic Beta oscillations. *J. Neurosci.* 32, 1791–1802. doi: 10.1523/jneurosci.4107-11.2012

Ghazanfar, A. A. (2013). Multisensory vocal communication in primates and the evolution of rhythmic speech. *Behav. Ecol. Sociobiol.* 67, 1441–1448. doi: 10.1007/s00265-013-1491-z

Gleiss, S., and Kayser, C. (2014). Oscillatory mechanisms underlying the enhancement of visual motion perception by multisensory congruency. *Neuropsychologia* 53, 84–93. doi: 10.1016/j.neuropsychologia.2013.11.005

Grahn, J. A. (2012a). See what I hear? Beat perception in auditory and visual rhythms. *Exp. Brain Res.* 220, 51–61. doi: 10.1007/s00221-012-3114-8

Grahn, J. A. (2012b). Neural mechanisms of rhythm perception: current findings and future perspectives. *Top Cogn. Sci.* 4, 585–606. doi: 10.1111/j.1756-8765.2012.01213.x

Haenschel, C., Vernon, D. J., Dwivedi, P., Gruzelier, J. H., and Baldeweg, T. (2005). Event-related brain potential correlates of human auditory sensory memory-trace formation. *J. Neurosci.* 25, 10494–10501. doi: 10.1523/jneurosci.1227-05.2005

Hein, G., Doehrmann, O., Müller, N. G., Kaiser, J., Muckli, L., and Naumer, M. J. (2007). Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. *J. Neurosci.* 27, 7881–7887. doi: 10.1523/jneurosci.1740-07.2007

Hove, M. J., Fairhurst, M. T., Kotz, S. A., and Keller, P. E. (2013a). Synchronizing with auditory and visual rhythms: an fMRI assessment of modality differences and modality appropriateness. *Neuroimage* 67, 313–321. doi: 10.1016/j.neuroimage.2012.11.032

Hove, M. J., Iversen, J. R., Zhang, A., and Repp, B. H. (2013b). Synchronization with competing visual and auditory rhythms: bouncing ball meets metronome. *Psychol. Res.* 77, 388–398. doi: 10.1007/s00426-012-0441-0

Iversen, J. R., Repp, B. H., and Patel, A. D. (2009). Top-down control of rhythm perception modulates early auditory responses. *Ann. N Y Acad. Sci.* 1169, 58–73. doi: 10.1111/j.1749-6632.2009.04579.x

Jeannerod, M. (2001). Neural simulation of action: a unifying mechanism for motor cognition. *Neuroimage* 14, S103–S109. doi: 10.1006/nimg.2001.0832

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Percept. Psychophys.* 14, 201–211. doi: 10.3758/bf03212378

Jones, M., and Boltz, M. (1989). Dynamic attending and responses to time. *Psychol. Rev.* 96, 459–491. doi: 10.1037//0033-295x.96.3.459

Jones, M. R., Johnston, H. M., and Puente, J. (2006). Effects of auditory pattern structure on anticipatory and reactive attending. *Cogn. Psychol.* 53, 59–96. doi: 10.1016/j.cogpsych.2006.01.003

Jones, M. R., Moynihan, H., MacKenzie, N., and Puente, J. (2002). Temporal aspects of stimulus-driven attending in dynamic arrays. *Psychol. Sci.* 13, 313–319. doi: 10.1111/1467-9280.00458

Kayser, C., Logothetis, N. K., and Panzeri, S. (2010). Visual enhancement of the information representation in auditory cortex. *Curr. Biol.* 20, 19–24. doi: 10.1016/j.cub.2009.10.068

Klemen, J., and Chambers, C. D. (2012). Current perspectives and methods in studying neural mechanisms of multisensory interactions. *Neurosci. Biobehav. Rev.* 36, 111–133. doi: 10.1016/j.neubiorev.2011.04.015

Koelewijn, T., Bronkhorst, A., and Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: a review of audiovisual studies. *Acta Psychol. (Amst)* 134, 372–384. doi: 10.1016/j.actpsy.2010.03.010

Lakatos, P., Chen, C. M., O'Connell, M. N., Mills, A., and Schroeder, C. E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron* 53, 279–292. doi: 10.1016/j.neuron.2006.12.011

Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I., and Schroeder, C. E. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science* 320, 110–113. doi: 10.1126/science.1154735

Lakatos, P., Musacchia, G., O'Connel, M. N., Falchier, A. Y., Javitt, D. C., and Schroeder, C. E. (2013). The spectrotemporal filter mechanism of auditory selective attention. *Neuron* 77, 750–761. doi: 10.1016/j.neuron.2012.11.034

Large, E. W., and Jones, M. R. (1999). The dynamics of attending: how people track time-varying events. *Psychol. Rev.* 106, 119–159. doi: 10.1037//0033-295x.106.1.119

Large, E. W., and Snyder, J. S. (2009). Pulse and meter as neural resonance. *Ann. N Y Acad. Sci.* 1169, 46–57. doi: 10.1111/j.1749-6632.2009.04550.x

Lee, H., and Noppeney, U. (2014). Temporal prediction errors in visual and auditory cortices. *Curr. Biol.* 24, R309–R310. doi: 10.1016/j.cub.2014.02.007

Marchant, J. L., and Driver, J. (2013). Visual and audiovisual effects of isochronous timing on visual perception and brain activity. *Cereb. Cortex* 23, 1290–1298. doi: 10.1093/cercor/bhs095

Meredith, M. A., and Stein, B. E. (1983). Interactions among converging sensory inputs in the superior colliculus. *Science* 221, 389–391. doi: 10.1126/science.6867718

Meyer, G. F., Harrison, N. R., and Wuerger, S. M. (2013). The time course of auditory-visual processing of speech and body actions: evidence for the simultaneous activation of an extended neural network for semantic processing. *Neuropsychologia* 51, 1716–1725. doi: 10.1016/j.neuropsychologia.2013.05.014

Meyer, G., and Wuerger, S. (2001). Cross-modal integration of auditory and visual motion signals. *Neuroreport* 12, 2557–2560. doi: 10.1097/00001756-200108080-00053

Miura, A., Kudo, K., Ohtsuki, T., and Kanehisa, H. (2011). Coordination modes in sensorimotor synchronization of whole-body movement: a study of street dancers and non-dancers. *Hum. Mov. Sci.* 30, 1260–1271. doi: 10.1016/j.humov.2010.08.006

Morein-Zamir, S., Soto-Faraco, S., and Kingstone, A. (2003). Auditory capture of vision: examining temporal ventriloquism. *Brain Res. Cogn. Brain Res.* 17, 154–163. doi: 10.1016/s0926-6410(03)00089-2

Musacchia, G., and Schroeder, C. E. (2009). Neuronal mechanisms, response dynamics and perceptual functions of multisensory interactions in auditory cortex. *Hear. Res.* 258, 72–79. doi: 10.1016/j.heares.2009.06.018

Naci, L., Taylor, K. I., Cusack, R., and Tyler, L. K. (2012). Are the senses enough for sense? Early high-level feedback shapes our comprehension of multisensory objects. *Front. Integr. Neurosci* 6:82. doi: 10.3389/fnint.2012.00082

Noppeney, U., Josephs, O., Hocking, J., Price, C. J., and Friston, K. J. (2008). The effect of prior visual information on recognition of speech and sounds. *Cereb. Cortex* 18, 598–609. doi: 10.1093/cercor/bhm091

Nozaradan, S., Peretz, I., and Mouraux, A. (2012). Steady-state evoked potentials as an index of multisensory temporal binding. *Neuroimage* 60, 21–28. doi: 10.1016/j.neuroimage.2011.11.065

Osborne, J. W. (2002). Notes on the use of data transformations. *Pract. Assess. Res. Eval.* 8.

Parise, C. V., and Spence, C. (2009). "When birds of a feather flock together": synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS One* 4:e5664. doi: 10.1371/journal.pone.0005664

Patel, A. D. (2014). Can nonlinguistic musical training change the way the brain processes speech? The expanded OPERA hypothesis. *Hear. Res.* 308, 98–108. doi: 10.1016/j.heares.2013.08.011

Petrini, K., Dahl, S., Rocchesso, D., Waadeland, C. H., Avanzini, F., Puce, A., et al. (2009a). Multisensory integration of drumming actions: musical expertise affects perceived audiovisual asynchrony. *Exp. Brain Res.* 198, 339–352. doi: 10.1007/s00221-009-1817-2

Petrini, K., Russell, M., and Pollick, F. (2009b). When knowing can replace seeing in audiovisual integration of actions. *Cognition* 110, 432–439. doi: 10.1016/j.cognition.2008.11.015

Prinz, W. (1997). Perception and action planning. *Eur. J. Cogn. Psychol.* 9, 129–154. doi: 10.1080/713752551

Recanzone, G. H. (2003). Auditory influences on visual temporal rate perception. *J. Neurophysiol.* 89, 1078–1093. doi: 10.1152/jn.00706.2002

Repp, B. H. (2010). Do metrical accents create illusory phenomenal accents? *Atten. Percept. Psychophys.* 72, 1390–1403. doi: 10.3758/app.72.5.1390

Repp, B., and Penel, A. (2004). Rhythmic movement is attracted more strongly to auditory than to visual rhythms. *Psychol. Res.* 68, 252–270. doi: 10.1007/s00426-003-0143-8

Repp, B. H., and Su, Y.-H. (2013). Sensorimotor synchronization: a review of recent research (2006–2012). *Psychon. Bull. Rev.* 20, 403–452. doi: 10.3758/s13423-012-0371-2

Rothermich, K., Schmidt-Kassow, M., and Kotz, S. A. (2012). Rhythm's gonna get you: regular meter facilitates semantic sentence processing. *Neuropsychologia* 50, 232–244. doi: 10.1016/j.neuropsychologia.2011.10.025

Sanabria, D., Spence, C., and Soto-Faraco, S. (2007). Perceptual and decisional contributions to audiovisual interactions in the perception of apparent motion: a signal detection study. *Cognition* 102, 299–310. doi: 10.1016/j.cognition.2006.01.003

Saygin, A. P., Driver, J., and de Sa, V. R. (2008). In the footsteps of biological motion and multisensory perception: judgments of audiovisual temporal relations are enhanced for upright walkers. *Psychol. Sci.* 19, 469–475. doi: 10.1111/j.1467-9280.2008.02111.x

Schouten, B., Troje, N. F., Vroomen, J., and Verfaillie, K. (2011). The effect of looming and receding sounds on the perceived in-depth orientation of pepth-ambiguous biological motion figures. *PLoS One* 6:e14725. doi: 10.1371/journal.pone.0014725

Schroeder, C. E., and Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends Neurosci.* 32, 9–18. doi: 10.1016/j.tins.2008.09.012

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., and Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends Cogn. Sci.* 12, 106–113. doi: 10.1016/j.tics.2008.01.002

Senkowski, D., Schneider, T. R., Foxe, J. J., and Engel, A. K. (2008). Crossmodal binding through neural coherence: implications for multisensory processing. *Trends Neurosci.* 31, 401–409. doi: 10.1016/j.tins.2008.05.002

Sevdalis, V., and Keller, P. E. (2010). Cues for self-recognition in point-light displays of actions performed in synchrony with music. *Conscious. Cogn.* 19, 617–626. doi: 10.1016/j.concog.2010.03.017

Spence, C. (2011). Crossmodal correspondences: a tutorial review. *Atten. Percept. Psychophys.* 73, 971–995. doi: 10.3758/s13414-010-0073-7

Stanislaw, H., and Todorov, N. (1999). Calculation of signal detection theory measures. *Behav. Res. Methods Instrum. Comput.* 31, 137–149. doi: 10.3758/bf03207704

Stein, B. E., and Stanford, T. R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nat. Rev. Neurosci.* 9, 225–266. doi: 10.1038/nrn2377

Stevenson, R. A., Ghose, D., Fister, J. K., Sarko, D. K., Altieri, N. A., Nidiffer, A. R., et al. (2014a). Identifying and quantifying multisensory integration: a tutorial review. *Brain Topogr.* 27, 707–730. doi: 10.1007/s10548-014-0365-7

Stevenson, R. A., Wallace, M. T., and Altieri, N. (2014b). The interaction between stimulus factors and cognitive factors during multisensory integration of audiovisual speech. *Front. Psychol.* 5:352. doi: 10.3389/fpsyg.2014.00352

Stevenson, R. A., Zemtsov, R. K., and Wallace, M. T. (2012). Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. *J. Exp. Psychol. Hum. Percept. Perform.* 38, 1517–1529. doi: 10.1037/a0027339

Su, Y.-H. (2014a). Audiovisual beat induction in complex auditory rhythms: point-light figure movement as an effective visual beat. *Acta Psychol. (Amst)* 151, 40–50. doi: 10.1016/j.actpsy.2014.05.016

Su, Y.-H. (2014b). Peak velocity as a cue in audiovisual synchrony perception of rhythmic stimuli. *Cognition* 131, 330–344. doi: 10.1016/j.cognition.2014.02.004

Su, Y.-H. (2014c). Visual enhancement of auditory beat perception across auditory interference levels. *Brain Cogn.* 90, 19–31. doi: 10.1016/j.bandc.2014.05.003

Talsma, D., Senkowski, D., Soto-Faraco, S., and Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends Cogn. Sci.* 14, 400–410. doi: 10.1016/j.tics.2010.06.008

Ten Oever, S., Sack, A. T., Wheat, K. L., Bien, N., and van Atteveldt, N. (2013). Audio-visual onset differences are used to determine syllable identity for ambiguous audio-visual stimulus pairs. *Front. Psychol.* 4:331. doi: 10.3389/fpsyg.2013.00331

Thomas, J. P., and Shiffrar, M. (2010). I can see you better if I can hear you coming: action-consistent sounds facilitate the visual detection of human gait. *J. Vis.* 10:14. doi: 10.1167/10.12.14

Thomas, J. P., and Shiffrar, M. (2013). Meaningful sounds enhance visual sensitivity to human gait regardless of synchrony. *J. Vis.* 13:8. doi: 10.1167/13.14.8

Toiviainen, P., Luck, G., and Thompson, M. R. (2010). Embodied meter: hierarchical eigenmodes in music-induced movement. *Music Percept.* 28, 59–70. doi: 10.1525/mp.2010.28.1.59

van Atteveldt, N. M., Blau, V. C., Blomert, L., and Goebel, R. (2010). fMR-adaptation indicates selectivity to audiovisual content congruency in distributed clusters in human superior temporal cortex. *BMC Neurosci.* 11:11. doi: 10.1186/1471-2202-11-11

van Atteveldt, N., Murray, M. M., Thut, G., and Schroeder, C. E. (2014). Multisensory integration: flexible use of general operations. *Neuron* 81, 1240–1253. doi: 10.1016/j.neuron.2014.02.044

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., and Theeuwes, J. (2008). Pip and pop: nonspatial auditory signals improve spatial visual search. *J. Exp. Psychol. Hum. Percept. Perform.* 34, 1053–1065. doi: 10.1037/0096-1523.34.5.1053

van Wassenhove, V., Grant, K. W., and Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proc. Natl. Acad. Sci. U S A* 102, 1181–1186. doi: 10.1073/pnas.0408949102

van Wassenhove, V., Grant, K. W., and Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia* 45, 598–607. doi: 10.1016/j.neuropsychologia.2006.01.001

Vatakis, A., and Spence, C. (2007). Crossmodal binding: evaluating the "unity assumption" using audiovisual speech stimuli. *Percept. Psychophys.* 69, 744–756. doi: 10.3758/bf03193776

Vroomen, J., and Keetels, M. (2010). Perception of intersensory synchrony: a tutorial review. *Atten. Percept. Psychophys.* 72, 871–884. doi: 10.3758/app.72.4.871

Vroomen, J., and Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *J. Cogn. Neurosci.* 22, 1583–1596. doi: 10.1162/jocn.2009.21308

Welch, R. B., and Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychol. Bull.* 88, 638–667. doi: 10.1037//0033-2909.88.3.638

Wuerger, S. M., Crocker-Buque, A., and Meyer, G. F. (2012). Evidence for auditory-visual processing specific to biological motion. *Seeing Perceiving* 25, 15–28. doi: 10.1163/187847611x620892

Wuerger, S. M., Hofbauer, M., and Meyer, G. F. (2003). The integration of auditory and visual motion signals at threshold. *Percept. Psychophys.* 65, 1188–1196. doi: 10.3758/bf03194844

Zion Golumbic, E., Cogan, G. B., Schroeder, C. E., and Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". *J. Neurosci.* 33, 1417–1426. doi: 10.1523/jneurosci.3675-12.2013

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 14 August 2014; paper pending published: 22 September 2014; accepted: 17 November 2014; published online: 08 December 2014.

*Citation: Su Y-H (2014) Content congruency and its interplay with temporal synchrony modulate integration between rhythmic audiovisual streams. Front. Integr. Neurosci. **8**:92. doi: 10.3389/fnint.2014.00092*

This article was submitted to the journal Frontiers in Integrative Neuroscience.

frontiers in
INTEGRATIVE NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 02 February 2015
doi: 10.3389/fnint.2015.00001

Multisensory perception of the six basic emotions is modulated by attentional instruction and unattended modality

Sachiko Takagi[1], Saori Hiramatsu[2], Ken-ichi Tabei[3] and Akihiro Tanaka[1]*

[1] *Tokyo Woman's Christian University, Tokyo, Japan*
[2] *Waseda Institute for Advanced Study, Tokyo, Japan*
[3] *Graduate School of Medicine, Mie University, Tsu, Japan*

Edited by:
Ruth Adam, Ludwig-Maximilian-University, Germany

Reviewed by:
Antje B. M. Gerdes, University of Mannheim, Germany
Fabrizio Leo, Istituto Italiano di Tecnologia, Italy

***Correspondence:**
Akihiro Tanaka, Tokyo Woman's Christian University, 2-6-1 Zempukuji, Suginami-ku, Tokyo 167-8585, Japan
e-mail: akihiro@lab.twcu.ac.jp

Previous studies have shown that the perception of facial and vocal affective expressions interacts with each other. Facial expressions usually dominate vocal expressions when we perceive the emotions of face–voice stimuli. In most of these studies, participants were instructed to pay attention to the face or voice. Few studies compared the perceived emotions with and without specific instructions regarding the modality to which attention should be directed. Also, these studies used combinations of the face and voice which expresses two opposing emotions, which limits the generalizability of the findings. The purpose of this study is to examine whether the emotion perception is modulated by instructions to pay attention to the face or voice using the six basic emotions. Also we examine the modality dominance between the face and voice for each emotion category. Before the experiment, we recorded faces and voices which expresses the six basic emotions and orthogonally combined these faces and voices. Consequently, the emotional valence of visual and auditory information was either congruent or incongruent. In the experiment, there were unisensory and multisensory sessions. The multisensory session was divided into three blocks according to whether an instruction was given to pay attention to a given modality (face attention, voice attention, and no instruction). Participants judged whether the speaker expressed happiness, sadness, anger, fear, disgust, or surprise. Our results revealed that instructions to pay attention to one modality and congruency of the emotions between modalities modulated the modality dominance, and the modality dominance is differed for each emotion category. In particular, the modality dominance for anger changed according to each instruction. Analyses also revealed that the modality dominance suggested by the congruency effect can be explained in terms of the facilitation effect and the interference effect.

Keywords: attentional instruction, audiovisual integration, unattended stimuli, modality dominance, congruency effect, emotion perception

INTRODUCTION

Human beings must perceive other people's emotions appropriately to facilitate successful social interactions. Emotions are expressed using different sensory channels, such as a face and voice, and are judged by integrating information from these channels in natural settings. Previous studies have shown that the perceptions of facial and vocal affective expressions are interactive. For instance, emotional judgments based on one modality are impaired by incongruent emotions and enhanced by congruent emotions expressed in other modalities (de Gelder and Vroomen, 2000; Kreifelts et al., 2007; Collignon et al., 2008). These findings have been confirmed by experiments using both static faces (Massaro and Egan, 1996; de Gelder et al., 1999; de Gelder and Vroomen, 2000) and dynamic faces (Collignon et al., 2008; Van den stock et al., 2008; Tanaka et al., 2010). Furthermore, integration of emotional information from the face and voice has been demonstrated in infants (Grossmann et al., 2006) and people with pervasive developmental disorder (Magnèe et al., 2008). Results from brain studies have shown that emotions from the face and voice interact with each other (Pourtois et al., 2005; Ethofer et al., 2006a,b; Kreifelts et al., 2007; Talsma et al., 2007, 2010). Specifically, neuroimaging data using functional magnetic resonance imaging (fMRI) on audiovisual integration of emotional information highlights stronger activation in the left middle temporal gyrus (Pourtois et al., 2005), left basolateral amygdala (Ethofer et al., 2006a), right fusiform gyrus (Ethofer et al., 2006a), right thalamus (Kreifelts et al., 2007), and posterior superior temporal sulcus (Ethofer et al., 2006b; Kreifelts et al., 2007).

In most behavioral studies, participants were instructed to pay attention to only one modality (i.e., face or voice) and judge the emotion shown in that modality. This method allows us to investigate whether the emotional information from the face and voice are integrated inevitably. This paradigm, the immediate cross-modal paradigm (Bertelson and de Gelder, 2004), has been widely used to examine the multisensory perception of the emotion (de Gelder and Vroomen, 2000; Vroomen et al., 2001; Collignon et al.,

2008). For instance, de Gelder and Vroomen (2000) asked participants to judge the emotion of congruent and incongruent face–voice stimuli expressing two opposing emotions (happiness and sadness) by instructing participants to attend to a certain modality. The results showed that accuracy of the emotion perception was higher for congruent stimuli than it was for incongruent stimuli. That is, though participants understood that they should pay attention to only one modality, the emotion perception was impaired by the emotion of the other modality in the presence of incongruent stimuli. Vroomen et al. (2001) examined whether integration of emotional information from the face and voice requires limited attentional resources using dual-task methodology. The results showed that emotional judgment from the face and voice was unconstrained by attentional resources. These findings imply that audiovisual integration of emotional information occurs as a mandatory process.

Is perceived emotion different according to whether participants are instructed to pay attention to either the face or voice or not? Few studies have compared the emotion perceptions with and without specific instructions regarding the modality to which attention should be directed. Collignon et al. (2008) conducted two experiments, one with instructions and one without instructions. They used stimuli denoting expressions of fear and disgust. Stimuli were faces and voices in which fear, disgust, or combinations of both were expressed. Experiment 1 without instructions used congruent multisensory stimuli and unisensory stimuli, and Experiment 2 with instructions used incongruent stimuli in addition to stimuli in Experiment 1. In both experiments, participants were asked to categorize fear and disgust. The results revealed that the performance on congruent multisensory stimuli was higher than that on unisensory stimuli in both experiments. That is, with regard to the emotion perception for unisensory and congruent multisensory stimuli, the results were consistent regardless of whether instructions as to attention were given. However, it is unclear whether the emotion perception for incongruent stimuli is modulated by attentional instructions.

Previous studies have also suggested that particular emotional channels dominate other channels when comparing the accuracy of emotion judgments in the unisensory condition against the multisensory condition. Collignon et al. (2008) suggest that performance of emotion judgment was better when participants attended to the face as opposed to the voice, at least for fear and disgust. Similarly, other research has reported higher accuracy for faces when compared to voices, even when only one of the two was presented (e.g., Johnstone and Scherer, 2000; Pell, 2002; Hawk et al., 2009).

Most studies investigating the audiovisual integration and modality dominance in judging emotions have evaluated a limited number of emotions (sometimes as few as two). However, recent studies have focused on more than two emotions and have shown face dominance in general. Paulmann and Pell (2011) used congruent face–voice pairs in which five emotions (anger, disgust, sadness, happiness, and surprise) and neutral were expressed. They investigated whether the emotion perception is more accurate for multi-channel stimuli by presenting stimuli with different combinations of the face and prosody of the voice. Participants were not given specific instructions with regard to attention. The emotion perception was better in response to multi-channel as opposed to single-channel stimuli. When stimuli contained only one emotional channel, perception tended to be higher for faces than for vocal prosody. However, this tendency was not uniform across emotion categories. Föcker et al. (2011) used both congruent and incongruent face–voice pairs in which three emotions (happiness, anger, and sadness) and neutral were expressed. They aimed to test whether participants were able to categorize the stimuli based on each emotion, expecting that the accuracy on incongruent stimuli would be lower than that on congruent stimuli. Participants were instructed to pay attention to one modality. The accuracy on congruent audiovisual stimuli was higher than that on unisensory stimuli, and the accuracy on unisensory stimuli was higher than that on incongruent audiovisual stimuli. However, the accuracies varied across emotion categories. That is, both studies showed that the emotion perception accuracy is not uniform across emotion categories. Therefore, it is necessary to closely examine the emotion perception for each emotion category.

As presented above, previous studies have revealed that emotional information from the face and voice demonstrated mandatory interaction, and that facial cues generally dominate vocal cues in judging emotions from the face and voice. However, it is unclear whether the interaction of emotional information from the face and voice is mandatory for any emotion categories. Also it remains unclear whether the modality dominance is the same across emotion categories. It is important to examine the emotion perception in terms of the impact of instructions and the unattended modality using the six basic emotions.

In the present study, we examined whether the emotion perception is modulated by instructions to pay attention to one of two modalities. We used faces and voices expressing the six basic emotions, and face–voice combinations in which the face and voice showed the same or different emotion.

MATERIALS AND METHODS

PARTICIPANTS

Twenty-six Japanese university students residing in Japan (13 male, 13 female; average age 20.3 ± SD 1.4) participated in the experiment. All participants provided written informed consent prior to participation. The study was approved by the local ethics committee and all subjects gave their written informed consent prior to inclusion in the study.

STIMULI SELECTION

Models

Twenty-one Japanese (10 male, 11 female) students demonstrated the six basic emotions for audiovisual speech stimuli.

Creation of the audiovisual speech stimuli

For the models' utterances, six short phrases with emotionally neutral meanings were chosen. The models were asked to say "*Soonandesuka?*" (Is that so?), "*Korenani?*" (What's this?), "*Sayoonara*" (Goodbye), "*Hai, moshimoshi*" (Hello), "*Doonatteruno?*" (What's going on?), and "*Daijoobu?*" (Are you okay?) in Japanese with angry, disgusted, fearful, happy, sad, and surprised expressions. While facially expressing each intended emotion, the models

uttered the six meaning-neutral phrases, filling them with the required emotion. Before starting the recording of each emotion, each model was instructed on how to facially and vocally perform the emotional expression. For facial expressions, instructions were given based on the Action Units of Ekman and Friesen (1978). For vocal expressions, samples by radio announcers were given and, when necessary, emotional context was provided to induce each emotion. After receiving these instructions, the models used a mirror to practice their expressions. The recording began when the model could adequately convey the emotion with simultaneous facial and vocal expressions. For the recording, they were asked to speak the phrases at three different speech rates—slow, normal, and fast—and to repeat them three times at each speech rate. Thus, 324 samples (6 emotions × 6 utterances × 3 speech rates × 3 repetitions) were recorded from each model.

A recording studio was used with sufficient lighting equipment on the ceiling. A digital video camera (SONY PVW−637 k) was used for recording video and a microphone (SONY ECM-77B) was used for recording audio. A gray background was used throughout the recording. The recordings took place per emotion type. All models wore a white cardigan and a pin microphone about 15 cm away from their mouths on their chests for recording audio. They sat about 2 m away from the camera and 30 cm in front of the background.

The recorded video was edited using Avid Xpress (Avid Technology, Inc.). For each model's performance, the onset and offset of the utterance was identified. Then, we extracted the clip including five frames before and five frames after the onset. From the 21 models, eight models were selected via agreement between two evaluators who judged that the facial and vocal expressions suitably expressed each emotion. The selection standards consisted mainly of whether the differences among emotions were clearly discriminated in facial expressions and whether there were minimal head movements and blinking. Regarding vocal expressions, models were selected according to whether differences among emotions were clearly expressed and whether the utterance was fluent and clear. Furthermore, for each combination of the emotion and utterance, a total of nine video clips were recorded from each model, having repeated each utterance three times at three speech rates. From the nine video clips recorded, three were selected that had approximately the same utterance duration, regardless of the instruction on speech rate. Two uttered phrases were eliminated. One phrase ("*Daijoobu?*") was eliminated because some participants pointed out that it does not have a neutral meaning. The other phrase ("*Hai, moshimoshi*") was eliminated because it has a pause between "*Hai*" and "*moshimoshi*" that makes it difficult to create incongruent stimuli. Finally, the uttered phrases were reduced to just four: "*Soonandesuka?*" (Is that so?), "*Korenani?*" (What's this?), "*Sayoonara*" (Goodbye), and "*Doonatteruno?*" (What's going on?).

Evaluation experiment

For the evaluation experiment, participants were 99 Japanese university students (47 male, 52 female; average age 20.7 ± SD 2.07). The experiment was divided into two rounds, with only facial expressions in the first round and only vocal expressions in the second round. The reason for keeping the order of rounds constant was that it was likely to preclude biased evaluations of the facial expressions due to lip-reading from the uttered phrase already being known. Both rounds were conducted in groups (10–20 participants), and participants were required to participate in both rounds. The experiment consisted of a total of eight sessions, corresponding to each of the eight models, for a total of 72 trials. The order of the sessions was counterbalanced. Images were projected onto a screen in the front of the classroom using a projector attached to a PC, and sound was presented through a loudspeaker. Participants were seated in a spot from which they could adequately see the entire screen. They were instructed to choose which the emotion was being expressed (or heard) from the six emotions (anger, disgust, fear, happiness, sadness, and surprise) and write on an answer form that was provided. Further, they were instructed to judge intuitively rather than think deeply about their decision.

The rates of matches between participants' responses and models' intended emotions were calculated per model and uttered phrase. Based on the results, the four models and two uttered phrases, "*Soonandesuka?*" (Is that so?) and "*Doonatteruno?*" (What's going on?), that generated the most matches were used in the main experiment. The mean accuracies of emotion judgment from faces and voices for selected stimuli with respect to the emotion category are shown in **Table 1**.

Based on the above result, faces and voices presented in the evaluation experiment were edited and processed so that emotions depicted by the facial expressions and vocal expressions were paired in a congruent or incongruent fashion. The former were congruent stimuli, and the latter were incongruent stimuli of the main experiment. There were 36 combinations of facial expression (6) and vocal expression (6) in total. In these combinations, six combinations were congruent and 30 combinations (6 × 5) were incongruent. Finally, there were 48 congruent stimuli (6 congruent combinations × 2 phrases × 4 actors) and 240 incongruent stimuli (30 incongruent combinations × 2 phrases × 4 actors).

PROCEDURE

The main experiment was conducted in a group setting. Visual and audio stimuli were presented in the same way as the evaluation experiment. Participants were seated such that they could see the entire screen and listen to the auditory stimuli. In the main experiment, there were unisensory and multisensory sessions. In the unisensory session, only faces or voices were presented. The multisensory session was divided into three blocks according to whether an instruction was given to pay attention to a given modality. In

Table 1 | Mean accuracies (%) of emotion judgment from faces and voices for selected stimuli with respect to the emotion category (SD in parentheses).

	Emotion category					
	Anger	**Disgust**	**Fear**	**Happiness**	**Sadness**	**Surprise**
Face	80.3	73.0	21.3	98.6	69.7	83.7
	(15.05)	(9.88)	(6.46)	(1.98)	(32.12)	(8.53)
Voice	64.4	49.3	31.9	62.4	62.4	73.6
	(24.28)	(9.79)	(20.22)	(28.98)	(33.33)	(13.80)

the no instruction (NI) block, participants were not instructed to pay attention to either of the two modalities. In the visual instruction (VI) block, participants were instructed to pay attention to the visual information (i.e., face). In the auditory instruction (AI) block, participants were instructed to pay attention to the audio information (i.e., voice). Participants were required to judge the emotion perceived from both the face and voice, the face only, and the voice only in NI, VI, and AI, respectively. They were required to ignore the voice in VI and the face in AI. The main experiment always began with the multisensory session. In the multisensory session, the first block was NI, followed by VI and AI. The order of VI and AI was counterbalanced. In the unisensory session, the order of the face and voice blocks was also counterbalanced. Participants answered the judged the emotion in handwriting by choosing one of the six options on the answer sheet.

Each block of the multisensory session consisted of 288 trials and each block of the unisensory session consisted of 192 trials (96 stimuli repeated two times). Stimuli were presented in random order in all blocks of both sessions. The main experiment took 2 h per day for 2 days. Participants were allowed to take a 10 min break every hour.

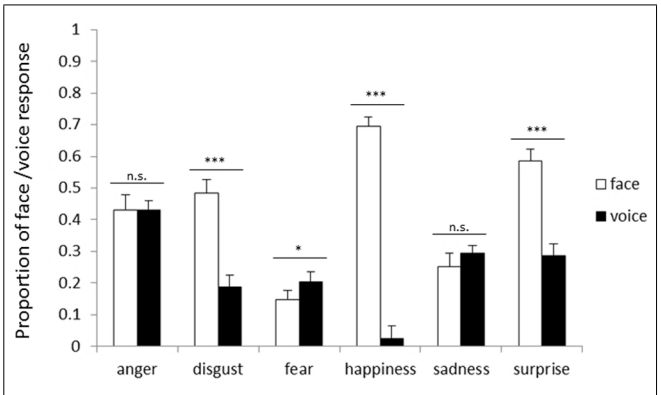


FIGURE 1 | Proportion of facial and vocal response on incongruent trials for each emotion category in the no instruction (NI) block. Error bars represent SE. Asterisks indicate significant differences between modalities (*$p < 0.05$, ***$p < 0.001$).

RESULTS

In this study, we examined whether the emotion perception is modulated by instructions to pay attention to one of two modalities. We used both congruent and incongruent stimuli expressing the six basic emotions. The modality dominance between the face and voice for each emotion was also examined. We describe the results according to these purposes in the following sections.

EMOTION PERCEPTION WITH AND WITHOUT INSTRUCTIONS ON INCONGRUENT TRIALS

Here, we focused on only incongruent trials in NI, VI, and AI. To examine the emotion perception when NI to pay attention to one modality were given, we performed a response modality (face or voice responses) × emotion (anger, disgust, fear, happiness, sadness, or surprise) two-way analysis of variance (ANOVA) on participants' responses on incongruent trials in NI. A significance level of $p < 0.05$ was used for all ANOVA to evaluate all effects. Face responses for each emotion were defined as the mean percentage of participants' responses for a given emotion when the face expressed that emotion. For example, when presented with stimuli in which the face showed anger but the voice showed a different emotion, we calculated the mean percentage of "anger" responses as the face response for anger. In contrast, voice responses for each emotion were defined as the mean percentage of responses for a given emotion when the voice expressed that emotion. For example, when presented with stimuli in which the voice expressed anger but the face expressed a different emotion, we calculated the mean percentage of "anger" responses as the voice response for anger. We used the face and voice responses for each emotion as the dependent variables.

Figure 1 shows the proportion of the face and voice responses for each emotion in NI. The main effect of response modality was significant [$F(1,25) = 106.86$, $p < 0.001$]. The proportion of the face responses (43.2%) was higher than that of voice responses (23.8%, $p < 0.001$), demonstrating that facial cues dominated vocal cues when MI was given to pay attention to one modality. The main effect of emotion was also significant [$F(5,125) = 21.57$, $p < 0.001$], and the proportion for fear was the lowest (17.6%, ps < 0.05). In addition, the two-way interaction between response modality and emotion was significant [$F(5,125) = 156.50$, $p < 0.001$]. Simple main effects analyses showed that the proportion of the face responses was higher than that of the voice responses for disgust (face 48.4%, voice 18.7%), happiness (face 69.4%, voice 2.4%), and surprise (face 58.6%, voice 28.7%; ps < 0.001). In contrast, the proportion of the voice responses (20.4%) was higher than that of the face responses for fear (14.8%, $p < 0.05$).

To examine the emotion perception when the instruction to pay attention to one modality was given, a similar analysis was applied to VI and AI. We performed an attended modality (VI or AI) × emotion (anger, disgust, fear, happiness, sadness, or surprise) two-way ANOVA on the accuracy on incongruent trials in VI and AI. The accuracy for each emotion was calculated according to the attended modality. For example, when presented with stimuli in which the face showed anger but the voice showed a different emotion in VI, we calculated the proportion of "anger" responses as the accuracy for anger in VI. In contrast, for stimuli in which the voice showed anger but the face showed a different emotion in AI, we calculated the proportion of "anger" responses as the accuracy for anger in AI. We used these accuracies for each emotion as the dependent variables.

Figure 2 shows the accuracies for each emotion in VI and AI. The main effect of the attended modality was significant [$F(1,25) = 48.16$, $p < 0.001$]. The accuracy in VI (61.9%) was higher than that in AI (46.3%, $p < 0.001$). This showed that facial cues dominated vocal cues when the instruction to pay attention to one modality was given. The main effect of emotion was also significant [$F(5,125) = 37.76$, $p < 0.001$], and the accuracy for fear was the lowest (27.1%, ps < 0.05). In addition, the two-way interaction between the attended modality and emotion was significant [$F(5,125) = 44.71$, $p < 0.001$]. Simple main effects analyses showed that the accuracy in VI was higher than that in AI for anger (VI 64.2%, AI 54.9%; $p < 0.05$), disgust (VI 66.7%, AI

FIGURE 2 | Accuracy for each emotion category in visual instruction (VI) block and auditory instruction (AI) block. Error bars represent SE. Asterisks indicate significant differences between modalities (*$p < 0.05$, ***$p < 0.001$).

28.4%), happiness (VI 93.3%, AI 41.2%), and surprise (VI 79.6%, AI 60.5%) (ps < 0.001). In contrast, the accuracy for fear was higher in AI (37.8%) than it was in VI (16.3%; ps < 0.001).

The comparison between **Figures 1** and **2** based on the analyses showed that facial cues generally dominate vocal cues, regardless of the presence of instructions. For anger, face dominance was shown only when paying attention to one modality.

IMPACT OF CONGRUENCY OF EMOTIONS BETWEEN FACES AND VOICES

In the previous section, we focused on only incongruent trials in NI, VI, and AI. Here, to examine the emotion perception in terms of congruency between facial and vocal cues, we focused on both congruent and incongruent trials when instructions to pay attention to one modality were given. We also investigated whether the channel being unisensory or multisensory had an effect. We performed an attended modality (face or voice) × presentation condition [multisensory_congruent (MC), unisensory (UNI), or multisensory_incongruent (MI)] × emotion (anger, disgust, fear, happiness, sadness, or surprise) three-way ANOVA on the accuracy in each presentation condition.

Figure 3 shows accuracies on all presentation conditions for each emotion category with respect to each modality. Results revealed a significant three-way interaction [$F(10,250) = 15.93$, $p < 0.001$]. There were significant simple interactions between the attended modalities and presentation conditions for anger [$F(2,300) = 12.67$, $p < 0.001$], disgust [$F(2,300) = 13.23$, $p < 0.001$], fear [$F(2,300) = 0.02$, $p < 0.05$], happiness [$F(2,300) = 41.33$, $p < 0.001$], and surprise [$F(2,300) = 8.86$, $p < 0.001$]. The subsequent analysis revealed simple–simple main effects of the presentation conditions for each attended modality and emotion categories. The simple–simple main effect of the presentation conditions was significant for anger [$F(2,600) = 17.06$, $p < 0.001$], fear [$F(2,600) = 6.15$, $p < 0.001$], sadness [$F(2,600) = 18.64$, $p < 0.001$], and surprise [$F(2,600) = 8.43$, $p < 0.001$] when the attended modality was the face. Also, the simple–simple main effect of the presentation conditions was significant for anger [$F(2,600) = 7.07$, $p < 0.001$], disgust [$F(2,600) = 26.65$, $p < 0.001$], happiness [$F(2,600) = 80.13$, $p < 0.001$], sadness [$F(2,600) = 8.71$, $p < 0.001$], and surprise [$F(2,600) = 28.18$, $p < 0.001$] when the attended modality was the voice. In the next sections, we describe the results of multiple comparisons for a simple–simple main effect of the presentation conditions for each attended modality and emotion category.

Besides three-way interaction, it should be noted that an attended modality × presentation condition two-way interaction was also significant [$F(2,50) = 6.13$, $p < 0.005$]. That is, the accuracies among presentation conditions were different by modalities. **Figure 4** shows accuracies on all presentation conditions for all emotion categories included with respect to modality. The difference in the accuracies between MC (71.9%) and UNI (68.7%) was not significant, while the accuracies in MC and UNI were higher than that in MI (61.9%, ps < 0.05) when the attended modality was the face. In contrast, the accuracy in MC (63.7%) was higher than that in UNI (56.2%, $p < 0.05$), and the accuracy in UNI was higher than that in MI (46.4%, $p < 0.05$) when the attended modality was the voice.

Impact of presentation conditions

To examine the impact of presentation condition on the accuracy of the emotion perception for each attended modality and emotion, we describe the results of the multiple comparisons for simple–simple main effects of three-way interaction. For this purpose, we define the *congruency effect* as the difference in the accuracies between MC and MI. The congruency effect included

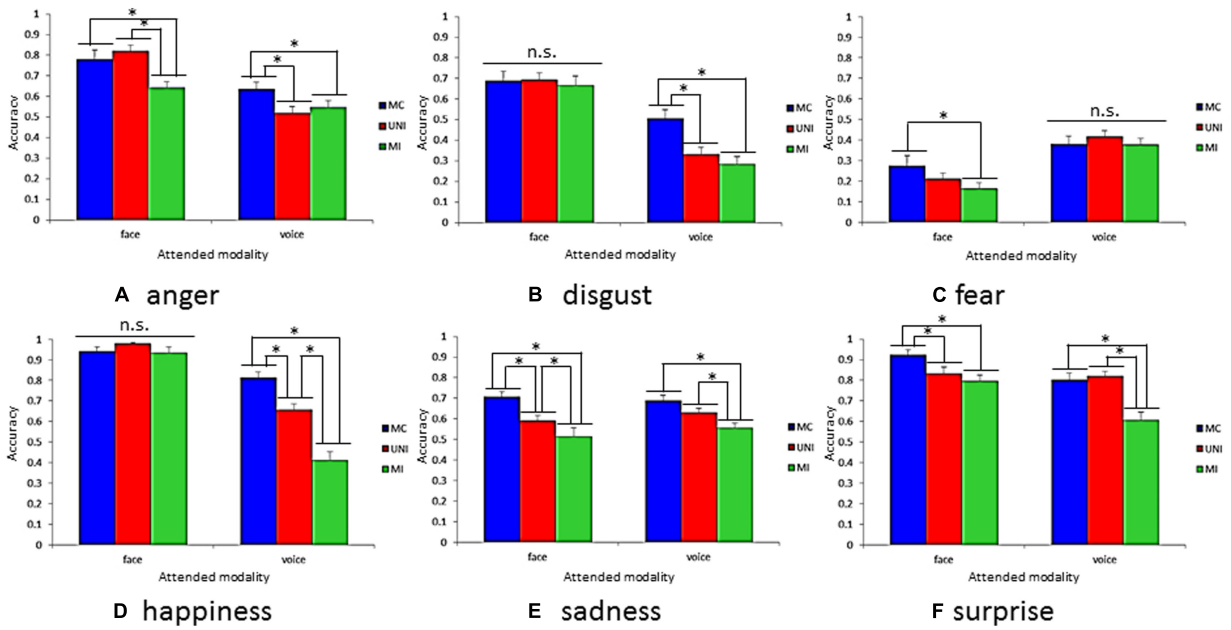


FIGURE 3 | Accuracy on all presentation conditions for each emotion category with respect to attended modality. (A) anger, **(B)** disgust, **(C)** fear, **(D)** happiness, **(E)** sadness, and **(F)** surprise. Error bars represent SE. Asterisk indicate significant differences between presentation conditions (*$p < 0.05$).

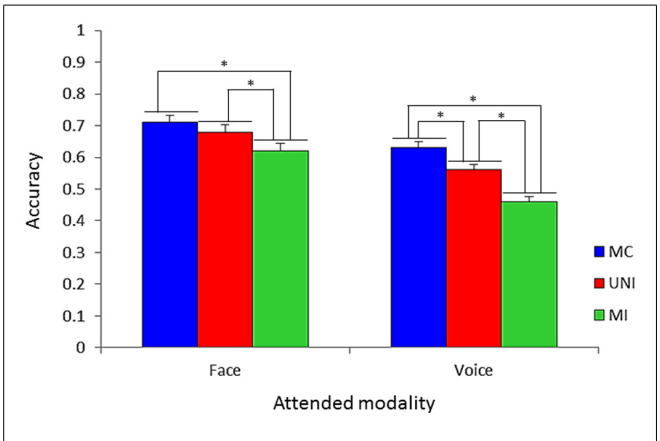


FIGURE 4 | Accuracy on all presentation conditions with respect to attended modality for all emotion categories. Error bars represent SE (*$p < 0.05$).

two dissociable effects. Then, we defined the *facilitation effect* as the difference in the accuracies between MC and UNI and the *interference effect* as the difference in the accuracies between UNI and MI. Congruency effects, facilitation effects, and interference effects for each emotion category are shown as the results of multiple comparisons for the simple–simple main effects of presentation conditions for each attended modality and emotion (**Table 2**).

Congruency effects are represented by the difference in the accuracies between the blue and green columns in **Figure 3**. Multiple comparisons for simple–simple main effects of presentation conditions revealed that there were congruency effects for anger, fear, sadness, and surprise when the attended modality was the face. The accuracies in MC were higher than those in MI for these emotions ($ps < 0.05$). Furthermore, congruency effects were observed for all emotions except for fear when the attended modality was the voice. The accuracies in MC were higher than those in MI ($ps < 0.05$).

Facilitation effects are represented by the difference in the accuracies between blue and red columns in **Figure 3**. Multiple comparisons for simple–simple main effects of presentation conditions revealed facilitation effects for sadness and surprise when the attended modality was the face. The accuracies in MC were higher than those in UNI for these emotions ($ps < 0.05$). Furthermore, facilitation effects were shown for anger, disgust, and happiness when the attended modality was the voice. The accuracies in MC were higher than those in UNI ($ps < 0.05$).

Interference effects are reflected in the difference in the accuracies between the red and green columns in **Figure 3**. Multiple comparisons for simple–simple main effects of presentation conditions revealed that there were interference effects for anger and sadness when the attended modality was the face. The accuracies in UNI were higher than those in MI for these emotions ($ps < 0.05$). Furthermore, interference effects were shown for happiness, sadness, and surprise when the attended modality was the voice. The accuracies in UNI were higher than those in MI ($ps < 0.05$).

Modality dominance in congruency effect, facilitation effect, and interference effect

We examined the modality dominance for each emotion category based on the congruency effect, facilitation effect, and interference effect. Specifically, we performed an attended modality (face or voice) × emotion (anger, disgust, fear, happiness, sadness, or

Table 2 | Congruency, facilitation, and interference effects for each emotion category with respect to the attended modality.

	Face			Voice		
	Congruency effect	**Facilitation effect**	**Interference effect**	**Congruency effect**	**Facilitation effect**	**Interference effect**
Anger	○	–	○	○	○	–
Disgust	–	–	–	○	○	–
Fear	○	–	–	–	–	–
Happiness	–	–	–	○	○	○
Sadness	○	○	○	○	–	○
Surprise	○	○	–	○	–	○

○, means significant, and –, means non significant in multiple comparisons for simple–simple main effect ($p < 0.05$).

surprise) two-way ANOVA for the congruency, facilitation, and interference effects. It was assumed that the modality in which each effect was smaller dominated the other modality.

To examine the congruency effect, an attended modality (face or voice) × emotion (anger, disgust, fear, happiness, sadness, or surprise) two-way ANOVA was performed. **Figure 5** shows congruency effects for each emotion category with respect to the attended modality. Two-way ANOVA revealed significant interaction between the attended modality and emotion [$F(5,125) = 21.91$, $p < 0.001$]. The simple main effects revealed face dominance for disgust and happiness ($ps < 0.001$). In contrast, voice dominance was shown for fear ($p < 0.05$).

To examine the facilitation effect, an attended modality (face or voice) × emotion (anger, disgust, fear, happiness, sadness, or surprise) two-way ANOVA was performed. **Figure 6** shows facilitation effects for each emotion category with respect to the attended modality. Two-way ANOVA revealed significant interaction between the attended modality and emotion [$F(5,125) = 10.07$, $p < 0.001$]. The simple main effects revealed face dominance for anger, disgust, and happiness ($ps < 0.001$). In contrast, voice dominance was shown for fear and surprise ($ps < 0.05$).

FIGURE 6 | Facilitation effect for each emotion category with respect to attended modality. Error bars represent SE. Asterisks indicate significant differences between attended modalities (*$p < 0.05$, ***$p < 0.001$). "V > A" in **Table 3** was the case that the bar length of the face was shorter than that of the voice, and "A > V" in **Table 3** was the case that the bar length of the voice was shorter than that of the face.

To examine the interference effect, an attended modality (face or voice) × emotion (anger, disgust, fear, happiness, sadness, or surprise) two-way ANOVA was performed. **Figure 7**

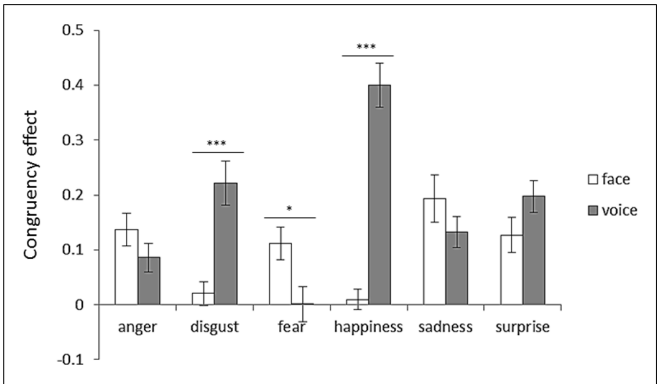


FIGURE 5 | Congruency effect for each emotion category with respect to attended modality. Error bars represent SE. Asterisks indicate significant differences between attended modalities (*$p < 0.05$, ***$p < 0.001$). "V > A" in **Table 3** was the case that the bar length of the face was shorter than that of the voice, and "A > V" in **Table 3** was the case that the bar length of the voice was shorter than that of the face.

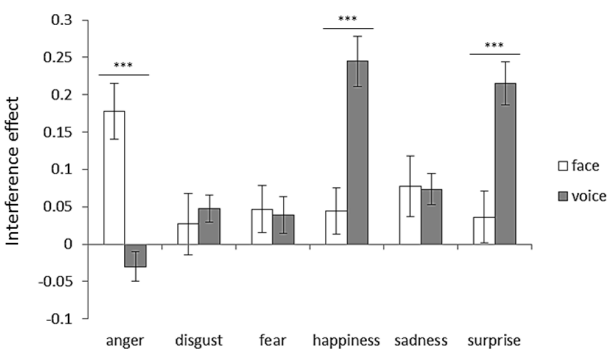


FIGURE 7 | Interference effect for each emotion category with respect to attended modality. Error bars represent SE. Asterisks indicate significant differences between attended modalities (***$p < 0.001$). "V > A" in **Table 3** was the case that the bar length of the face was shorter than that of the voice, and "A > V" in **Table 3** was the case that the bar length of the voice was shorter than that of the face.

shows interference effects for each emotion category with respect to the attended modality. Two-way ANOVA revealed significant interaction between the attended modality and emotion [$F(5,125) = 17.88$, $p < 0.001$]. The simple main effects revealed face dominance for happiness and surprise ($ps < 0.001$). In contrast, voice dominance was shown for anger ($p < 0.001$).

The modality dominances as shown by each effect for each emotion category are summarized in **Table 3**.

DISCUSSION

The present study examined the impact of attention on the emotion perception of facial and vocal stimuli. Our results revealed that instructions to pay attention to one modality and congruency of emotions between modalities modulated the modality dominance, and the modality dominance differed by each emotion category.

THE IMPACT OF ATTENTION ON EMOTION PERCEPTION OF FACES AND VOICES

Our results revealed face dominance in the audiovisual emotion perception. For most emotion categories, participants perceived the emotion from the face rather than from the voice, regardless of the modality to which participants were instructed to attend. This finding is in line with previous studies (Johnstone and Scherer, 2000; Pell, 2002; Collignon et al., 2008; Hawk et al., 2009). The emotion perception from faces is generally easier than from voices (Ekman and Friesen, 1971; Ekman et al., 1982; Russell et al., 1989; Russell, 1994; Scherer, 2003). In the previous studies examining the emotion perception from audiovisual stimuli when the attended instruction was given, the accuracies of emotion judgment from faces were generally higher than that from voices (Collignon et al., 2008; Föcker et al., 2011; Paulmann and Pell, 2011). Stimuli used in the current study showed similar tendency as shown in **Table 1**. Therefore, face dominance as shown in our results might have reflected this tendency.

More importantly, however, our results showed that the attentional instruction modulated the modality dominance for each emotion category. When the participants did not pay attention to any one modality, face dominance was shown for disgust, happiness, and surprise, while voice dominance was shown for fear. The modality dominance was not observed for anger or sadness. On the other hand, when participants paid attention to one modality, face dominance was shown for anger, disgust, happiness, and surprise, while voice dominance was shown for fear. The modality dominance was not observed for sadness. Therefore, the modality dominance for anger was modulated by the instruction. These results suggest that modality dominance was not consistent across emotion categories and that the modality dominance for anger was modulated by the instruction.

We speculate that this finding may be linked to the fact that the emotionally negative and threat information, especially anger and fear, is likely to capture the attention (Hansen and Hansen, 1988; Logan and Goetsch, 1993; Koster et al., 2004; Phelps et al., 2006; Bishop, 2008). In our study, face dominance was shown for anger when participants paid attention to one modality. This may simply reflect the fact that the emotion perception from faces is generally easier than that from voices (see **Table 1**). However, the modality dominance was not observed for anger when NI was given even though the same stimuli were presented. These findings might suggest that when NI was given, attention was automatically paid to one modality in which the threat information was expressed irrespective of whether it is the face or voice.

Table 3 | Modality dominance as shown by congruency, facilitation, and interference effects for each emotion category.

	Congruency effect	Facilitation effect	Interference effect
Anger	–	V > A	A > V
Disgust	V > A	V > A	–
Fear	A > V	A > V	–
Happiness	V > A	V > A	V > A
Sadness	–	–	–
Surprise	–	A > V	V > A

V > A means face dominance. A > V means voice dominance.

THE IMPACT OF FACE–VOICE CONGRUENCY ON EMOTION PERCEPTION

The accuracy in congruent trials was higher than it was in incongruent trials. These results are in line with previous studies showing that the emotion perception improves when more than one source of congruent information about the intended emotion is available (Massaro and Egan, 1996; de Gelder and Vroomen, 2000; Pell, 2002; Collignon et al., 2008). In order to directly examine the modality dominance for each emotion category in terms of the congruency effect, the size of congruency effects between modalities was compared. The results showed face dominance for disgust and happiness, and voice dominance for fear (**Figure 5**; **Table 3**). The modality dominance was not observed for anger, sadness, or surprise (**Figure 5**; **Table 3**).

The congruency effect includes two opposing effects. One is the facilitation effect, which occurred when the same emotion as the attended modality was expressed in the unattended modality. The other is the interference effect, which occurred when a different emotion from that presented in the attended modality was expressed in the unattended modality. By comparing the accuracy in the unisensory condition with the accuracy in multisensory congruent and incongruent conditions, we were able to examine more precisely the modality dominance suggested by the congruency effect in terms of the facilitation and interference effects.

Regarding the facilitation effect, with all emotions included, the facilitation effect occurred only for the voice (**Figure 4**). Thus, if the emotions in the attended and unattended modalities were congruent, the vocal emotion perception was enhanced by the emotion shown in the face, while the emotion perception of faces was not enhanced by emotions portrayed in voices. Again, these findings demonstrate that facial cues generally dominate vocal cues in the emotion perception. For each emotion category, face dominance was present for anger, disgust, and happiness, whereas voice dominance was present for fear and surprise (**Figure 6**; **Table 3**). No modality dominance was observed for sadness (**Figure 6**; **Table 3**).

Among all included emotions, the interference effect occurred both for facial and vocal cues (**Figure 4**). Thus, if the emotion expressed in the unattended modality was different from that in the attended modality, the emotion shown in the unattended modality interfered with the emotion perception. These findings confirmed that emotional information from the face and voice are subject to mandatory integration (Massaro and Egan, 1996; de Gelder et al., 1999; de Gelder and Vroomen, 2000; Kreifelts et al., 2007). For each emotion category, the results showed face dominance for happiness and surprise, and voice dominance for anger (**Figure 7**; **Table 3**). The modality dominance was not observed for disgust, fear, or sadness (**Figure 7**; **Table 3**).

As mentioned above, the modality dominances present in the congruency effect, facilitation effect, and interference effect did not coincide. The relations between these modality dominances will be discussed in the next section.

FACILITATION AND INTERFERENCE EFFECTS

As shown in the previous section and in **Table 2**, the modality dominance for each emotion was modulated not only by the instruction but also by congruency of emotions between modalities. Some of the modality dominance findings for each emotion category were in line with previous studies. Specifically, Paulmann and Pell (2011) showed face dominance for anger, disgust, and happiness. Collignon et al. (2008) also showed face dominance for disgust. Our data are consistent with these studies in that face dominance was suggested by the observed facilitation effects for anger, disgust, and happiness (see middle column of **Table 2**).

It is important that our results revealed that the modality dominance as reflected in the congruency effect can be classified into two patterns in terms of the facilitation effect and the interference effect. The first pattern is that the modality dominance suggested by the congruency effect occurs by adding up the modality dominances reflected by the facilitation effect and the interference effect. For instance, face dominance for happiness was suggested by both the facilitation effect and the interference effect. Face dominance in the congruency effect was shown for happiness by summing the modality dominances suggested by these effects. Another example is the modality dominance for sadness. The modality dominance for sadness was not suggested by both the facilitation and the interference effect. The modality dominance was not shown for sadness by summing these together. The second pattern was that the modality dominance suggested by the congruency effect did not occur by canceling out the facilitation and interference effects. For instance, for anger, face dominance was suggested by the facilitation effect and voice dominance was suggested by the interference effect. Consequently, the modality dominance in the congruency effect was not shown for anger by canceling out these opposing effects. Thus, the modality dominance for each emotion was elaborated by dividing the congruency effect into the facilitation and interference effects.

CONCLUSION AND ISSUES FOR FUTURE RESEARCH

In conclusion, our results revealed that instructions to pay attention to one of two modalities modulated the modality dominance for different emotion categories. In particular, the modality dominance for anger changed according to instructions. This finding was provided by comparing the emotion perception with and without instructions directly. It is important to give the instruction about the attention to set the participants' attitude and control the participants' understanding towards the task. This paradigm, the immediate cross-modal bias paradigm (Bertelson and de Gelder, 2004), has been widely used in the field of cross-modal perception. However, NI was given in the emotion perception in a natural environment. Therefore, the emotion perception when NI about the attention is given has still to be investigated.

Importantly, emotion congruency between the face and voice also modulated the modality dominance for each emotion category. That is, the emotion expressed in the unattended modality interacted with the emotion perception in a mandatory manner. Regarding the modality dominance, our results show that the modality dominance suggested by the congruency effect can be explained in terms of the facilitation effect and the interference effect. This methodology can provide additional perspective to behavioral and neuroscience study. By focusing on the facilitation and interference effects as well as the congruency effect, future research can examine the separable cognitive mechanisms and neural substrates of facilitation and interference effects.

We analyzed the accuracy on both congruent and incongruent trials when the instruction was given. The congruency effect was calculated by subtracting the accuracy in incongruent trials from that in congruent trials. Although it is possible to calculate the accuracy in congruent trials in NI, the accuracy in incongruent trials could not be calculated because it is impossible to define whether participants' responses were right or wrong in these trials. Therefore, we could not examine the congruency effect in NI. Instead, we analyzed the face responses and voice responses in incongruent trials when NI was given. In some trials, the reported emotion was neither the face nor voice. Although we eliminated these responses, the proportion of such responses was different among emotion categories. For example, the proportion of such responses for fear (64.8%) was higher than that for other emotions whereas the proportion of such responses for surprise (12.7%) was lower than that for other emotions. This difference might affect results on the face and voice responses particularly for those emotion characterized by a low hit rate (e.g., fear). Therefore, further research is required to examine such responses.

It remains to be investigated whether the emotional intensity, valence, and arousal of emotion expression affects the emotion perception. In this study, we did not manipulate these features to examine its effects on the emotion perception. If emotional intensity or arousal is strong in either modality, then that emotion will be perceived better from that modality. Therefore, it is possible that these features affected the emotion perception and modality dominance differently for each emotion category. Also the differences in the accuracies for each emotion category might have affected the results. Further study is necessary in which the emotional intensity, valence, and arousal of these stimuli are controlled.

It also remains to be investigated whether there are cultural differences with regard to the modality dominance for each emotion category. Regarding the modality dominance, Paulmann and Pell (2011) showed face dominance for fear, though our results demonstrated voice dominance. Other studies may provide

a potential answer to this issue. Tanaka et al. (2010) indicated that Japanese individuals are more attuned to voice processing than are Dutch individuals in the multisensory emotion perception. These findings suggest the need to examine cultural differences in the modality dominance for each emotion category.

ACKNOWLEDGMENTS

This work was supported by Strategic Information and Communications R&D Promotion Programme (SCOPE) (No. 102103011) from the Ministry of internal affairs and communications of Japan and JSPS KAKENHI Grant-in-Aid for Young Scientists (A) (No. 24680030).

REFERENCES

Bertelson, P., and de Gelder, B. (2004). "The psychology of multimodal perception," in *Crossmodal Space and Crossmodal Attentio,* eds C. Spence and J. Driver (Oxford: Oxford University Press), 151–177.

Bishop, S. J. (2008). Neural mechanisms underlying selective attention to threat. *Ann. N. Y. Acad. Sci.* 1129, 141–152. doi: 10.1196/annals.1417.016

Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., et al. (2008). Audio-visual integration of emotion expression. *Brain Res.* 25, 126–135. doi: 10.1016/j.brainres.2008.04.023

de Gelder, B., Böcker, K. B., Tuomainen, J., Hensen, M., and Vroomen, J. (1999). The combined perception of emotion from voice and face: early interaction revealed by human electric brain responses. *Neurosci. Lett.* 260, 133–136. doi: 10.1016/S0304-3940(98)00963-X

de Gelder, B., and Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cogn. Emot.* 14, 289–311. doi: 10.1080/026999300378824

Ekman, P., and Friesen, W. V. (1971). Constants across cultures in the face emotion. *J. Pers. Soc. Psychol.* 17, 124–129. doi: 10.1037/h0030377

Ekman, P., and Friesen, W. V. (1978). *Facial Action Coding System.* Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., Friesen, W. V., and Ellsworth, P. (1982). "What are the relative contributions of facial behavior and contextual information to the judgment of emotion?" in *Emotion in The Human Face,* ed. P. Ekman (Cambridge: Cambridge University Press), 111–127.

Ethofer, T., Anders, S., Erb, M., Droll, C., Royen, L., Saur, R., et al. (2006a). Impact of voice on emotional judgment of faces: an event-related fMRI study. *Hum. Brain Mapp.* 27, 707–714. doi: 10.1002/hbm.20212

Ethofer, T., Pourtois, G., and Wildgruber, D. (2006b). Investigating audiovisual integration of emotional signals in the human brain. *Prog. Brain Res.* 156, 345–361. doi: 10.1016/S0079-6123(06)56019-4

Föcker, J., Gondan, M., and Röder, B. (2011). Preattentive processing of audio-visual emotional signals. *Acta Psychol.* 137, 36–47. doi: 10.1016/j.actpsy.2011.02.004

Grossmann, T., Striano, T., and Friederici, A. D. (2006). Crossmodal integration of emotional information from face and voice in the infant brain. *Dev. Sci.* 9, 309–315. doi: 10.1111/j.1467-7687.2006.00494.x

Hansen, C. H., and Hansen, R. D. (1988). Finding the face in the crowd: an anger superiority effect. *J. Pers. Soc. Psychol.* 54, 917–924. doi: 10.1037/0022-3514.54.6.917

Hawk, S. T., van Kleef, G. A., Fischer, A. H., and van der Schalk, J. (2009). Worth a thousand words: absolute and relative decodability of nonlinguistic affect vocalizations. *Emotion* 9, 293–305. doi: 10.1037/a0015178

Johnstone, T., and Scherer, K. R. (2000). "Vocal communication of emotion," in *Handbook of Emotions,* eds M. Lewis and Haviland (New York, NY: Guilford Press), 220–235.

Koster, E. H., Crombez, G., Verschuere, B., and De Houwer, J. (2004). Selective attention to threat in the dot probe paradigm: differentiating vigilance and difficulty to disengage. *Behav. Res. Ther.* 42, 1183–1192. doi: 10.1016/j.brat.2003.08.001

Kreifelts, B., Ethofer, T., Grodd, W., Erb, M., and Wildgruber, D. (2007). Audiovisual integration of emotional signals in voice and face: an event-related fMRI study. *Neuroimage* 37, 1445–1456. doi: 10.1016/j.neuroimage.2007.06.020

Logan, A. C., and Goetsch, V. L. (1993). Attention to external threat cues in anxiety states. *Clin. Psychol. Rev.* 13, 541–559. doi: 10.1016/0272-7358(93)90045-N

Magnèe, M. J., de Gelder, B., van Engeland, H., and Kemmer, C. (2008). Atypical processing of fearful face–voice pairs in pervasive developmental disorder: an ERP study. *Clin. Neurophysiol.* 119, 2004–2010. doi: 10.1016/j.clinph.2008.05.005

Massaro, D. W., and Egan, P. B. (1996). Perceiving affect from the voice and the face. *Psychon. B. Rev.* 3, 215–221. doi: 10.3758/BF03212421

Paulmann, S., and Pell, M. D. (2011). Is there an advantage for recognizing for recognizing multi-modal emotional stimuli? *Motiv. Emot.* 35, 192–201. doi: 10.1007/s11031-011-9206-0

Pell, M. D. (2002). Evaluation of nonverbal emotion in face and voice: some preliminary findings on a new battery of tests. *Brain Cogn.* 48, 499–504.

Phelps, E. A., Ling, S., and Carrasco, M. (2006). Emotion facilitates perception and potentiates the perceptual benefits of attention. *Psychol. Sci.* 17, 292–299. doi: 10.1111/j.1467-9280.2006.01701.x

Pourtois, G., de Gelder, B., Bol, A., and Crommelinck, M. (2005). Perception of facial expressions and voices and of their combination in the human brain. *Cortex* 41, 49–59. doi: 10.1016/S0010-9452(08)70177-1

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression?: a review of the cross-cultural studies. *Psychol. Bull.* 115, 102–141. doi: 10.1037/0033-2909.115.1.102

Russell, J. A., Lewicka, M., and Niit, T. (1989). A cross-cultural study of a circumplex model of affect. *J. Pers. Soc. Psychol.* 57, 848–856. doi: 10.1037/0022-3514.57.5.848

Scherer, K. R. (2003). Vocal communication of emotion: a review of research paradigms. *Speech Commun.* 40, 227–256. doi: 10.1016/S0167-6393(02)00084-5

Talsma, D., Doty, T. J., and Woldorff, M. G. (2007). Selective attention and audiovisual integration: is attending to both modalities a prerequisite for early integration? *Cereb. Cortex* 17, 679–690. doi: 10.1093/cercor/bhk016

Talsma, D., Senkowski, D., Soto-Faraco, S., and Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends Cogn. Sci.* 14, 400–410. doi: 10.1016/j.tics.2010.06.008

Tanaka, A., Koizumi, A., Imai, H., Hiramatsu, S., Hiramoto, E., and de Gelder, B. (2010). I feel your voice: cultural differences in the multisensory perception of emotion. *Psychol. Sci.* 21, 1259–1262. doi: 10.1177/0956797610380698

Van den stock, J. B., Grèzes, J., and de Gelder, B. (2008). Human and animal sounds influence recognition of body language. *Brain Res.* 1242, 185–190. doi: 10.1016/j.brainres.2008.05.040

Vroomen, J., Driver, J., and de Gelder, B. (2001). Is cross-modal integration of emotional expressions independent of attentional resources? *Cog. Affect. Behav. Neurosci.* 1, 382–387. doi: 10.3758/CABN.1.4.382

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 October 2014; accepted: 07 January 2015; published online: 02 February 2015.

*Citation: Takagi S, Hiramatsu S, Tabei K and Tanaka A (2015) Multisensory perception of the six basic emotions is modulated by attentional instruction and unattended modality. Front. Integr. Neurosci. **9**:1. doi: 10.3389/fnint.2015.00001*

This article was submitted to the journal Frontiers in Integrative Neuroscience.

frontiers in
INTEGRATIVE NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 30 January 2015
doi: 10.3389/fnint.2015.00003

Audiovisual emotional processing and neurocognitive functioning in patients with depression

Sophie Doose-Grünefeld[1], Simon B. Eickhoff[1,2] and Veronika I. Müller[1,2]*

[1] Department of Clinical Neuroscience and Medical Psychology, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[2] Institute of Neuroscience and Medicine, Research Centre Jülich, Jülich, Germany

Edited by:
Ruth Adam, Ludwig-Maximilian-University, Germany

Reviewed by:
Hadas Okon-Singer, University of Haifa, Israel
Martin Klasen, RWTH Aachen University, Germany

***Correspondence:**
Veronika I. Müller, Institute of Neuroscience and Medicine, Research Centre Jülich, Wilhelm-Johnen-Straße, D-52428 Jülich, Germany
e-mail: v.mueller@fz-juelich.de

Alterations in the processing of emotional stimuli (e.g., facial expressions, prosody, music) have repeatedly been reported in patients with major depression. Such impairments may result from the likewise prevalent executive deficits in these patients. However, studies investigating this relationship are rare. Moreover, most studies to date have only assessed impairments in unimodal emotional processing, whereas in real life, emotions are primarily conveyed through more than just one sensory channel. The current study therefore aimed at investigating multi-modal emotional processing in patients with depression and to assess the relationship between emotional and neurocognitive impairments. Fourty one patients suffering from major depression and 41 never-depressed healthy controls participated in an audiovisual (faces-sounds) emotional integration paradigm as well as a neurocognitive test battery. Our results showed that depressed patients were specifically impaired in the processing of positive auditory stimuli as they rated faces significantly more fearful when presented with happy than with neutral sounds. Such an effect was absent in controls. Findings in emotional processing in patients did not correlate with Beck's depression inventory score. Furthermore, neurocognitive findings revealed significant group differences for two of the tests. The effects found in audiovisual emotional processing, however, did not correlate with performance in the neurocognitive tests. In summary, our results underline the diversity of impairments going along with depression and indicate that deficits found for unimodal emotional processing cannot trivially be generalized to deficits in a multi-modal setting. The mechanisms of impairments therefore might be far more complex than previously thought. Our findings furthermore contradict the assumption that emotional processing deficits in major depression are associated with impaired attention or inhibitory functioning.

Keywords: depression, emotional processing, neurocognitive functioning, audiovisual, executive deficits, congruent, symptom severity

INTRODUCTION

Major depression is a psychiatric disorder that is thought to represent one of the leading causes of disability worldwide (Ferrari et al., 2013). The disorder goes along with a range of symptoms including on the one hand emotional and social problems like low mood and loss of self-esteem as well as on the other hand cognitive impairments like poor concentration and indecisiveness (WHO, 2010). With regard to the former symptoms several theories have been postulated in order to gain a better understanding of the origins of these social and emotional problems in depression. The most influential theory suggests a negative bias for emotional but also neutral material, manifesting, for example, as more negative ratings of facial expressions or selective attention on negative stimuli. This has been supported by numerous studies (Gur et al., 1992; Bouhuys et al., 1999; Leppänen et al., 2004) reporting either a general negative bias or a bias specifically for neutral or ambiguous stimuli. However, most of these studies primarily investigated photographs or even schematic paintings of faces depicting emotions or neutrality (for a review, see Bourke et al., 2010). More recent studies now included on the one hand facial stimuli with varying intensity levels (Schaefer et al., 2010), and on the other hand other kinds of stimuli addressing different emotional channels such as voices (Schlipf et al., 2013) or music (Naranjo et al., 2011). Among these, fewer study outcomes indicated a clear negative bias in depression, but also an absence of a "healthy" positive bias. That is, while healthy non-depressed controls tend to interpret stimuli as positive, patients with depression do not. Joormann and Gotlib (2006) for example showed that when identifying emotions from faces, depressed individuals compared to controls needed significantly higher emotional intensity in order to correctly identify happy but not sad facial expressions. In addition, Loi et al. (2013) also only found labeling problems for happy body language depicted by photographs of body postures as well as frozen movie scenes and short clips of "Point-Light Walkers." Although Schlipf et al. (2013) report a negative bias in reference to judgments of neutral semantics, patients also rated positive semantics and positive prosody as less positive than healthy controls, thus also indicating an absence of a positive bias in the patient group.

To date, most previous studies have investigated emotional processing using only unimodal stimuli. In daily life, though, emotions are hardly ever conveyed through just one sensory modality but rather in a multimodal fashion, i.e., seeing a happy facial expression and concurrently hearing the sound of laughter. Thus, it is in question if findings from unimodal emotional processing reflect deficits in real life. However, there are very few studies, which used stimuli from more than just one modality. Schneider et al. (2012) for example presented short video clips of actors conveying emotions via facial expressions, semantics and prosody and found patients to be impaired in recognizing emotions, but did not find an overall negative bias. In addition, in an earlier study (Müller et al., 2014), we investigated audiovisual emotional integration in major depression using functional magnetic resonance imaging. There we demonstrated that impairments in emotional processing in patients with depression seem to be far more complex than a simple bias as we found patients to be impaired in the inhibition of auditory stimuli presented with emotionally congruent facial expressions. However, owing to the relatively low sample size in that imaging study, there were only tendencies toward a behavioral effect. Therefore the current study aims at complementing the previous study by investigating multi-modal emotional processing (on the behavioral level) in an extended sample of patients with depression.

Furthermore, besides the postulation of emotional biases, there is also an ongoing discussion if alterations in emotional processing result from the likewise prevalent executive deficits in these patients. Executive functions, however, rather ill defined, include amongst others inhibition, working memory as well as cognitive flexibility (Diamond, 2013). Furthermore executive skills as the basis for everyday life functioning also include cognitive domains like attention (as a precondition for inhibitory functions; DeBattista, 2005). Impairments in the mentioned cognitive functions have been shown to be present in depression. Airaksinen et al. (2004), for example, reported deficits in cognitive flexibility in individuals with depression, and Rose and Ebmeier (2006) described impairments in working memory. Furthermore, attentional deficits have been found to be present even in remitted states of depression (Paelecke-Habermann et al., 2005). Interestingly, Hoffstaedter et al. (2012) reported significantly worse performance on verbal vocabulary testing in patients with depression compared to controls, and they related these impairments to memory deficits. In addition, they found group differences in attention, cognitive flexibility, (visuo-) motor coordination, short-term and working memory, but not for basic motor speed. Overall, psychomotor retardation, however, has been described as a core feature of depression (Sobin and Sackeim, 1997) and Hoffstaedter et al. (2012) were able to show that patients were impaired in specific cognitive aspects of psychomotor functioning.

Regarding the treatment of cognitive aspects of depression, Owens et al. (2013) were able to show that working memory training in dysphoric individuals can improve inhibition of irrelevant information and thus lead to increased working memory capacity. Since poor inhibitory control has been shown to be related to problems in the interpretation of emotional information in depression (Joormann, 2004; Goeleven et al., 2006), cognitive functioning seems to be a valuable starting point in the therapy of depressive symptoms. In line with this view, Marazziti et al. (2010) hypothesized that decreased cognitive flexibility in patients with depression possibly prevents those individuals from being able to cope with life events which then leads to constant low mood due to increased stress exposition. All in all, there is much evidence that depression goes along with impaired cognitive performance, and symptom severity seems to be related especially to decreased episodic memory, executive functions, and processing speed (McDermott and Ebmeier, 2009). Thus, it seems reasonable that deficits in emotional and cognitive processing might be closely interrelated. However, studies in depression investigating the relationship between deficits in emotional perception of faces and sounds and impairments in cognitive functions, measured independently from emotional processing, are rather rare. Taken together, even though findings clearly indicate an impairment in emotional processing in major depression, authors do not agree on whether depression is associated with a general bias (present negative or absent positive bias), if it is possibly a result of cognitive deficits, or both.

The aims of the current study were therefore to first assess uni- and multi-modal emotional processing in patients with depression. Second, we explored executive functioning and related cognitive domains to be able to investigate the potential relationship of emotional and cognitive deficits. We chose a selection of different neurocognitive tests where patients with depression have been reported to be impaired (Hoffstaedter et al., 2012), i.e., that challenged the participants' attention, cognitive flexibility, (visuo-) motor speed and coordination, short-term as well as working memory, and verbal vocabulary.

Based on the previous literature regarding emotional processing in patients with depression, we hypothesized that for unimodal conditions, we would find a mood-congruent emotional bias in patients with depression (negative or absent positive), whereas in the multi-modal setting, impairments would probably appear in a manner different from a generalized bias as already described by Müller et al. (2014). Additionally we expected patients to perform worse than healthy controls on neurocognitive tests, especially those regarding cognitive flexibility and attention, and that these deficits would be associated with impairments in emotional processing, thus pointing in the direction that emotional problems in depression are related to likewise prevalent cognitive deficits.

MATERIALS AND METHODS

SUBJECTS

The current study is based on a previous study, which tested audiovisual emotional processing in depression by using fMRI (Müller et al., 2014). We now focus on the behavioral effects in an expanded sample of patients and healthy controls.

In total, 41 patients diagnosed with major depression (19 females, 22 males) and 41 healthy controls (19 females, 22 males) were now included. Data from 44 of these 82 subjects originated from the previous fMRI study, while the remaining 38 subjects conducted the paradigm outside the scanner. Both patients and controls gave informed consent into the study, which was approved by the ethics committee of the School of Medicine of

the RWTH Aachen University. In addition to gender matching, the two groups did not differ in their age or years of education (age: $T_{80} = -0.47$, $p = 0.64$; EDU: $T_{80} = 0.08$, $p = 0.80$; see **Table 1** for means). All subjects were right-handed according to the Edinburgh Handedness Questionnaire (Oldfield, 1971) and had normal or corrected-to-normal vision. Patients were recruited from the inpatient and outpatient units of the Department of Psychiatry, Psychotherapy and Psychosomatics, RWTH University Hospital. They were diagnosed by their treating psychiatrist with a depressive episode or a recurrent depressive disorder according to the criteria of the ICD-10 (WHO, 2010; see **Table 2** for the patients' clinical profiles). To confirm their diagnosis and to screen for possible psychiatric co-morbidities, the structured clinical interview for DSM-IV (SCID; Wittchen et al., 1997) was conducted. Furthermore, the Beck Depression Inventory (BDI-II; Hautzinger et al., 2006) as well as the Hamilton depression scale (Hamilton, 1960) were used to quantify depression-related symptoms and thus the illness severity. We only included patients without co-morbidities, i.e., without an indication of any psychiatric or neurological disease other than major depression, and without any kind of addiction or substance abuse in at least 6 months. Control subjects did not report any history of psychiatric or neurological disorders as well as any addiction in their past. Sub-clinical depressive symptoms in the control group were also assessed with the BDI-II (Hautzinger et al., 2006).

STIMULI

For a detailed description of stimulus material and procedure see Müller et al. (2011). In brief, the visual stimuli were color pictures obtained from the FEBA inventory (Gur et al., 2002) showing either a happy, neutral, or fearful facial expression. In total, 30 different faces were used, with five different female and five different male actors, each showing all three (happy/neutral/fearful) expressions. As pre-tests revealed that happy and fearful facial expressions were too clear in their emotionality to allow any contextual framing effects (Müller et al., 2011), they were made more ambiguous by merging them with the neutral mouths of the same actors. As auditory stimuli 10 laughs, 10 yawns, and 10 screams, each produced by five females and five males and lasting for 1500 ms, were used. Blurred versions of the neutral faces served as masks during the initial 1000 ms of sound presentation before the target faces were shown.

Table 1 | Demographic and clinical profile of patients and controls.

	Patients	Controls
Gender (female/male)	19/22	19/22
Age (SD)	36,49 (10,87)	37,61 (10,79)
EDU (SD)	12,41 (3,32)	12,59 (2,65)
BDI (SD)	23,07 (11,89)	1,90 (3,07)
HAMD (SD)	11,54 (5,98)	–

SD, standard deviation; EDU, years of education; BDI, Beck's depression inventory score; HAMD, Hamilton depression scale score.

AUDIOVISUAL PARADIGM

In total, 180 stimulus pairs, each consisting of a visual and an auditory stimulus, were used. Every face condition (happy/neutral/fearful) was paired with every sound condition (happy/neutral/fearful) resulting in a 3 × 3 design with nine different audiovisual conditions (fearful/scream, fearful/yawn, fearful/laugh, neutral/scream, neutral/yawn, neutral/laugh, happy/scream, happy/yawn, happy/laugh) and 20 individual audiovisual stimulus pairs per condition. The pairs were matched pseudo-randomly with regard to gender so that a female (male) face was always paired with a female (male) sound.

Figure 1 illustrates the experimental procedure. Every trial started with the presentation of a sound in combination with a blurred neutral face. After 1000 ms, the screen switched to a non-blurred picture of an emotional or neutral face (the target face), which was presented for another 500 ms with the ongoing sound. Participants had the task to ignore the sound and rate the facial expression on an eight-point rating scale (not including a neutral option and ranging from extremely fearful to extremely happy) as fast and as accurate as possible by pressing one of eight buttons on a response pad. To avoid expectation-led effects on the outcome of the experiment, the participants were told that the study focuses on attention processes. Stimuli were presented with the software Presentation 14.2 (http://www.neurobs.com/).

UNIMODAL VALENCE AND AROUSAL RATINGS OF FACES AND SOUNDS

After the audiovisual paradigm, patients and controls rated emotional valence and arousal of all faces and sounds used in the audiovisual paradigm individually. For that, two separate runs were conducted, one for the unimodal facial expression rating and one for the unimodal sound rating. Both valence and arousal of the stimuli had to be rated on a 9-point rating scale, i.e., including a neutral option and ranging from very fearful to very happy/not at all arousing to very arousing.

NEUROPSYCHOLOGICAL TESTING

To measure neurocognitive and psychomotor skills of patients and controls, diverse neurocognitive tests were conducted.

Visual attention/visuomotor speed

Trail making tests (versions A and B). The two trail making tests [TMT-A and TMT-B; Army Individual Test Battery (AITB), 1944] were used to assess attention and visuomotor speed. Participants had the task to accurately connect as fast as possible (a) a consecutive sequence of numbers from 1 to 25 (TMT-A) and (b) a sequence of numbers from 1 to 13 mixed with the first 12 letters of the alphabet (TMT-B), respectively. The TMT-B, where numbers and letters had to be connected alternately, also assessed cognitive flexibility. Measurement for the test results was the time it took the participants to accomplish the task. Additionally we calculated difference scores between performance in TMT-A and TMT-B (TMT-Diff).

Motor speed/coordination

Finger tapping test. The finger tapping test was used to determine basic motor speed. Participants were asked to tap on the table as fast as possible for 10 s with their left or right index finger. This procedure was conducted three times for both (left and right)

Table 2 | Patients' demographic data and clinical profile.

Gender	Age	Diagnosis	Medication	Age of onset	Duration of illness
Female	57	F 32.1	Reboxetine, Citalopram	33	24
Female	54	F 33.1	Venlafaxine	50	5
Female	33	F 33.2	Venlafaxine	30	3
Female	34	F 32.1	Citalopram	24	10
Female	49	F 33.2	Venlafaxine, Quetiapine, Lithium carbonate	25	24
Female	33	F 33.3	Venlafaxine, Quetiapine	12	21
Female	33	F 32.1	Fluoxetine	33	1
Female	51	F 33.1	Agomelatine, Duloxetine	46	5
Female	27	F 32.2	Venlafaxine, Mirtazapine	21	6
Female	22	F 32.1	Venlafaxine	13	9
Female	26	F 32.1	Duloxetine	23	3
Female	31	F 33.1	Citalopram	11	20
Female	50	F 33.2	Duloxetine, Quetiapine	47	3
Female	38	F 33.1	Reboxetine	34	4
Female	30	F 32.1	"none"	30	1
Female	51	F 33.1	Escitalopram	31	20
Female	41	F 32.1	Venlafaxine	33	8
Female	29	F 32.2	Citalopram	29	1
Female	42	F 33.1	Mirtazapine	26	16
Male	45	F 32.1	Venlafaxine, Trimipramine	43	2
Male	55	F 33.2	Mirtazapine, Quetiapine, Duloxetine, Pipamperone	49	6
Male	37	F 33.1	Venlafaxine	35	2
Male	46	F 32.1	Venlafaxine, Opipramol	40	6
Male	52	F 32.1	Venlafaxine	44	8
Male	43	F 32.1	Mirtazapine	19	24
Male	27	F 32.1	Reboxetine	24	3
Male	25	F 32.1	Sertraline	21	4
Male	40	F 32.1	Venlafaxine	38	2
Male	30	F 33.1	Citalopram, Lithium	21	9
Male	38	F 32.1	Venlafaxine	35	3
Male	29	F 32.1	Lithium carbonate	24	5
Male	19	F 32.1	Citalopram	16	3
Male	28	F 33.2	Venlafaxine	27	1
Male	30	F 33.1	Opipramol, Sertraline, Mirtazapine	29	1
Male	53	F 33.2	Venlafaxine, Mirtazapine	44	9
Male	34	F 33.2	Venlafaxine, Quetiapine	25	9
Male	22	F 32.1	Venlafaxine	18	4
Male	41	F 33.1	Venlafaxine	36	5
Male	29	F 33.1	Bupropion, Quetiapine	20	9
Male	18	F 32.2	Remergil, Venlafaxine	17	1
Male	24	F 33.1	Citalopram	16	8

Data for age of onset of depression and duration of illness is taken from self-reported information by the patients. Age, age of onset, and duration of illness are specified in years.

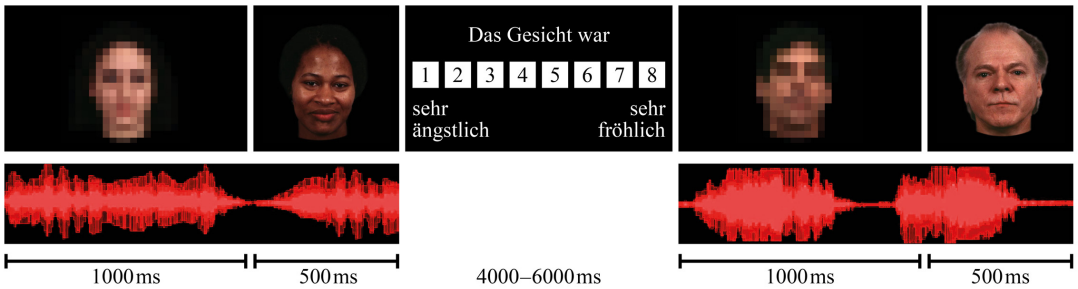


FIGURE 1 | Experimental procedure of the audiovisual paradigm. Every trial started with a sound in combination with a blurred neutral face, after 1000 ms the screen switched to the target face, which was presented with the ongoing sound for another 500 ms.

index fingers with short pauses in between to increase reliability but avoid muscular fatigue. For all six tapping runs, the number of taps was counted and the mean of all runs from both hands calculated (Halstead, 1947; Behrwind et al., 2011).

Pointing movements. To assess basic motor coordination, the participants were asked to perform pointing movements with their left or right index finger (Defer et al., 1999). On the table in front of the subject, a 30 cm long horizontal line was marked and the task was to point on the two ends 10 times alternately as fast and as precise as possible. The time the participants needed to accomplish the task was measured. Again, this test was conducted three times for left and right index finger and the mean time of all six runs from both hands was calculated.

Crystalline intelligence

Multiple choice vocabulary intelligence test (MWT). The multiple choice vocabulary intelligence test, version B (Lehrl, 1989), measured the participants' crystalline intelligence. There were 37 rows of five words from which the participant had to choose the only actual word by ruling out four pseudo-words. The number of correctly detected words provided the test result.

Short-term and working memory

Digit span subtest of the Wechsler Adult Intelligence Scale (Tewes, 1991). This test, in which verbally presented digit spans had to be repeated by the participant, consisted of two parts. In the first, used to measure short-term verbal memory and attention, the participant had to repeat the digit span in the same order as it was read to him (DS-F). For part two which assessed manipulation within working memory, the participant had to repeat the numbers backward (DS-B). For both test parts the number of correctly reproduced digit sequences was used as the test result.

Due to technical problems, the neuropsychological test results of one control subject could not be used and one patient did not take part in the TMT-A and TMT-B, while another one did not take part in the TAP10s and TAP10x30. The MWT was only conducted with native German speaking participants, and thus results from one of the patients and two control subjects are not available.

STATISTICAL ANALYSIS

We analyzed our data using SPSS Statistics 21 (IBM). Most data except for those of the multimodal emotional processing task were not normally distributed and tests were individually chosen, adapted to the particular conditions. The threshold for significance was set at $p < 0.05$, Bonferroni-corrected for multiple comparisons if appropriate. Data from unimodal valence and arousal ratings of sounds and faces were analyzed with Mann–Whitney-*U* tests for group comparison and Wilcoxon-tests for comparisons between conditions (corrected for multiple comparisons). Data from the audiovisual paradigm were analyzed calculating a MANOVA (due to violation of sphericity) with the factors face, sound and group and the dependent variable emotional valence rating of faces. Significant main effects and interactions were furthermore analyzed with *post hoc t*-tests (corrected for multiple comparisons). To test for possible incongruence effects, two additional ANOVAS (for happy and fearful faces) with the factors congruence and group and the dependent variable emotional valence rating of faces were calculated. Significant findings were again further analyzed with *post hoc t*-tests (corrected for multiple comparisons). Furthermore, we calculated Spearman-rank-correlations between findings in emotional processing and Beck's depression inventory score (BDI-scores) for the patient group.

Group differences in neurocognitive performance were tested with Mann–Whitney-*U* tests. Within the domains visual attention/visuomotor speed, motor speed/coordination and short-term/working memory, results were corrected for multiple comparisons (Bonferroni).

To investigate the relationship of emotional processing and neurocognitive functioning, we again calculated Spearman-rank-correlations.

RESULTS

UNIMODAL VALENCE AND AROUSAL RATING OF SOUNDS AND FACES

In total 12 Mann–Whitney-*U* tests were calculated (all results are Bonferroni-corrected for multiple comparisons). No significant group differences were found in the unimodal ratings of valence and arousal of faces and sounds (see **Table 3**).

Comparisons between conditions across groups demonstrated that happy and fearful faces were rated as more arousing than neutral ones and that all types of faces as well as all types of sounds

Table 3 | Group comparisons of unimodal valence and arousal ratings of sounds and faces – findings of Mann–Whitney-*U* tests.

		Arousal		Valence	
		U	***p***	***U***	***p***
Sounds	Positive	−2.018	0.044	−0.046	0.963
	Neutral	−0.191	0.848	−0.254	0.799
	Negative	−1.049	0.294	−0.599	0.549
Faces	Positive	−1.499	0.134	−0.803	0.422
	Neutral	−0.772	0.440	−1.545	0.122
	Negative	−0.603	0.546	−0.390	0.697

No significant group differences were found (Bonferroni-corrected cut-off $p < 0.017$).

Table 4 | Mean values and SD for emotional valence and arousal ratings of faces and sounds.

	Happy (SD)	Neutral (SD)	Fearful (SD)
Patients			
Emotional valence rating of faces	**7.049** (0.672)	**4.661** (0.536)	**3.437** (0.830)
Emotional valence rating of sounds	**7.742** (0.764)	**5.117** (0.483)	**1.649** (0.793)
Arousal rating of faces	**4.027** (1.521)	**3.698** (1.513)	**4.789** (1.635)
Arousal rating of sounds	**3.815** (1.640)	**3.129** (1.597)	**6.366** (2.095)
Controls			
Emotional valence rating of faces	**6.917** (0.816)	**4.481** (0.583)	**3.527** (0.658)
Emotional valence rating of sounds	**7.707** (0.847)	**5.117** (0.785)	**1.576** (0.806)
Arousal rating of faces	**4.539** (1.865)	**3.963** (1.507)	**5.034** (1.417)
Arousal rating of sounds	**4.642** (2.022)	**3.061** (1.646)	**6.868** (1.837)

There are no significant differences between patients and controls (Bonferroni-corrected cut-off $p < 0.017$).

differed from each other in their emotional valence rating (for mean values and SD see **Table 4**, all $p < 0.017$). In particular, fearful stimuli got the lowest ratings, followed by neutral faces and sounds, while the most positive ratings were given for happy stimuli.

AUDIOVISUAL PARADIGM

Valence ratings of faces in audiovisual setting

A MANOVA with the factors face (happy/neutral/fearful), sound (happy/neutral/fearful) and group (controls/patients), and the dependent variable emotional valence rating was calculated. Multivariate testing was chosen due to violation of sphericity of sounds [$\chi^2(2) = 6.267$, $p = 0.044$] and faces [$\chi^2(2) = 57.726$, $p < 0.001$]. Assumptions of equality of error variances and equality of covariance matrices were met, indicated by non-significant Box's M test and Levene's tests. Results revealed significant main effects of sound ($F_{2,79} = 12.89$, $p < 0.001$, effect size: partial $\eta^2 = 0.25$) and face ($F_{2,79} = 401.81$, $p < 0.001$, effect size: partial $\eta^2 = 0.91$), but no main effect of group ($F_{1,80} = 0.23$, $p = 0.63$). Furthermore an interaction between sound × group could be found ($F_{2,79} = 4.20$, $p = 0.018$, effect size: partial $\eta^2 = 0.10$), but no interactions of face × group ($F_{2,79} = 1.91$, $p = 0.155$), sound × face ($F_{4,77} = 2.40$, $p = 0.058$) or sound × face × group ($F_{4,77} = 0.96$, $p = 0.437$). To further analyze the significant interaction between sound × group *post hoc* t-tests were calculated (all results are Bonferroni-corrected for multiple comparisons). We found no significant global differences (across all types of concurrently presented faces) between groups, neither in the happy, nor neutral, nor fearful sound condition (faces paired with happy sounds: $T_{80} = -0.065$, $p = 0.948$; faces paired with neutral sounds: $T_{80} = 0.861$, $p = 0.392$; faces paired with fearful sounds: $T_{80} = 0.602$, $p = 0.549$). However, the *post hoc* tests demonstrated that while in healthy controls there was no difference in the rating of faces when concurrently hearing happy compared to neutral sounds ($T_{40} = -0.08$, $p = 0.939$), patients with depression rated faces as more fearful when presented with happy compared to neutral sounds ($T_{40} = 4.61$, $p < 0.001$). Furthermore, in controls the ratings of faces differed between the happy and fearful sound condition ($T_{40} = -3.36$, $p = 0.002$), whereas there was no difference between these two conditions in the patient group ($T_{40} = -1.28$, $p = 0.209$). Additionally, in both groups presentation of fearful compared to neutral sounds led to more fearful ratings of faces (controls: $T_{40} = -3.09$, $p = 0.004$; patients: $T_{40} = -4.22$, $p < 0.001$). These differences in the impact of sounds on the emotional valence ratings of faces between patients and controls are illustrated in **Figure 2**. For a more detailed analysis of the interaction, we further calculated difference scores in emotional valence ratings of faces between sound conditions, i.e., happy–neutral, happy–fearful, fearful–neutral. These difference scores were then compared between patients and controls using independent samples t-tests. The t-tests revealed a significant difference between patients and controls in the difference scores between the happy and neutral sound condition ($T_{80} = -2.723$, $p = 0.008$, effect size: Cohen's $d = 0.60$). For patients, the mean value of the difference scores happy minus neutral was −0.068, indicating more fearful ratings of faces when presented with happy compared to neutral sounds. In contrast, the mean value of the respective difference scores in controls was 0.002, indicating only slightly happier ratings of faces when presented with happy sounds compared to neutral sounds. In contrast, neither the difference score of the happy minus fearful (mean values: patients 0.024; controls 0.077) nor the one of the fearful minus neutral sound condition (mean values: patients −0.092; controls −0.076) revealed any significant group differences (happy–fearful: $T_{80} = -1.818$, $p = 0.073$; fearful–neutral: $T_{80} = -0.496$, $p = 0.621$).

In summary, the main difference in face ratings between patients and controls was that patients rated faces in combination with happy sounds as more fearful than in combination with neutral sounds while controls did not.

Spearman-rank-correlations of difference scores between the happy and neutral sound conditions with BDI-scores were calculated for the patient group. Results did not reveal any significant associations ($r = -0.087$, $p = 0.590$).

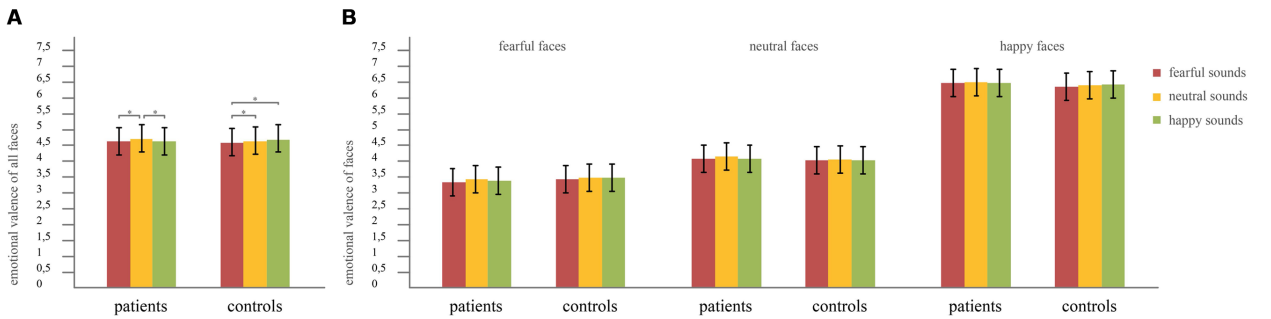


FIGURE 2 | Mean ratings of faces in the audiovisual setting. Error bars indicate the SD of mean. Significant interaction group × sound. **(A)** Mean ratings for each sound condition across all face conditions. Asterisks indicate significant differences revealed by *post hoc* tests at $p < 0.05$, Bonferroni-corrected for multiple comparisons. Patients rated faces significantly more fearful when presented with fearful or happy sounds than when presented with neutral sounds, while controls rated faces significantly more fearful when presented with fearful sounds than when presented with neutral sounds, but ratings did not differ between the neutral and the happy sound conditions. **(B)** Mean values of the ratings of fearful, neutral, and happy faces separately. No interaction sound × face × group was found in the MANOVA, therefore no further *post hoc* tests were calculated.

Incongruence effect

To investigate the interaction sound × group described above, we analyzed the audiovisual data more in detail. That is, we focused on the impact of emotional congruence/incongruence between concurrently presented sounds and faces on the valence ratings separately for happy and fearful faces. Therefore, we calculated two ANOVAs, one for happy faces and one for fearful faces. Both contained the factors congruence (congruent sound/incongruent sound) and group (controls/patients). In **Figure 3** the results are illustrated. For happy faces, we found a significant main effect of congruence ($F_{1,80} = 5.40$, $p = 0.023$, effect size: partial $\eta^2 = 0.06$) and an interaction congruence × group ($F_{1,80} = 6.15$, $p = 0.015$, effect size: partial $\eta^2 = 0.07$), but no main effect of group ($F_{1,80} = 1.59$, $p = 0.211$). In contrast, the ANOVA of fearful faces only revealed a significant main effect of congruence ($F_{1,80} = 11.81$, $p = 0.001$, effect size: partial $\eta^2 = 0.13$) but neither a significant interaction congruence × group ($F_{1,80} = 0.04$, $p = 0.834$) nor a main effect of group ($F_{1,80} = 1.23$, $p = 0.271$). To further analyze the significant interaction between congruence × group for happy faces, we calculated *post hoc* *t*-tests (all results are Bonferroni-corrected for multiple comparisons). Those did not reveal any significant differences in emotional valence rating of faces between patients and controls, neither for the congruent nor for the incongruent condition (congruent: $T_{80} = 0.76$, $p = 0.452$; incongruent: $T_{80} = 1.76$, $p = 0.083$; see **Figure 3**). However, patients did not rate happy faces significantly different when paired with happy sounds (congruent condition) compared to fearful sounds (incongruent condition, $T_{40} = -0.11$, $p = 0.910$). In contrast, controls rated happy faces significantly happier when paired with happy sounds compared to fearful sounds ($T_{40} = 3.29$, $p = 0.002$; mean value happy sound: 6.35/mean value fearful sound: 6.24; see **Figure 3**).

NEUROCOGNITIVE TESTS

Group comparisons

Mann–Whitney-*U* tests were calculated to compare the scores of the neurocognitive tests between the two groups (controls/patients). All results are Bonferroni-corrected for the

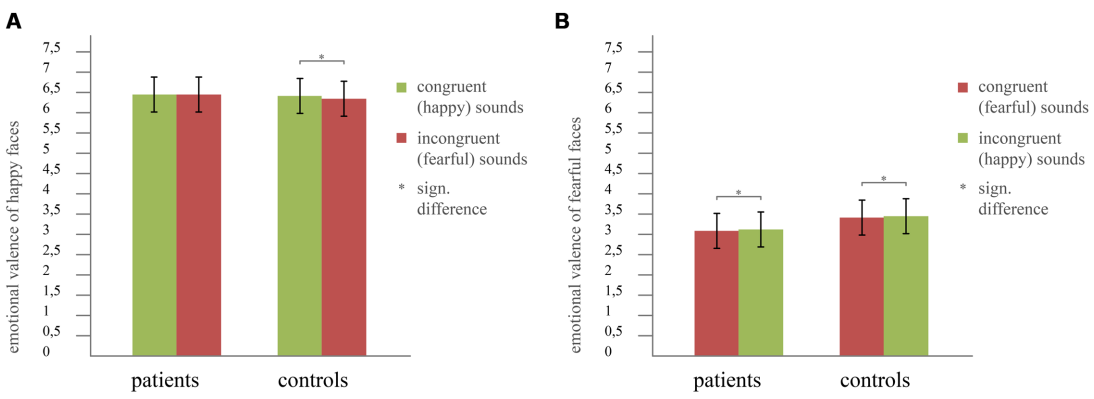


FIGURE 3 | Incongruence effect for (A) happy faces, revealing a significant interaction congruence × group and (B) fearful faces showing no group differences. Error bars indicate the SD of mean. Asterisks indicate significant differences revealed by *post hoc* tests at $p < 0.05$, Bonferroni-corrected for multiple comparisons.

number of tests within each test category. Significant differences between controls and patients were found for the TMT-A and TMT-B (TMT-A: $U = -2.723$, $p = 0.006$, effect size: $r = 0.30$; TMT-B: $U = -2.492$, $p = 0.013$, effect size: $r = 0.29$). In contrast, the other tests did not reveal any significant differences between groups (TMT-Diff: $U = -0.385$, $p = 0.700$; TAP10x30: $U = -1.304$, $p = 0.192$; TAP10s: $U = -1.411$, $p = 0.158$; DS-F:$U = -1.215$, $p = 0.224$; DS-B: $U = -0.739$, $p = 0.460$; MWT: $U = -1.542$, $p = 0.123$). **Table 5** shows the mean values and SD of all tests in patients and controls.

Correlations of findings in audiovisual emotional processing and neurocognitive functioning

To investigate the relationship between findings in audiovisual emotional processing and neurocognitive functioning in the patient group, Spearman-rank-correlations were calculated between neurocognitive test scores and the difference scores of the happy and neutral sound condition. All results are Bonferroni-corrected for the number of tests within each test category. No significant correlations between the difference scores and neurocognitive test scores were found (difference score between happy-neutral with: TMT-A: $r = 0.326$, $p = 0.040$; TMT-B: $r = 0.235$, $p = 0.145$; TMT-Diff: $r = -0.096$, $p = 0.554$; TAP10s: $r = -0.30$, $p = 0.856$; TAP10x30: $r = -0.263$, $p = 0.101$; DS-F: $r = 0.076$, $p = 0.637$; DS-B: $r = -0.015$, $p = 0.928$; MWT: $r = 0.044$, $p = 0.787$).

DISCUSSION

The aim of the current study was to investigate emotional processing in a more naturalistic setting by adding an auditory context to visual stimuli, and furthermore the relationship between emotional and neurocognitive deficits in patients with depression. For that purpose, an audiovisual paradigm was conducted, where happy, fearful, and neutral faces had to be rated whilst ignoring concurrently presented emotional or neutral sounds. In addition, patients and controls completed diverse neurocognitive tests challenging attention, working memory, (visuo-)motor speed, coordination, and crystalline intelligence. Results in audiovisual emotional processing revealed an aberrant integration of happy sounds in major depression as patients rated faces significantly more fearful when combined with happy as compared to neutral sounds. Conversely, controls showed no significant differences between these two conditions. Findings in audiovisual emotional processing in patients did not correlate significantly with depressive symptom severity as indicated by BDI-scores. We only found significant group differences for two of the neurocognitive tests. Despite the fact that patients with depression were impaired both in audiovisual emotional processing and neurocognitive performance, we found no significant correlations between these two fields.

Table 5 | Mean values and SD of the neuropsychological tests of all participants.

	Patients: **mean value (SD)**	***Controls:*** **mean value (SD)**
Visual attention/visuomotor speed		
TMT-A**	26.38 (15.61)	19.55 (5.68)
TMT-B**	47.32 (19.94)	37.54 (11.48)
TMT-Diff	20.94 (14.47)	18.00 (9.43)
Motor speed/coordination		
TAP10s	101.66 (15.97)	107.01 (11.65)
TAP10x30	16.76 (4.01)	15.63 (4.66)
Short-term and working memory		
DS-F	7.68 (1.93)	8.35 (1.98)
DS-B	6.80 (1.79)	7.18 (1.82)
Crystalline intelligence		
MWT	27.68 (4.45)	29.50 (3.80)

***significant ($p < 0.05$, Bonferroni-corrected).*

UNIMODAL VALENCE AND AROUSAL RATINGS OF FACES AND SOUNDS

Several studies investigating unimodal emotional processing in patients with depression reported that patients showed a general bias toward more negative ratings of emotional and neutral stimuli such as faces, prosody, and music (Leppänen et al., 2004; Douglas and Porter, 2010; Naranjo et al., 2011) as well as an attentional bias toward negative emotional material (Leung et al., 2009; Milders et al., 2010). In the current study, however, we found no group differences with regard to valence and arousal ratings, neither for faces, nor for sounds. Apart from a general negative bias, these findings also contradict previous studies in depression, which reported respective group differences for unimodal stimuli (Csukly et al., 2009; Schlipf et al., 2013). They are rather in line with findings showing no group differences in unimodal emotional processing (Kan et al., 2004; Bourke et al., 2012; Müller et al., 2014). Müller et al. (2014) already suggested that such discrepancies might arise from differences in methodology like varying stimulus presentation times (Surguladze et al., 2004) or different emotions (fear vs. sadness) used as negative stimuli (Hu et al., 2012).

Thus, we would argue that the current findings suggest that, contradictory to our hypothesis, patients did neither show a general negative bias nor an emotional blunting with regard to unimodal ratings of emotional and neutral stimuli. This result, however, has to be taken with caution since differences in methodology (as mentioned above) clearly have an impact on the outcome of investigations regarding emotional processing.

AUDIOVISUAL PROCESSING OF EMOTIONS IN PATIENTS WITH DEPRESSION

General group differences in multi-modal emotional processing

To date, only a few studies investigated audiovisual emotional processing in patients with depression. In our previous study (Müller et al., 2014), in which the same audiovisual task as in the current study was used, behavioral results revealed significant main effects of sounds and faces, but no main effect of group nor interactions. However, the further analysis in that paper indicated a deficit in patients when congruent emotional sound information had to be ignored. Since that previous sample was rather small and some of the reported behavioral findings

only showed a trend toward significance, we here aimed to further investigate these effects within an extended sample. Overall, our results across groups are similar to that in the previous smaller sample, but now we found a significant sound × group interaction. This indicates that the impact of sounds on the perception (and therefore ratings) of faces was different in patients compared to controls. In particular, *post hoc* calculations in the current study showed that while controls rated faces quite similar in combination with neutral and happy sounds, patients rated faces significantly more fearful when combined with happy rather than neutral sounds. Further *t*-tests support this result by showing that the difference scores between the happy and neutral sound conditions differed significantly between patients and controls. These results thus highlight the aberrant integration and inhibition of irrelevant and especially positive auditory information in patients with depression. Correlational analyses, however, indicate that this effect is not related to symptom severity.

With regard to existing theories of emotional processing in depression, the current findings can be explained in different ways.

On the one hand, the sound × group interaction could be interpreted as a missing influence of concurrently presented positive auditory stimuli on the ratings of facial expressions in patients with depression. In line with this view, Surguladze et al. (2004) reported that patients exhibited a decreased tendency to interpret happy (but also neutral) faces as happy. Likewise Douglas and Porter (2010) described patients with depression as being less likely to interpret neutral faces as happy, while Loi et al. (2013) reported a reduced ability of patients to appraise positive stimuli of emotional body language. Importantly, however, our *post hoc* tests did not reveal group differences for any stimulus combination and rather indicated that patients differed in the ratings of faces paired with a positive sound compared to a neutral one while controls did not. Furthermore we only found this effect for the audiovisual processing task but not in the unimodal setting. Thus, the current results cannot support the view of an absent positive bias in patients with depression.

On the other hand, our results could also be explained in terms of a negative bias. In particular, for patients with depression, the current results show that any emotional sound (positive or negative) led to more negative ratings of concurrently seen emotional or neutral faces. This suggests that in depressed patients increased emotional input and higher arousal (cf. see Unimodal Valence and Arousal Rating of Sounds and Faces, emotional stimuli were rated as more arousing than neutral stimuli) received from two different channels generally leads to a more negative perception. Nevertheless, although fitting into the view of a more negative perception of emotions going along with depression (Gur et al., 1992; Bouhuys et al., 1999; Leppänen et al., 2004; Naranjo et al., 2011; Kaletsch et al., 2014), it has to be noted that this negative bias is limited to processing of emotions in an audiovisual setting (not for unimodal ratings of faces and sounds). Furthermore it only appears when facial stimuli are paired with happy sounds as healthy individuals are likewise negatively influenced by fearful sounds as patients are. Therefore the negative bias theory also does not explain the results sufficiently.

Yet our results can also be related to another more recent bias theory: Everaert et al. (2014) investigated the combined cognitive bias theory that has been reported for individuals with social phobia (Hirsch et al., 2006) in depression. They found that emotional biases in attention, interpretation, and memory in subclinical depression are strongly interrelated which potentially influences how daily life events are perceived. In particular, participants with higher depression scores paid more attention to negative emotional stimuli, made more negative interpretations and remembered negative material more frequently. When relating these findings to our results, it can be argued that patients paid special attention to fearful sounds and therefore held them in memory during the whole paradigm. Subsequently, their ratings of faces during concurrent presentation of laughter were then influenced by negatively biased memory of previous stimuli which led to a more fearful interpretation of faces. Although fitting with the impaired integration of happy sounds, this interpretation can, however, also not completely explain our findings. If the more fearful rating of faces during laughter had been due to maintenance of screams in memory, neutral sound presentation should have been influenced by this effect as well. However, our patients rated faces in combination with laughter significantly more fearful than in combination with yawning, and this was the only difference compared to controls.

In summary, none of the three presented bias theories can fully explain the current findings in patients with depression. Thus, our results suggest that in a multi-modal setting, impairments in emotional processing in depression cannot be reduced to a specific bias but are far more complex than previously thought.

Incongruence effects for happy and fearful faces

With reference to Müller et al. (2014) and in order to investigate the general effects found for audiovisual emotional processing in patients with depression more in detail, we specifically analyzed incongruence effects in emotional valence separately for happy and fearful faces. In particular, Müller et al. (2014) described a trend toward significance for the congruence × group interaction when analyzing the happy face condition, whereas they did not find any significant effect for the fearful face condition. Furthermore, their neuroimaging data revealed that for the congruent happy condition (happy face paired with happy sound), controls showed stronger deactivation in both left inferior parietal cortex (IPC) and left inferior frontal gyrus (IFG) compared to the patient group. Our current findings confirm the behavioral findings in a larger sample, thus providing increased power of results: patients with depression did not rate happy faces significantly different when paired with congruent (= happy) sounds in contrast to incongruent (= fearful) sounds, while controls rated happy faces significantly happier when paired with happy sounds than paired with fearful sounds. These behavioral findings fit well with the dysregulation of left IPC and IFG, which was found especially for the congruent happy condition (Müller et al., 2014). Thus, a failure to deactivate those two regions during processing of congruent positive audiovisual information in patients with depression might be associated with the missing behavioral incongruence effect found in the current study.

Our results regarding incongruence in emotional valence also shed further light on the effect found in the overall calculations described above (sound × group interaction) by showing that the effect of impaired integration of happy sounds is strongly connected to the happy face condition. Ratings of happy faces indeed became more negative in both patients and controls by incongruent (= fearful) sounds. However, when happy faces were paired with congruent (= happy) sounds, it seems that controls were able to inhibit or were positively influenced by these, while this was not the case in patients. Our findings hence suggest that patients are impaired in the inhibition of positive information in an additional sensory channel, especially when the target information is also positive. This additional presentation of positive information then has a negative impact on the ratings of the positive target emotion, what might reflect a tendency to perceive actually positive emotional input as threat.

Apart from the study by Müller et al. (2014), findings of multimodal emotional processing in patients with depression are rare, but some studies exist, which investigate the impact of irrelevant (congruent/incongruent) emotional information on the perception and processing of emotional stimuli. Uekermann et al. (2008), for example, conducted several tests on identification and matching skills in depressed and healthy individuals, e.g., matching emotional/neutral prosody to semantics and faces. They reported deficits for all of these tests, except for conditions where information was congruent or in which sad stimuli were presented. These findings at first glance contradict the effect of impaired integration of congruent happy sounds in depression as found in the current study. However, this discrepancy may arise from the fact that the task in the current study was to rate the emotion of the faces on an intensity scale, while Uekermann et al. (2008) did not directly measure the influence of distracting information on perception. Rather they investigated differences in accuracy when matching/labeling concurrently presented information. Obviously patients are able to identify information from different channels as congruent, even when positive, but this does not necessarily mean that they perceive stimuli as equally positive as controls do.

Findings in the emotional Stroop task in depression shed further light on how irrelevant emotional information is processed by these patients. However, existing findings are quite inconsistent (Mogg and Bradley, 2005). A recent meta-analysis (Epp et al., 2012) thus quantified findings from behavioral studies investigating the (emotional) Stroop task and showed that depressed individuals exhibited a general attentional bias for emotional content, i.e., negative but importantly also positive words. With regard to positive stimuli, these findings are in line with the current results. The fact that we did not find an effect for the negative sound condition might again be due to the use of fearful rather than sad sounds as negative distractors (cf. see Unimodal Valence and Arousal Ratings of Faces and Sounds).

In summary, our results indicate that when confronted with audiovisual emotional information, patients with depression show in particular impairments when distracted by positive auditory information, especially when the visual information is also positive. This may – in line with studies showing that depression goes along with decreased responsiveness to reward (Henriques and Davidson, 2000) – suggest that depressed individuals are less likely to accept positive feedback as a kind of social reward from their environment, resulting in a more negative view on life and low mood.

NEUROCOGNITIVE PERFORMANCE

Deficits in neurocognitive functioning in depression have been reported by numerous studies (e.g., Breslow et al., 1980; Fisher et al., 1986; Veiel, 1997; Zakzanis et al., 1998; Stordal et al., 2004). It is suggested that these deficits account for or at least substantially contribute to problems in everyday life like occupational functioning deficits, which are experienced by individuals with depression (Evans et al., 2013). However, findings on the pattern, extent and specificity of cognitive deficits in depression are quite heterogeneous (Ottowitz et al., 2002; Naismith et al., 2003; Marazziti et al., 2010; Lee et al., 2012; Quinn et al., 2012). For example, there is no agreement whether impairments are limited to executive functioning (Baune et al., 2012) or also relate to other domains like attention (Godard et al., 2012), psychomotor speed, visual learning/memory, and others (Lee et al., 2012). The current study revealed significant differences between patients and controls only in (both versions of) the trail making test. Conversely, there were no significant differences in any other neurocognitive test, contradicting the findings by Hoffstaedter et al. (2012) who conducted the same tests and reported group differences for most of them. Since motor speed was also assessed by TAP10s and Tap10x30, which both did not reveal group differences, our results indicate relatively specific deficits of visual attention and cognitive flexibility. Also, group differences in TMT-B in particular may point in the direction of a deficit in dealing with distracting information in patients with depression. These results highlight the importance of the specificity of assessment instruments for neurocognitive performance in depression (cf. Trivedi and Greer, 2014). Executive functioning, for example, may be operationalized and then measured by a large number of different tests and study designs. The ensuing results, however, would all be interpreted under the domain of executive functioning. This might explain heterogeneity in study findings. In addition, findings are also influenced by the patient sample investigated. The patient status (inpatient/outpatient) has, for example, an impact on the severity of impairments (Burt et al., 1995), possibly due to the fact that inpatients exhibit generally worse symptomatology (or rather, patients with worse symptoms are more likely to receive inpatient treatment). Furthermore, subtypes of depression also have to be considered when investigating neurocognitive performance (Naismith et al., 2003; Lee et al., 2012), as co-morbidities like anxiety disorder or bipolar disorder might have crucial impact on the outcome of neurocognitive test batteries. Thus, the fact that the patient group tested in the current study consisted exclusively of patients with unipolar depression who were free of co-morbidities may explain inconsistencies to some previous findings where patient groups were variably mixed. At last, antidepressant medication also plays a role, since medicated patients have been reported to perform better on neurocognitive tests than unmedicated ones (Gualtieri et al., 2006). Wagner et al. (2012) investigated changes in neurocognitive functioning during antidepressant treatment and found that

performance of patients only improves in certain domains but not all. Especially test performance in TMT-B was excluded from improvement. This might provide an explanation for our results, as all but one patient of the current study were receiving antidepressant medication. However, subgroup analysis with regard to type of antidepressant medication was not possible due to the large variety in medication composition from patient to patient (see **Table 2**). Thus, future studies should deal with detailed analysis of the impact of certain antidepressant medication types on neurocognitive performance to be able to identify crucial factors on test results.

In summary, our findings show that when comparing patients with controls, differences are only found for attention and cognitive flexibility. This supports but also contradicts findings of other studies on neurocognitive functioning in depression and underlines the heterogeneity of depression.

RELATIONSHIP BETWEEN AUDIOVISUAL EMOTIONAL PROCESSING AND NEUROCOGNITIVE FUNCTIONING

Emotional and neurocognitive aberrations have already been reported for depressed patients in numerous studies. However, only few have yet investigated how deficits in these two domains are related to each other. We thus correlated the effects found in the emotional processing task and the performance in the different neurocognitive tests and found that the effects revealed for audiovisual emotional processing, i.e., the difference of valence ratings of faces paired with happy compared to neutral sounds, were not related to performance in the neurocognitive tests. However, when not correcting for multiple comparisons, correlational analysis of the TMT-A test findings with impairments in emotional processing would reach significance. This is interesting, because the only two tests that revealed group differences were the two trail making tests. One could hence argue that impairments in the processing of interpersonal stimuli are possibly related to general deficits in visual attention. This would expand the findings of studies reporting selective visual attention for negative emotional material in depressed individuals (Eizenman et al., 2003; Kellough et al., 2008), indicating that attentional biases toward negative emotions alone cannot sufficiently explain impairments in emotional processing in depression. Rather, general attentional deficits might also lead to problems in concentrating on visual emotional material. This would fit with our finding that patients with depression are distracted more easily than healthy controls by additional irrelevant (especially positive) auditory stimuli. All in all, however, as this result did not survive correction for multiple comparisons, our findings nevertheless argue against a (direct) relationship between emotional processing and neurocognitive functioning. This contradicts findings of one previous study (Uekermann et al., 2008) reporting correlations between perception of affective prosody with inhibition abilities, set shifting, and working memory. On the other hand, another study (Langenecker et al., 2005) reported no correlations between impairments in emotional perception and neurocognitive performance and is hence well in line with our findings. Discussing their results, the authors mentioned that executive impairments are a feature of several different disorders, indicating that cognitive disturbances may not necessarily account for emotional impairments in depression. Likewise, Bourke et al. (2012) found that patients with depression perform significantly worse on verbal memory and spatial working memory tasks, but they did not find differences to healthy controls on unimodal emotional face recognition tasks (going in line with our findings in unimodal emotional processing). Even though they did not directly correlate emotional with cognitive performance, their results indicate that neurocognitive impairments are largely independent from emotional deficits.

Other studies (Everaert et al., 2014), however, reported associations of emotion perception and cognition in the sense of cognitive biases for negative emotions. In particular, they could show that emotionally biased cognitive processes like attention, interpretation and memory are highly interrelated with each other. Under this aspect, it has to be pointed out that it seems to make a huge difference if, on the one hand, general cognitive functions in absence of emotional material are examined or if, on the other hand, cognitive processing of emotional material is assessed. Thus, with regard to our findings, we can only infer that there is no relationship between general cognitive performance and impairments in emotional processing.

In summary, our results thus indicate that deficits in audiovisual emotional processing in depression seem to be widely independent from general neurocognitive functioning. Thus they do not support the assumption that deficits in emotional processing in patients with major depression are the results of impaired general attention or inhibitory functioning. Nevertheless we cannot completely rule out a bias component especially toward emotional material.

SUMMARY

Our findings suggest that audiovisual integration of especially happy sounds is altered in patients with depression and that these alterations cannot be related directly to impairments in cognitive skills. Group differences in neurocognitive test performance were only revealed for measures of attention and cognitive flexibility. These results indicate that in real life, when emotions are processed in a multimodal fashion, deficits in depression cannot be reduced to an overall negative attitude toward emotional and neutral stimuli or a general absence of a positive bias. Rather, it is the influence of irrelevant positive stimuli, which plays a key role in emotion perception in depression. Though, impairments in audiovisual emotional processing do not change as a function of depressive symptom severity in patients. Furthermore there is no clear connection between emotional and neurocognitive impairments.

Although the current study did not directly investigate the role of attention in multisensory integration, our study adds further knowledge to this topic by investigating the relationship between both aspects in major depression and indicates that alterations in multi-modal emotional processing are not directly related to impaired attention.

AUTHOR CONTRIBUTIONS

Veronika I. Müller and Simon B. Eickhoff designed the study, Veronika I. Müller acquired the data, Sophie Doose-Grünefeld and Veronika I. Müller analyzed the data, Sophie

Doose-Grünefeld, Simon B. Eickhoff, and Veronika I. Müller wrote the paper.

ACKNOWLEDGMENTS

Simon B. Eickhoff is supported by the Deutsche Forschungsgemeinschaft (DFG, EI 816/4-1, EI 816/6-1, LA 3071/3-1) and the National Institute of Mental Health (R01-MH074457).

REFERENCES

Airaksinen, E., Larsson, M., Lundberg, I., and Forsell, Y. (2004). Cognitive functions in depressive disorders: evidence from a population-based study. *Psychol. Med.* 34, 83–91. doi: 10.1017/S0033291703008559

Army Individual Test Battery. (AITB). (1944). *Manual of Directions and Scoring.* Washington, DC: War Department, Adjutant General's Office.

Baune, B. T., Czira, M. E., Smith, A. L., Mitchell, D., and Sinnamon, G. (2012). Neuropsychological performance in a sample of 13-25 year olds with a history of non-psychotic major depressive disorder. *J. Affect. Disord.* 141, 441–448. doi: 10.1016/j.jad.2012.02.041

Behrwind, S. D., Dafotakis, M., Halfter, S., Hobusch, K., Berthold-Losleben, M., Cieslik, E. C., et al. (2011). Executive control in chronic schizophrenia: a perspective from manual stimulus-response compatibility task performance. *Behav. Brain Res.* 223, 24–29. doi: 10.1016/j.bbr.2011.04.009

Bouhuys, A. L., Geerts, E., and Gordijn, M. C. (1999). Depressed patients' perceptions of facial emotions in depressed and remitted states are associated with relapse: a longitudinal study. *J. Nerv. Ment. Dis.* 187, 595–602. doi: 10.1097/00005053-199910000-00002

Bourke, C., Douglas, K., and Porter, R. (2010). Processing of facial emotion expression in major depression: a review. *Aust. N. Z. J. Psychiatry* 44, 681–696. doi: 10.3109/00048674.2010.496359

Bourke, C., Porter, R. J., Carter, J. D., McIntosh, V. V., Jordan, J., Bell, C., et al. (2012). Comparison of neuropsychological functioning and emotional processing in major depression and social anxiety disorder subjects, and matched healthy controls. *Aust. N. Z. J. Psychiatry* 46, 972–981. doi: 10.1177/0004867412451502

Breslow, R., Kocsis, J., and Belkin, B. (1980). Memory deficits in depression: evidence utilizing the Wechsler Memory Scale. *Percept. Mot. Skills* 51, 541–542. doi: 10.2466/pms.1980.51.2.541

Burt, D. B., Zembar, M. J., and Niederehe, G. (1995). Depression and memory impairment: a meta-analysis of the association, its pattern, and specificity. *Psychol. Bull.* 117, 285–305. doi: 10.1037/0033-2909.117.2.285

Csukly, G., Czobor, P., Szily, E., Takacs, B., and Simon, L. (2009). Facial expression recognition in depressed subjects: the impact of intensity level and arousal dimension. *J. Nerv. Ment. Dis.* 197, 98–103. doi: 10.1097/NMD.0b013e3181923f82

DeBattista, C. (2005). Executive dysfunction in major depressive disorder. *Expert Rev. Neurother.* 5, 79–83. doi: 10.1586/14737175.5.1.79

Defer, G. L., Widner, H., Marie, R. M., Remy, P., and Levivier, M. (1999). Core assessment program for surgical interventional therapies in Parkinson's disease (CAPSIT-PD). *Mov. Disord.* 14, 572–584. doi: 10.1002/1531-8257(199907)14:4<572::AID-MDS1005>3.0.CO;2-C

Diamond, A. (2013). Executive functions. *Annu. Rev. Psychol.* 64, 135–168. doi: 10.1146/annurev-psych-113011-143750

Douglas, K. M., and Porter, R. J. (2010). Recognition of disgusted facial expressions in severe depression. *Br. J. Psychiatry* 197, 156–157. doi: 10.1192/bjp.bp.110.078113

Eizenman, M., Yu, L. H., Grupp, L., Eizenman, E., Ellenbogen, M., Gemar, M., et al. (2003). A naturalistic visual scanning approach to assess selective attention in major depressive disorder. *Psychiatry Res.* 118, 117–128. doi: 10.1016/S0165-1781(03)00068-4

Epp, A. M., Dobson, K. S., Dozois, D. J., and Frewen, P. A. (2012). A systematic meta-analysis of the Stroop task in depression. *Clin. Psychol. Rev.* 32, 316–328. doi: 10.1016/j.cpr.2012.02.005

Evans, V., Chan, S., Iverson, G. L., Bond, D., Yatham, L., and Lam, R. (2013). Systematic review of neurocognition and occupational functioning in major depressive disorder. *Neuropsychiatry* 3, 97–105. doi: 10.2217/npy.13.3

Everaert, J., Duyck, W., and Koster, E. H. (2014). Attention, interpretation, and memory biases in subclinical depression: a proof-of-principle test of the combined cognitive biases hypothesis. *Emotion* 14, 331–340. doi: 10.1037/a0035250

Ferrari, A. J., Charlson, F. J., Norman, R. E., Patten, S. B., Freedman, G., Murray, C. J., et al. (2013). Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. *PLoS Med.* 10:e1001547. doi: 10.1371/journal.pmed.1001547

Fisher, D. G., Sweet, J. J., and Pfaelzersmith, E. A. (1986). Influence of depression on repeated neuropsychological testing. *Int. J. Clin. Neuropsychol.* 8, 14–18.

Godard, J., Baruch, P., Grondin, S., and Lafleur, M. F. (2012). Psychosocial and neurocognitive functioning in unipolar and bipolar depression: a 12-month prospective study. *Psychiatry Res.* 196, 145–153. doi: 10.1016/j.psychres.2011.09.013

Goeleven, E., De Raedt, R., Baert, S., and Koster, E. H. (2006). Deficient inhibition of emotional information in depression. *J. Affect. Disord.* 93, 149–157. doi: 10.1016/j.jad.2006.03.007

Gualtieri, C. T., Johnson, L. G., and Benedict, K. B. (2006). Neurocognition in depression: patients on and off medication versus healthy comparison subjects. *J. Neuropsychiatry Clin. Neurosci.* 18, 217–225. doi: 10.1176/jnp.2006.18.2.217

Gur, R. C., Erwin, R. J., Gur, R. E., Zwil, A. S., Heimberg, C., and Kraemer, H. C. (1992). Facial emotion discrimination: II. Behavioral findings in depression. *Psychiatry Res.* 42, 241–251. doi: 10.1016/0165-1781(92)90116-K

Gur, R. C., Sara, R., Hagendoorn, M., Marom, O., Hughett, P., Macy, L., et al. (2002). A method for obtaining 3-dimensional facial expressions and its standardization for use in neurocognitive studies. *J. Neurosci. Methods* 115, 137–143. doi: 10.1016/S0165-0270(02)00006-7

Halstead, W. (1947). *Brain and Intelligence: A Quantitative Study of the Frontal Lobes.* Chicago: University of Chicago Press.

Hamilton, M. (1960). A rating scale for depression. *J. Neurol. Neurosurg. Psychiatry* 23, 56–62. doi: 10.1136/jnnp.23.1.56

Hautzinger, M., Keller, F., and Kühner, C. (2006). *Beck-Depressions-Inventar (BDI-II)*, Revision, 2nd Edn. Frankfurt: Harcourt Test Services.

Henriques, J. B., and Davidson, R. J. (2000). Decreased responsiveness to reward in depression. *Cogn. Emot.* 14, 711–724. doi: 10.1080/026999300501 17684

Hirsch, C. R., Clark, D. M., and Mathews, A. (2006). Imagery and interpretations in social phobia: support for the combined cognitive biases hypothesis. *Behav. Ther.* 37, 223–236. doi: 10.1016/j.beth.2006.02.001

Hoffstaedter, F., Sarlon, J., Grefkes, C., and Eickhoff, S. B. (2012). Internally vs. externally triggered movements in patients with major depression. *Behav. Brain Res.* 228, 125–132. doi: 10.1016/j.bbr.2011.11.024

Hu, Z., Liu, H., Weng, X., and Northoff, G. (2012). Is there a valence-specific pattern in emotional conflict in major depressive disorder? An exploratory psychological study. *PLoS ONE* 7:e31983. doi: 10.1371/journal.pone.0031983

Joormann, J. (2004). Attentional bias in dysphoria: the role of inhibitory processes. *Cogn. Emot.* 18, 125–147. doi: 10.1080/02699930244000480

Joormann, J., and Gotlib, I. H. (2006). Is this happiness I see? Biases in the identification of emotional facial expressions in depression and social phobia. *J. Abnorm. Psychol.* 115, 705–714. doi: 10.1037/0021-843X.115.4.705

Kaletsch, M., Pilgramm, S., Bischoff, M., Kindermann, S., Sauerbier, I., Stark, R., et al. (2014). Major depressive disorder alters perception of emotional body movements. *Front. Psychiatry* 5:4. doi: 10.3389/fpsyt.2014.00004

Kan, Y., Mimura, M., Kamijima, K., and Kawamura, M. (2004). Recognition of emotion from moving facial and prosodic stimuli in depressed patients. *J. Neurol. Neurosurg. Psychiatry* 75, 1667–1671. doi: 10.1136/jnnp.2004.036079

Kellough, J. L., Beevers, C. G., Ellis, A. J., and Wells, T. T. (2008). Time course of selective attention in clinically depressed young adults: an eye tracking study. *Behav. Res. Ther.* 46, 1238–1243. doi: 10.1016/j.brat.2008.07.004

Langenecker, S. A., Bieliauskas, L. A., Rapport, L. J., Zubieta, J. K., Wilde, E. A., and Berent, S. (2005). Face emotion perception and executive functioning deficits in depression. *J. Clin. Exp. Neuropsychol.* 27, 320–333. doi: 10.1080/13803390490490515720

Lee, R. S., Hermens, D. F., Porter, M. A., and Redoblado-Hodge, M. A. (2012). A meta-analysis of cognitive deficits in first-episode major depressive disorder. *J. Affect. Disord.* 140, 113–124. doi: 10.1016/j.jad.2011.10.023

Lehrl, S. (1989). *Mehrfach-Wortschatz-Intelligenztest MWT-B*, 2nd Edn. Erlangen: Perimed Fachbuch-Verlagsgesellschaft mbH.

Leppänen, J. M., Milders, M., Bell, J. S., Terriere, E., and Hietanen, J. K. (2004). Depression biases the recognition of emotionally neutral faces. *Psychiatry Res.* 128, 123–133. doi: 10.1016/j.psychres.2004.05.020

Leung, K. K., Lee, T. M., Yip, P., Li, L. S., and Wong, M. M. (2009). Selective attention biases of people with depression: positive and negative priming of depression-related information. *Psychiatry Res.* 165, 241–251. doi: 10.1016/j.psychres.2007.10.022

Loi, F., Vaidya, J. G., and Paradiso, S. (2013). Recognition of emotion from body language among patients with unipolar depression. *Psychiatry Res.* 209, 40–49. doi: 10.1016/j.psychres.2013.03.001

Marazziti, D., Consoli, G., Picchetti, M., Carlini, M., and Faravelli, L. (2010). Cognitive impairment in major depression. *Eur. J. Pharmacol.* 626, 83–86. doi: 10.1016/j.ejphar.2009.08.046

McDermott, L. M., and Ebmeier, K. P. (2009). A meta-analysis of depression severity and cognitive function. *J. Affect. Disord.* 119, 1–8. doi: 10.1016/j.jad.2009.04.022

Milders, M., Bell, S., Platt, J., Serrano, R., and Runcie, O. (2010). Stable expression recognition abnormalities in unipolar depression. *Psychiatry Res.* 179, 38–42. doi: 10.1016/j.psychres.2009.05.015

Mogg, K., and Bradley, B. P. (2005). Attentional bias in generalized anxiety disorder versus depressive disorder. *Cogn. Ther. Res.* 29, 29–45. doi: 10.1007/s10608-005-1646-y

Müller, V. I., Cieslik, E. C., Kellermann, T. S., and Eickhoff, S. B. (2014). Crossmodal emotional integration in major depression. *Soc. Cogn. Affect. Neurosci.* 9, 839–848. doi: 10.1093/scan/nst057

Müller, V. I., Habel, U., Derntl, B., Schneider, F., Zilles, K., Turetsky, B. I., et al. (2011). Incongruence effects in crossmodal emotional integration. *Neuroimage* 54, 2257–2266. doi: 10.1016/j.neuroimage.2010.10.047

Naismith, S. L., Hickie, I. B., Turner, K., Little, C. L., Winter, V., Ward, P. B., et al. (2003). Neuropsychological performance in patients with depression is associated with clinical, etiological and genetic risk factors. *J. Clin. Exp. Neuropsychol.* 25, 866–877. doi: 10.1076/jcen.25.6.866.16472

Naranjo, C., Kornreich, C., Campanella, S., Noel, X., Vandriette, Y., Gillain, B., et al. (2011). Major depression is associated with impaired processing of emotion in music as well as in facial and vocal stimuli. *J. Affect. Disord.* 128, 243–251. doi: 10.1016/j.jad.2010.06.039

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9, 97–113. doi: 10.1016/0028-3932(71)90067-4

Ottowitz, W. E., Dougherty, D. D., and Savage, C. R. (2002). The neural network basis for abnormalities of attention and executive function in major depressive disorder: implications for application of the medical disease model to psychiatric disorders. *Harv. Rev. Psychiatry* 10, 86–99. doi: 10.1080/10673220216210

Owens, M., Koster, E. H., and Derakshan, N. (2013). Improving attention control in dysphoria through cognitive training: transfer effects on working memory capacity and filtering efficiency. *Psychophysiology* 50, 297–307. doi: 10.1111/psyp.12010

Paelecke-Habermann, Y., Pohl, J., and Leplow, B. (2005). Attention and executive functions in remitted major depression patients. *J. Affect. Disord.* 89, 125–135. doi: 10.1016/j.jad.2005.09.006

Quinn, C. R., Harris, A., Felmingham, K., Boyce, P., and Kemp, A. (2012). The impact of depression heterogeneity on cognitive control in major depressive disorder. *Aust. N. Z. J. Psychiatry* 46, 1079–1088. doi: 10.1177/0004867412461383

Rose, E. J., and Ebmeier, K. P. (2006). Pattern of impaired working memory during major depression. *J. Affect. Disord.* 90, 149–161. doi: 10.1016/j.jad.2005.11.003

Schaefer, K. L., Baumann, J., Rich, B. A., Luckenbaugh, D. A., and Zarate, C. A. Jr. (2010). Perception of facial emotion in adults with bipolar or unipolar depression and controls. *J. Psychiatr. Res.* 44, 1229–1235. doi: 10.1016/j.jpsychires.2010.04.024

Schlipf, S., Batra, A., Walter, G., Zeep, C., Wildgruber, D., Fallgatter, A., et al. (2013). Judgment of emotional information expressed by prosody and semantics in patients with unipolar depression. *Front. Psychol.* 4:461. doi: 10.3389/fpsyg.2013.00461

Schneider, D., Regenbogen, C., Kellermann, T., Finkelmeyer, A., Kohn, N., Derntl, B., et al. (2012). Empathic behavioral and physiological responses to dynamic stimuli in depression. *Psychiatry Res.* 200, 294–305. doi: 10.1016/j.psychres.2012.03.054

Sobin, C., and Sackeim, H. A. (1997). Psychomotor symptoms of depression. *Am. J. Psychiatry* 154, 4–17. doi: 10.1176/ajp.154.1.4

Stordal, K. I., Lundervold, A. J., Egeland, J., Mykletun, A., Asbjornsen, A., Landro, N. I., et al. (2004). Impairment across executive functions in recurrent major depression. *Nord. J. Psychiatry* 58, 41–47. doi: 10.1080/08039480310000789

Surguladze, S. A., Young, A. W., Senior, C., Brebion, G., Travis, M. J., and Phillips, M. L. (2004). Recognition accuracy and response bias to happy and sad facial expressions in patients with major depression. *Neuropsychology* 18, 212–218. doi: 10.1037/0894-4105.18.2.212

Tewes, U. (1991). *HAWIE-R. Hamburg-Wechsler-Intelligenztest für Erwachsene. Revision 1991; Handbuch und Testanweisung.* Bern: Verlag Hans Huber.

Trivedi, M. H., and Greer, T. L. (2014). Cognitive dysfunction in unipolar depression: implications for treatment. *J. Affect. Disord.* 152–154, 19–27. doi: 10.1016/j.jad.2013.09.012

Uekermann, J., Abdel-Hamid, M., Lehmkamper, C., Vollmoeller, W., and Daum, I. (2008). Perception of affective prosody in major depression: a link to executive functions? *J. Int. Neuropsychol. Soc.* 14, 552–561. doi: 10.1017/S1355617708080740

Veiel, H. O. (1997). A preliminary profile of neuropsychological deficits associated with major depression. *J. Clin. Exp. Neuropsychol.* 19, 587–603. doi: 10.1080/01688639708403745

Wagner, S., Doering, B., Helmreich, I., Lieb, K., and Tadic, A. (2012). A meta-analysis of executive dysfunctions in unipolar major depressive disorder without psychotic symptoms and their changes during antidepressant treatment. *Acta Psychiatr. Scand.* 125, 281–292. doi: 10.1111/j.1600-0447.2011.01762.x

WHO. (2010). *International Statistical Classification of Diseases and Related Health Problems*, 10th Revision. Geneva: WHO.

Wittchen, H. U., Zaudig, M., and Fydrich, T. (1997). *Strukturiertes Klinisches Interview für DSM-IV*. Göttingen: Hogrefe.

Zakzanis, K. K., Leach, L., and Kaplan, E. (1998). On the nature and pattern of neurocognitive function in major depressive disorder. *Neuropsychiatry Neuropsychol. Behav. Neurol.* 11, 111–119.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Received: 25 July 2014; accepted: 08 January 2015; published online: 30 January 2015. Citation: Doose-Grünefeld S, Eickhoff SB and Müller VI (2015) Audiovisual emotional processing and neurocognitive functioning in patients with depression. Front. Integr. Neurosci. **9**:3. doi: 10.3389/fnint.2015.00003*

This article was submitted to the journal Frontiers in Integrative Neuroscience.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read, for greatest visibility

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

FAST PUBLICATION
Average 85 days from submission to publication (across all journals)

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

SUPPORT
By our Swiss-based editorial team

IMPACT METRICS
Advanced metrics track your article's impact

GLOBAL SPREAD
5'100'000+ monthly article views and downloads

LOOP RESEARCH NETWORK
Our network increases readership for your article

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on