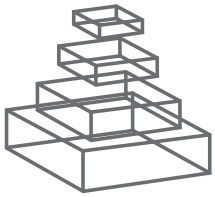
frontiers
RESEARCH TOPICS

MULTISENSORY PERCEPTION AND ACTION: PSYCHOPHYSICS, NEURAL MECHANISMS, AND APPLICATIONS

Topic Editors
Zhuanghua Shi and Hermann J. Müller

frontiers in
INTEGRATIVE NEUROSCIENCE



frontiers

Cover image provided by Ibbl sarl, Lausanne CH

ISSN 1664-8714
ISBN 978-2-88919-381-3
DOI 10.3389/978-2-88919-381-3

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

MULTISENSORY PERCEPTION AND ACTION: PSYCHOPHYSICS, NEURAL MECHANISMS, AND APPLICATIONS

Topic Editors:
Zhuanghua Shi, Ludwig-Maximilians-Universität München, Germany
Hermann J. Müller, Ludwig-Maximilians-Universität München, Germany

Table of Contents

frontiers in
INTEGRATIVE NEUROSCIENCE

EDITORIAL
published: 21 November 2013
doi: 10.3389/fnint.2013.00081

Multisensory perception and action: development, decision-making, and neural mechanisms

***Zhuanghua Shi*[1]* and Hermann J. Müller*[1,2]**

[1] *Department of Psychology, Experimental Psychology, Ludwig-Maximilians-Universität München, Munich, Germany*
[2] *Department of Psychological Science, Birkbeck College, University of London, London, UK*
**Correspondence: shi@psy.lmu.de*

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

Keywords: multisensory perception, multisensory timing, multisensory development, multisensory learning, multisensory neural mechanisms

Surrounded by multiple objects and events, receiving multisensory stimulation, our brain must sort through relevant and irrelevant multimodal signals to correctly decode and represent the information from the same and different objects and, respectively, events in the physical world. Over the last two decades, scientific interest has increased dramatically in how we integrate multisensory information and how we interact with a multisensory world, as evidenced by exponential growth of the relevant studies using behavioral and/or neuro-scientific approaches.

The Special Issue topic of "Multisensory perception and action: psychophysics, neural mechanisms, and applications" emerged from a scientific meeting dedicated to these issues: the *Munich Multisensory Perception Symposium* held in Holzhausen am Ammersee, Germany (June 24–26, 2011). This volume, which collects research articles contributed by attendees of the symposium as well as the wider community, is organized into three interrelated sections:

(I) Development, learning, and decision making in multisensory perception
(II) Multisensory timing and sensorimotor temporal integration
(III) Electrophysiological and neuro-imaging analyses of multisensory perception

DEVELOPMENT, LEARNING, AND DECISION-MAKING IN MULTISENSORY PERCEPTION

Many multisensory studies, ranging from spatial (e.g., Ernst and Banks, 2002; Alais and Burr, 2004) to temporal integration (e.g., Burr et al., 2009; Chen et al., 2010; Shi et al., 2013b), reveal that our brain combines multisensory signals if they are closely relevant to the task, in order to boost overall performance. Senses, however, are not the only source for decision-making. Prior, contextual, and symbolic cues can also contribute as an extra source of information to improve performance (Jazayeri and Shadlen, 2010; Petzschner and Glasauer, 2011; for a review, see Shi et al., 2013a). Accordingly, Petzschner et al. (2012) set out to examine how auxiliary contextual cues, such as symbolic "short" and "long" cues, are used optimally in a distance production-reproduction task. Their findings indicate that humans are capable of using symbolic cues for final estimates, even though the mapping of the symbolic cue onto the stimulus dimension has to be learned during the experiment.

With respect to learning, one prominent question in multisensory integration concerns when and how we acquire the capacity to optimally integrate multisensory cues. Some recent studies suggest that this capacity is not present at birth, but rather develops after about 8 years of age (e.g., Gori et al., 2008). Gori et al. (2012) expanded this line of research by examining audio-visual temporal and spatial bisection tasks in young children, finding that young children exhibit strong unisensory dominance over multisensory integration of audiovisual signals, with audition dominating audiovisual time perception and vision dominating space perception. Both dominance effects reflect a process of cross-sensory calibration of developing systems, where the more accurate sense calibrates or teaches the other, rather than fusing with it. In another study, Wismeijer et al. (2012) showed that our brain also exhibits remarkable ability to learn cue-associations, such as an arbitrary association of visual gloss and touch softness, and use the learned associate-cues for judgments—with learning being more efficient from touch-to-vision than from vision-to-touch, which is in line with earlier evidence of touch teaching vision for size discrimination in young children (Gori et al., 2008).

Multisensory signals, compared to separate unisensory signals, not only enhance overall performance, but also facilitate the speed of responses. Based on their previously developed framework of the time-window-of-integration (TWIN), Colonius and Diederich (2012) provided further qualitative and quantitative predictions of the TWIN model regarding how the probability of multisensory integration would affect response facilitation differently in the crossmodal-signals and the focused-attention paradigm. In the reverse direction Hong et al. (2012) examined response impairments arising from conflicting crossmodal stimuli or configurations that engender multisensory illusions, in particular, the hand-reversal illusion.

MULTISENSORY TIMING AND SENSORIMOTOR TEMPORAL INTEGRATION

Time perception is susceptible to a wide range of factors (Shi et al., 2013a), in particular with multisensory inputs. A number of authors examining this set of issues have attempted to pin down key factors in multisensory timing. With regard to the perception of multisensory durations, Shi et al. (2012) showed that high-arousal affective pictures have differential impacts on subsequent tactile duration judgments, with pictures that evoke threat meanings expanding subjective duration, whereas pictures that evoke disgust meanings exhibiting no effects on tactile temporal

judgments—indicative of the importance of crossmodal connections in the processing of multisensory timing. Ganzenmüller et al. (2012) further demonstrated that delayed onset of auditory signals generated by participants' manual button press immediately lengthens the reproduced duration, whereas offset delays did not—showing that multisensory timing relies differentially on sensory and motor signals in duration reproduction. Using apparent motion as an implicit measure of perceived duration, Zhang et al. (2012) reported another differential adaptation effect in multisensory timing: adaptation to a short auditory or visual interval resulting in a consistent negative aftereffect for Ternus apparent motion, whereas adaptation to a long interval yielded an aftereffect only for the auditory, and not the visual, condition.

Similar to multisensory duration, multisensory temporal-order processing is also influenced by many factors. For example, to identify key physical changes associated with the articulation of consonants and vowels that may influence the temporal integration window for audiovisual speech, Vatakis et al. (2012) examined the perception of audiovisual synchrony using video clips uttered by different speakers with differential audiovisual signal saliencies (with auditory saliency measured by a combination of three acoustic features: instantaneous energy of the most active filter, instantaneous amplitude, and frequency of the dominant filter's output; and visual saliency computed by intensity, color, and motion). They found that the (degree of) saliency of visual-speech signals can modulate the lead of visual over auditory signals that is necessary for them to be perceived as simultaneous, the lead typically found in audiovisual speech perception. These findings thus support the "information reliability hypothesis," on which the perception of a multisensory feature is dominated by the modality that provides the most reliable information (Welch and Warren, 1980; Ernst et al., 2004). Similarly, Hendrich et al. (2012) found that not only stimuli features, but also task requirements, such as dual tasks, could affect audiovisual temporal-order judgments, arguing that the influence of dual tasks on crossmodal temporal processing is mainly on the perceptual, rather than the response-selection, stage.

ELECTROPHYSIOLOGICAL AND NEURO-IMAGING ANALYSES OF MULTISENSORY PERCEPTION

The neural mechanisms underlying integrative and interactive functions are central to understanding multisensory perception. Quite a number of studies concerned with these functions have been designed to elucidate how information that comes from different sensory modalities are processed and integrated in the brain.

Several studies provide found evidence that multisensory signals are integrated at a very early stage. Naci et al. (2012), for example, found that higher-order regions in anterior temporal (AT) and inferior prefrontal cortex (IPC) performed audiovisual integration 100 ms earlier than a sensory-driven region in the posterior occipital (pO) cortex, suggesting the brain represents familiar and complex multisensory objects through early interactivity between higher-order, and sensory-driven regions. Stekelenburg and Vroomen (2012) also showed that spatial congruity between auditory and visual signals modulates audiovisual interactions reflected in early ERP components, namely, the N1 and P2. Early integration may boost the saliency of the multisensory signals, even when the multisensory signals are irrelevant distractors, causing an attentional shift toward the multisensory distractor, as measured by steady-state visual evoked potentials (SSVEP) in an audiovisual speech task (Krause et al., 2012). Instead of using multisensory signals, Töllner et al. (2012) presented separate auditory and visual signals in a dual-task paradigm requiring both auditory and visual discriminations, to investigate influences of task order predictability (TOP) and inter-task onset asynchrony (SOA) on perceptual, and motor processing stages, two stages indexed, respectively, by two EEG components: the Posterior-Contralateral- Negativity (PCN) and the Lateralized-Readiness-Potential (LRP). Töllner et al. found TOP to interact with inter-task SOA in determining the speed of perceptual processing, providing electrophysiological evidence of central capacity limitations in the processing of auditory and visual dual tasks.

Using functional MRI imaging techniques, two other studies examined brain regions involved in multisensory perception. Noesselt et al. (2012) investigated the neural basis of the perception of synchrony/asynchrony for audiovisual speech stimuli, and found a distinct pattern of modulations within the multisensory superior temporal sulcus complex (mSTS-c): "auditory leading (AL)" and "visual leading (VL) areas" lie closer to "synchrony areas" than to each other, suggesting the presence of distinct subregions within the human STS-c for the maintenance of temporal relations for audiovisual speech stimuli, with differential functional connectivity with prefrontal regions. Beer et al. (2013), on the other hand, found bimodal presentation of audiovisual speech and audiovisual movement stimuli, compared to unimodal stimulation, engaged a temporal-occipital brain network including the multisensory superior temporal sulcus (msSTS), the lateral superior temporal gyrus (ISTG), and the extrastriate body area (EBA). Moreover, brain areas involved in multisensory processing showed little direct connectivity with primary sensory cortices; rather these brain areas were connected to early sensory cortices via intermediate nodes of the STS and the inferior occipital cortex (IOC).

Taken together, this collection provides a broad-spectrum but overall coherent addition to the rapidly growing field of multisensory perception and action. Of course, more work needs to be carried out and many open questions and issues (some of which are identified in the present collection) remain to be addressed in order to achieve a full understanding the functions and neural mechanisms of multisensory perception and action. We would like to thank all the authors, the expert reviewers, and the Frontiers staff for helping to make this Special Issue possible. We hope this collection can act as a catalyst for some of the future work, and we look forward to further explorations of multisensory perception and action.

REFERENCES

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

Beer, A. L., Plank, T., Meyer, G., and Greenlee, M. W. (2013). Combined diffusion-weighted and functional magnetic resonance imaging reveals a temporal-occipital network involved in auditory-visual object processing. *Front. Integr. Neurosci.* 7:5. doi: 10.3389/fnint.2013.00005

Burr, D. C., Banks, M. S., and Morrone, M. C. (2009). Auditory dominance over vision in the perception of interval duration. *Exp. Brain Res.* 198, 49–57. doi: 10.1007/s00221-009-1933-z

Chen, L., Shi, Z., and Müller, H. J. (2010). Influences of intra- and crossmodal grouping on visual and tactile Ternus apparent motion. *Brain Res.* 1354, 152–162. doi: 10.1016/j.brainres.2010.07.064

Colonius, H., and Diederich, A. (2012). Focused attention vs. crossmodal signals paradigm: deriving predictions from the time-window-of-integration model. *Front. Integr. Neurosci.* 6:62. doi: 10.3389/fnint.2012.00062

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

Ernst, M. O., Bülthoff, H. H., 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

Ganzenmüller, S., Shi, Z., and Müller, H. J. (2012). Duration reproduction with sensory feedback delay: differential involvement of perception and action time. *Front. Integr. Neurosci.* 6:95. doi: 10.3389/fnint.2012.00095

Gori, M., Del Viva, M., Sandini, G., and Burr, D. C. (2008). Young children do not integrate visual and haptic form information. *Curr. Biol.* 18, 694–698. doi: 10.1016/j.cub.2008.04.036

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

Hendrich, E., Strobach, T., Buss, M., Müller, H. J., and Schubert, T. (2012). Temporal-order judgment of visual and auditory stimuli: modulations in situations with and without stimulus discrimination. *Front. Integr. Neurosci.* 6:63. doi: 10.3389/fnint.2012.00063

Hong, S. W., Xu, L., Kang, M.-S., and Tong, F. (2012). The hand-reversal illusion revisited. *Front. Integr. Neurosci.* 6:83. doi: 10.3389/fnint.2012.00083

Jazayeri, M., and Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nat. Neurosci.* 13, 1020–1026. doi: 10.1038/nn.2590

Krause, H., Schneider, T. R., Engel, A. K., and Senkowski, D. (2012). Capture of visual attention interferes with multisensory speech processing. *Front. Integr. Neurosci.* 6:67. doi: 10.3389/fnint.2012.00067

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

Noesselt, T., Bergmann, D., Heinze, H.-J., Münte, T., and Spence, C. (2012). Coding of multisensory temporal patterns in human superior temporal sulcus. *Front. Integr. Neurosci.* 6:64. doi: 10.3389/fnint.2012.00064

Petzschner, F. H., and Glasauer, S. (2011). Iterative Bayesian estimation as an explanation for range and regression effects: a study on human path integration. *J. Neurosci.* 31, 17220–17229. doi: 10.1523/JNEUROSCI.2028-11.2011

Petzschner, F. H., Maier, P., and Glasauer, S. (2012). Combining symbolic cues with sensory input and prior experience in an iterative bayesian framework. *Front. Integr. Neurosci.* 6:58. doi: 10.3389/fnint.2012.00058

Shi, Z., Church, R. M., and Meck, W. H. (2013a). Bayesian optimization of time perception. *Trends Cogn. Sci.* 17, 556–564. doi: 10.1016/j.tics.2013.09.009

Shi, Z., Ganzenmüller, S., and Müller, H. J. (2013b). Reducing bias in auditory duration reproduction by integrating the reproduced signal. *PLoS ONE* 8:e62065. doi: 10.1371/journal.pone.0062065

Shi, Z., Jia, L., and Müller, H. J. (2012). Modulation of tactile duration judgments by emotional pictures. *Front. Integr. Neurosci.* 6:24. doi: 10.3389/fnint.2012.00024

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

Töllner, T., Strobach, T., Schubert, T., and Müller, H. J. (2012). The effect of task order predictability in audio-visual dual task performance: just a central capacity limitation. *Front. Integr. Neurosci.* 6:75. doi: 10.3389/fnint.2012.00075

Vatakis, A., Maragos, P., Rodomagoulakis, I., and Spence, C. (2012). Assessing the effect of physical differences in the articulation of consonants and vowels on audiovisual temporal perception. *Front. Integr. Neurosci.* 6:71. doi: 10.3389/fnint.2012.00071

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

Wismeijer, D. A., Gegenfurtner, K. R., and Drewing, K. (2012). Learning from vision-to-touch is different than learning from touch-to-vision. *Front. Integr. Neurosci.* 6:105. doi: 10.3389/fnint.2012.00105

Zhang, H., Chen, L., and Zhou, X. (2012). Adaptation to visual or auditory time intervals modulates the perception of visual apparent motion. *Front. Integr. Neurosci.* 6:100. doi: 10.3389/fnint.2012.00100

Received: 28 October 2013; accepted: 04 November 2013; published online: 21 November 2013.

Citation: Shi Z and Müller HJ (2013) Multisensory perception and action: development, decision-making, and neural mechanisms. Front. Integr. Neurosci. 7:81. doi: 10.3389/fnint.2013.00081

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

frontiers in
INTEGRATIVE NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 23 May 2012
doi: 10.3389/fnint.2012.00024

Modulation of tactile duration judgments by emotional pictures

***Zhuanghua Shi*[1]*, *Lina Jia*[1] *and Hermann J. Müller*[1,2]**

[1] *Department of Psychology, Experimental Psychology, Ludwig-Maximilians-Universität München, Munich, Germany*
[2] *School of Psychological Science, Birkbeck College (University of London), London, UK*

Edited by:
Micah M. Murray, Université de Lausanne, Switzerland

Reviewed by:
Hao Zhang, Purdue University, USA
Domenica Bueti, Centre Hospitalier Universitaire Vaudois and University of Lausanne, Switzerland

***Correspondence:**
Zhuanghua Shi, Department of Psychology, Experimental Psychology, Ludwig-Maximilians-Universität München, 80802 Munich, Germany.
e-mail: shi@psy.lmu.de

Judging the duration of emotional stimuli is known to be influenced by their valence and arousal values. However, whether and how perceiving emotion in one modality affects time perception in another modality is still unclear. To investigate this, we compared the influence of different types of emotional pictures—a picture of threat, disgust, or a neutral picture presented at the start of a trial—on temporal bisection judgments of the duration of a subsequently presented vibrotactile stimulus. We found an overestimation of tactile duration following exposure to pictures of threat, but not pictures of disgust (even though these scored equally high on arousal), in a short-range temporal bisection task (range 300/900 ms). Follow-up experiments revealed that this duration lengthening effect was abolished when the range to be bisected was increased (1000/1900 ms). However, duration overestimation was maintained in the short-range bisection task regardless of whether the interval between the visual and tactile events was short or long. This pattern is inconsistent with a general arousal interpretation of duration distortion and suggests that crossmodal linkages in the processing of emotions and emotional regulation are two main factors underlying the manifestation of crossmodal duration modulation.

Keywords: duration estimation, emotion, threat, visual-tactile interaction, embodiment

INTRODUCTION

Judgments of time intervals are often distorted by the emotional state a person is in. For instance, when involved in an accident, such as car crash, people often report that they felt the world slow down. Although the phenomenon has been known for long, it has only been sparsely examined (Lang et al., 1961; Hare, 1963), with more systematic studies published only in recent years (Angrilli et al., 1997; Droit-Volet et al., 2004; Droit-Volet and Gil, 2009).

The most simple and classical explanation of interval timing is provided by the internal clock model (Treisman, 1963; Gibbon et al., 1984; Zakay and Block, 1996). This model assumes an internal pacemaker that emits pulses at regular intervals, and a switch that starts and stops the counting of pulses. The pulses recorded by an accumulator represent the subjective time. Studies on emotion and time have shown that emotion can influence the internal pacemaker and/or the switch and strongly distort perceived duration (see review Droit-Volet et al., 2004). For example, Angrilli and colleagues examined duration estimation for emotional pictures, taken from the international affective picture system (IAPS) (Lang et al., 2005), presented for 2, 4, or 6 s. They found that both emotional valence and arousal were important factors in duration judgments. For high-arousal stimuli, negative pictures (e.g., mutilated bodies) were perceived as longer in duration compared to positive pictures (e.g., erotic scenes). In contrast, for low-arousal stimuli, duration of negative pictures was judged shorter than that of positive pictures (Angrilli et al., 1997). Angrilli et al. argued that two different mechanisms, one attentional and the other emotional, play important roles in time judgment. Negative events themselves engage more attentional resources (as also indicated by lowered heart rates). As a result, less attention is devoted to time processing and the negative events' durations tend to be underestimated. For high-arousal stimuli, so they argued, the effect of attention is minimized, and an emotional mechanism triggered by the pictures dominates the time estimation. Since high-arousal negative pictures evoke a defense response (Bradley et al., 2001), the duration of negative pictures is overestimated. By contrast, positive pictures evoke an approach response and thus their durations are underestimated. Similarly, other studies have shown that angry faces were judged as longer than neutral faces (Droit-Volet et al., 2004; Droit-Volet and Meck, 2007). It has been argued that both anger and fear are arousing emotions (Phelps and LeDoux, 2005), which increase the internal pacemaker rates, leading to temporal overestimation. Besides the visual modality, emotional modulation of time perception has also been found in the auditory modality (Noulhiane et al., 2007). Emotional sounds (e.g., a woman crying) were often judged as longer than neutral ones; and negative sounds were perceived as longer than positive ones (e.g., laughs).

Although there is now ample evidence of how emotion distorts duration perception, most of the studies have focused on unisensory modulation only. Given this, to date, there is still only scant understanding of how emotion-induced from one sensory modality influences time judgments in another modality. The likely reason is that emotional effects are generally (and tacitly) assumed to be amodal in nature i.e., emotional arousal or anxiety exerts a general influence, not restricted to one sensory modality.

This implicit assumption can be clearly seen in early crossmodal duration studies. For example, Hare attempted to examine how electrical shock influences auditory interval judgments (Hare, 1963). Auditory intervals were defined by two successive clicks. In the shock condition, a moderately painful (tactile) shock was delivered to participants' fingers at the second click, to induce general anxiety. Hare found that anxiety did indeed lead to a greater overestimation of auditory intervals compared to the baseline condition. However, recent crossmodal studies have provided evidence that each sensory system may possess its own clock (see review Bueti, 2011) and time processing is distributed across brain regions (Matell and Meck, 2004). The sensory-specific clock model is supported by behavioral evidence, such as for modality-specific pacemaker rates (Wearden et al., 1998; Penney et al., 2000; Droit-Volet et al., 2007), as well as by neurophysiological evidence, for example, for separate brain regions underlying visual and auditory duration processing (Ghose and Maunsell, 2002; Bueti et al., 2008; Bueti, 2011). Studies on non-emotional crossmodal duration judgments have revealed rather complex and inconclusive results (Walker and Scott, 1981; van Wassenhove et al., 2008; Chen and Yeh, 2009; Chen et al., 2010, 2011; Shi et al., 2010). For example, van Wassenhove et al. (2008) examined influences of visual (and, respectively, auditory) inputs on duration judgments of auditory (visual) events using looming and receding stimuli. They found the duration of auditory events was lengthened or shortened by the presence of conflicting visual information, while the perceived duration of visual events was unaffected by auditory stimuli. However, other studies using static stimuli or implicit measures have reported the opposite effect, i.e., perceived visual duration was affected by auditory duration (Chen and Yeh, 2009; Shi et al., 2010). Interestingly, in order to explain the crossmodal duration interaction by looming stimuli, van Wassenhove et al. (2008) suggested that salient, looming stimuli might be treated as "threat" signals (i.e., as having a negative emotional valence), causing duration dilation within and across modalities. Again, as concerns emotion, the influence of emotion on duration judgments was implicitly assumed to reflect a sense-independent arousal effect.

However, as suggested by recent discrete emotion theory (Izard and Ackerman, 2000; Mikels et al., 2005), the arousal and valence dimensions may not provide a complete description of emotions. It is also conceivable that different types of emotion link to different behavioral functions and sensory modalities. For example, although both threat and disgust are categorized as high-arousal negative-valence emotions, they activate different processes. Threat activates our defensive system and biases motor responses (Bradley et al., 2001). Given that a threatening or dangerous event is most likely directed toward our body (e.g., the sight of a snake attacking), an association between what we see and what we feel in our body can be quickly established (Poliakoff et al., 2007). This, in turn, may increase the tactile pacemaker speed and/or shorten the latency of the switch. Disgust, by contrast, is more related to avoiding something detrimental to our health or something tasting bad (Rozin and Fallon, 1987; Droit-Volet and Gil, 2009). Given this, the linkage between the visual and the tactile system by disgust events might not be as strong as that by threat events. Consequently, visual disgust signals may have only a relatively weak, if any, influence on the internal clock of the tactile system.

Moreover, duration judgments may also be influenced by the strength of perception-action associations. Research on duration estimation of emotional faces has shown that angry or fearful faces are often perceived as longer than neutral faces (Droit-Volet et al., 2004; Effron et al., 2006). However, when participants in such a study held a pen in their mouth to inhibit imitation of emotional faces, the duration lengthening was abolished (Effron et al., 2006). This finding suggests that perception-action associations are one of the critical factors causing changes of the internal clock system. Crossmodal associations induced by emotional stimuli might have a similar impact on time judgments.

To examine whether crossmodal emotional modulation of perceived duration is a general arousal effect or an emotion-specific effect, we compared modulations induced by three types of emotional pictures (threat, disgust, and neutral) on subsequent judgments of vibrotactile duration (Experiment 1). We chose threat and disgust since both are categorized as high-arousal negative emotions. If crossmodal emotional modulation reflected a general arousal effect, both types of emotional picture would engender similar distortions of tactile duration judgments. On the other hand, images depicting threat or fear may have particularly strong associations with the defensive system, compared to disgusting images. As supported by studies on affective modulation of the human startle blink (Balaban and Taussig, 1994; Stanley and Knight, 2004), blink magnitude was significantly larger during the presentation of frightening pictures compared to disgusting pictures. Thus, an alternative prediction is that threatening pictures would influence perceived duration by related sensory systems, such as touch, more than disgusting pictures would.

To further investigate the mechanisms underlying crossmodal emotional modulation of the internal clock system, we explored effects of emotions by comparing their modulatory influences between short and long tactile durations (Experiment 2) as well as short and long inter-stimulus intervals (ISIs) between the emotional picture and the vibrotactile stimulus (Experiment 3). Analogous to unimodal studies, the rationale was to examine whether the internal pacemaker rate or/and the switch latency in the tactile modality are changed by emotional events from visual modality. If the tactile pacemaker rate is impacted, one would expect a slope effect (multiplicative effect) on short and long duration judgments (Wearden, 1992, 2006), i.e., the crossmodal emotional influence should be greater for long than for short durations. By contrast, if emotion influences only the switch latency, one would expect duration overestimation for both short and long duration conditions. However, if processes of emotional regulation supersede processes of activation during a late stage of processing, one might fail to observe duration overestimation in the long duration condition. Experiments 2 and 3 were designed to examine for these effect patterns.

MATERIALS AND METHODS

PARTICIPANTS

Fourteen (six female; mean age 28), 15 (10 female; mean age 25), and 16 volunteers (10 female; mean age 25) took part in Experiments 1, 2, and 3, respectively. All participants had

normal or corrected-to-normal vision, and none reported any somatosensory disorder. Written informed consent was obtained before the experiments.

MATERIALS

The experiments took place in a sound-isolated cabin, which was dimly lit with an ambient luminance of 0.76 cd/m^2. Visual stimuli were presented on a 21 inch Sony CRT monitor with a refresh rate of 100 Hz. The viewing distance was kept constant at 57 cm using a chin-rest. Tactile vibration (250 Hz) was produced by an *AEC* TACTAID *VBW32* vibrator (Audiological Engineering Corporation; Vibrating surface 1.6 × 2.4 cm), which was fixed to the index finger of the participant's right hand. The participant was asked to place her/his right hand, behind a short black curtain, on the table in front of her/him; the curtain ensured that the participant could not see her/his hand, while she/he had a free view of the display screen. Visual and tactile stimuli presentation was controlled by a Matlab program using the Psychophysics Toolbox (Brainard, 1997).

Three types of pictures were selected from the IAPS (Lang et al., 2005): threatening pictures evoking high-arousal (such as a snake, shark, etc.); disgusting pictures also classed as high on arousal (such as a burn victim, mutilation); and neutral pictures rated "neutral" in both valence and arousal. For Experiment 1, we used five pictures of disgust (mean valence 1.69; mean arousal 6.90), five pictures of threat (mean valence 3.28; mean arousal 6.73), and 10 neutral pictures (mean valence 4.82; mean arousal 2.47). For Experiments 2 and 3, we selected 10 threatening and 20 neutral pictures. Pictures were then evenly divided into two groups, each containing five attacking (mean valence 3.3; mean arousal 6.7) and 10 neutral pictures (mean valence 4.9; mean arousal 2.7); these were assigned randomly to one or the other of two test sessions (see details in Procedure section). Descriptions and IAPS numbers of the pictures are given in **Table A1**.

PROCEDURE

We used a temporal bisection task in all experiments. Participants were first trained with two anchor tactile durations: a short vibration (S) and a long vibration (L). Then, in the subsequent test sessions, several probe durations between S and L were presented to participants, who had to indicate whether the probe duration was closer to S or to L. In Experiments 1 and 3, S and L durations were 300 and 900 ms and the probe durations were 400, 500, 600, 700, and 800 ms, respectively. In Experiment 2, there were two different ranges of temporal bisection tasks: 300/900 ms and 1000/1900 ms. For the range of 300/900 ms, S, L, and probe durations were the same as in Experiment 1. For the range of 1000/1900 ms, S and L durations were 1000 and 1900 ms and the probe durations were 1150, 1300, 1450, 1600, and 1750 ms, respectively.

In the training session, an experimenter sat beside the participant to make sure that her/his anchor discrimination performance reached perfect level. Then, the experimenter left the cabin and the test session started. In the test session, a trial started with a "go" display which contained a central blue fixation dot (subtending 0.3° of visual angle) and the blue word "Ready!" just above fixation on a gray background. After the participant pressed a button, the "go" display disappeared and a blank display was shown randomly for 300–800 ms. Then a picture, randomly chosen from selected pictures, was presented for 1 s. In Experiments 1 and 2, after a short, random ISI of 400–600 ms, a vibration was delivered to the participant's index finger for a given probe duration (see above). In Experiment 3, the ISIs between picture and vibration were fixed to be either short (500 ms) or long (1500 ms). When the vibration had terminated, a question mark was displayed on the screen prompting the participant for a response: she/he had to judge, as accurately as possible, whether the duration of the vibration was closer to S or to L and indicate the choice by pressing keys labeled "short" and "long" on the keyboard. The inter-trial interval (ITI) was set to 4–6 s, in order to avoid potential inter-trial interference. There were four blocks, each of 25 trials. At the beginning of each block, both the S and L anchor durations were presented five times each, for refreshing the participant's memory of two anchors. Participants took rests of about 1 min between blocks.

After the test session, the participant was asked to rate the valence and arousal of the pictures using a sheet of paper with 9-point scales self-assessment-manikin (SAM) (Bradley and Lang, 1994).

RESULTS

ASSESSMENT OF EMOTIONS

For Experiment 1, a repeated-measures ANOVA revealed rated valence to differ significantly among threatening, disgusting, and neutral pictures, $[F_{(2,\ 26)} = 94.08,\ p < 0.01]$. Follow-up Bonferroni *t*-tests showed that the average valence was lower for disgusting pictures compared to both threatening and neutral pictures (both $p < 0.01$), and the mean valence of threatening pictures was lower than that of neutral pictures, $p < 0.01$. A further repeated-measure ANOVA revealed rated arousal, too, to differ significantly among conditions, $[F_{(2,\ 26)} = 112.89,\ p < 0.01]$. Follow-up Bonferroni *t*-tests showed that disgusting and threatening pictures were higher in arousal ratings than neutral pictures (both $p < 0.01$), whithout a difference between the former ($p > 0.1$).

The mean valence of threatening pictures was significantly lower than that of neutral pictures, in both Experiment 2 $[F_{(1,\ 14)} = 77.79, p < 0.01]$ and Experiment 3 $[F_{(1,\ 15)} = 116.94,\ p < 0.01]$. Furthermore, repeated-measures ANOVAs revealed the mean arousal to be significantly higher for threatening than for neutral pictures, $[F_{(1,\ 14)} = 86.30, p < 0.01]$ (Experiment 2) and $[F_{(1,\ 15)} = 125.88, p < 0.01]$ (Experiment 3).

Thus, the results of the subjective ratings were consistent with the rating of valence and arousal from the IAPS.

TEMPORAL BISECTION

The proportions of "long" responses were calculated for the five probe durations and fitted by a logistic function, for each condition and each subject. The temporal bisection point (TBP) was then calculated based on the 50% point of a given estimated logistic function (Treutwein and Strasburger, 1999). To measure the sensitivity of the temporal bisection task, we estimated the just noticeable difference (JND) of the temporal bisection using the half difference in duration between the 25% and 75% point

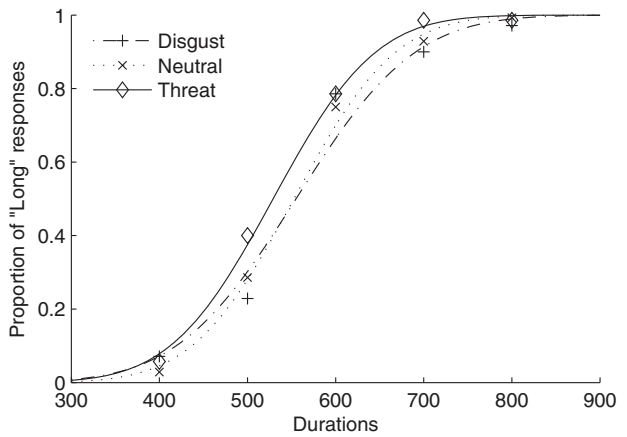


FIGURE 1 | Mean proportions of "Long" responses plotted against probe durations and fitted psychometric functions for three emotion conditions (neutral, threat, and disgust). The dotted curve and crosses represent the neutral condition, the dash-dotted curve and pluses the disgust condition, and the solid curve and diamonds the threat condition.

(Shi et al., 2008; Vroomen and Keetels, 2010). In addition, we measured the Weber fraction with the ratio of JND/TBP.

Figure 1 shows average psychometric curves for the three emotion (i.e., neutral, disgust, and threat) conditions in Experiment 1. The mean TBPs (±SE) for the tactile S/L duration pair 300/900 ms were 552 ± 14, 550 ± 19, and 529 ± 15 ms for the neutral, disgust, and threat conditions, respectively (**Figure 2A**). A repeated-measures ANOVA showed that the type of emotion picture significantly influenced the (subsequently performed) judgment of tactile duration, [$F_{(2,\ 26)} = 4.41$, $p < 0.05$]. Follow-on linear contrast tests revealed tactile TBP to be significantly lower in the threatening condition compared to both the neutral ($p < 0.01$) and disgust ($p = 0.05$) conditions, while there was no difference between the latter ($p > 0.1$). This pattern indicates that the modulatory influence of emotional pictures on tactile duration judgments was due mainly to the threatening condition. The lower TBP in this condition means that participants tended to overestimate the physical tactile duration of a vibratory stimulus preceded by a threatening picture. Interestingly, however, the subjective ratings of arousal (mean 7.41) were as high for disgusting pictures as for threatening pictures (mean 6.86). Given that duration overestimation only occurred in the threatening condition, arousal level alone cannot explain the crossmodal emotional modulation of time judgments.

The mean JNDs (±SE) were 72 ± 8, 73 ± 18, and 59 ± 12 ms for the neutral, disgust and threat conditions (**Figure 2B**). A one-way repeated-measures ANOVA on JNDs failed to reveal any difference among three types of emotion pictures, [$F_{(2,\ 26)} = 0.39$, $p = 0.68$]. The Weber fractions (±SE) were 0.13 ± 0.01, 0.14 ± 0.04, and 0.11 ± 0.02 for the neutral, disgust, and threat conditions, respectively. A repeated-measures ANOVA also indicated no difference among three conditions, [$F_{(2,\ 26)} = 0.32$, $p = 0.73$]. Both results suggested that emotional pictures did not influence the sensitivity of the subsequent tactile duration judgments.

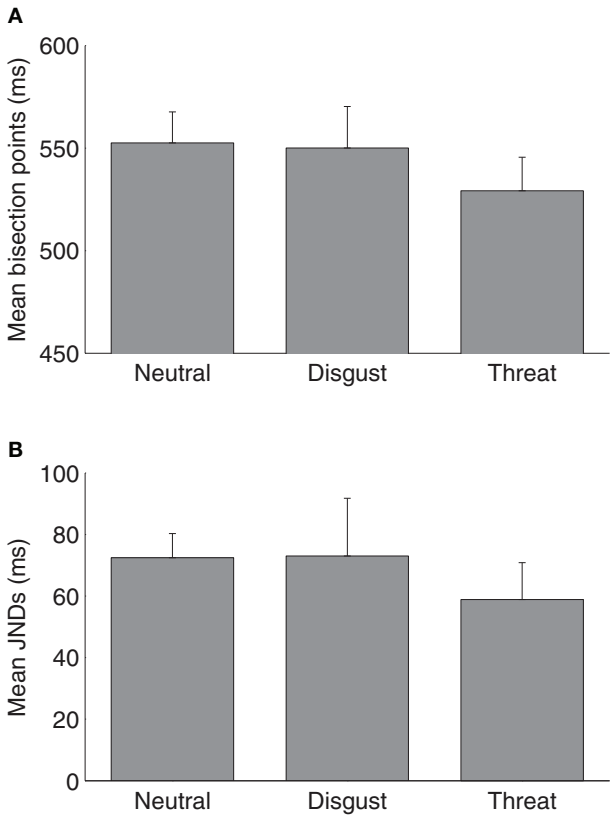


FIGURE 2 | (A) Mean temporal bisection points and **(B)** mean JNDs (and associated standard errors) for three emotional conditions ($n = 14$).

Experiment 2 was designed to examine how threatening pictures influence performance on short-range (300/900 ms) and long-range (1000/1900 ms) tactile temporal bisection tasks. In the short-range task, the mean TBPs (±SE) were 550 ± 12 and 573 ± 12 ms for the threat and neutral conditions, respectively, and in the long-range task, the points were 1399 ± 16 (threat) and 1385 ± 18 ms (neutral), respectively (**Figure 3A**). A two-way repeated-measures ANOVA with the factors temporal bisection range (300/900 vs. 1000/1900 ms) and emotional picture type (threat vs. neutral) revealed the main effect of temporal bisection range, [$F_{(1,\ 14)} = 2136.55$, $p < 0.01$], and the interaction, [$F_{(1,\ 14)} = 6.18$, $p < 0.05$], to be significant; the main effect of emotional picture type was non-significant, [$F_{(1,\ 14)} = 0.14$, $p = 0.71$]. Follow-up simple contrast tests showed that the TBP was lower with threatening pictures (indicative of a duration overestimation) in the short-range task, [$F_{(1,\ 14)} = 5.17$, $p < 0.05$], but not in the long-range task, [$F_{(1,\ 14)} = 0.71$, $p = 0.42$]. Thus, while the results from the short-range condition are consistent with those of Experiment 1, there was no evidence of crossmodal duration lengthening in the long-range condition.

The mean JNDs (±SE) were 55 ± 12 and 69 ± 9 ms for the threat and neutral conditions in the short-range task, and 119 ± 21 and 152 ± 12 ms for the threat and neutral conditions, respectively, in the long-range task (**Figure 3B**). A two-way repeated-measures ANOVA revealed that JND was larger in the

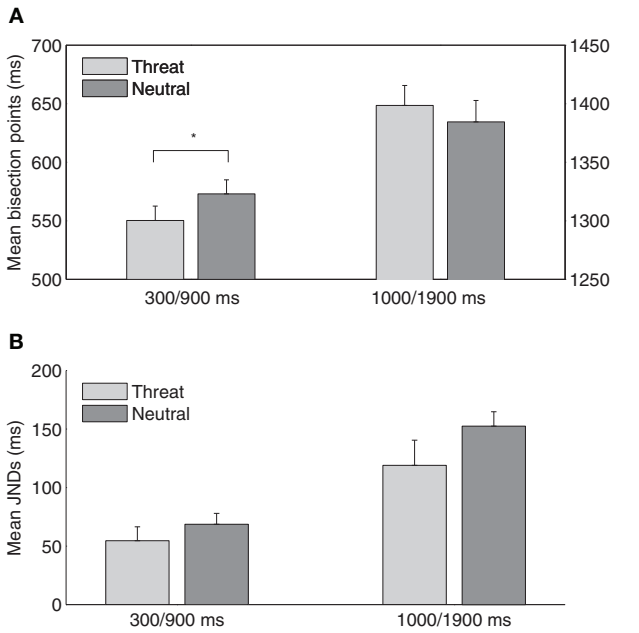


FIGURE 3 | (A) Mean temporal bisection points and **(B)** mean JNDs (and associated standard errors) plotted against threat and neutral picture conditions, for the short-range (300/900 ms) and the long-range (1000/1900 ms) sessions ($n = 15$).

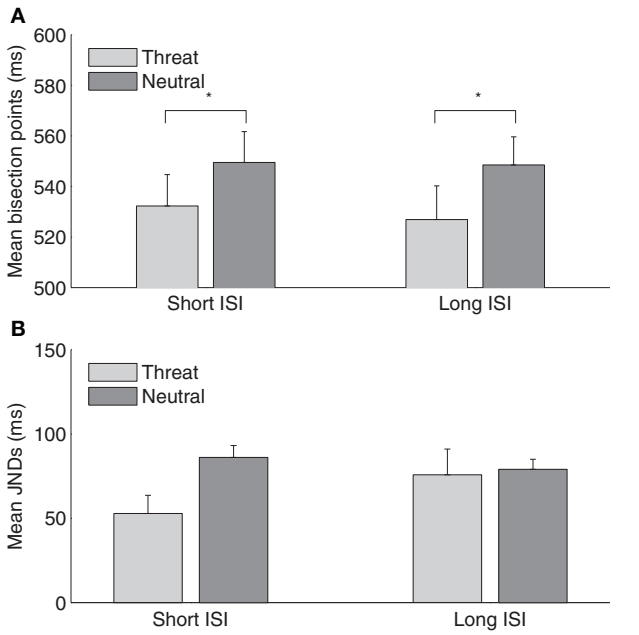


FIGURE 4 | (A) Mean temporal bisection points and **(B)** mean JNDs (and associated standard errors) plotted against threat and neutral picture conditions, for the short-ISI (500 ms) and the long-ISI (1000 ms) session ($n = 16$).

long- than the short-range task, $[F_{(1,\ 14)} = 32.61, p < 0.01]$, and was marginally smaller in the threat than the neutral condition, $[F_{(1,\ 14)} = 4.54,\ p = 0.05]$. This indicated that the threatening picture might increase the sensitivity of the temporal bisection for the subsequent tactile duration task. However, the interaction between the duration range and the emotion type was not significant, $[F_{(1,\ 14)} = 0.80,\ p = 0.39]$. To compare task difficulties, we further calculated the Weber fractions. They were 0.10 ± 0.02, 0.12 ± 0.02, 0.09 ± 0.01, and 0.11 ± 0.01 for the threat and neutral conditions in the short- and long-range tasks, respectively. A repeated-measures ANOVA revealed non-significant main effects and interaction (duration range: $[F_{(1,\ 14)} = 0.58,\ p = 0.46]$; emotion: $[F_{(1,\ 14)} = 3.62,\ p = 0.08]$; interaction: $[F_{(1,\ 14)} = 0.01, p = 0.97]$), which suggested task difficulties were relative similar among different conditions (e.g., the short- vs. long-range task).

However, it remains unclear from Experiment 2 whether the absence of a crossmodal duration overestimation following threatening pictures in the long-range temporal bisection task (1000/1900 ms) is due to the modulatory effect of emotion passively dissipating over time. Experiment 3 was designed to examine this question by comparing the effects of short (500 ms) and long (1500 ms) ISIs between the emotional picture and the tactile stimulus using the short-range temporal bisection task (300/900 ms). The intervals from the onset of the emotional picture to the offset of the tactile stimulus in the long ISI condition were then similar to that in the long-range condition (Experiment 2). **Figure 4A** depicts the mean tactile TBPs for the neutral and threat picture conditions for short and long visual-tactile ISIs, respectively. The average TBPs ($\pm$SE) were 532 ± 12 and 549 ± 11 for threatening and neutral pictures in the short-ISI condition, and 527 ± 12 (threatening) and 549 ± 10 ms (neutral) in the long-ISI condition. A two-way repeated-measures ANOVA with main terms of ISI and emotional picture type revealed the bisection points to be significantly lower in the threatening compared to the neutral condition, for both short and long visual-tactile ISIs. There were no effects involving ISI (main effect, $[F_{(1,\ 15)} = 0.26, p = 0.62]$; interaction, $[F_{(1,\ 15)} = 0.11,\ p = 0.74]$). This indicates that the modulatory effect of threatening picture in the short-range condition did not simply lessen over time, i.e., as a function of merely lengthening the ISI between the emotional picture and the tactile stimulus.

Figure 4B depicts the mean JNDs ($\pm$SE) for the neutral and threat conditions in the short and long visual-tactile ISIs. A two-way repeated-measures ANOVA revealed that JNDs were not influenced by the visual-tactile ISI, $[F_{(1,\ 15)} = 0.58,\ p = 0.46]$, but modulated by the type of pictures, $[F_{(1,\ 15)} = 4.91, p < 0.05]$. However, there was no interaction between the visual-tactile ISI and the emotion picture, $[F_{(1,\ 15)} = 3.07,\ p = 0.1]$. The significant smaller JNDs in the threat than the neutral condition confirmed the finding in Experiment 2. Both results suggest that threating pictures might increase the sensitivity of subsequent tactile temporal bisection task. Weber fractions were 0.10 ± 0.02, 0.16 ± 0.01 for the threat and neutral conditions in the short ISI and 0.15 ± 0.03, 0.15 ± 0.01 for the correspondent conditions in the long ISI. A repeated-measures ANOVA, however, showed no effects of the visual-tactile ISI, $[F_{(1,\ 15)} = 0.61,\ p = 0.45]$, the type of pictures, $[F_{(1,\ 15)} = 3.19, p = 0.10]$, and their interaction, $[F_{(1,\ 15)} = 2.48, p = 0.14]$.

DISCUSSION

The present study was designed to investigate the effect of viewing visual emotional stimuli on the subsequent estimation of the duration of non-emotional tactile events. We compared the effects of viewing three types of emotional pictures (neutral, threat, and disgust) in a short-range (300/900 ms) tactile temporal bisection task in Experiment 1. The results revealed the processing of threatening pictures to lengthen, relative to the neutral baseline, the subsequent judgments of tactile duration, as evidenced by a lowered mean TBP in the threat compared to the neutral condition. Interestingly, the lengthening effect was not simply due to the high-arousal induced by the threatening pictures: both threat and disgust pictures were rated as high in arousal negative in valence in the subjective ratings (using SAM sheets) of the participants in the present study as well as in the IAPS norms. Yet, no lengthening effect was evident in the disgust condition. This is clearly inconsistent with the predictions deriving from the assumption of a general arousal effect.

Previous studies of judged durations of emotional events themselves have shown that arousal and valences are two main factors for duration distortions (Angrilli et al., 1997; Droit-Volet et al., 2004; Noulhiane et al., 2007; Grommet et al., 2011). Using IAPS pictures, Angrilli and colleagues observed that the durations of high-arousal negative-valence pictures were overestimated (Angrilli et al., 1997). A similar effect has been reported for the auditory modality, with high-arousal negative sounds being judged as longer in duration than positive ones (Noulhiane et al., 2007). Moreover, a recent study suggests that negative high-arousal activation, such as produced by a frightening movie, can also influence the subsequent time judgment of a neutral visual event (Droit-Volet et al., 2011). However, it is not clear from those studies whether arousal activation from one modality can influence time perception in another modality. In contrast to these earlier studies on the temporal perception of emotional events themselves, in the present study, we focused on duration distortions induced by crossmodal emotional linkages. We found that viewing a rather threatening (e.g., a snake attacking), but not a disgusting (e.g., a mutilated body), picture expanded the subsequent tactile duration, although both threat and disgust emotions induced high-arousal. Our findings suggest that the crossmodal modulatory effect of emotion depends on the type of emotional stimuli. This is consistent with the "discrete emotion" theory (Izard and Ackerman, 2000; Mikels et al., 2005), which posits that different core emotions (such as disgust, fear, anger, etc.) link different behavioral functions. Studies of the affective modulation of the startle blink reflex (Balaban and Taussig, 1994; Stanley and Knight, 2004) and duration estimation of emotional faces (Droit-Volet et al., 2007; Droit-Volet and Gil, 2009) suggest that the emotion of disgust has less salience than that of threat. A threatening picture often portrays an attack signal, which invokes the anticipation (or fear) of potential damage to perceiver's body. Thus, the perceiver is put in a state in which she/he needs to react as quickly as possible to the threatening signal (e.g., fight or flight). Indeed, it has been found that automatic defense systems come into operation within an "eye blink" for biologically relevant threat events (e.g., snakes, spiders), with their activation being based mainly on preattentive coding mechanisms (Öhman, 1997; Öhman and Mineka, 2001). The threatening event also establishes a strong association between the visual and tactile modalities, as suggested by several neuro-imaging studies (Dong et al., 1994; Gray and Tan, 2002; Keysers et al., 2004; Lloyd et al., 2006). For example, posterior parietal cortex has been shown to play an important role in the early integration of visual information with somatosensory, proprioceptive signals. Lloyd and colleagues found an increase in posterior parietal cortex activity in response to observing a sharp (painful) stimulus, vs. a non-painful stimulus, touching a rubber hand in peripheral space, in the absence of any direct tactile stimulation (Lloyd et al., 2006). Consistent with reports such as these, our findings provide further behavioral evidence of visual-tactile associations elicited by threat-type emotional pictures.

The asymmetrical crossmodal modulation of duration judgments by pictures of threat versus those of disgust would also argue in favor of multiple clock models (Ivry and Richardson, 2002; Buhusi and Meck, 2005; Bueti, 2011). On this notion, time processing is "distributed" to different sensory-specific brain regions, with each of the multiple clocks operating separately. Within this framework, our results complement, rather than being in conflict with, previous, unimodal studies of emotional modulations of duration judgments. These studies have shown that the durations of emotional pictures themselves are overestimated, likely due to the "visual" clock being modulated by the pictures' arousal and valence signals. Our results go beyond this by showing that emotions induced via the visual modality may influence the "tactile" clock, depending on the strength of the emotional association induced between the visual and tactile modalities.

How does visual threat influence the tactile clock? Does exposure to threatening pictures subsequently speed up the tactile pacemaker or/and shorten the switch latency? Using a short- and a long-range temporal bisection task in Experiment 2, we observed a crossmodal duration lengthening by the threatening pictures in the short-range temporal bisection task (300/900 ms), replicating the finding of Experiment 1; by contrast, no such lengthening was observed in the long-range task (1000/1900 ms). The lacking crossmodal modulation in the long-range condition suggests that the tactile pacemaker is unlikely to be speeded up by preceding high-arousal visual stimuli. Otherwise, one would have expected to see a general slope effect, i.e., a larger duration expansion in the long-range condition. A recent study (Grommet et al., 2011) of the time estimation of visual fear cues using two different duration ranges (250/1000 ms, 400/1600 ms) concluded that the fear effects were mediated mainly by the switch latency, rather than the speeding up of the internal pacemaker. In the study of Grommet et al., the duration expansion of the fear cue itself was of a similar magnitude in both the short- and the long-range condition.

If the tactile switch latency is shortened by the presentation of threatening images in the present study, then why did we fail to observe a duration lengthening in the long-range condition? No difference on Weber fractions between the short- and long-range conditions suggests that the task difficulty cannot be the reason for the non-effect in the long-range condition. Furthermore, the mean standard errors of the TBPs were not significant different

between the short and long duration conditions (12 vs 17 ms, $p > 0.1$). This could rule out the potential cause by large variations for long duration estimations. We suggest that the absence of such an effect is due to a dynamic shifting of attention from emotional activation to emotional regulation mechanisms (Zakay, 1989; Macar et al., 1994; Casini and Macar, 1997; Fortin, 2003). Emotional activation is often followed by emotional regulation, in line with the existence of two emotional pathways, one subcortical and one cortical (LeDoux, 1995). The former is rapidly activated by potentially dangerous or survival-relevant stimuli—even though the stimuli are not fully processed, facilitating the preparation of (physiologically autonomous) response programs for avoidance (flight) or fight. The cortical pathway, by contrast, processes information more precisely, though this takes more time. Precise cortical stimulus analysis in turn can help to inhibit or correct "erroneous" early responses elicited by the subcortical pathway, thus readjusting the subsequent behavior. When participants in the present study are exposed to threatening pictures, attentional resources may first be rapidly directed to the defensive system, including the somatosensory system, for preparing a reaction. Possibly, the strong visual-tactile linkage reduces the latency of the tactile switch at the beginning. Consistent with this, tactile duration was overestimated in the short-range temporal bisection tasks of the present Experiments 1 and 2. While the same would apply to the long-range condition, participants (in this condition) would eventually realize that the tactile vibration is not a threat event. Accordingly, attentional resources would be increasingly redirected to processes of emotional regulation. As a consequence, some pulses may be lost in the time accumulation, leading to an underestimation of the tactile duration. The absence of an (overt) emotional modulation in the long-range condition may then arise from the overestimation brought about by the shortened switch latency being nulled by an underestimation owing to the emotional regulation.

It is interesting to note, however, in both short- and long-range conditions the sensitivity of temporal bisection task increased in the threat condition compared to the neutral condition. The higher sensitivity (smaller JND) in the threat condition is further confirmed in Experiment 3 and shown a trend in Experiment 1. These results may well reflect the general alerting effect induced by threatening pictures. However, the alerting could not account for the differential effects in the short- and long-range conditions.

One alternative explanation: general emotional attenuation, might account for the absence of duration lengthening in the long-range condition. As reported in previous unimodal studies (Angrilli et al., 1997; Noulhiane et al., 2007), the duration lengthening induced by emotional stimuli disappeared for the judgment of long durations (usually above 4 s). The absence of an emotion effect in these studies has been attributed to dynamic pacemaker changes by emotional attenuation: the pacemaker rate would be increased by the onset of the emotion event and would then gradually return to baseline when emotion attenuates over time. Note, however, that the emotional attenuation could also be the result of emotional regulation—which are the two faces of one and the same coin.

One interesting question, though, is at what point in time emotional regulation takes over. The results of the present Experiment 3 suggest that emotional regulation is unlikely to occur prior to the subsequent (tactile) event. Recall that in the long-ISI condition of Experiment 3, the time interval from the onset of the emotional picture to the offset of the tactile vibration was the same as that in the long duration condition of Experiment 2. If emotional regulation (or emotional attenuation) took place immediately after the onset of the emotional event, one would predict both conditions to yield the same crossmodal emotional modulation of duration judgments. However, on the opposite (and unlike the nulling effect in the long duration condition), the tactile duration lengthening effect evoked by threatening pictures was almost as large in the long-ISI as in the short-ISI condition. This suggests that the crossmodal linkage activated by threatening events was not attenuated before the subsequent event, at least within the time rage of our study (3 s). The defensive system appears to be still highly activated and dominant for reacting to the external world after the threatening events. Only when the subsequent event is identified to be non-threatening (as under the long duration condition of Experiment 2) does emotional regulation become dominant and the emotion-induced defensive bias dissipates gradually.

In summary, the present results indicate that the crossmodal subjective-duration lengthening effect is emotion-specific: tactile duration is overestimated following exposure to pictures of threat, but not to pictures of disgust of the same high-arousal potential. However, the duration lengthening disappears for long-range durations. This pattern may be best explained by the latency of the tactile (clock's) switch being shortened by crossmodal emotional activation, while emotional regulation takes over after the subsequent (tactile) event is identified as a non-threatening signal.

REFERENCES

Angrilli, A., Cherubini, P., Pavese, A., and Mantredini, S. (1997). The influence of affective factors on time perception. *Percept. Psychophys.* 59, 972–982.

Balaban, M. T., and Taussig, H. N. (1994). Salience of fear/threat in the affective modulation of the human startle blink. *Biol. Psychol.* 38, 117–131.

Bradley, M. M., Codispoti, M., Cuthbert, B. N., and Lang, P. J. (2001). Emotion and motivation I: defensive and appetitive reactions in picture processing. *Emotion* 1, 276–298.

Bradley, M. M., and Lang, P. J. (1994). Measuring emotion: the self-assessment manikin nd the semantic differential. *J. Behav. Ther. Exp. Psychiatry* 25, 49–59.

Brainard, D. H. (1997). The psychophysics toolbox. *Spat. Vis.* 10, 433–436.

Bueti, D. (2011). The sensory representation of time. *Front. Integr. Neurosci.* 5, 1–3.

Bueti, D., Bahrami, B., and Walsh, V. (2008). Sensory and association cortex in time perception. *J. Cogn. Neurosci.* 20, 1054–1062.

Buhusi, C. V., and Meck, W. H. (2005). What makes us tick? Functional and neural mechanisms of interval timing. *Nat. Rev. Neurosci.* 6, 755–765.

Casini, L., and Macar, F. (1997). Effects of attention manipulation on judgments of duration and of intensity in the visual modality. *Mem. Cognit.* 25, 812–818.

Chen, K. M., and Yeh, S. L. (2009). Asymmetric cross-modal effects in time perception. *Acta Psychol. (Amst.)* 130, 225–234.

Chen, L., Shi, Z., and Müller, H. J. (2010). Influences of intra- and crossmodal grouping on visual and tactile ternus apparent motion. *Brain Res.* 1354, 152–162.

Chen, L., Shi, Z., and Müller, H. J. (2011). Interaction of perceptual grouping and crossmodal temporal capture in tactile apparent-motion.

PLoS ONE 6:e17130. doi: 10.1371/journal.pone.0017130

Dong, W. K., Chudler, E. H., Sugiyama, K., Roberts, V. J., and Hayashi, T. (1994). Somatosensory, multisensory, and task-related neurons in cortical area 7b (PF) of unanesthetized monkeys. *J. Neurophysiol.* 72, 542–564.

Droit-Volet, S., Brunot, S., and Niedenthal, P. M. (2004). Perception of the duration of emotional events. *Cogn. Emotion* 18, 849–858.

Droit-Volet, S., Fayolle, S. L., and Gil, S. (2011). Emotion and time perception: effects of film-induced mood. *Front. Integr. Neurosci.* 5, 1–9.

Droit-Volet, S., and Gil, S. (2009). The time - emotion paradox. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 1943–1953.

Droit-Volet, S., and Meck, W. H. (2007). How emotions colour our perception of time. *Trends Cogn. Sci.* 11, 504–513.

Droit-Volet, S., Meck, W. H., and Penney, T. B. (2007). Sensory modality and time perception in children and adults. *Behav. Processes* 74, 244–250.

Effron, D. A, Niedenthal, P. M., Gil, S., and Droit-Volet, S. (2006). Embodied temporal perception of emotion. *Emotion* 6, 1–9.

Fortin, C. (2003). "Attentional time-sharing in interval timing," in *Functional and Neural Mechanisms of Interval Timing*, ed W. H. Meck (London, UK: CRC Press), 235–260.

Ghose, G. M., and Maunsell, J. H. (2002). Attentional modulation in visual cortex depends on task timing. *Nature* 419, 616–620.

Gibbon, J., Church, R. M., and Meck, W. H. (1984). Scalar timing in memory. *Ann. N.Y. Acad. Sci.* 423, 52–77.

Gray, R., and Tan, H. Z. (2002). Dynamic and predictive links between touch and vision. *Exp. Brain Res.* 145, 50–55.

Grommet, E. K., Droit-Volet, S., Gil, S., Hemmes, N. S., Baker, A. H., and Brown, B. L. (2011). Time estimation of fear cues in human observers. *Behav. Processes* 86, 88–93.

Hare, R. D. (1963). The estimation of short temporal intervals terminated by shock. *J. Clin. Psychol.* 19, 378–380.

Ivry, R. B., and Richardson, T. C. (2002). Temporal control and coordination: the multiple timer model. *Brain Cogn.* 48, 117–132.

Izard, C. E., and Ackerman, B. P. (2000). "Motivational, organizational, and regulatory functions of discrete emotions," in *Handbook of Emotions,* Vol. 2, eds M. Lewis and J. M. Haviland-Jones (New York, NY: The Guilford Press), 253–264.

Keysers, C., Wicker, B., Gazzola, V., Anton, J. L., Fogassi, L., and Gallese, V. (2004). A touching sight: SII/PV activation during the observation and experience of touch. *Neuron* 42, 335–346.

Lang, P. J., Bradley, M. M., and Cuthbert, B. N. (2005). *International Affective Picture System (IAPS): Affective Ratings of Pictures and Instruction Manual. Emotion*. Gainesville, FL: NIMH, Center for the Study of Emotion and Attention.

Lang, P. J., Wapner, S., and Werner, H. (1961). The effect of danger upon the experience of time. *Am. J. Psychol.* 74, 94–97.

LeDoux, J. E. (1995). Emotion: clues from the brain. *Annu. Rev. Psychol.* 46, 209–235.

Lloyd, D., Morrison, I., and Roberts, N. (2006). Role for human posterior parietal cortex in visual processing of aversive objects in peripersonal space. *J. Neurophysiol.* 95, 205–214.

Macar, F., Grondin, S., and Casini, L. (1994). Controlled attention sharing influences time estimation. *Mem. Cognit.* 22, 673–686.

Matell, M., and Meck, W. H. (2004). Cortico-striatal circuits and interval timing: coincidence detection of oscillatory processes. *Cogn. Brain Res.* 21, 139–170.

Mikels, J. A., Fredrickson, B. L., Larkin, G. R., Lindberg, C. M., Maglio, S. J., and Reuter-Lorenz, P. A. (2005). Emotional category data on images from the international affective picture system. *Behav. Res. Methods* 37, 626–630.

Noulhiane, M., Mella, N., Samson, S., Ragot, R., and Pouthas, V. (2007). How emotional auditory stimuli modulate time perception. *Emotion* 7, 697–704.

Öhman, A. (1997). "As fast as the blink of an eye: evolutionary preparedness for preattentive processing of threat," in *Attention and Orienting: Sensory and Motivational Processes,* eds P. J. Lang, R. F. Simons, and M. T. Balaban (Mahwah, NJ: Lawrence Erlbaum Associates), 165–184.

Öhman, A., and Mineka, S. (2001). Fears, phobias, and preparedness: toward an evolved module of fear and fear learning. *Psychol. Rev.* 108, 483–522.

Penney, T. B., Gibbon, J., and Meck, W. H. (2000). Differential effects of auditory and visual signals on clock speed and temporal memory. *J. Exp. Psychol. Hum. Percept. Perform.* 26, 1770.

Phelps, E. A., and LeDoux, J. E. (2005). Contributions of the amygdala to emotion processing: from animal models to human behavior. *Neuron* 48, 175–187.

Poliakoff, E., Miles, E., Li, X., and Blanchette, I. (2007). The effect of visual threat on spatial attention to touch. *Cognition* 102, 405–414.

Rozin, P., and Fallon, A. E. (1987). A perspective on disgust. *Psychol. Rev.* 94, 23–41.

Shi, Z., Chen, L., and Müller, H. J. (2010). Auditory temporal modulation of the visual ternus effect: the influence of time interval. *Exp. Brain Res.* 203, 723–735.

Shi, Z., Hirche, S., Schneider, W. X., and Müller, H. J. (2008). *Influence of visuomotor action on visual-haptic simultaneous perception: a psychophysical study. 2008 Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems.* IEEE, 65–70.

Stanley, J., and Knight, R. G. (2004). Emotional specificity of startle potentiation during the early stages of picture viewing. *Psychophysiology* 41, 935–940.

Treisman, M. (1963). Temporal discrimination and the indifference interval: implications for a model of the "internal clock." *Psychol. Monogr.* 77, 1–31.

Treutwein, B., and Strasburger, H. (1999). Fitting the psychometric function. *Percept. Psychophys.* 61, 87–106.

van Wassenhove, V., Buonomano, D. V., Shimojo, S., and Shams, L. (2008). Distortions of subjective time perception within and across senses. *PLoS ONE* 3:e1437. doi: 10.1371/journal.pone.0001437

Vroomen, J., and Keetels, M. (2010). Perception of intersensory synchrony: a tutorial review. *Atten. Percept. Psychophys.* 72, 871–884.

Walker, J. T., and Scott, K. J. (1981). Auditory-visual conflicts in the perceived duration of lights, tones and gaps. *J. Exp. Psychol. Hum. Percept. Perform.* 7, 1327–1339.

Wearden, J. H. (1992). Temporal generalization in humans. *J. Exp. Psychol. Anim. Behav. Process.* 18, 134–144.

Wearden, J. H. (2006). When do auditory visual differences in duration judgments occur? *Q. J. Exp. Psychol.* 59, 1709–1724.

Wearden, J. H., Edwards, H., Fakhri, M., and Percival, A. (1998). Why "sounds are judged longer than lights": application of a model of the internal clock in humans. *Q. J. Exp. Psychol. B* 51, 97–120.

Zakay, D. (1989). "Subjective time and attentional resource allocation: an integrated model of time estimation," *Time and Human Cognition: A Life-Span Perspective,* Vol. 59, eds I. Levin and D. Zakay. (Amsterdam: Elsevier), 365–397.

Zakay, D., and Block, R. A. (1996). The role of attention in time estimation processes. *Adv. Psychol.* 115, 143–164.

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: 08 February 2012; accepted: 07 May 2012; published online: 23 May 2012.

*Citation: Shi Z, Jia L and Mueller HJ (2012) Modulation of tactile duration judgments by emotional pictures. Front. Integr. Neurosci. **6**:24. doi: 10.3389/fnint.2012.00024*

APPENDIX

Table A1 | IAPS stimuli used in the current study.

Category of pictures	IAPS number	Picture description
Mutilation pictures	3030	Mutilation
	3053	Burn victim
	3060	Mutilation
	3071	Mutilation
	3120	Dead body
Animal or human attacking pictures	1052	Snake
	1120	Snake
	1201	Spider
	1300	Pit bull
	1321	Bear
	1930	Shark
	6250	Aimed gun
	6260	Aimed gun
	6300	Knife
	6510	Attack
Neutral pictures	2840	Chess
	5500	Mush room
	7000	Rolling pin
	7009	Mug
	7035	Mug
	7041	Baskets
	7050	Hair driver
	7059	Key ring
	7090	Book
	7140	Bus
	7150	Umbrella
	7161	Pole
	7185	Abstract art
	7224	File cabinets
	7233	Plate
	7235	Chair
	7490	Window
	7700	Office
	7705	Cabinet

frontiers in
INTEGRATIVE NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 26 September 2012
doi: 10.3389/fnint.2012.00082

Are the senses enough for sense? Early high-level feedback shapes our comprehension of multisensory objects

***Lorina Naci*[1,2]*, *Kirsten I. Taylor*[2,3], *Rhodri Cusack*[1,4] *and Lorraine K. Tyler*[2]**

[1] *Department of Psychology, The Brain and Mind Institute, Western University, London, ON, Canada*
[2] *Department of Experimental Psychology, Centre for Speech and Language, University of Cambridge, Cambridge, UK*
[3] *Memory Clinic-Neuropsychology Center, University Hospital Basel, Basel, Switzerland*
[4] *Medical Research Council Cognition and Brain Sciences Unit, Cambridge, UK*

Edited by:
Zhuanghua Shi, Ludwig-Maximilians-Universität München, Germany

Reviewed by:
Edmund C. Lalor, Trinity College Dublin, Ireland
Clara Suied, Institut de Recherche Biomédicale des Armées, France

***Correspondence:**
Lorina Naci, Department of Psychology, The Brain and Mind Institute, Western University, London, ON N6A 5B7, Canada. e-mail: lorina.clare@gmail.com

A key question in cognitive neuroscience is how the brain combines low-level features processed in remote sensory cortices to represent *meaningful* multisensory objects in our everyday environment. Models of visual object processing typically assume a feedforward cascade through the hierarchically organized ventral stream. We contrasted this feedforward view with an alternate hypothesis in which object processing is viewed as an interactive, feedforward and feedback process. We found that higher-order regions in anterior temporal (AT) and inferior prefrontal cortex (IPC) performed audio-visual (AV) integration 100 ms earlier than a sensory-driven region in the posterior occipital (pO) cortex, and were modulated by semantic variables (congruency), from as early as 50–100 ms. We propose that the brain represents familiar and complex multisensory objects through early interactivity between higher-order and sensory-driven regions. This interactivity may underpin the enhanced behavioral performance reported for semantically congruent AV objects.

Keywords: familiar, multisensory, integration, object, semantic, top–down, feedback

INTRODUCTION

To recognize a familiar object in our everyday environment (e.g., an animal or a tool), the brain effortlessly integrates inputs from different sensory modalities into a coherent meaningful representation. While multisensory integration responses have been consistently reported at numerous sites across the cortex, a key, unresolved question in cognitive neuroscience concerns the temporal mechanism that combines multisensory integration responses to familiar object features (e.g., the visual percept and roar of a lion) that occur across the brain into an object representation. fMRI studies have consistently reported audio-visual (AV) integration responses to complex stimuli in both auditory (A) and visual (V) sensory regions, and in higher-order anterior ventral regions. For example, regions in primary and association auditory and visual cortices (Calvert et al., 1999), traditionally thought to be sensory-specific, show AV integration responses during multisensory speech perception and object processing. AV responses to meaningful multisensory object stimuli have been reported in regions higher-up in the object processing hierarchy, including the lateral temporal (Beauchamp et al., 2004; Hein et al., 2007), anterior temporal (AT), and in particular the antero-medial temporal cortex (AMTC) (Taylor et al., 2006), prefrontal cortex (Laurienti et al., 2003), and inferior prefrontal cortex (IPC) (Hein et al., 2007). It remains unclear when AV integration responses in higher-order and sensory-driven regions interact, and whether integration responses from anterior ventral regions feed back to affect integration responses in posterior occipital (pO) regions from early stages of multisensory object processing.

Research on *visual* object recognition provides two models that may explain how multisensory integration responses in sensory-driven and higher-order regions are combined during the early stages of multisensory object processing. The traditional, feedforward model (Riesenhuber and Poggio, 2000) claims that object recognition is achieved through a feedforward, bottom–up processing cascade from sensory-driven to higher-order regions, where the evaluation of the meaning of an object is carried out at the final stages of object processing. However, an important architectural aspect of the visual system, i.e., the anatomical back projections between almost all ventral stream sites (Felleman and Van Essen, 1991), appears incompatible with a strictly feedforward view. Indeed, studies on visual object recognition (Barceló et al., 2000; Bar et al., 2006) instead support an interactive feedforward and feedback (top–down) model of object recognition, whereby processes in higher-order regions influence those in sensory-driven regions from the earliest stages, prior to recognition, and before a fine-grained meaningful representation has been achieved (Lamme and Roelfsema, 2000; Bullier, 2001; Bar et al., 2006; Clarke et al., 2011).

Nevertheless, models of the temporal dynamics of multisensory integration have been largely influenced by the feedforward view of visual object recognition. According to feedforward accounts of multisensory integration, auditory (A) and visual (V) inputs are analyzed within separate, hierarchically structured sensory processing streams, whose outputs finally become integrated in higher-order, multisensory sites (Felleman and Van Essen, 1991; Stein and Meredith, 1993). One such site is the AMTC (Simmons and Barsalou, 2003), where polymodal neurons bind

inputs from the different sensory modalities together (Murray and Richmond, 2001). Thus, this view assumes that multisensory integration responses in higher-order regions occur at later stages, i.e., after extensive processing has taken place in sensory-specific streams, and does not allow for early interactions between higher-order and sensory-driven integration responses (Felleman and Van Essen, 1991; Calvert, 2001).

The strict independence of unisensory processing has been called into question by reports of rapid AV integration responses (i.e., below 100 ms) in auditory (Foxe et al., 2000) and visual (Giard and Peronnet, 1999) cortex. Based in part on findings of direct connections between visual areas V1 and V2, and the core belt and parabelt auditory areas in the macaque monkey (Falchier et al., 2002), these early effects have been attributed to direct interactions between sensory cortices, independent of top–down triggers (Foxe and Schroeder, 2005). Thus, the feedforward, hierarchical account has been modified to allow for early interactions between unisensory cortices. The multisensory responses from the sensory-driven regions are proposed to feed forward to higher-order regions, where recognition is accomplished.

An alternative, interactive account of multisensory integration across the brain would claim that early AV integration responses, which may result from direct interactions between the sensory cortices, are modulated in a top–down fashion by ongoing AV integration responses in higher-order regions, and that these interactions occur before multisensory object recognition. To determine whether AV integration involves early top–down feedback, as suggested for visual object processing (Barceló et al., 2000; Lamme and Roelfsema, 2000; Miyashita and Hayashi, 2000; Bar et al., 2006), and what its role might be, we exploited the temporal sensitivity of EEG recordings and investigated the time-course of AV integration responses in sensory-driven and higher-order regions.

Two candidate regions for early top–down feedback exist within the ventral object processing system, and both appear critical for processing meaningful aspects of AV object stimuli: the ventral portion of the orbitofrontal cortex (OFC) located within the IPC region, and the AMTC located in the AT region. The AMTC and OFC are among the most heteromodal cortical regions, receiving afferents from all sensory modalities (Kringelbach, 2005). Both show multisensory responses in monkeys (Murray and Richmond, 2001; Romanski, 2007) and humans (Taylor et al., 2006). Importantly, activation in both the AT and IPC regions is modulated by high-level object information in monkeys (Sugase et al., 1999; Freedman et al., 2001), and by semantic variables in humans (Moss et al., 2005; Hein et al., 2007). Within the AMTC, the perirhinal cortex, located at the culmination of the occipito-temporal portion of the ventral object processing stream, is specifically involved in differentiating objects that share many properties and are therefore ambiguous (Moss et al., 2005; Barense et al., 2007; Clarke et al., 2011). The IPC is involved in processing visual object identity (Ranganath, 2006) and is thus hypothesized to represent the prefrontal extent of the object processing stream (Ungerleider and Haxby, 1994). Within the IPC, the ventral OFC plays a multifaceted role in object processing. This includes context-dependent semantic processing of objects to determine their behavioral meaning (Miller and Cohen, 2001), and context-independent processing of low visual spatial frequencies to determine the form of visual objects, starting from as early as 150 ms (Bar et al., 2006). Although the time-course of human AMTC involvement in object processing has not been investigated, findings of its direct connections with the pO cortex via the inferior longitudinal fasciculus (Catani et al., 2002) and strong bilateral connections with the OFC (Kringelbach, 2005) suggest that it may play an early top–down role in AV object processing.

To investigate the spatiotemporal profile of AV integration responses in a set of theoretically and empirically motivated regions of interest (ROIs), we performed source analyses of EEG data. Two higher-order regions, one in AT and one in IPC, were defined as the sites onto which the activity from our medial ROIs, the AMTC and OFC, respectively, would most likely be localized by the distributed source modeling method. The more lateral AT (Mummery et al., 2000) and IPC (Wagner et al., 2001) regions have inherent semantic processing capacities. In addition, we defined a sensory-driven/auditory region in the lateral superior temporal (ST) cortex, and a sensory-driven/visual region in the lateral pO cortex.

We used pairs of A, V, and AV stimuli (i.e., two image parts, two sounds, or a sound and an image) to represent familiar objects (e.g., animals and tools), and manipulated object meaning via the variable of semantic congruency. EEG data were recorded while participants made semantic congruency decisions in each unisensory (A, V) and cross-sensory (AV) trial. Stimuli in congruent trials represented the same object (e.g., a complete picture of a lion and the sound "roar"), whereas stimuli in incongruent trials represented different objects (e.g., a complete telephone picture and the sound "woosh"). By measuring responses to stimuli that could be either meaningfully integrated (congruent) or not (incongruent), we were able to evaluate each region's response to the semantic relationship between A and V stimuli, over time.

We asked two related questions. First, we tested whether the AT and IPC regions are involved in *early* stages of familiar AV object processing (<150 ms), i.e., prior to the onset of the EEG components correlated with object recognition (Johnson and Olshausen, 2003). Second, we tested whether early AV responses in AT and IPC reflected semantic processing. We predicted that semantic congruency would modulate AV integration responses in the AT and IPC regions, based on reports of AV semantic congruency effects in the AMTC (Taylor et al., 2006) and IPC (Hein et al., 2007). If the emergence of a familiar object representation is underpinned by early top–down feedback, then semantic congruency will modulate early AV integration responses in these regions.

MATERIALS AND METHODS

PARTICIPANTS

Eighteen healthy volunteers (age-range 18–40 years; 13 males) with normal or corrected-to-normal vision participated. Participants had no history of neurological disorders and did not take any psychotropic or drowsiness-inducing medication. All were right-handed, as determined by the Edinburgh Handedness Inventory (Oldfield, 1971), and gave informed consent. The study

was approved by the Cambridge Psychology Research Ethics Committee.

MATERIALS

The stimuli were naturalistic color photographs (**Figure 1**) and environmental sounds of living and non-living things (e.g., animals and tools/appliances). *All* conditions (auditory baseline, visual baseline, crossmodal condition) used concepts from the *same* living and non-living categories, i.e., animals and tools/appliances. Each category had an equal number of living and non-living things, and within each domain, an equal number of congruent and incongruent stimuli.

One of the unique contributions of this experiment, as compared to other experiments investigating cross-modal integration, is that here we look at the effects of *integration* that are unique to *cross-modal* (combined auditory and visual) stimuli, as compared to effects of integration of *unisensory* (auditory or visual) stimuli. In order to be able to investigate unisensory and cross-modal integraton within the same paradigm, we created conditions where the integration of the stimuli could take place in each sensory modality, as well as across modalities, independently of one another. To avoid any priming effects across conditions, we created stimuli that were unique within each condition.

Another motivation for the stimuli selection was to keep them the same as those used in a previous fMRI experiment (Taylor et al., 2006, *PNAS*). This would enable us to compare the effects of cross-modal integration across neuroimaging modalities (see "Discussion" section).

The unisensory visual (V) trials ($n = 100$) consisted of two picture halves, the unisensory auditory (A) trials ($n = 100$) consisted of two sound parts, and the AV trials ($n = 100$) consisted of whole pictures and whole environmental sounds. All 100 AV and all 200 unimodal stimuli were unique. In half of the trials of each condition the two stimulus parts were congruent, and in the other half of the trials, in each condition, they were incongruent. Specifically, for the congruent unisensory visual (V) trials we used two halves of the same objects (i.e., from the same image), whereas for the incongruent visual trials, we used two halves from pictures of different objects (within-domain) (e.g., congruent V: left half of a cat picture on left, right half of a cat picture on right; incongruent V: left part of a dog picture on left and right part of a cow picture on right; congruent A: the sound "jjj" followed by its other sound half "jjj"; incongruent A: the sound "moo" followed the part of another sound "clack"; congruent AV: a complete picture of a lion and the whole environmental sound "roar"; incongruent AV: a complete picture of a telephone and the whole environmental sound "woosh"). Within each congruency condition, half of the trials represented living and half non-living things.

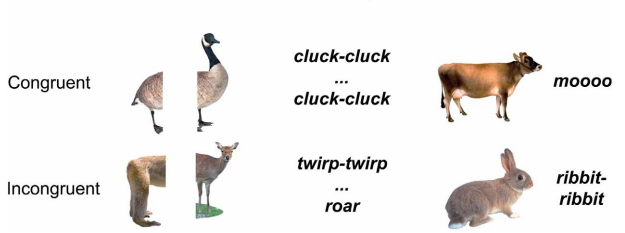


FIGURE 1 | Example stimuli belonging to the living things domain.

Critically, continuity between stimuli pairs in the incongruent trials was addressed by making sure that incongruent trials presented stimuli (images/sounds) from the same semantic category (i.e., they were both animals, or tools, etc.). Thus, we were able to avoid confounding the effects of semantic congruency with effects due to semantic domain.

Images were presented on a grey background of a 21-inch computer monitor placed 45 cm in front of the participant, and with a screen resolution of 1024 × 768 pixels and refresh rate of 60 Hz. The sounds were matched for peak amplitude (−7.7 dB). They were truncated to have the same length (1185 ms) for all the AV trials. For the A trials, the environmental sounds were divided into two halves with length of 593 ms and 592 ms. We also included control trials consisting of pairs of visual noise picture halves (Vscrambled), pink noise filtered environmental sound halves (Anoise), and visual noise whole pictures with pink noise transformed environmental whole sounds (AVnoise and scrambled) ($n = 52$ in each condition), to control for the effects of low-level visual and auditory information processing on meaningful unisensory and multisensory object integration. In order to create the Anoise stimuli, the environmental sounds were transformed into pink noise by using the "generate noise" (pink) option in the Audacity software (http://audacity.sourceforge.net/). The noise filter was applied for the entire duration of the sound. The Vscrambled stimuli were created in Photoshop (Adobe Photoshop CS5, Version 12.0 × 64) by applying the "noise" filter to each image. This filter added 100% Gaussian distributed noise to the image.

TASK

Participants were presented with an environmental sound and a picture (e.g., the sound "roar" and a picture of a lion) in the AV condition, and two parts of a sound and two parts of a picture in the unimodal A and V conditions, respectively. Participants decided, for every trial, whether the two items were congruent or incongruent by pressing different response keys. This design allowed us to isolate the processes unique to meaningful integration of object features across sensory modalities, as different from integration *per se* and associated decision-making processes, by contrasting neural responses to the AV integration conditions with the sum of the responses due to unimodal (A + V) integration.

PROCEDURE

E-Prime (Psychology Software Tools) was used to present and control the timing of the stimuli, and to communicate with the data acquisition software (Net Station; Electrical Geodesics, Inc.). In the unimodal conditions, participants were presented with two halves of stimuli. In the visual condition, the two image parts were presented simultaneously, and in the auditory condition, the two parts of the auditory object/two sounds were presented sequentially, separated by 750 ms of silence. The V and AV trials

were 1185 ms long, whereas the A trials were 1835 ms, including the silence. The stimuli were pseudo-randomly presented in fours blocks of 114 trials each. Within each block, the trial types were pseudo-randomized and the SOA jittered, between 1000 and 3200 ms. The order of block presentation was counterbalanced across subjects. Participants pressed a key to indicate whether the two stimuli were congruent or incongruent, and did not respond during the control trials. To avoid the motor response overlaying on the electrical activity due to integration processes, participants were instructed to not respond as soon they knew the answer, but, rather, to wait until the end of the trial before making a response. The resulting RT data were considered inadequate for analysis.

DATA ACQUISITION AND PRE-PROCESSING

Continuous EEG was acquired from 128 scalp electrodes (impedances $<50\,k\Omega$), band pass filtered between 0.01 and 100 Hz and digitized at 250 Hz, using a Geodesic EEG System 250 (Electrical Geodesics, Inc.). The data were band-pass filtered offline with a 0.1–40 Hz forward filter to remove low frequency drifts as well as high frequency noise, including line noise. The continuous EEG was divided into epochs from −200 ms pre- to 800 ms post-stimulus presentation. Trials contaminated by blinks and horizontal eye movements were rejected off-line on the basis of vertical and horizontal electro-oculograms. In addition, exclusion criteria for amplitude $>100\,\mu V$ and gradient $>70\,\mu V$ were used to reject trials with excessive EMG and other noise transients. Participants with artifacts in more than 20% of the object trials were excluded from further analysis ($n = 3$), to ensure adequate power in the source localization analysis. Average referenced EEG data were submitted to ERP analyses and source modeling.

ERP ANALYSIS

EEG epochs were sorted according to each condition and averaged for each subject to compute individual subject ERPs. Group averaged ERPs for each condition were calculated for display and analysis purposes. Consistent with previous studies, AV integration responses were defined as AV > (A + V). (When calculating the sum A + V, we used the second part of the auditory trial, as the unimodal auditory objects gradually unfolded in time and all the auditory stimulus information would be available during the second sound.) The stringent criterion of super-additivity [AV > (A + V)] was used to avoid false positives when measuring AV responses, or responses due to the concurrent processing and integration of A and V stimuli (Giard and Peronnet, 1999; Foxe et al., 2000; Molholm et al., 2002). The latency window and electrode sites for the visual N1 in the AV condition were defined based on the unisensory V condition, before assessing the effect of AV integration processes. The mean ERP values (N1 interval) were averaged across montages of electrodes from the left and right pO regions (four per hemisphere), and entered into repeated measures ANOVA with the factors *Hemisphere* (2), and *Condition* (2: AV, A + V).

DISTRIBUTED SOURCE MODELING

To investigate the cortical generators that underlie AV integration, and in particular to reveal the time-course of their responses, Minimum-Norm Current Estimates (MNCEs) were calculated. L2 minimum norm was computed using Brain Electric Source Analysis software (BESA 5.1, MEGIS Software GmbH, Munich). The 128 electrode positions were transformed to head coordinates using the standard BESA 5.1 brain. An idealized four-shell ellipsoidal head model (Berg and Scherg, 1994) with a radius of 92.5 mm, and scalp, skull and CSF thickness of, respectively, 6 mm, 7 mm, and 1 mm were used to calculate the EEG forward solution, before the inverse solution was computed. BESA modeled the neural activity from medial and lateral sources by projecting it on the lateral surface of the cortex. In total, there were 1426 evenly distributed regional sources (713 per hemisphere), each consisting of three orthogonally oriented dipoles, which modeled the electrical activity across the cortex at each time sample (4 ms). To account for the contribution of deep sources, the L2 minimum norm was computed for a source configuration consisting of two layers of regional sources 10 and 30% below the cortical surface. Thus, for each location on the lateral surface of the cortex, the minimum norm was computed for two regional sources below it. The larger activity of the two sources was projected onto the lateral surface of the cortex. This source placement is a standard feature of the BESA software.

ROI ANALYSES

Based on the MNCEs, ROI waveforms (group and individual data) were extracted for four ROIs bilaterally, located on the pO, ST, AT, and IPC. The particular location of the AT and IPC ROIs on the lateral surface of the cortex was chosen to optimize the detection of the response from the medial sources of interest (AMTC and OFC). ROI waveforms were computed by averaging, at each time sample, the strength of sources within the boundaries of each ROI, defined by Brodmann (BA) areas in MRIcro (www.mricro.com) (pO: BA 17, 18; ST: BA 41, 42; IPC: BA 45, 47; AT: BA 38). For statistical comparisons, the data was averaged along empirically and theoretically latency regions, based on 100 ms or 50 ms time-intervals locked to stimulus presentation, thus avoiding biasing the statistical results (Vul et al., 2009). Within each condition (A, V, and AV), the ROI activity was investigated by entering averaged ROI responses (100 ms time-bin) into repeated measures ANOVAs with factors *Time*, *Hemisphere* and *ROI*. The Huynh-Feldt correction was applied to spherical-data. Planned paired *T*-tests or independent sample *T*-tests were used to explore significant effects of ANOVA, or test *a priori* hypotheses.

REGIONAL RESPONSE ANALYSES

In the A trials, the two parts of the auditory object/two sounds were presented sequentially, separated by 750 ms of silence. The first and second parts of the auditory object (separate sounds) were averaged and analyzed independently, as the unimodal auditory objects gradually unfolded in time, and the underlying neural processes were expected to differ. Specifically, in the context of the semantic congruency task, no integration could take place during the first sound. Separate repeated measures ANOVAs with factors *Hemisphere* (2), *ROI* (3), and *Time* (4) were run on responses from each sound. Significant effects were explored

further with planned paired *T*-test comparisons of the responses in the ST to those in the AT/IPC regions. Different repeated measures ANOVA with factors *Hemisphere* (2), *ROI* (3), and *Time* (4) were performed separately for each set of Vscrambled, V, and AV trials. Significant effects were explored further with planned paired *T*-test comparisons of the responses in the pO to those in the AT/IPC regions. Regional responses were collapsed across hemispheres, to limit the number of comparisons.

ANALYSIS OF AV INTEGRATION

Similarly to the ERP analysis, the criterion of super-additivity [AV > (A + V)] was used to calculate AV integration responses in the source-localized data. As mentioned above, when calculating the sum A + V, we used the second part of the auditory trial. Fifty millisecond time-intervals were used when testing the difference between conditions [e. g., AV − (A + V)], to ensure adequate temporal resolution of subtle effects. The effect of semantic processing on AV integration responses was investigated across semantic congruency trials by comparing AV integration responses in congruent and incongruent trials [[AV congruent > (A + V) congruent] − [AV incongruent > (A + V) incongruent]]. Effects of semantic congruency were explored pre-150 ms, in two intervals (50–100 ms, 100–150 ms), determined by orthogonal analysis, with repeated measures ANOVA with factors *Hemisphere* (2), *ROI* (3), and *Time* (2 or 3). Significant effects of the ANOVAs were explored further by planned independent sample *T*-tests.

RESULTS

ERP ANALYSIS

Initially, we tested for early AV integration responses on the scalp-based visual ERPs. We found an enhancement of the visual N1 component during AV trials compared to the sum of unisensory (A + V) trials, at pO sensors, in the latency-window 150–200 ms (**Figure 2**). A repeated measures ANOVA with the factors *Hemisphere* (2) and *Condition* (2) showed a significant main effect of *Condition* [$F_{(1, 14)} = 23$, $p < 0.001$], with the AV response significantly more negative-going than the sum of unisensory responses (A + V). This finding replicates earlier reports (Giard and Peronnet, 1999; Molholm et al., 2002; Molholm, 2004).

SOURCE MODELING ANALYSIS

We then analyzed the source-localized unimodal (A/V) responses for proof of principle that the ST and pO regions were primarily driven by sensory processes, whereas the AT and IPC regions were driven by higher-order processes. Although sensory/bottom-up processes and higher-order/top–down processes may occur throughout the time-course of object processing, we expected the former to dominate early (0–200 ms), and the latter to dominate subsequent (200–400 ms) processing stages. Thus, we expected unimodal ST/pO responses to be stronger than AT/IPC responses between 0–200 ms, and the reverse to be true from 200–400 ms.

Initially, we tested regional responses during unisensory auditory integration. In this context, we refer to *auditory integration* as the process by which two sounds naturally merge into a longer, coherent auditory percept. For instance, each congruent auditory trial presented two halves of the same environmental sound (e.g., the sound "jjj" followed by its other sound half "jjj"). Therefore, in a congruent auditory trial, during the presentation of the second sound, integration occurred naturally, as the two sounds merged into one percept. By contrast, each incongruent auditory trial presented two halves of different sounds (e.g., the sound "moo" followed the part of another sound, e.g., "clack"). As a result, in an incongruent auditory trial, the two sounds clearly did not go together and did not form a coherent whole. At the end of the trial, the two sounds were still perceived as two separate items.

The responses during each sound were analyzed separately, as integration could not yet occur during the 1st sound. (We did not use the Anoise sounds in this analysis. The 1st sound served as a low-level baseline for the 2nd sound, when all the auditory information could be integrated into a familiar auditory object.) **Figure 3** displays regional response waveforms for each sound. **Figure 4** displays these responses time-binned (100 ms) for statistical analyses. For the 1st sound, repeated measures ANOVA with factors *Hemisphere* (2), *ROI* (3), and *Time* (4) showed (a) a significant effect of *ROI* and (b) a significant *ROI by Time* interaction. These were driven by large response fluctuations in the ST region compared to the relatively small changes in the AT/IPC responses over time [a: $F_{(2, 82)} = 11.7$; $p < 0.001$; b: $F_{(6, 84)} = 13$; $p < 0.001$]. Paired *T*-tests of regional responses showed (a) ST > AT and (b) ST > IPC, from 0–100 ms [a: $t_{(14)} = 4.2$; $p = 0.001$; b: $t_{(14)} = 2.8$; $p < 0.05$] (**Figure 4A**). Regional dominance was reversed during the 2nd sound. Repeated measures ANOVA with factors *Hemisphere* (2), *ROI* (3), and *Time* (4) showed (a) a significant effect of *ROI* and (b) a significant *ROI by Time* interaction. These were driven both by the decrease of ST and increase of AT/IPC responses over time [a: $F_{(2, 82)} = 10.3$; $p < 0.001$; b: $F_{(6, 84)} = 5.2$; $p < 0.001$]. Paired *T*-tests of regional responses showed (a) AT > ST, and (b) IPC > ST, from 300 to 400 ms

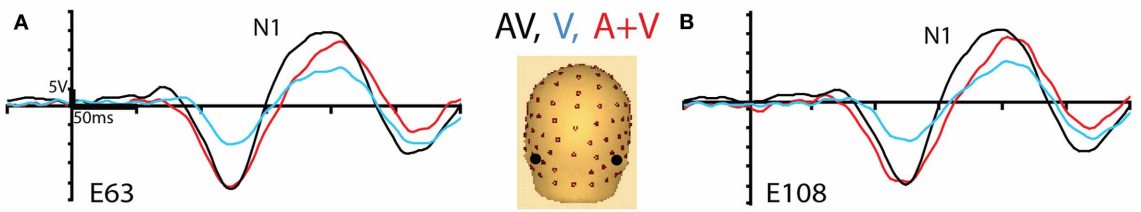


FIGURE 2 | Group averaged ERPs for AV trials (black trace), V trials (blue trace), and the sum of unisensory (A + V) trials (red trace), (A) at left and (B) right posterior occipital (pO) electrode sites, indicated in black on the 3D electrode layout.

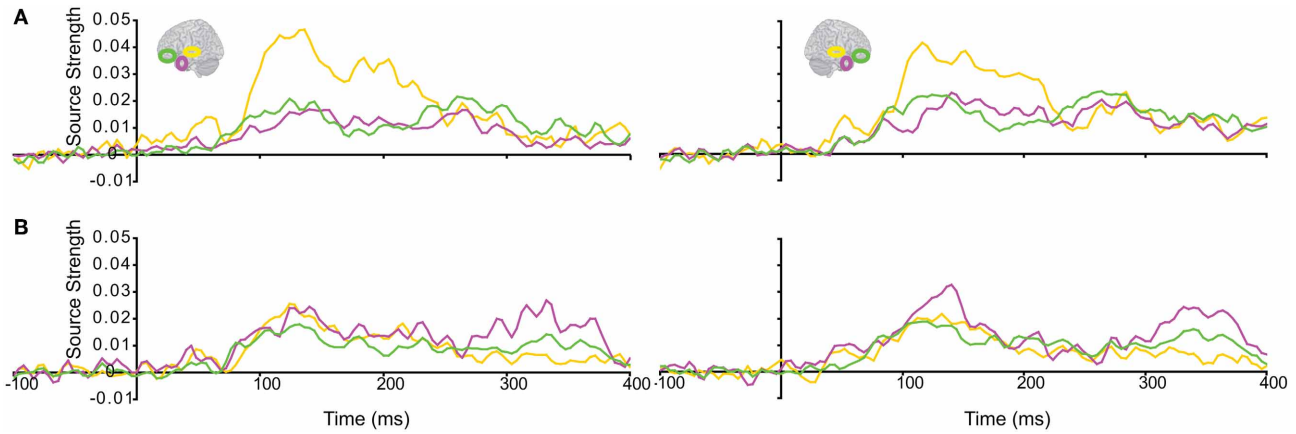


FIGURE 3 | Regional response waveforms averaged across participants during (A) the first sound, when integration could not yet occur, and (B) the second sound, when participants attempted to integrate the two sounds into a meaningful auditory object. Left/right panel displays left/right hemisphere.

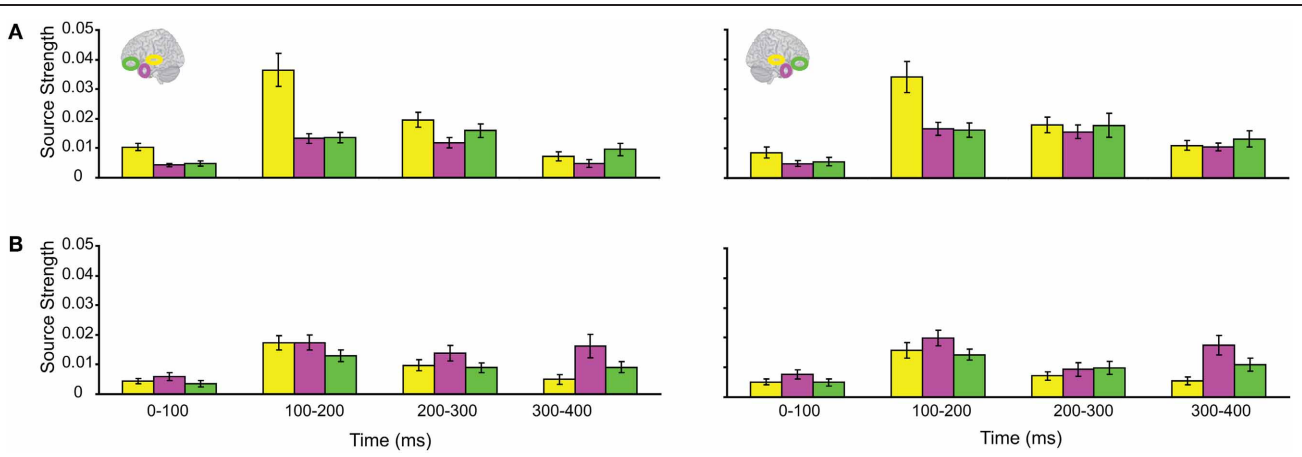


FIGURE 4 | Time-binned (100 ms) responses to auditory trials (same as Figure 3).

[a: $t_{(14)} = 4.3$; $p = 0.001$; b: $t_{(14)} = 3.2$; $p < 0.01$] (**Figure 4B**). In summary, we found dominance of ST over AT and IPC responses during the 1st sound, and the reverse effect during the later stages of the 2nd sound processing (time >200 ms), when integration between the two sounds could take place. This suggested that the ST region had greater involvement than higher-order regions in sensory processes. By contrast, the AT and IPC regions had greater involvement than sensory regions in the integration of the two sounds into a familiar auditory object.

Subsequently, we tested regional responses during unisensory visual processing. We compared responses to phase-scrambled visual stimuli, with those to intact visual objects. **Figure 5** displays regional response waveforms for scrambled objects (**Figure 5A**) and intact objects (**Figure 5B**). **Figure 6** displays these responses time-binned (100 ms) for statistical analyses. For scrambled objects, repeated measures ANOVA with factors *Hemisphere* (2), *ROI* (3), and *Time* (4) showed (a) a significant effect of *ROI* and (b) a significant *ROI by Time* interaction. These were driven by the peaking and subsiding pattern of pO responses versus the relatively small changes in the AT/IPC responses over time [a: $F_{(2,\ 82)} = 28$; $p < 0.001$; b: $F_{(6,\ 84)} = 28.2$; $p < 0.001$]. Paired *t*-tests of ROI strength showed (a) pO > AT and (b) pO > IPC, from 0–100 ms [a: $t_{(14)} = 5$; $p < 0.001$; b: $t_{(14)} = 5.4$; $p < 0.001$] (**Figure 6A**); similarly, from 100–200 ms, (a) pO > AT and (b) pO > IPC [a: $t_{(14)} = 5.3$; $p < 0.001$; b: $t_{(14)} = 4.7$; $p < 0.001$]; also, from 200 to 300 ms, (a) pO > AT and (b) pO > IPC [a: $t_{(14)} = 3.3$; $p < 0.005$; b: $t_{(14)} = 3.5$; $p < 0.005$]. By contrast, for visual objects, repeated measures ANOVA with the factors *Hemisphere* (2), *ROI* (3), and *Time* (4) showed a significant *ROI by Time* interaction. This was driven by interleaved (peaking and subsiding) responses in the pO and AT/IPC regions over time [$F_{(6,\ 84)} = 13.3$; $p < 0.001$]. Paired *t*-tests of ROI strength showed that, from 0–100 ms, (a) pO > AT, and (b) pO > IPC, and from 300–400 ms, (c) AT > pO [a: $t_{(14)} = 4.6$; $p < 0.001$; b: $t_{(14)} = 4.9$; $p < 0.001$; c: $t_{(14)} = 2.5$; $p < 0.05$] (**Figure 6B**). In summary, during scrambled object trials, we found dominance

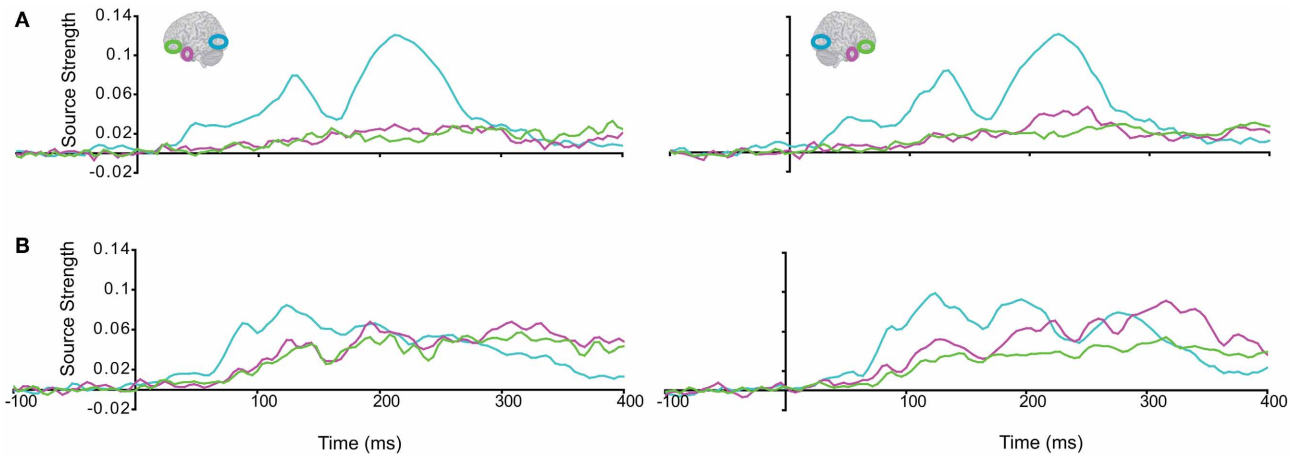


FIGURE 5 | Regional response waveforms to visual trials, averaged across participants, during (A) the Vscrambled trials, presenting phase-scrambled visual images, and (B) the Vobject trials, presenting two object (s) parts.

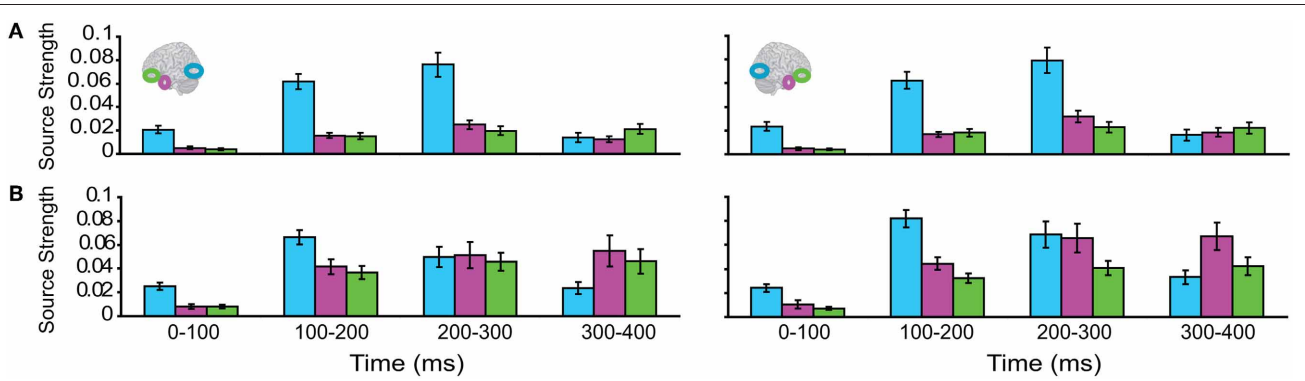


FIGURE 6 | Time binned (100 ms) responses to visual trials (same as Figure 5).

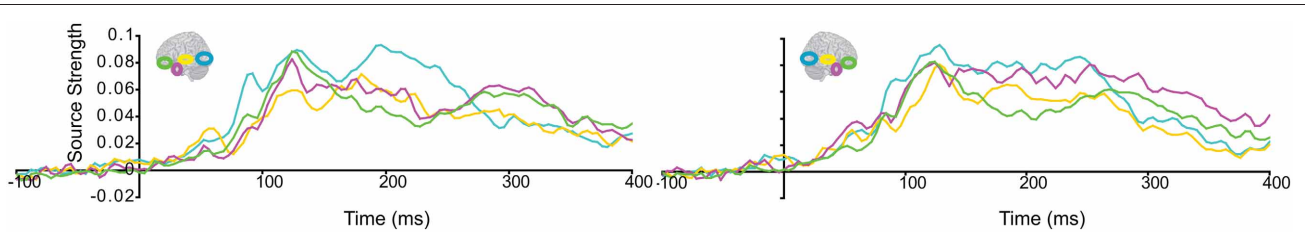


FIGURE 7 | Regional response waveforms to intact audio-visual object trials, averaged across participants.

of pO over AT/IPC responses. By contrast, during intact object trials, we found dominance of pO over AT/IPC responses, only at early stages (0–200 ms). This pattern reversed for the AT region in the latter stages (300–400 ms) of object processing. These results suggested that (similarly as for unisensory auditory trials) during unisensory visual object trials, the AT and IPC regions had greater involvement than sensory regions, when stimuli could be integrated into familiar objects (intact objects), as compared to when stimuli could not be integrated (scrambled objects). Responses during AV trials showed a similar bilateral pattern of decreasing responses in pO, accompanied by increasing responses in the AT and IPC regions, during 0–400 ms (**Figure 7**). Repeated Measures ANOVA with factors *Hemisphere* (2), *ROI* (4), and *Time* (4) showed a significant *ROI by Time* interaction. This was driven by the interleaved pattern (peaking and subsiding) of responses in the pO and AT/IPC regions [$F_{(9,\ 126)} = 5$; $p < 0.001$] (**Figure 8**). We analysed the AV trials further, to test for responses unique to AV integration [AV > (A + V)].

MULTISENSORY INTEGRATION

We then turned to our main question of whether the AT and IPC regions were involved early (pre-150 ms) in AV integration. All trials (collapsed across semantic domains and semantic congruency categories), except for the AVscrambled (noise), were used for this analysis. (AV scrambled trials were not needed, as the variable of semantic congruency was used to investigate the effect of *semantic processing* during AV integration). **Figure 9** displays regional waveforms for AV integration responses. **Figure 10** displays the time-binned (50 ms) responses used in the statistical analyses. The ST region was not included in these analyses, as it did not exhibit AV integration responses (i.e., AV $<$ A + V) (**Figures 9**, **10**). Repeated measures ANOVA on the AV integration responses with factors *hemisphere*, *ROI* (pO, AT, and IPC), and *time* (50 ms steps from 50 to 400 ms) showed a significant *ROI by time* interaction, which was driven by the sequential pattern of AV responses in the AT/IPC and pO regions [$F_{(12,\ 168)} = 3.2$; $p < 0.005$]. Planned comparisons of AV responses were performed for the anterior regions in the time-intervals 50–100 ms and 100–150 ms. We found early, bilateral AV integration responses in (a) the left AT, from 100 to 150 ms, (b) the right AT, from 50 to 100 ms, (c) the left IPC, from 50 to 100 ms and (d) the right IPC, from 50 to 100 ms ROIs [a: $t_{(14)} = 4.03$; $p = 0.001$; b: $t_{(14)} = 3.6$; $p < 0.005$; c: $t_{(14)} = 4.25$; $p = 0.001$; d: $t_{(14)} = 4.53$; $p < 0.001$] (**Figure 10**). The early AT and IPC responses cannot be explained by increased eye movements during the AV trials, as trials contaminated by blinks and eye movements were removed prior to localization analysis. We also found significant AV integration responses in the left pO region, peaking at 200–250 ms [$t_{(14)} = 2.3$; $p < 0.05$]. These regional results were corroborated by the whole-brain, unconstrained localization (L2 minimum norm) of the integration responses across the entire lateral cortical surface (top panel in **Figure 10**). In summary, AV integration responses in the AT and IPC preceded by 100 ms those in the pO region (200–250 ms), suggesting early top–down feedback from these regions during AV integration.

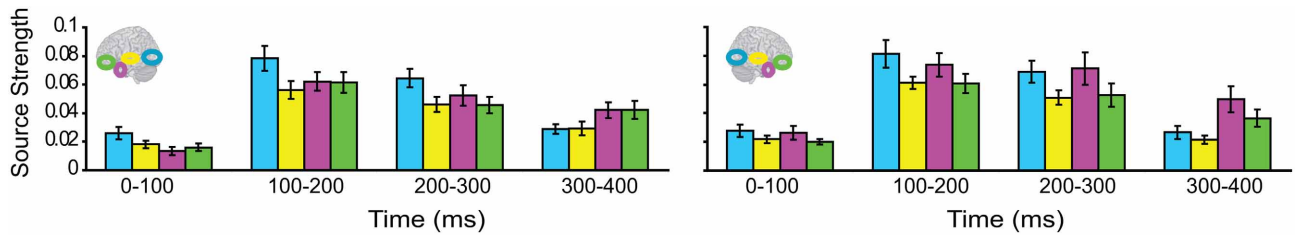


FIGURE 8 | Time binned (100 ms) responses to audio-visual trials (same as Figure 7).

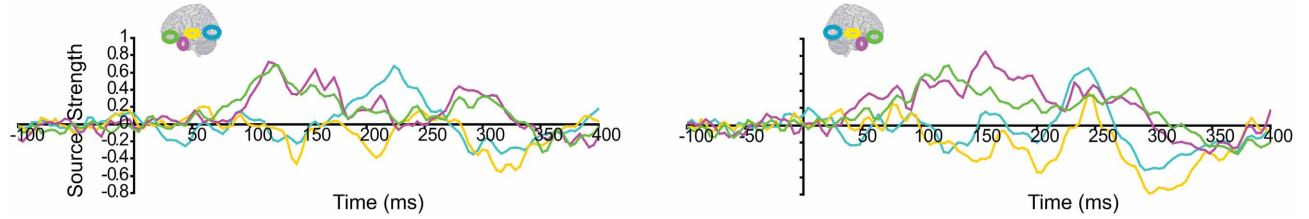


FIGURE 9 | Regional waveforms showing audio-visual integration responses [AV – (A + V)], averaged across participants, in the left and right hemispheres.

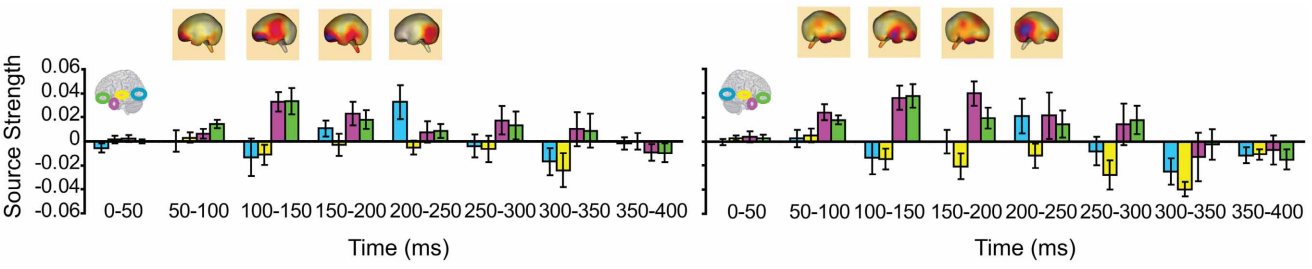


FIGURE 10 | Time-binned (50 ms) audio-visual integration responses (same as in Figure 9). The top panel shows L2 minimum norm images from the whole-brain, which represent unconstrained localization of the group-averaged responses, for each respective time-interval. Goodness of fit = 96.3%.

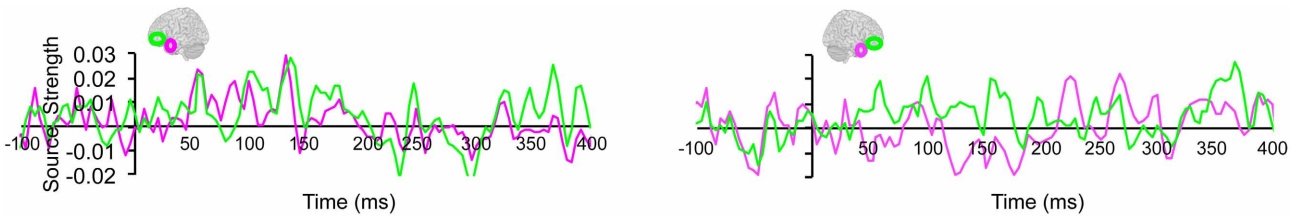


FIGURE 11 | Regional waveforms showing the difference between integration responses in congruent and incongruent AV object trials, averaged across participants, in the left and right hemispheres.

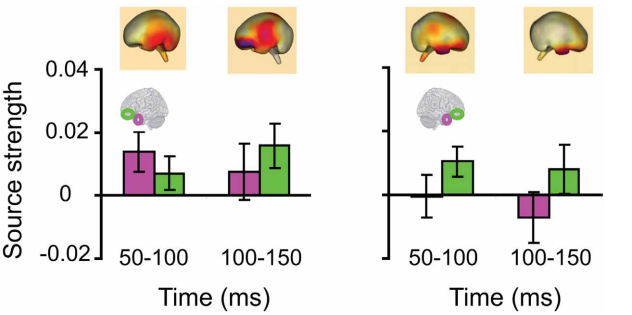


FIGURE 12 | Time-binned (50 ms) *semantic* audio-visual integration responses (same as in Figure 11), averaged across participants, at 50–100 ms and 100–150 ms. The top panel shows L2 minimum norm images from the whole-brain, which represent unconstrained localization of the group-averaged responses, for each respective time-interval. Goodness of fit = 91.8%.

SEMANTIC MULTISENSORY INTEGRATION

Lastly, we tested whether the *early* integration responses in AT and IPC regions was modulated by semantic integration. We compared integration responses in congruent and incongruent trials, during two intervals (50–100 ms, 100–150 ms), determined from the previous orthogonal analysis (**Figures 11**, **12**). Repeated measures ANOVA with factors *hemisphere*, *ROI* (AT, IPC) and *time* (50–100 ms; 100–150 ms) showed a significant main effect of *hemisphere* reflecting stronger responses in the left hemisphere [$F_{(1,\ 14)} = 4.23; p = 0.05$]. Planned post-hoc comparisons revealed significantly stronger AV integration responses for congruent than incongruent AV stimuli (a) in the left AT, between 50 and 100 ms, and (b) in the left IPC, between 100 and 150 ms [a: $t_{(14)} = 2.2; p < 0.05$; b: $t_{(14)} = 2.3; p < 0.05$]. These regional results were corroborated by the whole-brain, unconstrained localization (L2 minimum norm) of the semantic effect across the entire lateral cortical surface (top panel in **Figure 12**). This result suggested that AT and IPC regions play an early role in the semantic integration of auditory and visual object features, from as early as 50–100 ms. No early effects of semantic domain were observed.

DISCUSSION

We tested whether the AT and IPC play an early (<150 ms) role in the semantic integration of auditory and visual features of familiar objects. Initially, we replicated previous findings by observing AV enhancement of early visual (N1) event related potentials (Giard and Peronnet, 1999; Molholm et al., 2002; Molholm, 2004). Subsequent source modeling of the electrical signals revealed early AV integration responses in AT and IPC, starting from 50 to 100 ms, and preceding integration responses in the pO cortex (200–250 ms). This pattern of temporally interleaved integration responses supported an interactive account of multisensory integration, where early top–down feedback from higher-order regions biases or changes the inputs processed in the sensory-driven regions (Simons and Spiers, 2003). Although beyond the scope of this paper, determining the precise onset of these multisensory integration effects would be worthwhile investigating in future studies. The early effects (i.e., in the range 50–100 ms) should be interpreted cautiously, as their nature and functional significance may vary widely, depending the onset time (e.g., 50 ms or 80 ms).

Critically, these integration responses in the AT and IPC were modulated by semantic congruency, with enhanced responses for congruent stimuli. These early semantic effects suggested that AV integration responses in these regions differentially modulated integration in lower-level regions, as a function of the meaningfulness of the AV stimuli. Our findings agree with previous studies, which have reported semantic effects in similar regions, but have not investigated the time-course of their involvement in AV integration (Taylor et al., 2006; Hein et al., 2007). Unlike the present findings, stronger responses for incongruent than congruent trials have been reported in the IPC (Hein et al., 2007). This discrepancy may be explained by the different methodologies (fMRI vs. EEG), which do not yield directly comparable neural measures. Another important difference is the experimental task. The semantic congruency decision used here may drive stronger responses to stimuli that were felicitous with respect to the task description, than the passive viewing task used by Hein et al. (2007). The effect we observe is consistent with a context-specific role of the IPC, in particular to do with assessment of an object's meaning based on behavioral outcome (Miller and Cohen, 2001).

The early involvement of the AT region may be part of the mechanism for the rapid integration of stimuli from the auditory and visual modalities, and may underpin the enhanced behavioral performance reported for semantically congruent AV stimuli (Doehrmann and Naumer, 2008), including faster reaction times (Molholm, 2004) and improved

target detection (Molholm, 2004). The AMTC has been found to integrate object features from different sensory modalities (Murray and Richmond, 2001; Taylor et al., 2006) and is modulated by AV semantic congruency (Taylor et al., 2006). In addition, the signal strength in the rhinal region has been found to correlate with the degree of an object's familiarity (Ranganath et al., 2004), and the perirhinal cortex has been implicated in familiarity-based object recognition (Aggleton and Brown, 1999). The rapid integration of AV properties in the AT region may involve simultaneous referencing of the sensory-specific representations in lower-level regions. When the stimuli are congruent, or can be integrated into a coherent, familiar object representation, a wide network of associations, strengthened through repeated exposures to the familiar object, may be activated to support the emergence of the AV object representation.

ACKNOWLEDGMENTS

This work was supported by a graduate scholarship from the Jack Kent Cooke Foundation to Lorina Naci, a Marie Curie Intra-European and Swiss National Science Foundation Ambizione Fellowships to Kirsten I. Taylor, and a Medical Research Council Programme Grant [grant number: 75000] and a British Academy Larger Research Grant [grant number: LRG-45583] to Lorraine K. Tyler. We thank Emmanuel Stamatakis at the University of Cambridge, and Friedemann Pulvermuller, and Yuri Shtyrov at the MRC Cognition and Brain Sciences Unit, for their contributions.

REFERENCES

Aggleton, J. P., and Brown, M. W. (1999). Episodic memory, amnesia, and the hippocampal-anterior thalamic axis. *Behav. Brain Sci.* 22, 425–489.

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., Hämäläinen, M. S., Marinkovic, D. L., Schacter, D. L., Rosen, B. R., and Halgren, E. (2006). Top-down facilitation of visual recognition. *Proc. Natl. Acad. Sci. U.S.A.* 103, 449–454.

Barceló, F., Suwazono, S., and Knight, R. T. (2000). Prefrontal modulation of visual processing in humans. *Nat. Neurosci.* 3, 399–403.

Barense, M. D., Gaffan, D., and Graham, K. S. (2007). The human medial temporal lobe processes online representations of complex objects. *Neuropsychologia* 45, 2963–2974.

Beauchamp, M. S., Lee, K. E., Argall, B. D., and Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron* 41, 809–823.

Berg, P., and Scherg, M. (1994). A fast method for forward computation of multiple-shell spherical head models. *Electroencephalogr. Clin. Neurophysiol.* 90, 58–64.

Bullier, J. (2001). Feedback connections and conscious vision. *Trends Cogn. Sci.* 5, 369–370.

Calvert, G. A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cereb. Cortex* 11, 1110–1123.

Calvert, G. A., Brammer, M. J., Bullmore, E. T., Campbell, R., Iversen, S. D., and David, A. S. (1999). Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport* 10, 2619–2623.

Catani, M., Howard, R. J., Pajevic, S., and Jones, D. K. (2002). Virtual *in vivo* interactive dissection of white matter fasciculi in the human brain. *Neuroimage* 17, 77–94.

Clarke, A., Taylor, K. I., and Tyler, L. K. (2011). The evolution of meaning: spatiotemporal dynamics of visual object recognition. *J. Cogn. Neurosci.* 23, 1887–1899.

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

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

Felleman, D. J., and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1, 1–47.

Foxe, J. J., Morocz, I. A., Murray, M. M., Higgins, B. A., Javitt, D. C., and Schroeder, C. E. (2000). Multisensory auditory-somatosensory interactions in early cortical processing revealed by high-density electrical mapping. *Brain Res.* 10, 77–83.

Foxe, J. J., and Schroeder, C. E. (2005). The case for feedforward multisensory convergence during early cortical processing. *Neuroreport* 16, 419–423.

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* 291, 312–316.

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.

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.

Johnson, J. S., and Olshausen, B. A. (2003). Timecourse of neural signatures of object recognition. *J. Vis.* 3, 499–512.

Kringelbach, M. L. (2005). The human orbitofrontal cortex: linking reward to hedonic experience. *Nat. Rev. Neurosci.* 6, 691–702.

Lamme, V. A., and Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* 23, 571–579.

Laurienti, P. J., Wallace, M. T., Maldjian, J. A., Susi, C. M., Stein, B. E., and Burdette, J. H. (2003). Cross-modal sensory processing in the anterior cingulate and medial prefrontal cortices. *Hum. Brain Mapp.* 19, 213–223.

Miller, E. K., and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* 24, 167–202.

Miyashita, Y., and Hayashi, T. (2000). Neural representation of visual objects: encoding and top-down activation. *Curr. Opin. Neurobiol.* 10, 187–194.

Molholm, S. (2004). Multisensory visual-auditory object recognition in humans: a high-density electrical mapping study. *Cereb. Cortex* 14, 452–465.

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. *Brain Res.* 14, 115–128.

Moss, H. E., Rodd, J. M., Stamatakis, E. A., Bright, P., and Tyler, L. K. (2005). temporal cortex supports fine-grained differentiation among objects. *Cereb. Cortex* 15, 616–627.

Mummery, C. J., Patterson, K., Price, C. J., Ashburner, J., Frackowiak, R. S., and Hodges, J. R. (2000). A voxel-based morphometry study of semantic dementia: relationship between temporal lobe atrophy and semantic memory. *Ann. Neurol.* 47, 36–45.

Murray, E. A., and Richmond, B. J. (2001). Role of perirhinal cortex in object perception, memory, and associations. *Curr. Opin. Neurobiol.* 11, 188–193.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh handedness inventory. *Neuropsychologia* 9, 97–113.

Ranganath, C. (2006). Working memory for visual objects: complementary roles of inferior temporal, medial temporal, and prefrontal cortex. *Neuroscience* 139, 277–289.

Ranganath, C., Yonelinas, A. P., Cohen, M. X., Dy, C. J., Tom, S. M., and D'Esposito, M. (2004). Dissociable correlates of recollection and familiarity within the medial temporal lobes. *Neuropsychologia* 42, 2–13.

Riesenhuber, M., and Poggio, T. (2000). Models of object recognition. *Nat. Neurosci.* 3, 1199–1204.

Romanski, L. M. (2007). Representation and integration of auditory and visual stimuli in the primate ventral lateral prefrontal cortex. *Cereb. Cortex* 17, i61–i69.

Simons, J. S., and Spiers, H. J. (2003). Prefrontal and medial temporal lobe interactions in long-term memory. *Nat. Rev. Neurosci.* 4, 637–648.

Simmons, W. K., and Barsalou, L. W. (2003). The similarity-in-topography principle: reconciling theories of conceptual deficits. *Cogn. Neuropsychol.* 20, 451–486.

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

Sugase, Y., Yamane, S., Ueno, S., and Kawano, K. (1999). Global and fine information coded by single neurons in the temporal visual cortex. *Nature* 400, 869–873.

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.

Ungerleider, L. G., and Haxby, J. V. (1994). "What" and "where" in the human brain. *Curr. Opin. Neurobiol.* 4, 157–165.

Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspet. Psychol. Sci.* 4, 274–290.

Wagner, A. D., Paré-Blagoev, E. J., Clark, J., and Poldrack, R. A. (2001). Recovering meaning: left prefrontal cortex guides controlled semantic retrieval. *Neuron* 31, 329–338.

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: 13 May 2012; accepted: 07 September 2012; published online: 26 September 2012.

*Citation: Naci L, Taylor KI, Cusack R and Tyler LK (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*

frontiers in
INTEGRATIVE NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 17 September 2012
doi: 10.3389/fnint.2012.00077

Development of visuo-auditory integration in space and time

***Monica Gori*[1]*, *Giulio Sandini*[1] *and *David Burr*[2,3]

[1] *Robotics, Brain and Cognitive Sciences Department, Istituto Italiano di Tecnologia, Genoa, Italy*
[2] *Department of Psychology, University of Florence, Florence, Italy*
[3] *Institute of Neuroscience, National Research Council, Pisa, Italy*

Edited by:
Zhuanghua Shi, Ludwig-Maximilians-Universität München, Germany

Reviewed by:
David Alais, University of Sydney, Australia
Tino Just, University of Rostock, Germany

***Correspondence:**
Monica Gori, Robotics, Brain and Cognitive Sciences Department, Istituto Italiano di Tecnologia, via Morego 30, 16163 Genoa, Italy.
e-mail: monica.gori@iit.it

Adults integrate multisensory information optimally (e.g., Ernst and Banks, 2002) while children do not integrate multisensory visual-haptic cues until 8–10 years of age (e.g., Gori et al., 2008). Before that age strong unisensory dominance occurs for size and orientation visual-haptic judgments, possibly reflecting a process of cross-sensory calibration between modalities. It is widely recognized that audition dominates time perception, while vision dominates space perception. Within the framework of the cross-sensory calibration hypothesis, we investigate visual-auditory integration in both space and time with child-friendly spatial and temporal bisection tasks. Unimodal and bimodal (conflictual and not) audio-visual thresholds and PSEs were measured and compared with the Bayesian predictions. In the temporal domain, we found that both in children and adults, audition dominates the bimodal visuo-auditory task both in perceived time and precision thresholds. On the contrary, in the visual-auditory spatial task, children younger than 12 years of age show clear visual dominance (for PSEs), and bimodal thresholds higher than the Bayesian prediction. Only in the adult group did bimodal thresholds become optimal. In agreement with previous studies, our results suggest that also visual-auditory adult-like behavior develops late. We suggest that the visual dominance for space and the auditory dominance for time could reflect a cross-sensory comparison of vision in the spatial visuo-audio task and a cross-sensory comparison of audition in the temporal visuo-audio task.

Keywords: audio, bisection, development, integration, multisensory, space, time, visual

INTRODUCTION

Multisensory integration is fundamental for our interaction with the world. Many recent studies show that our brain is able to integrate unisensory signals in a statistically optimal fashion, weighting each sense according to its reliability (Clarke and Yuille, 1990; Ghahramani et al., 1997; Ernst and Banks, 2002; Alais and Burr, 2004; Landy et al., 2011). However, children do not integrate unisensory information optimally until late (Gori et al., 2008; Nardini et al., 2008, 2010). We recently showed that in a visual-haptic integration task (similar to that used by Ernst and Banks, 2002) children younger than 8 years of age show unisensory dominance rather than bimodal integration and the modality that dominates is task specific: the haptic modality dominates bimodal size perception and the visual modality dominates orientation bimodal perception (Gori et al., 2008). This dominance could reflect a process of cross-sensory calibration, where in the developing brain the most robust modality is used to calibrate the others (see Burr and Gori, 2011 for a discussion of this idea). It has been suggested that vision calibrates touch for orientation judgments, and touch calibrates vision for size judgments. A good deal of evidence suggests that the calibration process may be fundamental to acquire specific perceptual concepts: in particular we have shown that the impairment of the system that should calibrate the other impacts on the modality that needs calibration (Gori et al., 2010, 2012).

If the communication between sensory modalities has a fundamental role in the development of multisensory function, then we should find different forms of calibration for different dimensions, such as space and time. For example the visual system is the most accurate sense for space judgments and it should be the more influential modality for cross-modal calibration of spatial perception during development. Many studies in adults support this idea, showing that when the spatial locations of audio and visual stimuli are in conflict, vision usually dominates, resulting the so called "ventriloquist effect" (Warren et al., 1981; Mateeff et al., 1985). In adults the ventriloquist effect has been explained as the result of optimal cue-combination where each cue is weighted according to its statistical reliability. Vision dominates perceived location because it specifies location more reliably than audition does (Alais and Burr, 2004). The auditory system, on the other hand, is the most precise sense for temporal judgments (Burr et al., 2009), so it seems reasonable that it should be the more influential in calibrating the perception of temporal aspects of perception during development. In agreement with this idea, studies in adults show that when a flashed spot is accompanied by two beeps, it appears to flash twice (Shams et al., 2000). Furthermore, the apparent multiple flashes actually had lower discrimination thresholds (Berger et al., 2003). Also the apparent frequency of a flickering visual stimulus can be driven up or down by an accompanying auditory stimulus presented at a different rate (Gebhard and Mowbray,

1959; Shipley, 1964), audition dominates in audio-visual time bisection task (Burr et al., 2009), and in general audition seems to affect the interpretation of a visual stimulus also under many other conditions (e.g., see Sekuler and Sekuler, 1999; Shams et al., 2001).

All these results suggest that in the adult visual information has a fundamental role for multisensory space perception, and that audition is fundamental for temporal perception. Like adults, children are immersed in a multisensory world but, as mentioned above, unlike adults they do not integrate optimally across senses until fairly late in development, about 8 years of age (Gori et al., 2008) and some unisensory information seems to be strongly relevant for the creation of specific perceptual aspects (Gori et al., 2008, 2010, 2011; Burr and Gori, 2011; Burr et al., 2011). If the cross-sensory calibration process is necessary for development, then the auditory modality should calibrate vision in a bimodal temporal task, and the visual modality should calibrate audition in a bimodal spatial task. To test this idea we measured visual-auditory integration during development in both the temporal and the spatial domains. To compare the results between the two domains we used a bisection task both in space and in time to study the relative contributions of visual and auditory stimuli to the perceived timing and space of sensory events. For the spatial task we reproduced in 48 children and adults a child-friendly version of the ventriloquist stimuli used by Alais and Burr (2004). For the temporal task we reproduced in 57 children and adults a child-friendly version of the stimulus used by Burr et al. (2009). We also test whether and at which age the relative contributions of vision and audition can be explained by optimal cue-combination (Ernst and Banks, 2002; Alais and Burr, 2004; Landy et al., 2011).

FIGURE 1 | (A) Representation of the setup used for the temporal bisection task while a subject is tested. **(B)** Image reporting the setup used for the temporal bisection task. Three lights are presented in front and two speakers are present behind. **(C)** Representation of the setup used for the space bisection task. The blurring panel was positioned in front of the speakers so that the subject could not see the speakers behind it. For illustrative purposes this has been replaced with a transparent panel to show the speakers.

MATERIALS AND METHODS

AUDIO-VISUAL TEMPORAL BISECTION TASK

Fifty-seven children and adults performed the unimodal and bimodal temporal bisection tasks (illustrated in **Figure 2A**). All stimuli were delivered within a child-friendly setup (**Figures 1A,B**). The child was positioned in front of the setup and observed a sequence of three lights (red, green, and yellow, positioned in the nose of a clown cartoon **Figure 1B**), listened to a sequence of sounds (produced by speakers spatially aligned with the lights **Figure 1B**), or both. Three stimuli (visual, auditory, or both) were presented in succession for a total duration of 1000 ms, and the observer reported whether the middle stimulus appeared closer in time to the first or the third stimulus. To help the children to understand the task and the response, they were presented a cartoon with a schematic representation of the two possible responses to be indicated. In the visual task the subject perceived a sequence of three lights: the first one was always red, the second yellow, and the third green. The subject had to respond whether the yellow light appears closer in time to the first or the last one (**Figure 2A** upper panel). In the auditory task the subject had to respond if the second sound was presented closer in time to the first or the third one (**Figure 2A** panel in the middle). In the bimodal task the subject perceived a sequence of three lights associated with three sounds (**Figure 2A** bottom panel). The sequence of the lights presentation was identical to the visual task. The visual and the auditory stimuli could be presented in conflict or not ($\Delta = -100$; 0; 100 ms). The procedure was similar to that used by Burr et al. (2009). In the bimodal condition, all stimuli had an audio-visual conflict, where the auditory stimulus preceded or followed the visual stimulus. For the second stimulus, the conflict was Δ ms ($\Delta = -50$; 0; 50 ms), while for the first and the third stimulus the offset was inverted in sign (-Δ ms).

The visual stimuli were 1°diameter LEDs displayed for 74 ms. Auditory stimuli were tones (750 Hz) displayed for 75 ms. Accurate timing of the visual and auditory stimuli was ensured by setting system priority to maximum during stimulus presentation, avoiding interrupts from other processes (and checking synchrony by recording with microphone and light sensor). The presentation program waited for a frame-synchronization pulse then launched the visual and auditory signals. Before collecting data, subjects were familiarized with the task with two training sessions of 10 trials each (one visual and one audio). Subjects indicated after each presentation of the three stimuli whether the second appeared earlier or later than the midpoint between the first and third stimuli. We provided feedback during these training sessions so observers could learn the task and minimize errors in their responses. No feedback was given after the training sessions. During the experiment proper, five different conditions were intermingled within each session: vision only, auditory only, and three audio-visual

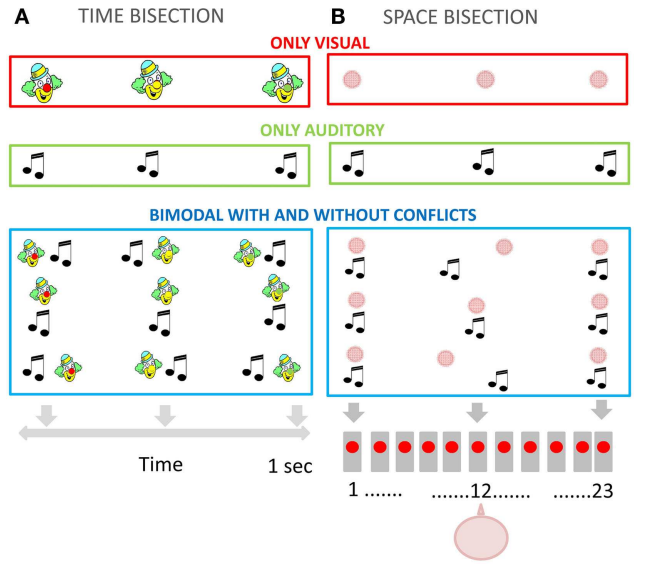


FIGURE 2 | (A) Temporal bisection task. Representation of the visual stimulation (upper panel), auditory stimulation (middle panel), and bimodal conflictual and not conflictual visual-auditory stimulation (bottom panel). **(B)** Spatial bisection task. Representation of the visual stimulation (upper panel), auditory stimulation (middle panel), and bimodal conflictual and not conflictual visual-auditory stimulation (bottom panel). The subject was aligned with the speaker in the middle (number 12).

conditions. The total single session comprised 150 trials (30 for each condition). The time of presentation of the probe was varied by independent QUEST routines (Watson and Pelli, 1983). Three QUESTs were run simultaneously in the conflict conditions (and one in each of the unisensory conditions). The timing of the second stimulus was adjusted with Quest algorithm (Watson and Pelli, 1983) to home in on the perceived point of bisection of the first and third stimuli. The timing for each trial was given by this quest estimate, plus a random offset drawn from a Gaussian distribution. This procedure ensured that the psychometric function was well sampled at the best point for estimating both the PSE and slope of the functions, as well as giving observers a few "easy" trials from time to time. Also, as the Gaussian offset was centered at zero, it ensured equal responses of closer to first and to third. Data for each condition were fitted by cumulative Gaussians, yielding PSE and threshold estimates from the mean and standard deviation of the best-fitting function, respectively. Standard errors for the PSE and threshold estimates were obtained by bootstrapping (Efron and Tibshirani, 1993). One hundred iterations of bootstrapping were used and the standard error was the standard deviation of the bootstrap distribution. All conflict conditions were used to obtain the two-cue threshold estimates. Both unimodal and bimodal (conflict or not) audio-visual thresholds and PSEs were compared with the prediction of the Bayesian optimal-integration model.

AUDIO-VISUAL SPATIAL BISECTION TASK

Forty-eight children and adults performed the unimodal and bimodal spatial bisection tasks (illustrated in **Figure 2B**). Stimuli were presented with a child-friendly setup (**Figure 1C**) which displayed a sequence of three red light, three sounds, or both. The setup comprised 23 speakers, with a red LED in front of each, which projected onto a white screen in front of the speaker array, yielding a blurred blob of 14° diameter at half height (see **Figure 1C**). The room was otherwise completely dark. The audio stimulus was identical to that used for the temporal bisection task (see previous section). The subject was seated 75 cm from the screen, causing the speaker array to subtend 102° (each speaker suspended about 4.5°). The child was positioned in front of the central speaker (number 12). Three stimuli (visual, auditory, or both) were presented in succession for a total duration of 1000 ms (identical to the duration used in the temporal bisection task), with the second stimulus occurring always 500 ms after the first. Observers reported whether the middle stimulus appeared closer in space to the first or the third stimulus (corresponding to the speakers at the extreme of the array: see **Figure 1C**).

In the unisensory visual and auditory task subjects were presented with a sequence of three lights or sounds (**Figure 2B** upper panel and panel in the middle). In the bimodal task they were presented with a sequence of three lights associated with three sounds (**Figure 2B** bottom panel). The second stimulus was presented in conflict, the standard now comprised visual and auditory stimuli positioned in different locations: the visual stimulus was the central stimulus $+\Delta^\circ$ and the auditory stimulus was the central stimulus $-\Delta^\circ$ ($\Delta = 0$ or $\pm 4.5^\circ$ or $\pm 9^\circ$). The first and the last stimuli, the auditory, and visual components were presented aligned, with no spatial conflict. The position of the second stimulus was adjusted with Quest algorithm as for the temporal task. The durations of the auditory and visual stimulations were both 75 ms.

Before collecting data, subjects were familiarized with the task with two training sessions of 10 trials each (one visual and the other audio). To facilitate the understanding of the task and the response in the training phase was presented at the child the image of two monkey cartoons (one red and one green) positioned the red on the left, in proximity of the first speaker and the green on the right, in proximity of the speaker (number 23). The child had to report if the second light was closer to the position of the red or green monkey. Subjects indicated after each presentation of the three stimuli whether the second appeared closer in space to the first or to the third stimulus. We provided feedback during these training sessions so observers could learn the task and minimize errors in their responses. No feedback was given after the training sessions.

During the experiment proper, seven different conditions were intermingled within each session: vision only, auditory only, and five two-cue conditions. The total single session comprised 210 trials (30 for each condition). As before data for each condition were fitted with cumulative Gaussians, yielding PSE and threshold estimates from the mean and standard deviation of the best-fitting function, respectively. Standard errors for the PSE and threshold estimates were obtained by bootstrapping (Efron and Tibshirani, 1993). All conflict conditions were used to obtain the bimodal threshold estimates. Both unimodal and bimodal (conflictual or not) audio-visual thresholds and PSEs were compared with the prediction of the Bayesian optimal-integration model.

In bisection tasks, there are often constant biases, particularly for temporal judgments: the first interval tends to appear longer

than the second (Rose and Summers, 1995; Tse et al., 2004). These constant biases were of little interest to the current experiment, so we eliminated them by subtracting from the estimates of each PSE the PSE for the zero conflict condition.

No children with hearing and vision impairments participated to the two tests. We excluded for data recording the children that were not able to perform correctly at least 7 of 10 trials in the training condition (in which the distance between the standard and the comparison were maximal and the test was presented in the simplest version).

BAYESIAN PREDICTIONS

The MLE prediction for the visuo-auditory threshold σ_{VA} is given by:

$$\sigma_{VA}^2 = \frac{\sigma_V^2 \sigma_A^2}{\sigma_V^2 + \sigma_A^2} \leq \min\left(\sigma_V^2, \sigma_A^2\right) \quad (1)$$

where σ_V and σ_A are the visual and auditory unimodal thresholds. The improvement is greatest ($\sqrt{2}$) when $\sigma_V = \sigma_A$.

The MLE calculation assumes also that for time and space judgments, the optimal bimodal estimate of PSE $\left(\hat{S}_{AV}\right)$ is given by the weighted sum of the independent audio and visual estimates $\left(\hat{S}_V \text{ and } \hat{S}_A\right)$.

$$\hat{S}_{VA} = w_V \hat{S}_V + w_A \hat{S}_A \quad (2)$$

Where weights w_V and w_A 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 visual and auditory judgments:

$$w_V = \frac{\sigma_A^2}{\left(\sigma_A^2 + \sigma_V^2\right)}, \; w_A = \frac{\sigma_V^2}{\left(\sigma_A^2 + \sigma_V^2\right)} \quad (3)$$

To calculate the visual and auditory weights from the PSEs (**Figure 6**), we substituted the actual spaces or times (relative to standard) into Eq. 2:

$$\hat{S}\left(\Delta\right) = \left(w_V \Delta - w_A \Delta\right) = \left(1 - 2w_A\right)\Delta \quad (4)$$

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

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

Rearranging:

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

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

The data of **Figure 5** show as a function of age the proportion of the variance of the PSE data explained by the MLE model. The explained variance R^2 was calculated by:

$$R^2 = 1 - \frac{1}{\hat{\sigma}^2 + \sigma^2} \cdot \frac{1}{N} \cdot \sum_{i=1}^{N} \left(S_i - \hat{S}_i\right)^2 \quad (7)$$

Where N is the total number of PSE values for each specific age group (all children and all values of Δ), S_i the individual PSEs for time and space, $\hat{S}_i$ is the predicted PSE for each specific condition, $\hat{\sigma}^2$ is the variance associated with the predicted PSEs and σ^2 the variance associated with the measured PSEs. $R^2 = 1$ implies that the model explains all the variance of the data, $R^2 = 0$ implies that it does no better (or worse) than the mean, and $R^2 < 0$ implies that the model is worse than the mean.

RESULTS

Figure 3 reports the PSEs for both temporal bisection (**Figure 3A**) and space bisection (**Figure 3B**). In both Figures we adjusted the PSEs for constant errors in bias by subtracting for each conflictual PSE the PSE obtained in the not conflictual condition. In the temporal bisection task (**Figure 3A**), PSEs tend to follow the green line, suggesting auditory dominance over vision. As may be expected, the results for the 5–7 age-group are noisier than the others, but the tendency is similar at all ages, particularly the older age-groups. In the audio-visual spatial bisection task (**Figure 3B**) PSEs follow the visual standard (indicated by the red line) especially until 12 years of age.

To observe how much this behavior is predicted by the MLE model, we plotted in **Figures 4A,B** the PSEs measured against the PSEs predicted by the Bayesian model (Eq. 2). Superimposition of the dots on the black line (equality line) would suggest that the behavior of the group is well predicted by the Bayesian model. From this graph we can observe that for the temporal bisection task (**Figure 4A**) the behavior becomes adult-like at about 8–9 years of age when the dots lie close to (but not entirely superimposed on) the equality black line as occurs in the adult groups. On the other hand, for the space bisection task, the dots lie on the equality line only in the adult group (**Figure 4B**).

Figure 5 summarizes how visuo-auditory integration develops with age. It plots the amount of variance (R^2) in PSEs explained by MLE model. A value of 1 means that all the variance was explained by the model, 0 that the model performed as well as the mean, and less than 0 that it performed worse than the mean (see Eq. 7). For both the spatial and temporal tasks, the MLE model explains a large proportion of the variance at all ages except the youngest (6-year-olds). For both space and time in the 6 years old group $R^2 \simeq 0$, suggesting that the model performed as well as the mean. The 8-year-old group shows a larger proportion of explained variance ($R^2 > 0.5$) but interestingly, there is a dip in the curve at 10–12 years showing less explained variance, especially for the space bisection test ($R^2 < 0.5$). In the adult group a larger amount of variance is explained by the MLE model in the space bisection task than in the time bisection task suggesting better integration for the first task.

We then calculated the audio and visual weights required for the Bayesian sum (Eq. 2), separately from the estimates of PSEs (Eqs 4–6) and from the estimates of unimodal thresholds (Eq. 3). The results are plotted in **Figure 6**, showing auditory weights on the left ordinate and visual weights on the right (the two sum to unity). In general, for the time bisection (**Figure 6A**), the auditory weight for the PSE was more than that predicted by thresholds (points tend to fall to the right of the bisector). This occurred at all ages, but was clearest for the adults. Conversely, for the space bisection

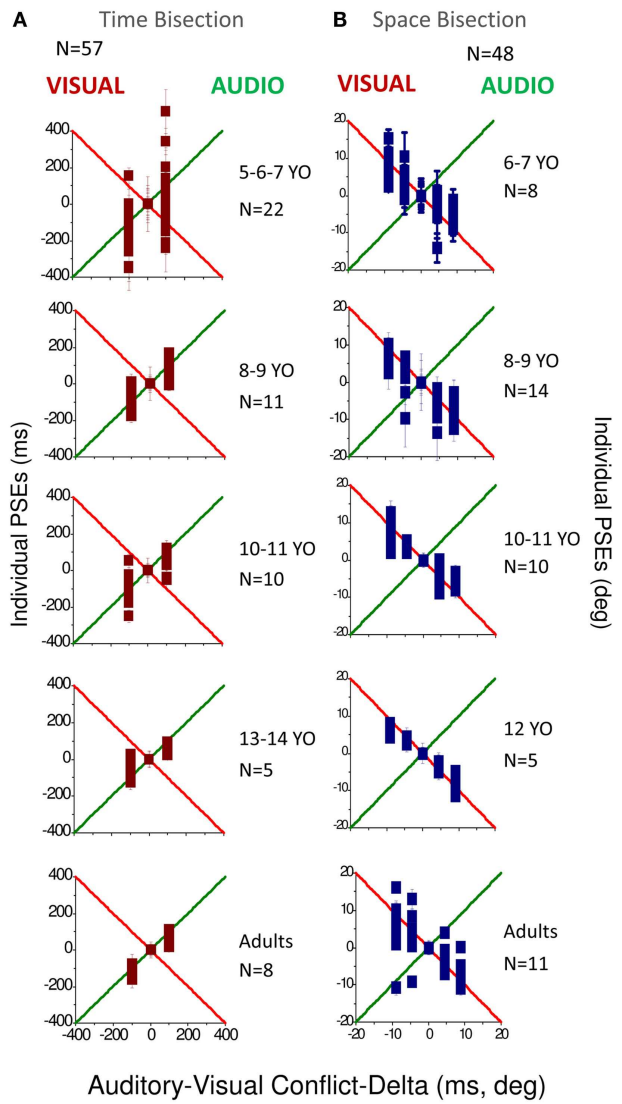


FIGURE 3 | (A) PSEs measured for the different conflictual condition in the temporal bisection task. **(B)** PSEs measured for the different conflictual condition in the spatial bisection task. In both panels the green line represents total auditory dominance and the red line total visual dominance. Different ages are reported in different panels. The number of subjects who participated is indicated in each panel for each age and condition.

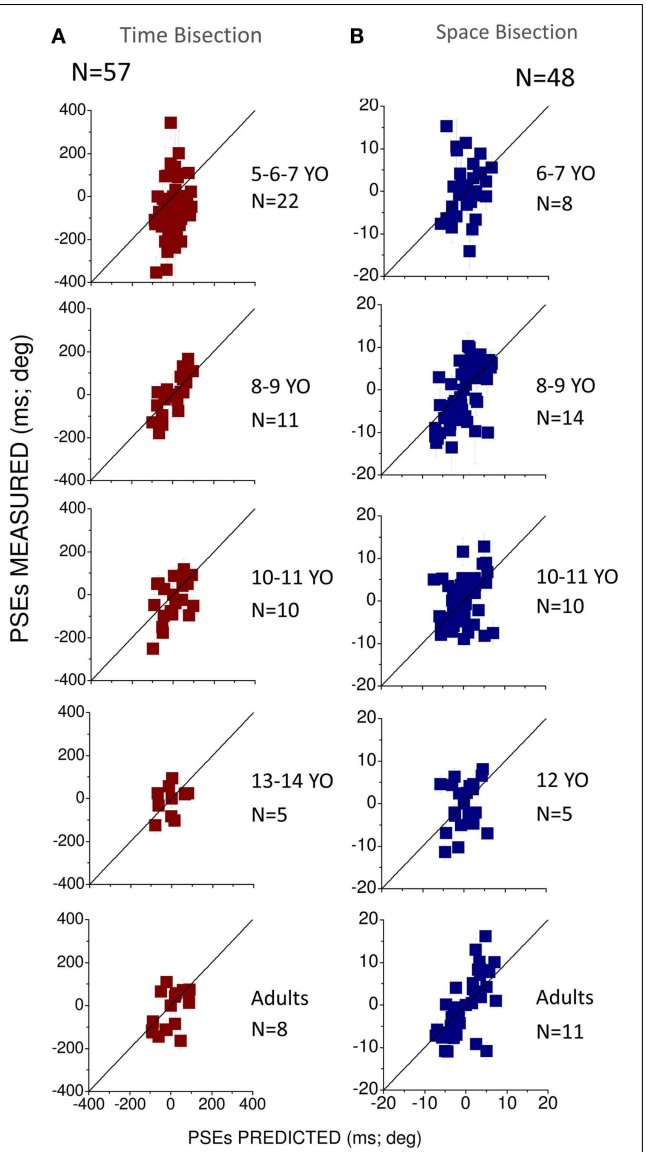


FIGURE 4 | (A) Measured against predicted PSEs for the different conflictual conditions in the temporal bisection task. **(B)** Measured against predicted PSEs for the different conflictual condition in the spatial bisection task. In both panels the black line represents the prediction of the Bayesian model and suggests optimal integration. Different ages are reported in different panels. The number of subjects who participated is indicated in each panel for each age and condition.

(**Figure 6B**), the PSE has less auditory weight (more visual weight) than predicted by thresholds until adulthood.

Figure 7 plots average theoretical auditory and visual weights as a function of age: gray lines show the MLE-predicted weights (Eq. 3), and blue lines the weights calculated from the PSE vs. conflict functions (Eq. 6). These graphs tell a similar story to **Figure 6**. For temporal judgments (**Figure 7A**), the PSEs show a greater auditory weight than predicted by thresholds while for spatial judgments (**Figure 7B**) the PSEs show a greater visual weight than predicted. The only exception is the spatial judgments for adults, where PSE and thresholds estimates are very similar (both heavily biased toward vision).

The strong test of optimal integration is an improvement in bimodal thresholds (given by the standard deviation of the cumulative Gaussian fits). **Figure 8** shows the results. For the temporal bisection task (blue dots in **Figures 8A–C**), the improvement in thresholds for bimodal presentations was less than predicted at all ages (see stars in **Figure 8C** and caption), if compared with the Bayesian prediction (gray symbols in **Figures 8A–C**). In the youngest group of children (5–7 years of age), bimodal thresholds follow the poorer modality (the visual one, red and blue dots in **Figure 8A**). Interestingly, at this age the bimodal PSEs also are

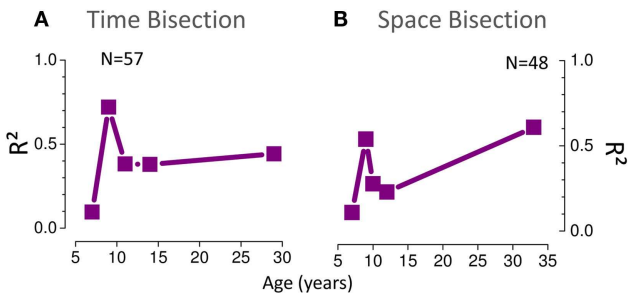


FIGURE 5 | (A) Proportion of variance (R^2) of the PSE data (**Figure 3**) for the time bisection task explained by the MLE model. A value of 1 means that all the variance was explained by the model, 0 that the model performed as well as the mean, and less than 0 that it performed worse than the mean (see Eq. 7). **(B)** The same for the space bisection task.

much noisier than the older groups (see **Figure 4A**). After 7 years of age, when also PSEs become less noisy and adult-like, bimodal thresholds become identical to the auditory thresholds and remain equal to the auditory one also in the older groups (green dots in **Figure 8A**). Also for the space bisection task, PSEs and thresholds show related behaviors: when PSEs show less inter-subject variability (in the adult group), the bimodal thresholds become well predicted by the Bayesian model (blue and gray dots in **Figure 8B**, see stars in **Figure 8D**). In the younger groups they follow the poorer sense (the auditory one, blue and green dots in **Figure 8B**).

DISCUSSION

AUDIO-VISUAL SPACE AND TIME BISECTION IN ADULTS

In this study we investigated audio-visual integration in space and in time perception during development. The goal was to examine the roles of the visual and auditory systems in the development of spatial and temporal aspects. To compare these two aspects, similar tasks were used to study space and time, requiring subjects to bisect temporal or spatial intervals. In adults, optimal multisensory integration, which has been reported for many tasks (Clarke and Yuille, 1990; Ghahramani et al., 1997; Ernst and Banks, 2002; Alais and Burr, 2004; Landy et al., 2011), is not evident in our temporal bisection task at any age tested and is evident in our spatial bimodal task only for the adult group. The absence of integration obtained in our temporal task is in agreement with other studies (e.g., Tomassini et al., 2011) that show that multisensory integration is sub-optimal also for a visual-tactile time reproduction tasks. It is also in agreement with previous studies that show auditory dominance over vision rather than optimal integration in adults (Shams et al., 2000; Burr et al., 2009) for temporal localization. In particular, Burr et al. (2009) examined audio-visual integration in adults using a bisection task (similar to the one we used), and found that sound does tend to dominate the perceived timing of audio-visual stimuli. Our stimulus is for the most part similar to the stimulus used by Burr et al. (2009) with few exceptions. One difference was the larger temporal conflicts and the fact that all the three stimuli presented in the conflictual conditions contained conflict information, while in the Burr et al. (2009) stimuli the conflict was only in the first and last stimuli. Overall, if

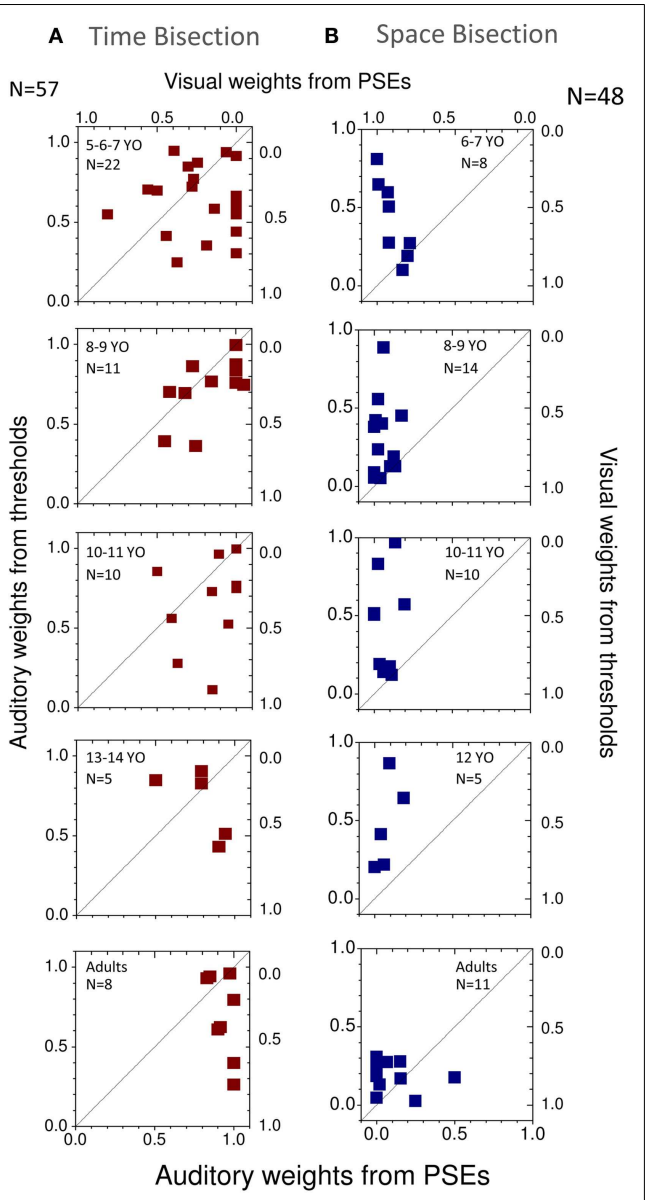


FIGURE 6 | (A) Individual weights predicted from thresholds plotted against those predicted from PSEs for different ages for the time bisection task. The black line shows the equality line. **(B)** Same for the space bisection task.

some differences between these two experiments were present, our results are mostly in agreement with those of Burr et al. (2009), particularly for the fact that auditory dominance of PSEs was not well predicted by the Bayesian model, with more weight to audition than predicted from thresholds. This audio dominance can be specific to the audio stimulus used. Burr et al. (2009) reported that bimodal prediction of thresholds was less successful for higher auditory tones (1700 Hz) than for lower tones (200 Hz) and in agreement with this finding we found auditory dominance rather than optimal integration by using a high auditory tone (750 Hz).

Our results on audio-visual space integration in adults agree well with previous studies. Like Alais and Burr (2004), we found

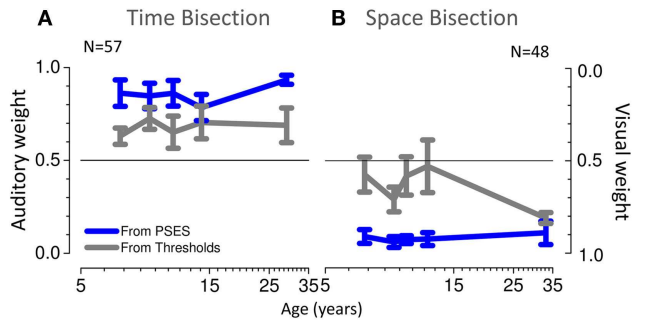


FIGURE 7 | (A) Average weights as a function of age, predicted from thresholds in gray and from PSEs in blue, for the time bisection task. **(B)** Same for the space bisection task.

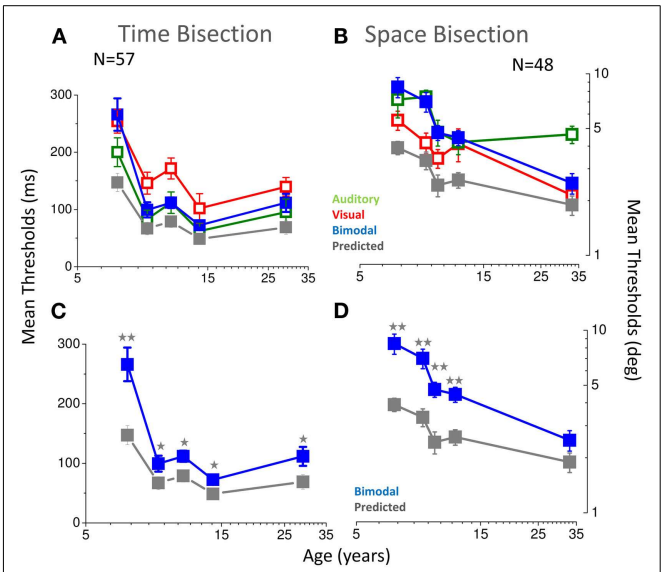


FIGURE 8 | (A) Thresholds as a function of age for the temporal bisection task. Visual thresholds are reported in red, auditory in green, bimodal in blue, and predictions of the Bayesian model in gray. **(B)** Same for the space bisection task. **(C)** Same as A, showing for clarity only bimodal thresholds (blue) and Bayesian prediction (gray). **(D)** Same as C for the space bisection task. In all cases, two stars represent a significance level of less than 0.01 and one star a significance level of less than 0.05 in a one tailed one sample *t*-tests.

optimal integration of bimodal thresholds, shown by an increment in precision compared with the unisensory performances. Both visual and multisensory thresholds (considering a similar visual blurred condition) were similar to those obtained by Alais and Burr (2004). Our auditory thresholds were better than those obtained by Alais and Burr (2004), possibly because of the different audio stimulation. Indeed in their experiment the audio stimulus was defined by only one cue (interaural timing difference), while our stimuli were real speakers in space, thereby providing many cues to localization, binaural and monaural. On the other hand our results suggest sub-optimal integration for PSEs, for which the proportion of the variance of the PSEs data is not completely explained by the MLE model (see **Figure 5**) and the weights predicted from thresholds are not completely superimposed to those computed from PSEs (see **Figure 7**). A possible explanation for this difference could be that the task in our experiment was a bisection task rather than the discrimination task as used by Alais and Burr (2004). Another difference could be that Alais and Burr's subjects were trained extensively on the auditory task and were instructed to attend to both visual and auditory aspects of the stimuli. Given the limited time available to test children (and not wanting differences between children and adults), all subjects had the same 20 trials of training without particular attention to the auditory or bimodal aspects.

AUDIO-VISUAL SPACE AND TIME BISECTION IN CHILDREN

In agreement with our previous results (Gori et al., 2008), we found that for both tasks the bimodal adult-like behavior emerges only late in development. For the time bisection the adult-like behavior occurs after 8 years of age while for the space bisection task, it was fully mature only in our adult group. Like the visual-haptic studies (Gori et al., 2008), children show strong unisensory dominance rather than multisensory integration of audio and visual space and time perception. In the child, audition dominates visual-auditory time perception and vision dominates visual-auditory space perception. This result is in agreement with our prediction and in line with our cross-sensory calibration theory (Burr and Gori, 2011). The auditory dominance can reflect a process of cross-sensory calibration in which the auditory system could be used to calibrate the visual sense of time since it is the most accurate sense for temporal judgments. This result is also in agreement with many experiments performed with adults that show a dominant role of the auditory system for time (Gebhard and Mowbray, 1959; Sekuler and Sekuler, 1999; Shams et al., 2000, 2001; Berger et al., 2003; Burr et al., 2009). Why the auditory dominance of both PSEs and bimodal thresholds persists into adulthood is not clear. A possible explanation is that for this kind of task the cross-sensory calibration process is still occurring since audition is too accurate with respect to the visual modality, and the precision of the visual system for this kind of task prevents the transition from unisensory dominance to multisensory integration. This dominance may however not be apparent with a different kind of stimulation. For example it would be interesting to observe whether auditory dominance in children occurs in other visual-auditory temporal integration tasks for which a strong multisensory integration in adults has been reported (as for example reducing the auditory tone from 750 to 200 Hz).

Similarly, the visual dominance of space during development could reflect a process of cross-sensory calibration in which the visual system is used to calibrate the auditory system for space perception, since it is the most accurate spatial sense. In agreement with this idea, many studies in adults show that the visual system is the most influential in determining the apparent spatial position of auditory stimuli (Pick et al., 1969; Warren et al., 1981; Mateeff et al., 1985; Alais and Burr, 2004). Only after 12 years of age, visual-auditory integration seems to occur in this spatial task suggesting a very late development. Audio-visual space integration seems to mature later than visual-haptic spatial integration (that develops

after 8–10 years of age, Gori et al., 2008) and also visual-auditory temporal integration. This could be related to the time of maturation of the individual sensory systems. Indeed, our previous work (Gori et al., 2008) suggested that multisensory integration occurs after the maturation of each unisensory system. The unisensory thresholds of **Figure 8** suggest that both visual and auditory thresholds continue to improve over the school years, particularly for the spatial task. For the space bisection task, the unisensory thresholds are still not mature at 12 years of age, and nor is integration optimal at this age. For the temporal task, unisensory thresholds become adult-like after 8–9 years of age, and at this age the auditory dominance appears. A delay in the development of unisensory systems seems to be related to the delay in the development of multisensory adult-like behavior.

These results support the idea that in children the use of one sense to calibrate the other precludes useful combination of the two sources (Gori et al., 2008; Burr and Gori, 2011). On the other hand, given the strong variability between subjects and also the noise in the developing system we cannot exclude the possibility that these results reflect the greater noise in the sensory system of the developing child. The fact that the weights derived from thresholds lie at the midpoint between auditory and visual dominance do not allow us to exclude this hypothesis.

To examine further whether this dominance reflects a process of cross-sensory calibration it would be interesting to measure how the impairment of the dominant system impacts on the non-dominant modality that may need calibration (as we did in Gori et al., 2010, 2012). In particular, it would be interesting to see how auditory spatial perception is impaired in children and adults with visual disabilities and how visual time perception is impaired in children and adults with auditory disabilities by using stimuli and procedures similar to those used in this study. If this dominance really reflects a process of a cross-sensory calibration it should allow clear and important predictions about spatial and temporal deficits in children and adults with visual and auditory disabilities.

ACKNOWLEDGMENTS

We would like to thank the school "Dante Alighieri" of Bolzaneto, the school "De Amicis" of Voltri, the school "ISG" of Genoa and all the children that participated at this study. We would also like to thank Elisa Freddi and Marco Jacono for their important contribution for this work.

REFERENCES

Alais, D., and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Curr. Biol.* 14, 257–262.

Berger, T. D., Martelli, M., and Pelli, D. G. (2003). Flicker flutter: is an illusory event as good as the real thing? *J. Vis.* 3(6), 406–412.

Burr, D., Banks, M. S., and Morrone, M. C. (2009). Auditory dominance over vision in the perception of interval duration. *Exp. Brain Res.* 198, 49–57.

Burr, D., Binda, P., and Gori, M. (2011). "Combining information from different senses: dynamic adjustment of combination weights, and the development of cross-modal integration in children,"in *Book of Sensory Cue Integration*, eds J. Trommershauser, K. Körding, and M. S. Landy (New York: Oxford University Press), 73–95.

Burr, D., and Gori, M. (2011). "Multisensory integration develops late in humans," in *Frontiers in the Neural Bases of Multisensory Processes,* eds M. Wallace and M. Murray (Boca Raton: Taylor & Francis Group), 345–363.

Clarke, J. J., and Yuille, A. L. (1990). *Data fusion for Sensory Information Processing.* Boston: Kluwer Academic.

Efron, B., and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* New York: Chapman & Hall.

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

Gebhard, J. W., and Mowbray, G. H. (1959). On discriminating the rate of visual flicker and auditory flutter. *Am. J. Psychol.* 72, 521–529.

Ghahramani, Z., Wolpert, D. M., Jordan, M. I. (1997). "Computational models of sensorimotor integration," in *Self-organization, Computational Maps and Motor Control*, eds V. Sanguineti and P. G. Morasso (Amsterdam: Elsevier Science Publication), 117–147.

Gori, M., Del Viva, M., Sandini, G., and Burr, D. (2008). Young children do not integrate visual and haptic form information. *Curr. Biol.* 18, 694–698.

Gori, M., Sandini, G., Martinoli, C., and Burr, D. (2010). Poor haptic orientation discrimination in nonsighted children may reflect disruption of cross-sensory calibration. *Curr. Biol.* 20, 223–225.

Gori, M., Sciutti, A., Burr, D., and Sandini, G. (2011). Direct and indirect haptic calibration of visual size judgments. *PLoS ONE* 6, e25599. doi:10.1371/journal.pone.0025599

Gori, M., Tinelli, F., Sandini, G., Cioni, G., and Burr, D. (2012). Impaired visual size-discrimination in children with movement disorders. *Neuropsychologia* 50, 1838–1843.

Landy, M. S., Banks, M. S., 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: Oxford University Press), 5–30.

Mateeff, S., Hohnsbein, J., and Noack, T. (1985). Dynamic visual capture: apparent auditory motion induced by a moving visual target. *Perception* 14, 721–727.

Nardini, M., Bedford, R., and Mareschal, D. (2010). Fusion of visual cues is not mandatory in children. *Proc. Natl. Acad. Sci. U.S.A.* 107, 17041–17046.

Nardini, M., Jones, P., Bedford, R., and Braddick, O. (2008). Development of cue integration in human navigation. *Curr. Biol.* 18, 689–693.

Pick, H. L., Warren, D. H., and Hay, J. C. (1969). Sensory conflict in judgements of spatial direction. *Percept. Psychophys.* 6, 203–205.

Rose, D., and Summers, J. (1995). Duration illusions in a train of visual stimuli. *Perception* 24, 1177–1187.

Sekuler, A. B., and Sekuler, R. (1999). Collisions between moving visual targets, what controls alternative ways of seeing an ambiguous display? *Perception* 28, 415–432.

Shams, L., Kamitani, Y., and Shimojo, S. (2000). Illusions. What you see is what you hear. *Nature* 408, 788.

Shams, L., Kamitani, Y., Thompson, S., and Shimojo, S. (2001). Sound alters visual evoked potentials in humans. *Neuroreport* 12, 3849–3852.

Shipley, T. (1964). Auditory flutter-driving of visual flicker. *Science* 145, 1328–1330.

Tomassini, A., Gori, M., Burr, D., Sandini, G., and Morrone, M. C. (2011). Perceived duration of visual and tactile stimuli depends on perceived speed. *Front. Integr. Neurosci.* 5:51. doi:10.3389/fnint.2011.00051

Tse, P., Intriligator, J., Rivest, J., and Cavanagh, P. (2004). Attention and the subjective expansion of time. *Percept. Psychophys.* 66, 1171–1189.

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.

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

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: 04 May 2012; paper pending published: 04 June 2012; accepted: 29 August 2012; published online: 17 September 2012.

Citation: 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*

frontiers in
INTEGRATIVE NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 06 September 2012
doi: 10.3389/fnint.2012.00067

Capture of visual attention interferes with multisensory speech processing

Hanna Krause[1]*, Till R. Schneider[1], Andreas K. Engel[1] and Daniel Senkowski[1,2]*

[1] *Department of Neurophysiology and Pathophysiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany*
[2] *Department of Psychiatry and Psychotherapy Charité, University Medicine Berlin, St. Hedwig Hospital, Berlin, Germany*

Edited by:
Zhuanghua Shi, Ludwig-Maximilians-Universität München, Germany

Reviewed by:
Roberto Martuzzi, Ecole Polytechnique Fédérale de Lausanne, Switzerland
Erik Van Der Burg, Vrije Universiteit, Netherlands

****Correspondence:***
Hanna Krause, Department of Neurophysiology and Pathophysiology, University Medical Center Hamburg-Eppendorf, Martinistr. 52, 20246 Hamburg, Germany.
e-mail: hkrause@uke.uni-hamburg.de

Daniel Senkowski, Department of Psychiatry and Psychotherapy Charité, University Medicine Berlin, St. Hedwig Hospital, Große Hamburger Str. 5-11, 10115 Berlin, Germany.
e-mail: daniel.senkowski@charite.de

Attending to a conversation in a crowded scene requires selection of relevant information, while ignoring other distracting sensory input, such as speech signals from surrounding people. The neural mechanisms of how distracting stimuli influence the processing of attended speech are not well understood. In this high-density electroencephalography (EEG) study, we investigated how different types of speech and non-speech stimuli influence the processing of attended audiovisual speech. Participants were presented with three horizontally aligned speakers who produced syllables. The faces of the three speakers flickered at specific frequencies (19 Hz for flanking speakers and 25 Hz for the center speaker), which induced steady-state visual evoked potentials (SSVEP) in the EEG that served as a measure of visual attention. The participants' task was to detect an occasional audiovisual target syllable produced by the center speaker, while ignoring distracting signals originating from the two flanking speakers. In all experimental conditions the center speaker produced a bimodal audiovisual syllable. In three distraction conditions, which were contrasted with a no-distraction control condition, the flanking speakers either produced audiovisual speech, moved their lips, and produced acoustic noise, or moved their lips without producing an auditory signal. We observed behavioral interference in the reaction times (RTs) in particular when the flanking speakers produced naturalistic audiovisual speech. These effects were paralleled by enhanced 19 Hz SSVEP, indicative of a stimulus-driven capture of attention toward the interfering speakers. Our study provides evidence that non-relevant audiovisual speech signals serve as highly salient distractors, which capture attention in a stimulus-driven fashion.

Keywords: crossmodal, EEG, bimodal, SSVEP, oscillatory

INTRODUCTION

In everyday life, speech signals from a person that we are listening to are often accompanied by distracting other sensory input, such as auditory and visual stimuli from surrounding people. These distracting stimuli can capture attention and interfere with the recognition of speech. How exactly distracting auditory and visual speech stimuli affect the recognition and processing of attended speech is, to date, not well understood.

Speech recognition, in particular in noisy conditions, is considerably improved when matching visual inputs, i.e., lip movements, are presented (Sumby and Pollack, 1954; Ross et al., 2007a,b). Moreover, a recent functional magnetic resonance imaging study showed that attending to lip movements that match a stream of auditory sentences leads to an enhanced target detection rate and to stronger activity in a speech-related multisensory network compared to attending to non-matching lip movements (Fairhall and Macaluso, 2009). This suggests an important role of top-down attention for multisensory processing of speech (Koelewijn et al., 2010; Talsma et al., 2010).

This notion is consistent with an electroencephalographic (EEG) study, in which we examined the influence of task relevant and task irrelevant visual speech stimuli on audiovisual speech processing in a multiple speaker scenario (Senkowski et al., 2008). In this study, participants were instructed to detect an occasional audiovisual target syllable by a speaker (i.e., a speaking face) who was presented centrally and surrounded by two flanking speakers. The study comprised of *no interference* trials, in which a syllable was produced by the relevant central speaker only, and *interference trials*, in which different audiovisual syllables were produced by three speakers simultaneously. Using steady-state visual evoked potentials (SSVEP) as a real-time index of deployment of visual attention, we observed that visual attention toward the task irrelevant flanking speakers interferes with the recognition of task relevant audiovisual signals. The main open question raised by this study is whether the interference effect is specific for the processing of naturalistic audiovisual speech or whether similar effects would occur when the flanking speakers produce other distracting stimuli, like moving their lips without a sound or when they produce noise instead of syllables.

Using an extended setup of our previous study (Senkowski et al., 2008), we addressed this question by examining behavioral data and SSVEPs in three *interference* conditions and one control condition. In the interference conditions, the flanking speakers produced either naturalistic audiovisual syllables, lip movements alone, or lip movements in combination with acoustic noise. In line with our previous study (Senkowski et al., 2008), we

expected distraction effects in behavioral data that are paralleled by enhanced SSVEPs to flanking speakers when these speakers produced naturalistic audiovisual speech. Given the salience of naturalistic audiovisual speech, we predicted that the interference effects of lip movements alone and lip movements accompanied by auditory noise would be much weaker or even vanished.

MATERIALS AND METHODS

PARTICIPANTS

Twenty volunteers, who reported no history of neurologic or psychiatric illness, participated in the study. Four participants were excluded from the analysis on the basis of extensive eye movements. Additional three participants were excluded because their hit rate (HR) was lower than 50% in the "Speech Interference" condition (see below). The remaining 13 participants (all right handed, mean age 22.92 years, range 21–29 years, 6 females) had normal hearing, as assessed by a hearing test in which 30 dB HL sinusoidal tones of varying frequencies had to be detected. Participants had normal or corrected-to-normal vision, as ensured by the Landolt test of visual acuity (visus ≥ 0.9). The Institutional Review Board of the Medical Association of Hamburg approved the experimental procedures, and each subject provided written informed consent and was paid for participation.

PROCEDURE AND STIMULI

A continuous stream of four stimulation conditions was presented (**Figure 1**). Two of the conditions were identical to those used in our previous study (Senkowski et al., 2008). This previous study comprised of a "No Interference" control condition, in which only the center speaker produced a syllable, and a "Speech Interference" condition, in which all three speakers produced syllables (a short clip of this experiment is provided at: http://www.sciencedirect.com/science/article/pii/S1053811908007933). In the present study, two conditions were added to examine in further detail how visual attention toward flanking speakers interferes with audiovisual speech processing. In one of these conditions the flanking speakers produced acoustic non-speech noise instead of syllables. Non-speech noise samples were directly derived from the original syllables by phase-scrambling the auditory syllables, thereby maintaining basic properties like stimulus power. We will refer to this condition as "Auditory Noise Interference." In the other condition the flanking speakers moved their lips without producing an acoustic syllable. We will refer to this condition as "Lip Movement Interference" condition. Thus, the study comprised of four conditions: "No Interference," "Speech Interference," "Auditory Noise Interference," and "Lip Movement Interference." The center speaker produced one of the syllables /ta/, /da/, /ga/, or /ba/ in all conditions, whereas the flanking speakers could produce the syllables /ta/, /da/, or /ga/ in the "Speech Interference" condition. The four conditions and the different syllables were presented in random order. Participants were instructed to focus their attention to the center speaker and to ignore the signals from the flanking speakers. Furthermore, they had to indicate the occasional appearance of the target syllable /ba/ by the center speaker with a button press of their right index finger. The target syllable occurred in 20% of all trials. The three speakers never produced the same syllable in a trial and syllable combinations that could evoke the McGurk illusion (McGurk and MacDonald, 1976), like the combination /ba/ and /ga/ were excluded.

On average 76 targets and 300 non-target stimuli were presented for each condition. One trial consisted of 120 visual frames of 6.67 ms each, resulting in a trial duration of 792 ms. Two fixed cycles of 24 frames were added per trial. Moreover, a variable number of 1–5 cycles (average: 3 cycles) was added, resulting in a total average trial duration of 1592 ms. During the inter-trial interval the faces of the three speakers were presented on the screen without producing any lip movements or speech sounds, but the 19 Hz flicker of the flanking speakers and the 25 Hz flicker of the center speaker continued. An additional number of 645 (about 30% of all trials) "omitted trial" periods (Busse

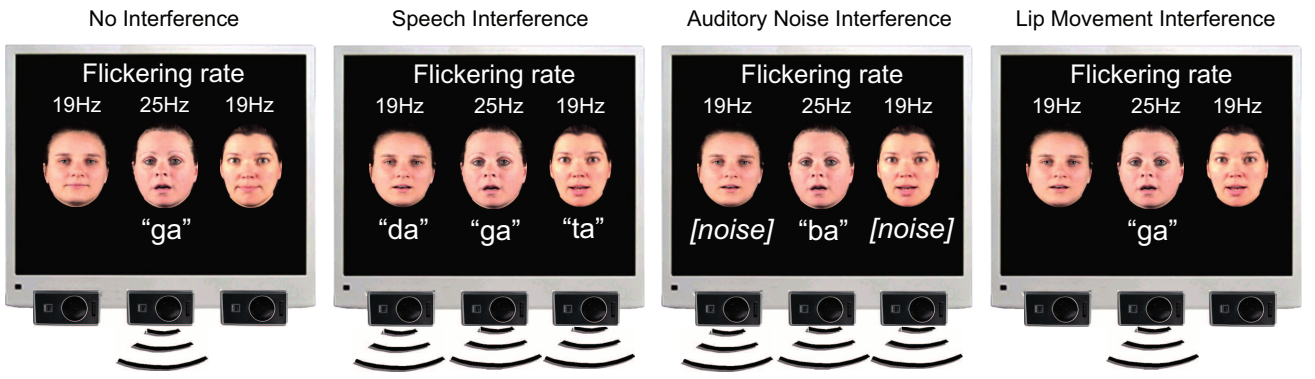


FIGURE 1 | Stimulus setup. Stimuli consist of three horizontally aligned speakers on a black background. In all experimental conditions, the center face is visually presented in an on-off fashion so that a 25 Hz flicker was elicited. The center speaker produces natural auditory and visible syllables ('ta', 'da', 'ga', 'ba'), whereas the two flanking speakers are always presented with a flicker frequency of 19 Hz. The subject's task is to detect the syllable 'ba' by the center speaker. In the No Interference condition, the flanking speakers produce neither visual lip movements nor speech sounds, whereas they produce natural speech syllables ('ta', 'da', and 'ga') simultaneous with the syllables of the center speaker in the Speech Interference condition. The Auditory Noise Interference condition consists of phase-scrambled versions of the original syllables 'ta', 'da', and 'ga' produced by the flanking speakers. In the Lip Movement Interference condition the flanking speakers produce lip movements of the original syllables without any accompanying auditory signal.

and Woldorff, 2003) were randomly inserted into the continuous stream of stimuli, further reducing the predictability of the experimental stimuli. During omitted trial periods, the faces of the three speakers were presented for a time interval that was identical to the interval of regular experimental events (i.e., 792 ms) but without any lip movements or speech sounds. Each participant underwent 18 experimental blocks with 120 trials each.

Recordings of syllables from the three speakers were obtained at frame rates of 30/s. Each syllable consisted of 20 frames of 33 ms duration, which results in a total duration of 660 ms for each syllable. The visual angle of the speakers subtended 7° between adjacent speakers (from mouth to mouth) and the width of the speakers' faces subtended an angle of 4.8° each. The characters of the flanking speakers switched their location (i.e., left or right of the center speaker) after every block, while the center speaker character remained the same throughout the experiment. The monitor was set to a refresh rate of 150 Hz, i.e., the refresh rate duration for one frame was 6.67 ms. To induce SSVEPs, the continuous stream of pictures was dissected in an on–off fashion, i.e., pictures of the continuous stream ("on") were presented alternately with blank screens ("off"). Pictures of the continuous stream and blank frames alternated every 20 ms. Thus, the flicker frequency (i.e., on–off cycle) was 25 Hz for the center speaker. For the two flanking speakers, the on–off periods alternated every 26.6 ms simultaneously for both speakers, corresponding to a flicker frequency of about 19 Hz. In the EEG the time-frequency (TF) transformed activity of a sustained visual on–off flicker is reflected in event-related activity that corresponds to the presented flicker frequency (Herrmann, 2001).

Both the 19 Hz flicker and the 25 Hz flicker were presented continuously and all trials started with an "on" period. The average stimulus duration of the acoustic syllables was 295 ms and the onset of these syllables followed the onset of visual lip movements on average by 230 ms. To eliminate overlapping event-related responses to the sounds, a relative stimulus onset jitter of 110 ms (more than two times the duration of a 19 Hz and a 25 Hz cycle) was used by adding or subtracting a random time interval between ±55 ms to the real acoustic sound onset in each trial (Woldorff, 1993; Senkowski et al., 2007). This jitter prevented overlapping event-related 19 and 25 Hz responses to the acoustic inputs. A spline curve FFT filter between 400 and 4000 Hz was applied to all syllables to align the voice characteristics between the three speakers.

DATA ACQUISITION

The EEG was recorded from 124 scalp sites using an active electrode system (EASYCAP, Herrsching, Germany). In addition, the electrooculogram was recorded by two electrodes. One of these electrodes was placed below the eye and the other one was placed at the lateral bridge of the nose. The nose tip was used as reference during recording and data were off-line re-referenced to average reference. Data were digitized at a sampling rate of 1000 Hz using BrainAmp amplifiers (BrainProducts, Munich, Germany), filtered from 0.3 to 120 Hz and downsampled to 250 Hz for the off-line analysis. Epochs were cut around the visual motion onset (0 indicates the first frame of the visible movement) from –1000 ms before to 1200 ms after visual motion onset. Trials containing artifacts in EEG data resulting from eyeblinks, horizontal eye movements, or muscle activity were removed from the further analysis. Noisy channels were linearly interpolated. Finally, an automatic threshold was applied, excluding all trials in which the EEG amplitude exceeded 100 μV.

DATA ANALYSIS

Reaction times (RTs) to target stimuli were calculated by averaging all trials in which subjects responded between 230 and 1000 ms after visual motion onset and in which the RT did not exceed 2 standard deviations from the mean RT within each participant and condition. For the statistical analysis of RTs, HR, and false alarms (FA), an ANOVA or Friedman test (if the assumption of gaussianity was violated) with the factor experimental condition (No Interference, Speech Interference, Auditory Noise Interference, Lip Movement Interference) was calculated. A Kolmogorov-Smirnov test was computed to test for gaussianity of RT, HR, and FA distributions. Moreover, three planned contrasts were computed: Speech Interference vs. No Interference, Auditory Noise Interference vs. No Interference, and Lip Movement Interference vs. No Interference.

EEG data were analyzed using MATLAB (Version 7.10), EEGLAB 5.03 (http://www.sccn.ucsd.edu/eeglab), and the FIELDTRIP toolbox (Oostenveld et al., 2011). For the analysis of SSVEPs, event-related activity was calculated by averaging the epochs of each condition. For the averaged activity, TF analyses were calculated using wavelet transformation with Morlet wavelets spanning a range of 10–30 Hz with a length of 12 cycles. The TF analysis was computed in 0.25 Hz steps. In agreement with our previous study (Senkowski et al., 2008), we analyzed SSVEPs for three predefined regions of interest (ROIs): an occipital ROI, comprising of 7 electrodes that were located at midline-occipital scalp, and two symmetric bilateral ROIs that were located at lateral temporal scalp, comprising of 6 electrodes each. In line with the observed SSVEP response pattern, the analysis was done for the time window of 230–550 ms after visual motion onset. To investigate how visual inputs of the center speaker and the flanking speakers were processed in the different experimental conditions, wavelet transformed data were analyzed for those frequencies that corresponded to the visual stimulation frequencies of the speakers. The length of the wavelet was 480 ms for the analysis of 25 Hz activity and 632 ms for the analysis of 19 Hz activity, with a wavelet length of 12 cycles. Repeated measures ANOVAs with the within-subject factors Condition (No Interference, Speech Interference, Auditory Noise Interference, Lip Movement Interference) and ROI (left temporal, right temporal, and occipital) were conducted. Furthermore, planned contrasts between each of the three interference conditions (Speech Interference, Auditory Noise Interference, and Lip Movement Interference) and the no-interference condition were computed. In case of non-sphericity, as tested by Mauchly's sphericity test, the degrees of freedom were adjusted in the ANOVAs.

RESULTS

BEHAVIORAL DATA

The ANOVA for RTs with the factor Condition (No Interference, Speech Interference, Auditory Noise Interference, and

Lip Movement Interference) revealed a significant effect [$F_{(2.07,\ 24.85)} = 16.169$, $p < 0.0001$; **Figure 2B**]. The analysis of planned contrasts revealed significant longer RTs in Speech Interference Condition (731 ms) compared to the No Interference condition (673 ms; $t_{12} = -6.557$, $p < 0.001$). No other significant effects were observed for RTs.

Since the Kolmogorov–Smirnov tests indicated violations of gaussianity in the distributions of HR and FA data, non-parametric Friedman tests were computed for the analysis of effects in HR and FA rate. For the HR, this test revealed a significant difference between conditions ($p < 0.0001$). The analysis of pair-wise planned contrasts (using non-parametric Wilcoxon tests) revealed significant differences between the No Interference and the Speech Interference Condition ($p = 0.001$) and the No Interference and the Auditory Noise Interference Condition ($p = 0.003$). For both comparisons the HR was higher in the No Interference condition. There was no significant difference between the No Interference and the Lip Movement Interference Condition ($p = 0.128$). For the three Interference Conditions, a significant difference was found between the Lip Movement Interference and the Auditory Noise Interference Condition ($p = 0.011$), due to a higher HR in the Lip Movement Interference Condition. Furthermore, there were significant differences between the Lip Movement Interference and the Speech Interference Conditon ($p = 0.001$) as well as between the Speech Interference and the Auditory Noise Interference Condition ($p = 0.001$). The HR was higher in the Lip Movement and the Auditory Noise Interference conditions compared to the Speech Interference Condition.

The Friedman test for FA rate revealed a significant result ($p = 0.019$; **Figure 2B**). Pairwise Wilcoxon tests revealed significantly larger FA rates in the Speech Interference Condition (0.799%) compared to the No Interference Condition (0.287%, $p = 0.021$). However, the differences between the No Interference compared to the Auditory Noise Interference Condition (0.835%) and the Lip Movement Interference Condition (0.257%) were not significant.

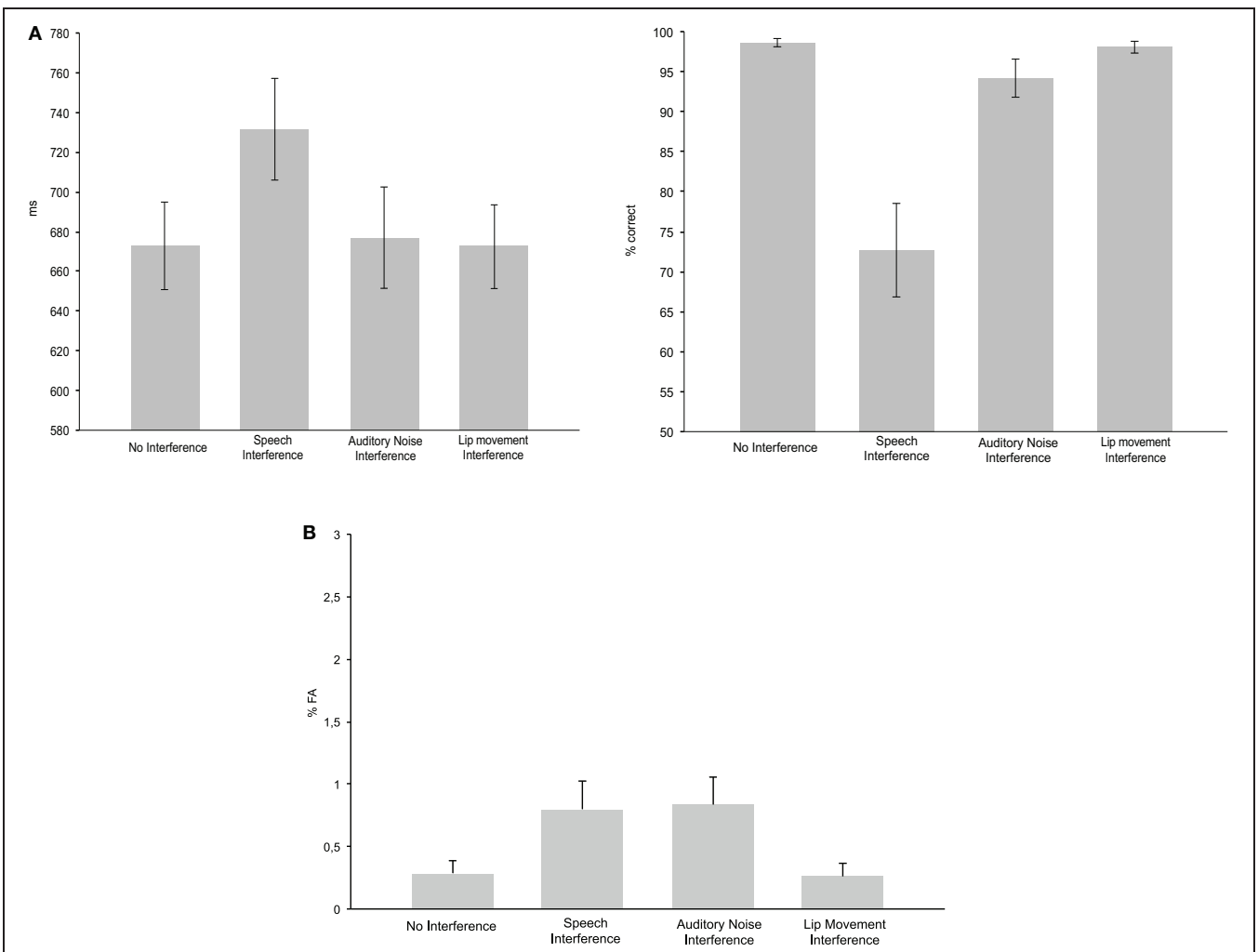


FIGURE 2 | Behavioral performance. (A) Reaction times (left panel) and hit rates (right panel) for the No Interference control condition as well as for the three Interference conditions. **(B)** False alarm rates in the four experimental conditions.

STEADY-STATE VISUAL EVOKED POTENTIALS

The spectral analysis revealed occipital SSVEPs that corresponded to the flicker frequency of flanking speakers (19 Hz) and the center speaker (25 Hz, **Figure 3**). The Two-Way ANOVA for flanking speakers' 19 Hz SSVEPs using the factors Condition (No Interference, Speech Interference, Auditory Noise Interference, and Lip Movement Interference) and ROI (left temporal, right temporal, and occipital) revealed significant main effects of Condition [$F_{(3,\ 12)} = 4.123$, $p < 0.05$] and ROI [$F_{(2,\ 12)} = 12.780$, $p < 0.001$], and a significant interaction between these factors [$F_{(6,\ 72)} = 2.770, p < 0.05$]. Follow-up analyses were performed separately for the three ROIs. Whereas no significant effects were observed for the bilateral temporal ROIs (all p's > 0.1), a significant main effect of Condition was found for the occipital ROI [$F_{(3,\ 12)} = 3.777, p < 0.05$, **Figure 4**]. The analysis of planned contrasts revealed a significant effect for the contrast between the Speech Interference and the No Interference condition [$F_{(1,\ 12)} = 5.996$, $p < 0.05$], due to larger flanking speaker SSVEPs in the Speech Interference condition. Moreover, a trend toward significance was found for the contrast between the Lip Movement Interference and the No Interference condition [$F_{(1,\ 12)} = 4.488$, $p < 0.1$]. SSVEPs tended to be larger in the Lip Movement than in the No Interference condition. No other significant effects were found, neither in the 19 Hz nor in the 25 Hz SSVEPs.

The present finding of a occipital modulation of the 19 Hz SSVEPs differs from our previous study, which found relevant effects at a left temporal ROI. To ensure that the differences in the topographic distribution of the maximum SSVEP power between our studies are not due to a technical malfunction, we tested the original stimulation setup as used in our previous study (Senkowski et al., 2008) as well as the stimulation files which we used in the present study with a photodiode but found no deviations in visual stimulation frequencies.

DISCUSSION

The present study demonstrates that processing of distracting audiovisual speech signals interferes with the recognition of attended audiovisual speech. Comparing speech recognition performance in three interference conditions with a no-interference control condition, we observed a decrease in response speed primarily when the distracting signals comprised

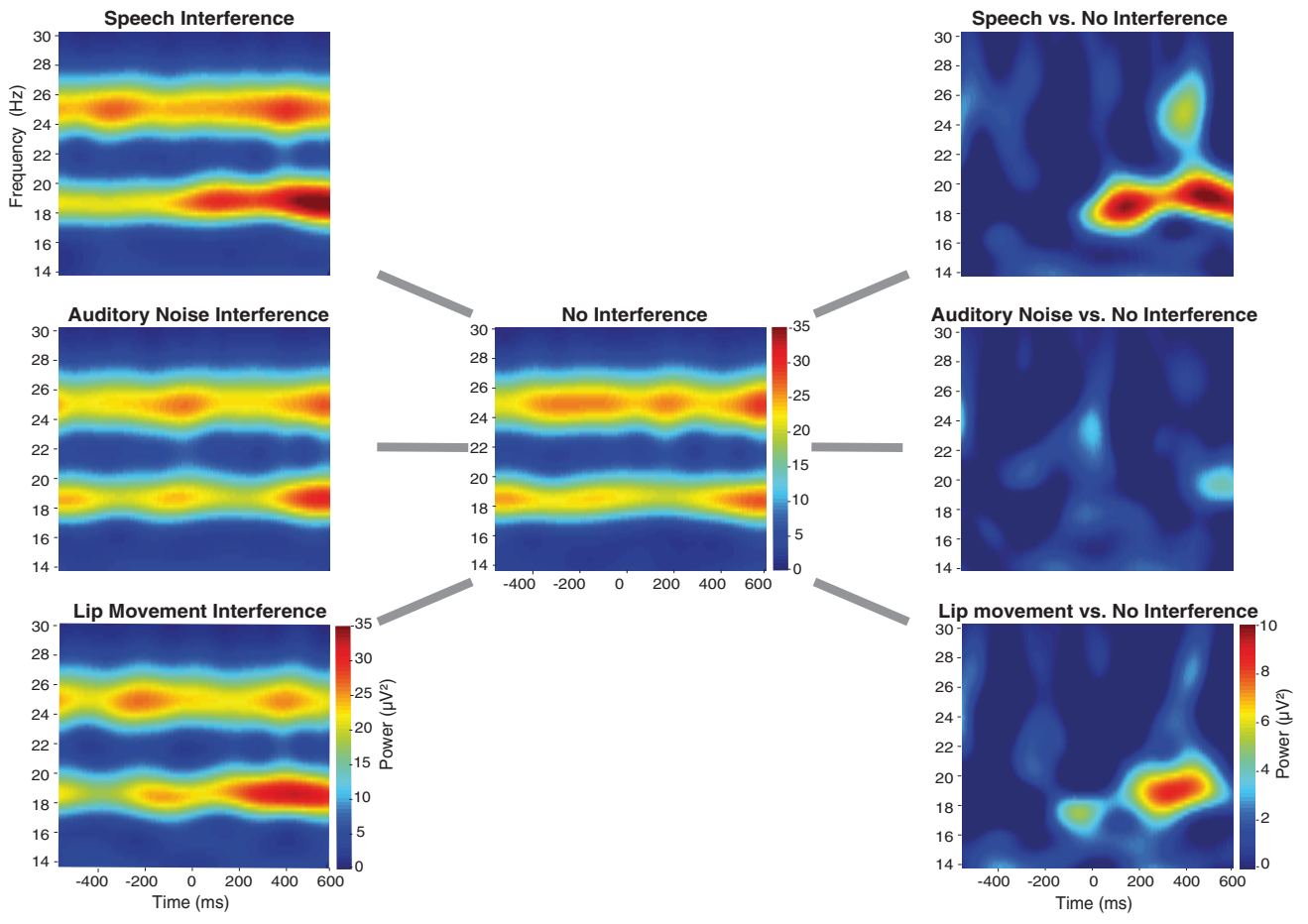


FIGURE 3 | Time-frequency-plots of SSVEPs for the three Interference conditions (left panel), the No Interference condition (middle panel) as well as for the differences between Interference, and No Interference conditions (right panel) for the occipital ROI (see Figure 4B). For the statistical analysis a time-frequency window of 230–550 ms and 19 Hz was selected.

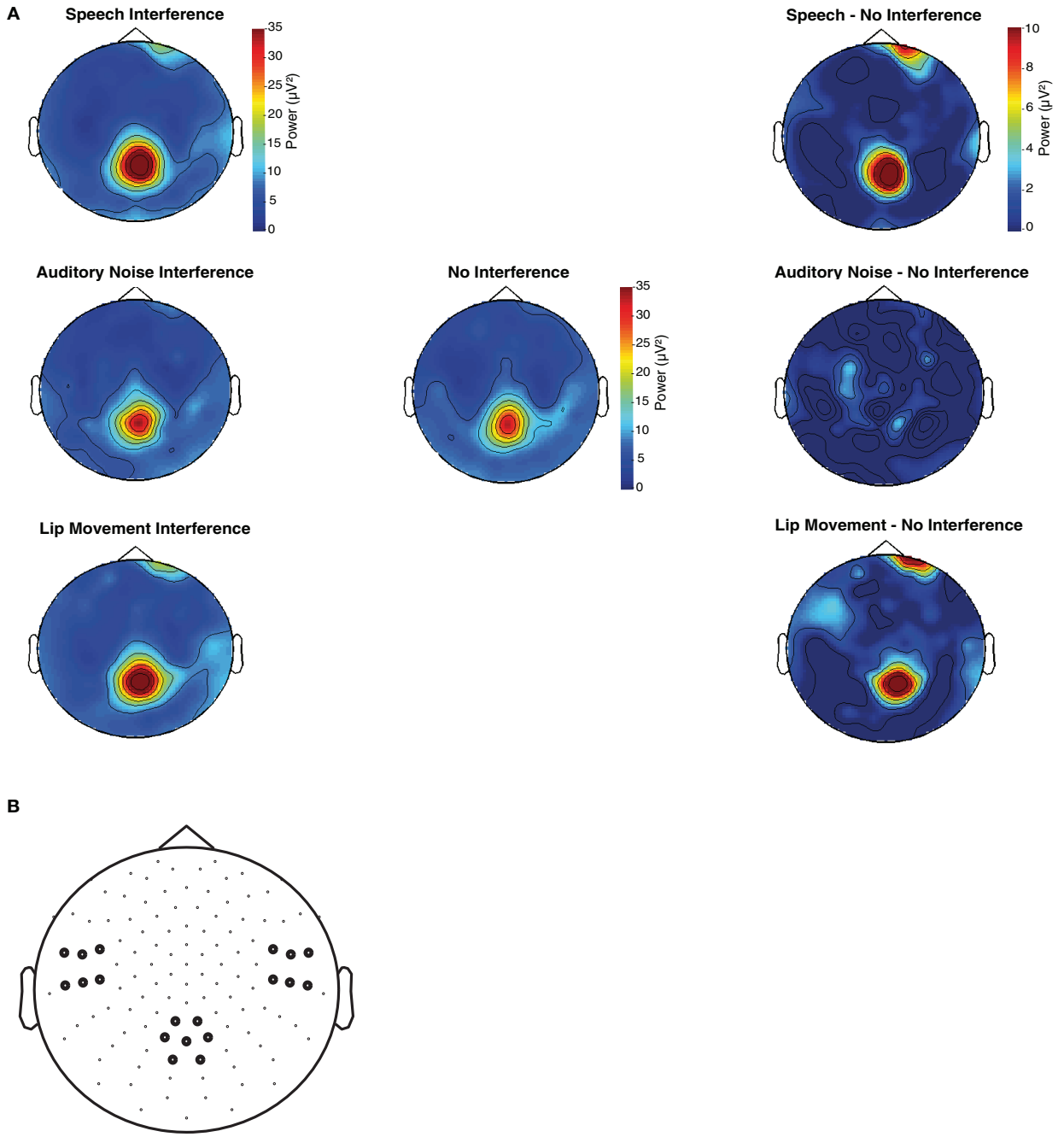


FIGURE 4 | (A) Topographies of flanking speaker's induced SSVEPs (19 Hz) for the time window of 230–550 ms after visual motion onset. **(B)** Left temporal, right temporal, and occipital sensors were pooled in three regions of interest and used for the statistical analysis of SSVEPs.

of naturalistic audiovisual speech. This finding was paralleled by an enhancement of flanking speakers SSVEPs over the occipital lobe.

BEHAVIORAL DATA

From the three distraction conditions (**Figure 1**), an interference effect in RT data was found particularly in the naturalistic audiovisual speech interference condition. Although we also found a significant difference in the HR between the Auditory Noise Interference Condition and the No Interference Condition, the most robust interference effects on RT, FA, and HR were observed in the Speech Interference Condition (see **Figure 2A**). Previous studies have shown that synchronously presented auditory and visual stimuli can serve as salient distractors, which can, for instance, bias temporal order judgements and simultaneity judgements of visual stimuli (Van der Burg et al., 2008a). Furthermore,

it has been demonstrated that task irrelevant auditory signals can facilitate visual search (Van der Burg et al., 2011), in particular when the auditory signal is presented synchronously with the visual target (Van der Burg et al., 2008b) and when it is transient (Van der Burg et al., 2010). Using a spatial cueing paradigm, another study showed a stronger attentional capture for bimodal audiovisual compared to unimodal visual distractors (Matusz and Eimer, 2011). All of these studies have used basic, semantically meaningless, auditory, and visual stimuli. A study in which participants were asked to detect or localize a naturalistic face (out of up to four faces) that matches in its lip movements with a simultaneously presented auditory speech, showed a decrease in accuracy and an increase in search times with increasing set size in the localization task (Alsius and Soto-Faraco, 2011). This suggests that the faces were processed in a serial fashion (Wolfe, 2003).

Alsius and Soto-Faraco (2011) conducted another experiment, in which the task was to detect or to localize an auditory stream (out of up to four auditory streams) matching the lip movements of a face. In this experiment, RTs and accuracy did not depend on set size in the detection task, supporting the assumption of parallel processing of the auditory streams. Together, these studies show that auditory speech represents a salient input and that auditory stimuli can strongly bias the processing of concurrently presented visual stimuli.

In the present study two bimodal audiovisual interference conditions were examined: one consisted of natural audiovisual speech signals and the other of lip movements and auditory noise. In agreement with the above-described studies, our finding of distraction effects in the naturalistic speech interference condition suggests that auditory speech stimuli serve as salient inputs in our environment, even if they are unattended. Taken together, our study demonstrates that irrelevant naturalistic audiovisual speech signals have a much stronger interference effect on RTs than visual lip movements alone or lip-movements that are accompanied by acoustic noise. This highlights the unique relevance of speech signals in our environment.

INTERFERENCE EFFECTS IN SSVEP

The finding of enhanced flanking speaker induced SSVEPs for naturalistic audiovisual speech stimuli fits with our previous study (Senkowski et al., 2008), which had only two experimental conditions (Audiovisual Speech Interference and No Interference). Importantly, the present observations extend our previous findings by demonstrating that the enhancement of flanking speaker induced SSVEP occurs primarily when the flanking speakers produced naturalistic audiovisual speech but this enhancement is weaker (in the Lip Movement Interference condition) or even vanished (in the Auditory Noise Interference condition) in the other distraction conditions. In contrast to our previous study (Senkowski et al., 2008), the present results allow a more specific interpretation of the interfering effects of naturalistic audiovisual speech signals, since no interfering effects on RTs were found when auditory noise, which resembled the naturalistic syllables in its basic properties, like stimulus power, was presented. As shown in previous visual attention studies, SSVEP enhancement likely reflects an increased processing of the respective visual flicker stimuli and thus can serve as an electrophysiological measure for the allocation of visual attention (Morgan et al., 1996; Müller et al., 2003; Martens et al., 2011). Therefore, we suggest that the enhanced flanking speaker's SSVEPs reflect a capture of visual attention by the non-relevant audiovisual speech signals.

Another interesting observation was the trend toward a significant enhancement of the flanking speaker's SSVEPs in the Lip Movement Interference condition. Since there were no behavioral interference effects of viewing lip movements alone, the enhanced SSVEPs in this condition do not appear to reflect a behaviorally relevant capture of visual attention. An explanation for the observed trend could be that the lip movements of the flanking speakers were not accompanied by an acoustic stimulus, which may have led to a crossmodal mismatch detection (Arnal et al., 2011) that enhanced visual processing of the flanking speakers.

The absence of interference effects on SSVEPs induced by the center speaker is in line with our previous study (Senkowski et al., 2008). It may be that the capture of attention observed in the Speech Interference Condition involves a split of the attentional focus when the flanking speakers produced bimodal audiovisual syllables. Previous studies have shown that the attentional spotlight can be split (Müller et al., 2003; McMains and Somers, 2004). Specifically, these studies have shown that visual input presented at multiple locations can be monitored in parallel by our attentional system. In the current study, however, such a possible split of the attentional spotlight did not substantially affect the processing of visual input from the attended center speaker.

While the finding that the distraction effects are particularly reflected in flanking speakers SSVEP is in agreement with our previous study (Senkowski et al., 2008), there are also some differences in results. The main difference is that the effects on flanking speakers SSVEPs in our previous study were found at left lateral temporal electrode sites, whereas in the present study we observed modulations at occipital sites. The differences between our previous study and the present work may emerge from differences in experimental setups. The paradigm in the present study consisted of four experimental conditions (including three distraction conditions) compared to two conditions (with only one distraction condition) in our previous study. It is possible that these differences contributed to the differences in results (i.e., topography of effects). Notably, however, the effects in both studies were found particularly for flanking speaker SSVEPs. Interpreting the results in terms of a capture of visual attention, the observation of effects at occipital electrodes in the present study fits well with previous studies showing attention related effects on SSVEPs at postero-occitipal scalp (e.g., Müller et al., 2003).

CONCLUSION

Our study demonstrates that non-relevant audiovisual speech stimuli serve as highly salient distractors in the processing of audiovisual speech. The enhanced attentional capture in the naturalistic audiovisual speech interference condition is reflected

by a decrease in behavioral performance and an enhancement of flanking speaker induced SSVEPs. The interference effects in the other distraction conditions, comprising of visual lip movements alone and lip movements accompanied by auditory noise, were much weaker or even vanished, respectively. Taken together, our study provides evidence that non-relevant audiovisual speech in particular leads to stronger distraction in speech interference situations as compared to other sensory signals.

ACKNOWLEDGMENTS

We would like to thank K. Saha for recruitment of participants and help with data recording and C. Hipp for her thoughtful comments about this work. This research was supported by a NARSAD young investigator award (Daniel Senkowski), and grants from the German Research Foundation (SE 1859/1-2, Daniel Senkowski) and the EU (IST-2005-27268, HEALTH-F2-2008-200728, ERC-2010-AdG-269716, Andreas K. Engel; ERC-2010-StG-20091209, Daniel Senkowski).

REFERENCES

Alsius, A., and Soto-Faraco, S. (2011). Searching for audiovisual correspondence in multiple speaker scenarios. *Exp. Brain Res.* 213, 175–183.

Arnal, L. H., Wyart, V., and Giraud, A.-L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nat. Neurosci.* 14, 797–801.

Busse, L., and Woldorff, M. G. (2003). The ERP omitted stimulus response to "no-stim" events and its implications for fast-rate event-related fMRI designs. *Neuroimage* 18, 856–864.

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.

Herrmann, C. S. (2001). Human EEG responses to 1–100 Hz flicker: resonance phenomena in visual cortex and their potential correlation to cognitive phenomena. *Exp. Brain Res.* 137, 346–353.

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.

Martens, U., Trujillo-Barreto, N., and Gruber, T. (2011). Perceiving the tree in the woods: segregating brain responses to stimuli constituting natural scenes. *J. Neurosci.* 31, 17713–17718.

Matusz, P. J., and Eimer, M. (2011). Multisensory enhancement of attentional capture in visual search. *Psychon. Bull. Rev.* 18, 904–909.

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748.

McMains, S. A., and Somers, D. C. (2004). Multiple spotlights of attentional selection in human visual cortex. *Neuron* 42, 677–686.

Morgan, S. T., Hansen, J. C., and Hillyard, S. A. (1996). Selective attention to stimulus location modulates the steady-state visual evoked potential. *Proc. Natl. Acad. Sci. U.S.A.* 93, 4770–4774.

Müller, M. M., Malinowski, P., Gruber, T., and Hillyard, S. A. (2003). Sustained division of the attentional spotlight. *Nature* 424, 309–312.

Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.-M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* 2011, 156869.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., and Foxe, J. J. (2007a). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cereb. Cortex* 17, 1147–1153.

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

Senkowski, D., Saint-Amour, D., Gruber, T., and Foxe, J. J. (2008). Look who's talking: the deployment of visuo-spatial attention during multisensory speech processing under noisy environmental conditions. *Neuroimage* 43, 379–387.

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.

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215.

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.

Van der Burg, E., Cass, J., Olivers, C. N. L., Theeuwes, J., and Alais, D. (2010). Efficient visual search from synchronized auditory signals requires transient audiovisual events. *PLoS ONE* 5:e10664. doi: 10.1371/journal.pone.0010664

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., and Theeuwes, J. (2008a). Audiovisual events capture attention: evidence from temporal order judgments. *J. Vis.* 8, 2.1–2.10.

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

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.

Woldorff, M. G. (1993). Distortion of ERP averages due to overlap from temporally adjacent ERPs: analysis and correction. *Psychophysiology* 30, 98–119.

Wolfe, J. M. (2003). Moving towards solutions to some enduring controversies in visual search. *Trends. Cogn. Sci.* 7, 70–76.

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: 31 May 2012; accepted: 20 August 2012; published online: 06 September 2012.

Citation: Krause H, Schneider TR, Engel AK and Senkowski D (2012) Capture of visual attention interferes with multisensory speech processing. Front. Integr. Neurosci. 6:67. doi: 10.3389/fnint.2012.00067

frontiers in
INTEGRATIVE NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 28 August 2012
doi: 10.3389/fnint.2012.00064

Coding of multisensory temporal patterns in human superior temporal sulcus

Tömme Noesselt[1,2]*†, Daniel Bergmann[3,4†], Hans-Jochen Heinze[3], Thomas Münte[5] and Charles Spence[6]

[1] Department of Biological Psychology, Otto-von-Guericke-Universität Magdeburg, Magdeburg, Germany
[2] Center of Behavioral Brain Sciences, Otto-von-Guericke-Universität Magdeburg, Magdeburg, Germany
[3] Department of Neurology, Otto-von-Guericke-Universität Magdeburg, Magdeburg, Germany
[4] Psychosomatic Medicine, Asklepios Westklinikum Hamburg, Hamburg, Germany
[5] Department of Neurology, UniversitätzuLübeck, Lübeck, Germany
[6] Crossmodal Research Laboratory, Department of Experimental Psychology, University of Oxford, Oxford, UK

Edited by:
Zhuanghua Shi, Ludwig-Maximilians-Universität München, Germany

Reviewed by:
Emiliano Macaluso, Fondazione Santa Lucia, Italy
Mark W. Greenlee, University of Regensburg, Germany

***Correspondence:**
Tömme Noesselt, Department of Biological Psychology, Otto-von-Guericke-Universität Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany.
e-mail: toemme@med.ovgu.de

† *These authors equally contributed to this work.*

Philosophers, psychologists, and neuroscientists have long been interested in how the temporal aspects of perception are represented in the brain. In the present study, we investigated the neural basis of the temporal perception of synchrony/asynchrony for audiovisual speech stimuli using functional magnetic resonance imaging (fMRI). Subjects judged the temporal relation of (a)synchronous audiovisual speech streams, and indicated any changes in their perception of the stimuli over time. Differential hemodynamic responses for synchronous versus asynchronous stimuli were observed in the multisensory superior temporal sulcus complex (mSTS-c) and prefrontal cortex. Within mSTS-c we found adjacent regions expressing an enhanced BOLD-response to the different physical (a)synchrony conditions. These regions were further modulated by the subjects' perceptual state. By calculating the distances between the modulated regions within mSTS-c in single-subjects we demonstrate that the "auditory leading (A_L)" and "visual leading (V_L) areas" lie closer to "synchrony areas" than to each other. Moreover, analysis of interregional connectivity indicates a stronger functional connection between multisensory prefrontal cortex and mSTS-c during the perception of asynchrony. Taken together, these results therefore suggest the presence of distinct sub-regions within the human STS-c for the maintenance of temporal relations for audiovisual speech stimuli plus differential functional connectivity with prefrontal regions. The respective local activity in mSTS-c is dependent both upon the physical properties of the stimuli presented and upon the subjects' perception of (a)synchrony.

Keywords: audiovisual, temporal perception, fMRI, speech, human

INTRODUCTION

When observers are confronted with incongruent auditory and visual information, that information is often fused into a congruent multisensory percept. Spatial, semantic, and temporal factors have all been shown to contribute to this perceptual fusion (see e.g., Driver and Noesselt, 2008, for a review). The temporal relationship between inputs from different senses plays a particularly important role in multisensory integration (Köhler, 1947; Dennett, 1991; Spence and Squire, 2003; Kelly, 2005) and the perceived synchrony declines when the audio-visual asynchrony exceeds a certain temporal delay. When simple auditory beeps and visual flashes are being judged, subjects' temporal synchrony window spans approximately 100 ms (Slutsky and Recanzone, 2001; Vatakis and Spence, 2006a) becoming broader/wider when stimuli are more complex (consisting of semantic content; Dixon and Spitz, 1980; McGrath and Summerfield, 1985; Spence and Squire, 2003; Miller and D'Esposito, 2005; Vatakis and Spence, 2006b, see also Vroomen and Keetels, 2010 for review).

Several brain structures have been implicated in the multisensory integration of auditory and visual stimuli. Among them are the superior colliculi (Stein and Meredith, 1993), the superior temporal sulcus complex (STS-c), the intraparietal sulcus (IPS), the insular cortex, the claustrum and prefrontal areas (e.g., Calvert et al., 2000; Bushara et al., 2001; Calvert, 2001; Driver and Noesselt, 2008). Within the STS-c, areas within or close to the upper bank have been identified as key regions governing multisensory integration in both humans (Wright et al., 2003; Beauchamp, 2005a; Noesselt et al., 2007) and non-human primates (Benevento et al., 1977; Desimone and Gross, 1979; Bruce et al., 1981; Hikosaka et al., 1988; Barraclough et al., 2005). Direct neuronal recordings from the superior temporal polysensory (STP) region in monkeys have revealed that neurons can respond to both visual and auditory stimuli in both the upper (Bruce et al., 1981; Hikosaka et al., 1988) and lower banks (Benevento et al., 1977). Barraclough et al. (2005) reported neurons within the STS-c that respond to action-related congruent

audiovisual stimulation. When focusing on complex, speech-related animal communication, results from studies in macaques suggest that temporal regions in the macaque brain (including in the STS-c) are activated by audiovisual species-specific vocalizations (Gil-da-Costa et al., 2004; Ghazanfar et al., 2008). In humans, using linguistic stimuli, van Atteveldt et al. (2004) found regions in the STS-c that responded to visually presented letters, spoken single letters, or their combination. As in the study by Wright et al. (2003) employing lip-movements plus spoken syllables, the STS-c response was greatest for audiovisual stimuli. van Atteveldt and colleagues (2004) reported that multisensory enhancement was seen for congruent but not for incongruent stimuli. However, other studies reported enhancements in functional magnetic resonance imaging (fMRI)-responses for incongruent stimuli within STS-c (e.g., van Atteveldt et al., 2007). These findings suggest that the STS-c is involved in the temporal binding of audiovisual stimuli. However, it still needs to be established whether congruent or incongruent audiovisual stimuli elicit a higher fMRI-signal in STS-c, or whether different subregions within the STS-c may differentially code multisensory temporal relations.

Hence, the aim of the present study was to investigate the functional neuroanatomy of the multisensory regions including STS-c and prefrontal cortex when perceiving a temporal (mis-)alignment of ecologically-valid long speech sequences; and to examine whether audiovisual temporal relationships may subdivide multisensory regions functionally. Subjects were shown videos of temporally aligned and misaligned video streams [either auditory leading (A_L) or visual leading (V_L) and reported whether those were perceived as being synchronous or asynchronous. Importantly, they also reported changes of perceived timing *during* the presentation of each stimulus. This design enabled us to dissociate those neural processes that were related to perceptual switches and those related to stable perceptual states during the presentation of audiovisual speech sequences. To anticipate, we found differential BOLD-effects for the different temporal percepts (A_L, V_L, and synchrony (AV_S)] within adjacent subregions in human STS-c, plus differential interregional connectivity with prefrontal cortex.

METHODS

A temporal-threshold experiment was conducted prior to scanning, to account for any individual differences in temporal perception. By choosing bistable stimuli for each subject we maximized the number of trials per condition during the fMRI-experiment (see below). Subjects ($n = 14$, 7 female) were placed in a dark, sound-attenuated chamber after providing written informed consent in accord with local ethics. They had to report the perceived synchrony or direction of asynchrony of auditory and visual information of video sequences by pressing one of three buttons (thereby indicating A_L, AV_S, V_L). Importantly, subjects could change their judgements *during* each video presentation. The stimuli consisted of 20 video clips (length 23.7 s), depicting the face of a trained female speaker reading sentences (see **Figure 1**). Stimuli were randomized with MATLAB 6.1 and presented using Presentation 9.11

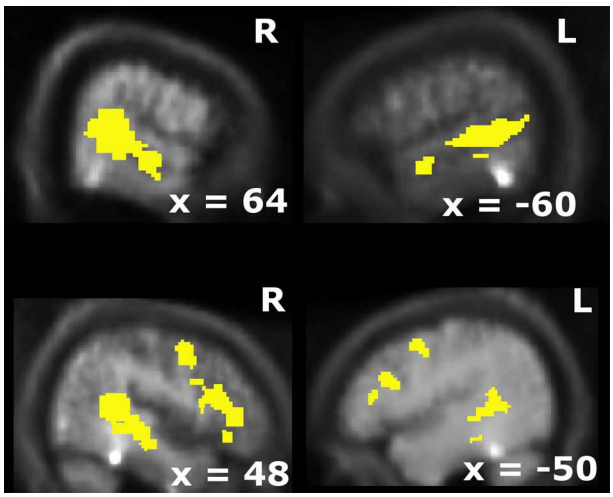


FIGURE 1 | Overlap of visual and auditory BOLD-modulations for unisensory stimulus presentations ($p < 0.005$; $k > 10$). This activation map was used as the search volume for the fMRI-analysis in the main experiment.

(Neurobehavioral Systems, Inc., CA). Initially, 20 synchronous sequences plus 80 temporally shifted sequences were presented (−130 ms, −60 ms (A_L) and 200 ms/400 ms (V_L), 20 video clips each, see **Figure 2A**). These asynchronies for threshold-determination were chosen in accord with previous reports (Dixon and Spitz, 1980). For the fMRI-experiment, those stimuli were chosen for each subject that had a similar number of synchrony and asynchrony judgments (called near-threshold below).

fMRI-DATA ACQUISITION

fMRI-data was acquired on a whole body Siemens 3 T Trio-scanner (Siemens, Erlangen, Germany) using a circular-polarized whole-head coil (BrukerBioSpin, Ettlingen, Germany). Subjects performed the same task as they had outside the scanner, reporting their responses with their right index, middle, and ring finger. Within the scanner subjects were presented three conditions: near-threshold V_L, near-threshold A_L plus the AV_S condition. All other stimulus parameters were kept as in the behavioral experiment outside the scanner except for the following: first, a baseline period of 20 s was introduced after each video clip. Second, eye movements were monitored using an fMRI-compatible infrared recording system (Kanowski et al., 2007) plus evaluation software (PupilTracker, HumanScan, Erlangen, Germany). The eye movement data was analysed with MATLAB 6.5. Third, before the main fMRI-experiment, a functional localizer was run in which only unimodal auditory or unimodal visual stimuli from the videos were presented (331 volumes covering the whole head, TR 2 s, TE 30 ms, flip 80°, resolution $64 \times 64 \times 32$ at $3.5 \times 3.5 \times 4$ mm). The derived overlapping audio-visual activation map was then used to identify candidate multisensory areas (see below). Fourth, subjects wore earplugs; perceived loudness and balance were adapted individually to ensure easy comprehension of the auditory speech sequences despite the scanner

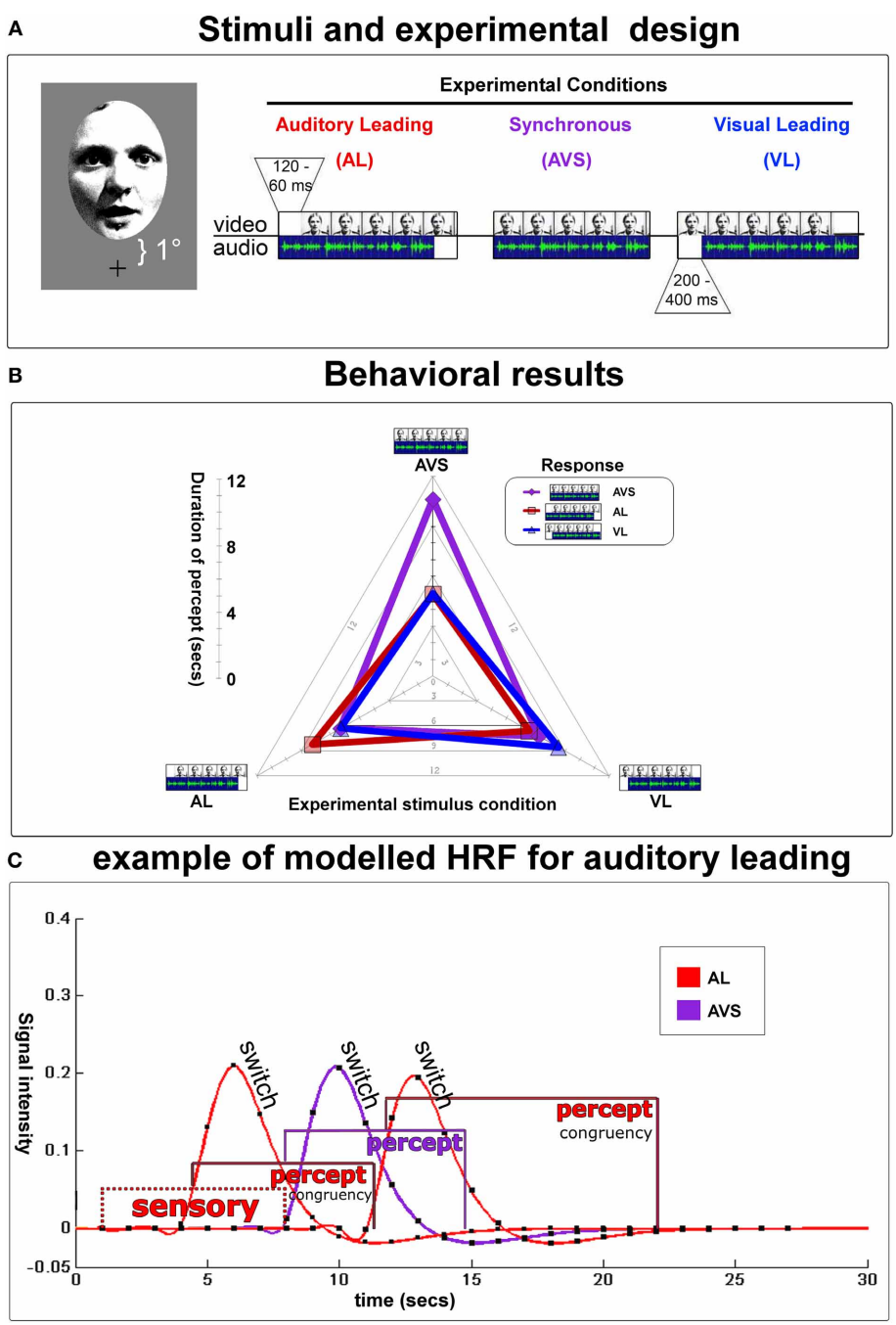


FIGURE 2 | Experimental design and behavioral results. (A) Depicts an example of a video-clip presented in three conditions [i.e., auditory leading (top left, temporal lag from 60–120 ms), auditory and video synchronous (top middle), or visual leading (top right, temporal lag from 200–400 ms)]. Auditory and visual lags were determined in a preliminary threshold-determination-experiment. Stimuli were presented at 1° visual angle above fixation (lower boundary) up to 7° (upper boundary). The duration of all 20 video-clips was 23.7 s, the interstimulus interval was 20 s. Partcipants indicated whether they perceived the auditory stream leading, the visual stream leading, or the 2 streams as being synchronous. They were encouraged to report any changes in their perception during the presentation of each video. Note that the physical lag was fixed within each video clip near the individual's synchrony/asynchrony-threshold. **(B)** Radar graph depictsmean durations (time from one keypress to the next) of subjects' (a)synchrony-percepts for each experimental condition during fMRI-scanning: perceptual states were longest when perception of (a)synchrony was congruent with physical stimulation. Therefore, in the fMRI-analysis, hemodynamic response functions (HRF) could be specifically modeled and extracted for each stable percept and perceptual switches using a mixed model (see below). **(C)** An example trial modeled with hemodynamic response functions for an auditory leading-stimulus (A_L). Gamma-curves depict perceptual switches/decisions, whereas box-car functions illustrate the sensory procesing prior to the first decision and perceptual states. Purple curves stand for AV_S, red for A_L. Note that each box-car function was individually specified based on the trial-by-trial inter-button-press duration.

noise. The stimuli were presented using MR-compatible, electrodynamic headphones (MRconfon, Magdeburg, Germany).

During the main experiment functional volumes were collected in four sessions (331 volumes each, covering the whole head, TR 2 s, TE 30 ms, flip 80°, resolution $64 \times 64 \times 32$ at $3.5 \times 3.5 \times 4$ mm). Additionally, for anatomical localization an inversion-recovery EPI was acquired (TR 2 s, TE 30 ms, TI:1450 ms, resolution $64 \times 64 \times 32$ at $3.5 \times 3.5 \times 4$ mm, same slice orientation and distortions as the functional volumes). The first five volumes from each session were excluded from further analysis. The remaining volumes were acquisition-corrected to the first acquired slice of each volume, motion-corrected, normalized at 2 mm^3 voxel size and smoothed (6 mm), using SPM2 (Wellcome Department of Cognitive Neurology, London, UK).

GROUP-LEVEL STATISTICS

After pre-processing the data from a localizer run were modeled with two box-car functions convolved with the hemodynamic response function (HRF) for the auditory and visual trials. For the localizer runs, blocks were compared to the baseline during which no stimulus was present ($p < 0.005$; $k > 10$). An audiovisual mask (i.e., overlap of unisensory visual and auditory activations) was computed to identify candidate multisensory structures (see **Figure 1**; cf. Beauchamp et al., 2004b; Beauchamp, 2005b; Noesselt et al., 2007; Szycik et al., 2008).

Next, all experimental conditions were modeled with the HRF with variable durations when appropriate (mixed model; see **Figure 2C**). In particular, 21 conditions were defined in a mixed model: three perceptual switches (subjects' button press, event-related), three perceptual states (time after button press, variable block) and the initial stimulation (time before the first button press, variable block) for every stimulus condition (A_L, V_L, and AV_S). To test condition effects, linear contrasts were used for each subject and condition and masked inclusively with the audiovisual overlap from the functional localizer. The resulting contrast images were applied to perform random effects second-level analyses. The statistical parametric maps of the *t*-statistics at each voxel were thresholded at $p < 0.05$ (small-volume-corrected) and the spatial extent threshold was set at $k >$ 5 voxels.

The following contrasts were computed: First, we identified regions that responded to physical synchrony and asynchronous conditions. Second, we identified regions that showed differential fMRI-signals for perceived synchrony vs. asynchrony conditions. Finally, we computed interaction effects for differential perceptual states with identical physical stimulation (i.e., asynchronous vs. synchronous percepts separately for A_L, V_L, and AV_S stimulation).

SINGLE-SUBJECT STATISTICS

We also analysed the data from individual subjects in order to confirm our group-level results and to test the interaction between stimulation and percepts formally. We identified for each subject regions within STS-c using the identical contrasts as in the group analysis above: for A_L stimulation: veridical A_L percept > non-veridical synchronous percept; for AV_S stimulation: veridical synchrony percept > both non-veridical percepts; for V_L stimulation: veridical V_L percept > non-veridical synchronous percept. Subject-specific regions of interest (ROI) were identified by searching for significant clusters of the three contrasts of interest within subregions of the STS-c (anatomical criterion) which expressed unisensory responses to both modalities (additional functional criterion). We extracted the beta-weights of all experimental conditions from these three local maxima for each subject and tested whether these local maxima would express significantly different results across stimulations. Note that this analysis is non-trivial and provides additional information, since any BOLD-modulation of different perceptual states to the AV_S-stimulation was left unspecified in the A_L and V_L stimulation contrasts and vice versa.

ANALYSIS OF INTERREGIONAL CONNECTIVITY

Complementary to the analysis of local modulations of the BOLD-response we investigated the effects of interregional connectivity in the context of perception of AV_S, V_L, and A_L as described above (Friston et al., 1997). We seeded our analyses in the subject-specific local maxima in STS-c and analyzed which other regions showed enhanced functional coupling in the context of A_L percepts in the A_L condition (relative to non-veridical synchronous percept in the A_L condition), in the context of V_L percepts in the V_L condition (relative to the non-veridical synchronous percept in the A_L condition) and in the context of synchronous percepts in the synchronous condition (relative to the non-veridical asynchronous percept in the synchronous condition) using a model with 21 regressors (see above) plus the physiological response and the psychophysiological interaction as two additional regressors (see e.g., Noesselt et al., 2007 for a similar approach) to reveal differential functional interregional connections in the psychological context of synchronous or asynchronous percepts. Three models were calculated separately for each STS-local maximum (corresponding to veridical A_L-percepts, veridical V_L-percepts, and veridical AV_S percepts).

Differential group-level effects were calculated with an analysis of variance (ANOVA) pertaining the three PPIs from the three connectivity analysis.

ANALYSIS OF CONSISTENT PATTERNING OF SUBREGIONS

Finally, distances between single subject maxima in STS-c were computed and analysed to reveal any systematic anatomical distribution of subjects' local maxima for the A_L, V_L, and AV_S representation. For this we used a three step approach: normalization of MNI-coordinates, calculation of distances by subtracting the normalized MNI-coordinates and calculation of Euclidian distances in three-dimensional space. In particular, for the calculation of distances, the MNI coordinates (in millimeters) of the three contrasts and their respective local maxima were scaled by adding the maximum negative value to all coordinates of one dimension so that all values were positive. This procedure was applied for the *y* and *z* extension/dimension; *x* coordinates were converted into their absolute value. Second, coordinate values of the same dimension but different local maxima were subtracted from each other (A_L/V_L minus synchrony and A_L minus V_L). Finally, we computed Euclidean distances for the

difference measures: following Pythagoras' Theorem, difference values of the *x* and *y* dimension (cathetuses) were squared and added together and the resulting value (hypotenuse) added to the squared *z* dimension difference. The square roots of the resulting values (again hypotenuse) represent the reported distances between voxels.

RESULTS

BEHAVIORAL RESULTS

The results of the behavioral experiment outside the scanner revealed that subjects' judgments became more consistent with stimulation as the audiovisual delay increased. For the auditory stream leading condition, the mean delay for near-threshold stimuli was 105 ms (±35 ms) while for the visual stream leading condition it was 227 ms (±47 ms). Inside the scanner, subjects again judged temporal relations of the video clips while fMRI-data were acquired. The eye-movement data were analysed using both deviations from fixation and eye blinks (Kanowski et al., 2007). Three subjects who showed extensive eye movements or blinking were excluded from further analysis. In the remaining 11 subjects, neither "real" eye movements nor eye blinks showed any differential effect across the experimental conditions (i.e., eye movements $< 1°$).

During each video subjects ($n = 11$) switched 5.72 (2.34 SD) times toward a "congruent" perceptual state [i.e., one during which perception and the physical stimulus were identical] vs. 3.97 (2.0) times toward a non-veridical one. Moreover, subjects maintained veridical percepts for 9.13 (3.38) s on average, whereas non-veridical percepts lasted 6.04 (2.02) s (see **Figure 2B** for length of stable durations as a function of the stimulus type). This allowed for an unbiased mixed model design (see **Figure 2C** and Kleinschmidt et al., 1998; Dosenbach et al., 2006 for similar approaches).

NEUROIMAGING RESULTS

Voxel-based group results

First, we computed candidate multisensory structures (i.e., the overlap of activation patterns found with unisensory visual and auditory stimuli before the main experiment; see Beauchamp et al., 2004b; Noesselt et al., 2007; Szycik et al., 2008, for similar approaches). These candidate multisensory structures comprised of bilateral superior temporal sulcus, bilateral anterior insula extending into prefrontal cortex plus bilateral premotor cortex.

When comparing stable perceptual states with switches we found stronger fMRI-responses in bilateral STS-c and lateral prefrontal cortex for the maintenance of perceptual states relative to switches whereas perceptual switches engaged posterior parietal regions plus anterior cingulate in accord with earlier studies (e.g., Heekeren et al., 2008). Since perceptual switches did not significantly modulate voxels within temporal regions, we then focused on the experimental effects of the different stimulus types and of stable perceptual states (i.e., inter-response intervals) within multisensory regions.

First, comparisons of AV_S vs. (V_L+A_L) perceptual states (collapsed over stimulus types) revealed modulations in adjacent subregions of bilateral multisensory STS-c, in right insular cortex, and in bilateral prefrontal areas (see **Figure 3A** and **Tables 1A,B**); note that both asynchronous and synchronous perceptual states modulated regions within STS-c, whereas only asynchronous perceptual states additionally modulated the anterior insula and prefrontal cortex (see **Table 1**). Second, comparisons of the physically AV_S minus (V_L+A_L) stimuli (regardless of perceptual states) revealed right-lateralised modulations in middle and posterior STS-c plus prefrontal cortex (see **Figure 3B**, purple spots). A_L and V_L stimuli (relative to synchronous stimuli; see **Figure 3B**, red and blue spots, respectively) showed enhanced BOLD-responses in bilateral STS-c, prefrontal cortex, and anterior insula (see **Tables 2A–C** for local maxima). Please note, that the time-related modulations are more widespread in the left hemisphere, which might be a reason for the left-sided dominance of synchronous representation reported in other studies (e.g., Miller and D'Esposito, 2005; Marchant et al., 2012).

Finally, we compared different perceptual states separately for each stimulus type (and not collapsed across stimulus type as above). Note that these stimulus-type-specific comparisons were designed to reveal perceptual effects for identical physical stimuli. Differential non-overlapping BOLD-modulations were again found in anterior insula, prefrontal cortex, and STS-c; with only asynchronous perceptions expressing higher activations in the insula and prefrontal cortex (see **Figure 4**, plus **Tables 3A–C**). Within STS-c, distinct regions for synchronous and asynchronous perceptions were observed as a function of stimulus type. BOLD-modulations for AL and VL conditions (veridically perceived as asynchronous) enclosed a region with an enhanced BOLD-response for veridically perceived AVS stimuli within the left hemisphere (see **Figure 4**, middle row and lower left panel). In the right hemisphere, regions within the STS-c responded to veridically perceived AVS and VL stimuli (see **Figure 4**, middle and bottom row). We also investigated whether we would find modulations in the fMRI-signal for the main effects of stimulus type, perception and perceptual states for each stimulus type outside the multisensory ROI. However no significant modulations were observed ($p_{FWE-corrected} < 0.05$, since we did not have any *a priori* hypothesis).

Single-subject region-of-interest approach

Because of the possible anatomical differences between subjects within the STS-c (Ochiai et al., 2004), a ROI analysis was performed within single subjects to confirm and extend voxel-based group-level responses to physical and/or perceptual (a)synchrony.

For this ROI analysis, three differential temporal percepts were evaluated for each subject with the following contrasts: veridical (asynchronous) minus non-veridical synchronous perception for A_L and V_L speech stimuli; plus synchronous minus asynchronous perception for AV_S stimulus trains. This analysis was again applied within subjects' audiovisual masks. Mean beta weights responses (proportional to percent signal change) for the subjects' perceptual states in every experimental condition were assessed for the three (a)synchrony areas and their respective local maxima. (Note, that these local maxima were identified by conducting comparisons of a limited number of perceptual states, regardless of any other differential effects between conditions.

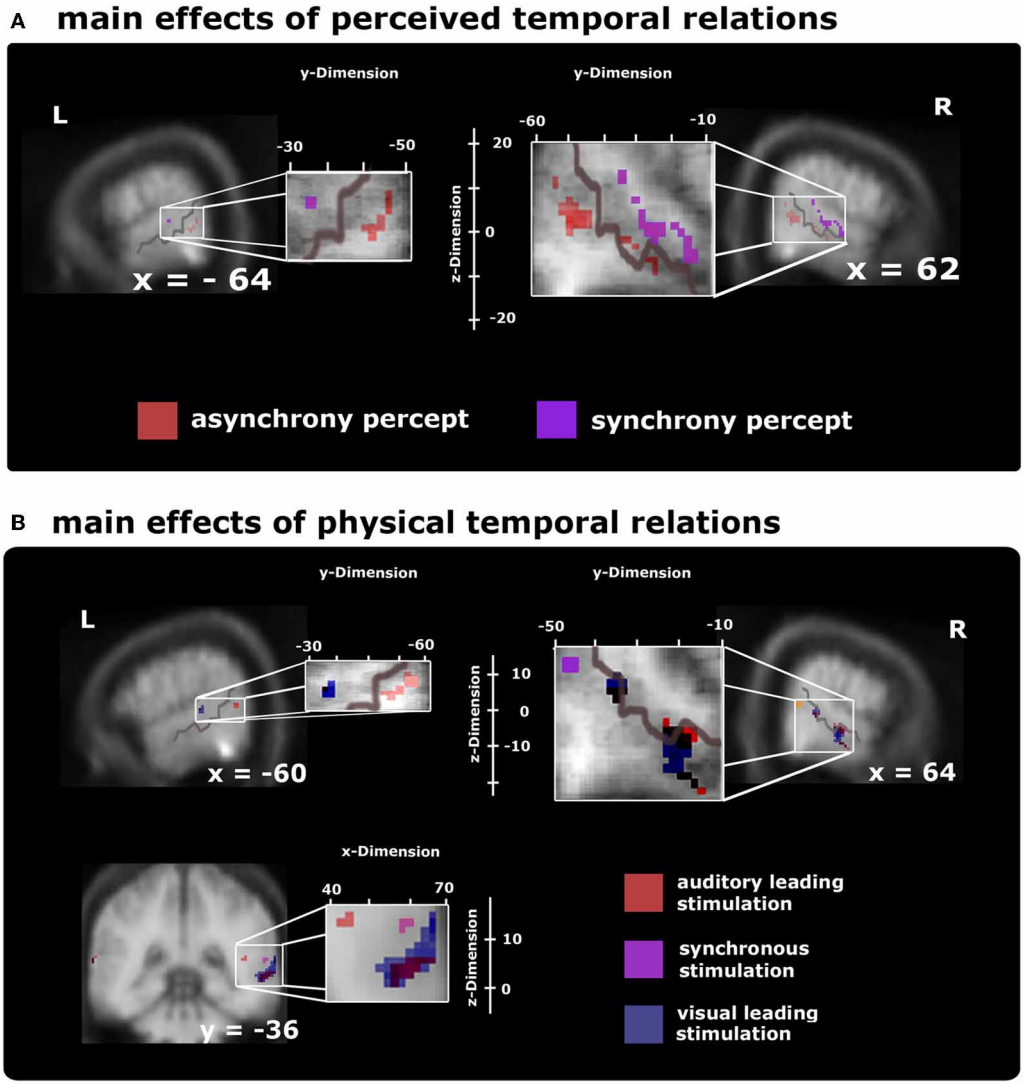


FIGURE 3 | Voxel-based group BOLD-effect of subjects' stable perceptual states (time from one keypress to the next, upper row) and the effects of the different stimulus types (lower row) within audiovisual activation maps (as defined by the overlap of unisensory stimuli) thresholded at $p < 0.05$ (small-volume-corrected). Note that the distribution of time-sensitive regions differed in the left and right hemisphere, with the left hemisphere showing a more widespread pattern than the right hemisphere (as evidenced by the formatting). **(A)** Comparison of synchrony > asynchrony percepts collapsed over stimulus type (purple spots) highlights modulations reaching from posterior to middle STS-c. Adjacent regions within STS-c were also found to be relevant for stable asynchrony percepts > synchronous ones (red spots; additionally, the asynchrony > synchrony percepts-contrast produced significant modulations in prefrontal areas; not shown, see **Table 1**). **(B)** Differential BOLD-responses for the three stimulus types collapsed over peceptual state show significant effects at the right posterior STS-c (purple spots; plus premotor regions; not shown, but see **Table 2**) for synchronous relative to asynchronous stimulation; at both STS-c (blue spots) and prefrontal areas (not shown) for visual leading relative to synchronous stimuli, and at left posterior STS-c and right anterior/posterior STS-c (red spots plus modulations at precentral gyrus and prefrontal areas; not shown, see **Table 2**) for auditory leading relative to synchronous stimulation.

Thus, the analysis of BOLD-effects reported below, will provide additional information concerning the overall response patterns within the STS-c-subregions):

We extracted the beta weights for all perceptual states (3 states × 3 stimulus types) from the three local maxima within STS-c and conducted a 2 × 3 × 3 × 3 repeated measures ANOVA with the factors of hemisphere, type of (a)synchrony area, percept, and stimulus type (see **Figure 5B**). As no effect of hemisphere was found [$F_{(1,\ 10)} < 1$; n.s.], beta weights averaged over hemispheres are displayed in **Figure 5C**. Interaction effects occurred between type of area, percept, and stimulus type [$F_{(8,\ 80)} = 3.1$; $p < 0.01$] suggesting that, within each (a)synchrony area, beta weights change as a function of the subjects' percept and stimulus type. Main effects were observed for type of (a)synchrony area [$F_{(2,\ 20)} = 4.9$; $p < 0.05$] and percept [$F_{(1.33,\ 13.26)} = 10.9$; $p < 0.01$]. Although *post-hoc t*-tests showed no significant effects, responses within the "V_L areas" were lower than in the other two areas. BOLD responses to synchronous

Table 1 | Local maxima ($p < 0.05$, $k > 5$ small-volume-corrected) for (A) synchrony minus asynchrony perception within multisensory regions (see Figure 3A, purple spots) and (B) asynchrony minus synchrony percepts (see Figure 3A, red spots) collapsed across physical stimulation.

Anatomical structure	Hemisphere	Cluster size (voxels)	*t*-value	MNI coordinates		
				x	*y*	*z*
A.SYNCHRONY PERCEPT > ASYNCHRONY PERCEPT						
Temporal Lobe						
Anterior STS	R	96	4.95 (0.001)	60	−22	−2
Anterior STS	–	22	3.68 (0.005)	62	−10	−10
Posterior/middle STS	R	32	2.84 (0.01)	48	−38	8
Posterior/middle STS	L	13	3.47 (0.007)	−58	−34	−4
B. ASYNCHRONY PERCEPT > SYNCHRONY PERCEPT						
Temporal Lobe						
Posterior/middle STS	R	206	5.63 (0.000)	66	−34	−6
Posterior/middle STS/MTG	R	10	2.98 (0.007)	58	−40	−8
Posterior/middle STS	R	14	2.92 (0.008)	54	−44	18
Posterior/middle STS	L	17	3.56 (0.006)	−66	−50	2
Frontal Lobe						
Anterior insula	R	14	2.66 (0.01)	42	36	−10
Prefrontal cortex	R	644	10.09 (0.000)	56	24	22
Prefrontal cortex	L	9	2.44 (0.02)	−54	30	14

MNI, Montreal Neurological institute; L, left; R, right.

stimuli were significantly lower than to asynchronous stimuli [$t_{(20)} = -3.53$; $p < 0.01$]. Interaction effects occurred between hemisphere and type of area [$F_{(2,\ 20)} = 8.04$; $p < 0.01$], type of area and percept [$F_{(4,\ 40)} = 3.48$; $p < 0.05$], type of area and stimulus type [$F_{(1.73,\ 17.3)} = 9.17$; $p < 0.01$], percept and stimulus type [$F_{(1.7,\ 17.06)} = 4.7$; $p < 0.05$].

Further analysis of the ANOVA-data (*post-hoc t*-tests) revealed that for each stimulus category, subjects' BOLD responses were highest when a veridical judgment was made. Within the "A_L area" (red), the mean BOLD response was highest when subjects perceived an A_L stimulus as A_L (veridical percept). The according beta weight differed statistically from the two other beta weights and their respective perceptual states [$t_{(10)} = 3.12$; $p < 0.05$], whereas the beta weights of the non-veridical percepts did not differ statistically from each other. The same pattern of results was also observed for the AV_S region (yellow) [$t_{(10)} = 4.76$; $p < 0.001$] and V_L percepts (blue) [$t_{(10)} = 2.72$; $p < 0.05$]. Since, in the AV_S area, veridical responses were not significantly different from BOLD-responses for other stimulus types, this region may serve additional sub-functions on top of the maintenance of synchrony perception. In general, these ROI-results reaffirm the functional micro-compartmentalization of the STS-c found in the voxel-based group results into areas specialized for the perception of distinct audiovisual temporal patterns.

Interregional connectivity of STS-c-regions

Moreover, we assessed whether the subregions within STS-c that consistently expressed differential local activity (see **Figure 5**) would also be functionally linked to other multisensory regions. We used the assumption-free "psychophysiological interaction" (PPI; Friston et al., 1997) and seeded our analysis in subject-specific STS-c maxima. We analysed whether the strength of functional coupling of these adjacent STS-c-regions with other multisensory regions would differ. We found that both A_L and V_L-regions in bilateral STS-c showed a significantly stronger coupling with right prefrontal regions than did the AV_S-region (see **Figure 6** and **Table 4**). Moreover, synchronous patches with the middle STS-c expressed a stronger functional connection with posterior STS-c regions in the left hemisphere, whereas asynchronous patches showed a stronger coupling with posterior STS-c in the right hemisphere (see **Table 4**).

Spatial configuration of STS-c-subregions

In addition, we evaluated whether the spatial configuration of the identified sub-regions within bilateral STS-c showed a systematic spatial distribution across subjects: the analysis revealed that perceived asynchrony (A_L or V_L) and synchrony modulated distinct regions along the STS-c which were situated adjacent to one another (with asynchrony enclosing synchrony modulations). For every subject, this specific pattern differed in its position along STS-c but occurred regularly (see **Figure 5A** for average, **Figure 5B** for illustrative subjects). Distances between the areas modulated by an interaction of stimulus type and perception were calculated. We found that, on average, the local maxima of the "A_L" and "V_L areas" were situated closer to "synchrony areas" (12.1 and 11.1 mm) than to each other (17.6 mm).

A 2 × 3 repeated measures ANOVA with the factors hemisphere and distance showed an effect of distance [$F_{(2,\ 20)} = 10.2$; $p < 0.001$]. The distance between the "asynchrony areas" was statistically different from their respective distance to the "synchrony area" [A_L: $t_{(10)} = 3.77$; $p < 0.05$; V_L: $t_{(10)} = 3.40$; $p < 0.05$]; the distances between the asynchrony areas and the "synchrony area" were similar [$t_{(10)} = 0.63$; $p = 0.55$]. There was no effect of hemisphere [$F_{(1,\ 10)} < 1$; n.s.], nor any interaction between hemisphere and distance [$F_{(2,\ 20)} < 1$; n.s.].

Table 2 | Local maxima ($p < 0.05$, $k > 5$ small-volume-corrected) for (A) AV_S minus (A_L+V_L) stimulation within multisensory regions (see Figure 3B, purple spots); (B) V_L minus synchrony stimulation (see Figure 3B, blue spots); and (C) A_L minus synchrony stimulation (see Figure 3B, red spots) collapsed across perceptual states.

Anatomical structure	Hemisphere	Cluster size(voxels)	*t*-value	MNI coordinates		
				x	*y*	*z*
A. PHYSICAL SYNCHRONY > PHYSICAL ASYNCHRONY						
Temporal cortex						
STS	R	9	2.39 (0.03)	54	−46	14
STS	R	8	2.31 (0.04)	62	−50	10
Frontal lobe						
Prefrontal cortex	R	14	2.92 (0.008)	50	36	12
B. PHYSICAL VISUAL LEADING ASYNCHRONY > PHYSICAL SYNCHRONY						
Temporal lobe						
Anterior STS	R	370	3.81 (0.001)	64	−20	−12
Middle STS	L	9	3.05 (0.005)	−68	−38	14
Middle STS	L	14	2.95 (0.007)	−60	−30	8
Frontal lobe						
Prefrontal cortex	R	41	2.78 (0.01)	38	18	26
Prefrontal cortex	L	11	2.73 (0.01)	−46	20	28
Anterior insula	R	8	2.99 (0.006)	50	42	2
Anterior insula/IFG	L	41	3.88 (0.001)	−36	38	−16
C. PHYSICAL AUDITORY LEADING ASYNCHRONY > PHYSICAL SYNCHRONY						
Temporal Lobe						
Anterior STS	R	177	3.55 (0.002)	62	−14	−8
Posterior/middle STS	R	122	3.32 (0.002)	54	−46	−2
Middle STS	L	12	4.07 (0.001)	−68	−38	14
Posterior STS	L	57	2.86 (0.009)	−54	−54	8
Frontal Lobe						
Precentral gyrus	R	17	3.27 (0.003)	44	0	40
Precentral gyrus	R	7	2.47 (0.02)	48	6	44
Anterior insula/IFG	L	6	3.31 (0.002)	−36	40	−18
Prefrontal cortex	L	19	2.46 (0.02)	−46	22	24

MNI, Montreal neurological institute; L, left; R, right.

DISCUSSION

The present study investigated the neural basis of both the processing of physical properties and subjective perception of the temporal relationship between auditory and visual speech stimuli, thereby pinpointing the functional neuroanatomy of audiovisual temporal processing and perception in multisensory cortex in humans. We found that sub-regions within the superior temporal sulcus have a distinct response pattern during the maintenance of perceptual states and for the processing of physical stimulus differences regardless of subjects' perceptual state. Within lateral prefrontal regions and anterior insula only the perception of asynchrony was consistently linked to an increase in BOLD-response. A ROI-based single-subject analysis corroborated and extended this pattern: three subregions within the STS-c showed a differential response for the different physical stimuli (AL, VL, and AVS). Responses were further enhanced if subjects' perceptual states were congruent to the physical stimulus being presented. Further, analyses of interregional connectivity suggest that during the perception of asynchronous stimuli AL and VL regions within the STS-c are coupled more strongly to lateral prefrontal regions, whereas connectivity within posterior STS-c was lateralized with stronger connections of the middle with posterior STS-c in the left hemisphere for synchrony patches and with posterior STS-c in the right hemisphere for asynchronous patches. Finally, analysis of the anatomical patterning of these regions suggests that they are distributed regularly within the STS-c with a synchrony region being enclosed by asynchrony regions.

Previous neuroimaging studies have reported that the STS-c (among other structures) is involved in audiovisual temporal processing and synchrony perception (Calvert, 2001; Macaluso et al., 2004; Miller and D'Esposito, 2005; Dhamala et al., 2007; Noesselt et al., 2007; Stevenson et al., 2010; Marchant et al., 2012; see Driver and Noesselt, 2008, for a review). However, most of these studies investigating the crossmodal binding of semantically meaningful stimuli (Calvert et al., 2000; Calvert and Campbell, 2003; Macaluso et al., 2004) did not separate task- and perception-related effects; their reported modulations may therefore reflect a mixture of stimulus-, decision-, and perception-related processing.

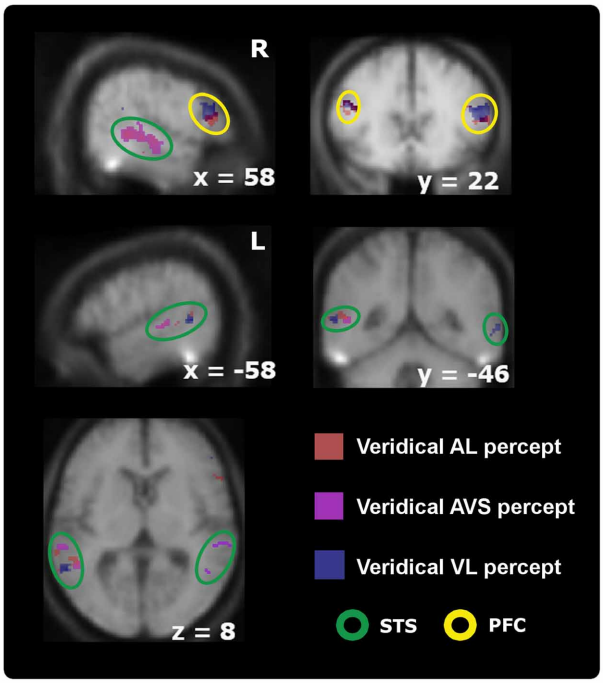


FIGURE 4 | Activation maps comparing participants' stable veridical percepts (i.e., identical with the physical stimulus) against non-veridical percepts within audiovisual regions thresholded at $p < 0.05$; $k > 5$ (small-volume-corrected). Whereas the right-sided STS-c is only modulated by synchronous > asynchronous percepts (orange spots, upper left panel) when processing synchronous stimuli, the results also show higher activations for asynchronous judgments during asynchronous stimulation [both auditory (red spots) and visual leading (blue spots); represents coherence] compared to synchronous judgments during asynchronous stimulation (incoherence) within prefrontal regions (top right) and the left STS-c region (middle and lower row). Note that within this area analysis revealed distinct spots for each veridical percept. Prefrontal regions were only modulated by veridical percepts during asynchrony stimulation (see **Table 3**).

Previous research (Miller and D'Esposito, 2005; Stevenson et al., 2010) reported effects of the temporal fusion of short AV-syllables using event-related fMRI. Stevenson and his colleagues (2010) reported functional subregions within STS-c, that preferentially processed asynchronous or synchronous speech. Miller and D'Esposito (2005) reported left-hemispheric modulations within STS-c for perceptual fusion and right hemispheric effects for perceptual segregation. However, the differences in stimulus materials used in the various conditions may explain the different activation maps reported there. Nonetheless, while we did not find lateralized effects of the local fMRI-signal, our interregional connectivity analysis revealed a lateralized pattern, that accord with Miller and D'Esposito.

Other studies have investigated the effects of audiovisual timing with streams of simple stimuli: Calvert et al. (2001) investigated multisensory interactions using simple synchronized and desynchronized audiovisual stimulus sequences. Synchronous or asynchronous bimodal inputs showed non-linear enhancements or suppressions (respectively) of BOLD-responses in multisensory areas, including STS-c, plus frontal regions. Noesselt et al. (2007) reported effects in contralateral STS-c for the processing of lateralized non-semantic synchronous audiovisual stimuli, but did not report effects for asynchronous audiovisual stimuli. In a related study, Marchant et al. (2012) observed left-sided synchrony representations in left STS-c. Meanwhile, van Atteveldt et al. (2004, 2007) identified lateral temporal areas (PT, STP, and STS-c) as major integration sites whenever audiovisual grapheme-morpheme pairs were being processed. While the intensity of modulations increased in auditory areas for semantically congruent conditions, the location of modulations within the STS-c changed as a function of the temporal distance/delay between vision and sound: asynchrony was predominantly processed at the eccentricity of the STS-c activation pattern, whereas smaller temporal delays were related to the activation's core region. However, no effect of synchrony was reported for synchronous audiovisual letters in the STS-c and the reported activations for different audiovisual lags overlapped substantially.

In the present study, asynchronous percepts engaged the posterior STS-c, the anterior insula, and the prefrontal cortex bilaterally. Our results accord with previous imaging studies on temporal asynchrony which reported right-sided effects within the STS-c, supplementary motor areas (Miller and D'Esposito, 2005) and prefrontal (MFG, IFG) cortices (Bushara et al., 2001; Dhamala et al., 2007) in the perception of asynchrony. Our findings corroborate previous results and suggest that audiovisual prefrontal areas and the STS-c are functionally linked during the maintenance of the perception of audiovisual asynchrony. There is also corroborating anatomical evidence that the STS-c is reciprocally linked to prefrontal regions (see e.g., Yeterian et al., 2012). We speculate that the perception of asynchronous percepts may be more demanding than synchrony perception and requires the on-line updating of two separate working memory representations in prefrontal cortex with input from the STS-c. Alternatively, the separation of auditory and visual input may be processed by prefrontal cortical regions (in line with the notion of a hierarchical multisensory processing model, see e.g., Noppeney et al., 2010) and fed back into the STS-c. Future research in non-human primates or in humans using transcranial magnetic stimulation/transcranial direct current stimulation is needed to disentangle these two possibilities.

Most remarkably of all, our results indicate that the multisensory superior temporal sulcus complex (mSTS-c) can be further differentiated into subregions that process particular audiovisual temporal patterns. Anatomical studies in non-human primates that have investigated the anatomical texture of TPO (the likely homologue to the human STS-c; Beauchamp, 2005a) have provided evidence for three caudal-to-rostral subdivisions within this region (Cusick et al., 1995). Those subdivisions are distinct in terms of their chemoarchitecture. Seltzer and Pandya (1991) provided evidence that TPO consists of cytoarchitectonic subdivisons of which particularly the rostral part is directly connected to the insula. Further chemoarchitectonic results support the view that the upper bank of TPO in the rhesus monkey contains several different anatomical and functional zones

Table 3 | Local maxima ($p < 0.05$, $k > 5$ small-volume-corrected) for (A) Auditory leading minus synchrony percepts during A_L stimulation within multisensory regions (see Figure 4, red spots); (B) synchrony minus ($A_L + V_L$) percepts during AV_S stimulation (see Figure 4, purple spots); and (C) V_L minus synchrony percepts during V_L stimulation (see Figure 4, blue spots).

Anatomical structure	Hemisphere	Cluster size (voxels)	*t*-value	MNI coordinates		
				x	*y*	*z*
A. COHERENT AUDITORY LEADING PERCEPT > COHERENT SYNCHRONY PERCEPT						
Temporal lobe						
Posterior/middle STS	L	57	2.62 (0.01)	−52	−42	4
Posterior/middle STS	L	18	2.96 (0.01)	−64	−38	10
Posterior/middle STS	L	6	2.85 (0.01)	−64	−36	−8
Middle STS	R	8	3.43 (0.005)	−54	−30	−14
Anterior/middle STS	R	15	2.95 (0.007)	−64	−42	12
Frontal lobe						
Anterior insula	L	20	3.64 (0.001)	−32	28	−6
anterior insula	R	66	6.52 (0.000)	42	32	−6
Precentral gyrus	R	86	3.98 (0.002)	48	10	40
precentral gyrus	L	28	4.67 (0.000)	−36	8	60
Precentral gyrus	L	11	3.41 (0.006)	−40	8	38
Prefrontal cortex	R	191	3.25 (0.003)	54	28	12
Prefrontal cortex	L	12	2.97 (0.01)	−50	6	44
B. COHERENT SYNCHRONY PERCEPT > COHERENT ASYNCHRONY PERCEPT						
Anterior STS	R	447	4.24 (0.001)	62	−16	−4
Anterior STS	L	62	4.05 (0.002)	−62	−24	4
Posterior STS	R	5	2.66 (0.02)	48	−52	10
Posterior/middle STS	L	29	3.28 (0.007)	−50	−46	8
Frontal lobe						
Anterior Insula/Prefrontal	L	14	2.63 (0.01)	−34	40	−14
C. COHERENT VISUAL LEADING PERCEPT > COHERENT SYNCHRONY PERCEPT						
Temporal lobe						
Posterior STS	R	18	3.5 (0.002)	64	−50	2
Posterior STS	R	5	2.41 (0.02)	62	−50	14
Posterior STS	L	29	3.47 (0.002)	−58	−50	8
Frontal lobe						
Anterior insula	R	72	4.77 (0.000)	44	40	−10
Anterior insula	L	20	2.71 (0.008)	−34	30	−2
Precentral gyrus	R	21	3.42 (0.002)	42	8	46
Prefrontal cortex	R	451	3.76 (0.001)	50	24	24
Precentral cortex	L	134	3.56 (0.002)	−42	16	26

MNI, Montreal Neurological institute; L, left; R, right.

(Padberg et al., 2003). They demonstrated that within those distinct neurochemical/connectional modules the STS-c shows a patchy organization of connections toward other cerebral regions. Those patches within the STS-c may have functional relevance. In a functional imaging study, Beauchamp et al. (2004a) reported that STS-c can be parcellated into unisensory auditory, visual, and multisensory patches. Our imaging analysis extends these findings and reveals distinct multisensory patches along the STS-c that encode separate audiovisual temporal patterns when the synchrony/asynchrony of continuous speech is being judged. Given that the identified synchrony patches lie in-between auditory- and visual-leading audiovisual patches, these modulations build up a chronological array that suggests the existence of a "time line." Moreover, another publication (Fairhall and Macaluso, 2009) also reported a modulation of the fMRI-signal due to attention within middle but not posterior STS-c, when subjects processed congruent audiovisual speech, thereby suggesting a large-scale segregation of the STS-c along the anterior-posterior axis (though asynchronous representations seem to be more variable; see **Tables 1**–**3**). Moreover, Marchant et al. (2012) investigated the correspondence of an audiovisual behavioral benefit on BOLD-modulations in the cerebrum and found significant effects in middle but not posterior STS-c for synchronous stimulus trains. The results from our study—revealing an interaction effects in middle STS-c specific for temporal patterns and their perception plus an enhanced connectivity with more posterior regions—are in accord with this proposition (though note that our results did not reveal a clear anterior-posterior

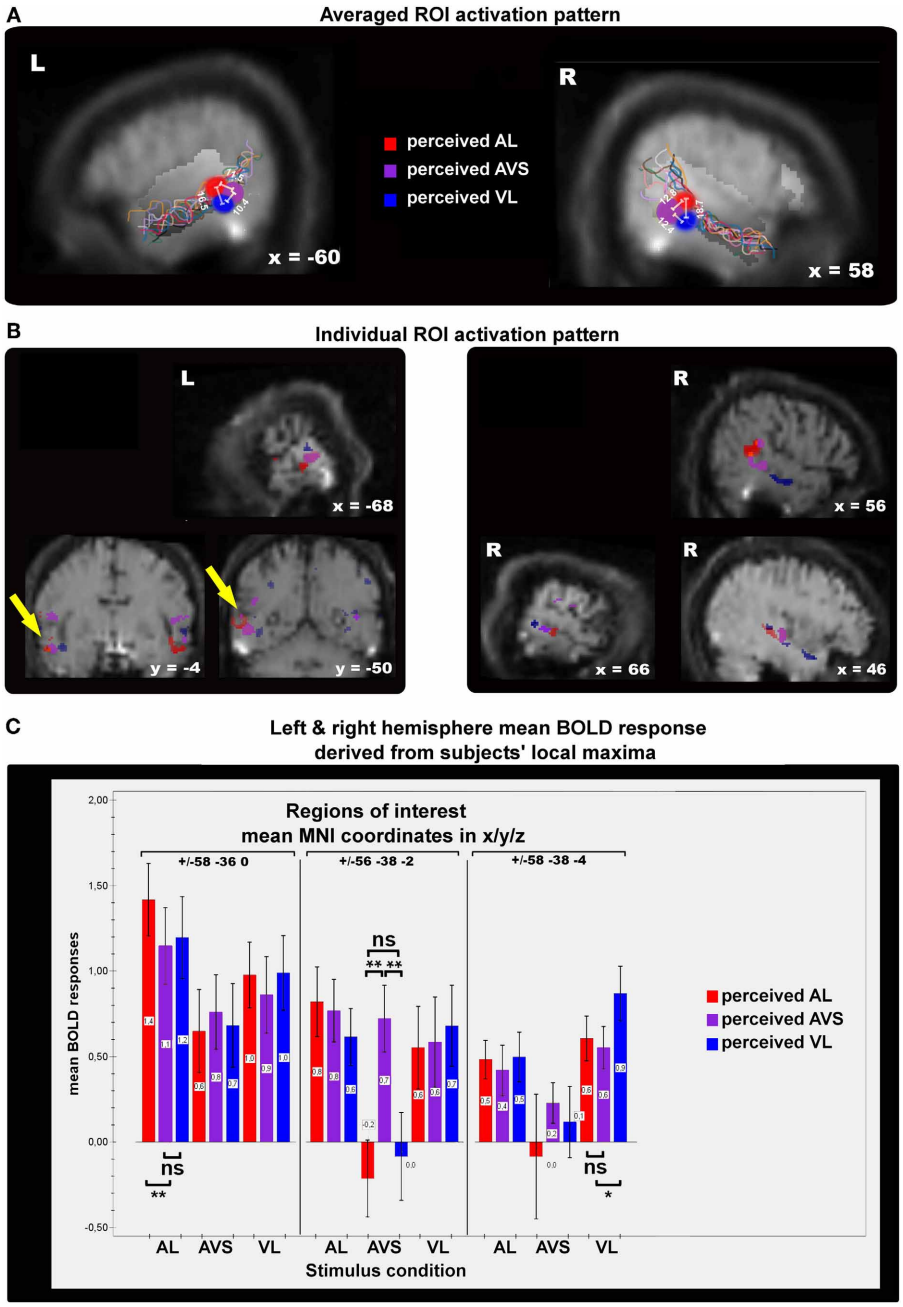


FIGURE 5 | Panels showing the results of single-subject analyses. The contrasts displayed here represent subjective perceptions that were congruent with physical stimulation > incogruent perceptions for auditory/visual leading (AL,VL, red/blue spots) and synchronous stimuli (AVS, purple spots). **(A)** The colored spots indicate average local maxima (11 subjects) of areas that express higher activations for veridical percepts (see main text for contrast definitions) than for non-veridical ones within the STS-c region (region of interest). The white lines and their corresponding numbers display averaged distances in millimeters from one activation spot to the two others (see "Methods" section for details). Note that asynchrony spots are always more distant from each other than from synchrony activation. Colored lines show the individual anatomical curvatures of STS-c of the all subjects after normalization. **(B)** The middle row depicts the activation maps of three individual subjects for the above-described contrasts. Note that synchrony spots (purple) are enclosed by two asynchrony spots [blue and red spots; see also distances in panel **(A)**]. Such activation patterns were found in both left and right hemispheres. **(C)** Mean beta-weights (proportional to % signal change) for the local maxima in panel **(A)** were collapsed over hemispheres. Bars show the height of the BOLD-effect (y-axis) for each stable percept [auditory leading (red bars), synchronous (orange bars), and visual leading (blue bars)] for the three stimulus types (auditory leading, visual leading, and synchrony, x-axis) within each each of the local maxima shown in panel **(A)** [auditory leading percept maximum (left graph section), synchrony percept maximum (middle section of graph), and visual leading percept (right graph section)]. BOLD-responses to asynchrony percepts *within* asynchrony percept maxima were always higher (outer left and right bars) than to any other percept for the different stimulus types. Within the synchrony percept maximum BOLD-responses to synchrony percepts were higher than asynchrony percepts whenever synchronous video clips were presented.

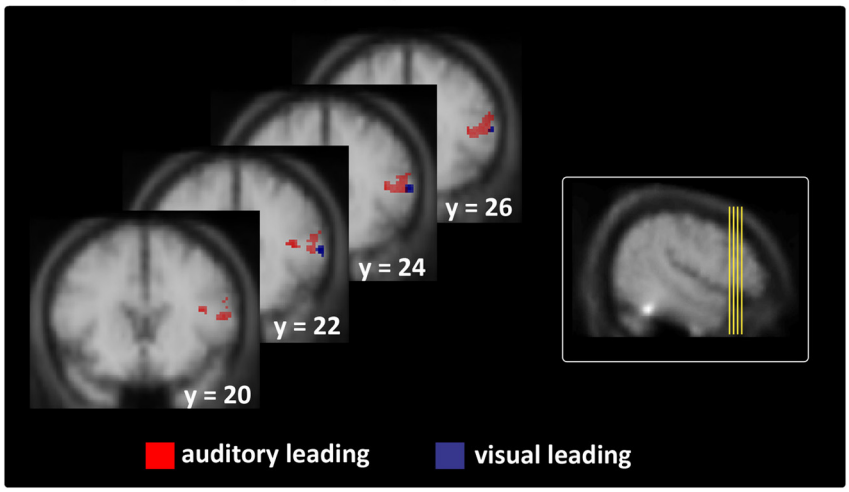


FIGURE 6 | Interregional connectivity of subjects' stable veridical percepts (i.e., identical with the physical stimulus) during asynchronous stimulation thresholded at $p < 0.05$; $k > 5$ (small-volume-corrected). Left column: right prefrontal regions only expressed stronger coupling with temporal regions in the context of veridical asynchronous > non-veridical synchronous perceptions during A_L and V_L stimulation (see **Table 4** for all maxima). Right column shows the origin of the brain sections depicted on the left on a lateral group mean view.

Table 4 | Local Maxima ($p < 0.05$, $k > 5$ small-volume-corected) of interregional connectivity in the context of (A) veridical synchrony percepts (relative to non-veridical asynchrony percepts) during AV_S stimulation; (B) veridical auditory leading percepts (relative to non-veridical synchrony percepts) during A_L stimulation; (C) veridical visual leading percepts (relative to non-veridical synchrony percepts) during V_L stimulation.

Anatomical structure	Hemisphere	Cluster size (voxels)	*p*-value	MNI coordinates		
				x	*y*	*z*
A. PSYCHOPHYSICAL INTERACTION OF SYNCHRONY PERCEPTS						
Temporal regions						
Posterior STS	L	31	4.32 (0.001)	−54	−54	12
B. PSYCHOPHYSICAL INTERACTION OF AUDITORY LEADING PERCEPTS						
Temporal regions						
Anterior STS	R	73	3.57 (0.001)	64	−12	−8
Posterior STS	R	22	3.32 (0.001)	−50	−46	16
Frontal regions						
Middle/inferior frontal gyrus	R	168	2.94 (0.002)	40	22	16
precentral gyrus	L	54	2.68 (0.005)	−48	−2	46
C. PSYCHOPHYSICAL INTERACTION OF VISUAL LEADING PERCEPTS						
Temporal regions						
Posterior STS/STG	R	44	2.42 (0.009)	52	−46	0
Frontal regions						
Inferior frontal gyrus	R	21	2.51 (0.008)	58	22	14

distinction for the main effects of physical vs. perceptual states). Finally, our results could be applied to nonhuman primates to enable more invasive measures [combined with fMRI (see Tsao et al., 2006)] to identify the pathways and neural mechanisms involved. A study in non-human primates on audiovisual face-voice integration (Ghazanfar et al., 2008) reported enhanced coupling of STS-c-neurons with auditory areas when processing audiovisual stimuli (Schroeder et al., 2008). Our results would predict the existence of distinct patches within mSTS-c that may differentially engage unisensory cortices via feedback connections (Driver and Noesselt, 2008).

In conclusion, we found a distinct pattern of modulations within mSTS-c reflecting an interaction between perceptual state and the physical properties of audiovisual speech stimuli. Our data therefore suggest that there is an aligned spatial representation of audiovisual temporal patterns parcellating the multisensory STS-c in humans, with differential functional connections to multisensory prefrontal regions.

ACKNOWLEDGMENTS

Tömme Noesselt, Daniel Bergmann, and Hans-Jochen Heinze were supported by DFG-SFB-TR31/TPA8; Thomas Münte by DFG-SFB-TR31/TPA7 and Charles Spence by the Alexander von Humboldt Foundation, Germany. Tömme Noesselt, Daniel Bergmann, and Charles Spence planned the experiment. Daniel Bergmann collected and analysed the data. All authors were involved in writing the MS.

REFERENCES

Barraclough, N. E., Xiao, D., Baker, C. I., Oram, M. W., and Perrett, D. I. (2005). Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *J. Cogn. Neurosci.* 17, 377–391.

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

Beauchamp, M. S. (2005b). Statistical criteria in fMRI studies of multisensory integration. *Neuroinformatics* 3, 93–113.

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., and Martin, A. (2004a). Unraveling multisensory integration: patchy organization within human STS-c multisensory cortex. *Nat. Neurosci.* 7, 1190–1192.

Beauchamp, M. S., Lee, K. E., Argall, B. D., and Martin, A. (2004b). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron* 41, 809–823.

Benevento, L. A., Fallon, J., Davis, B. J., and Rezak, M. (1977). Auditory-visual interaction in single cells in the cortex of the superior temporal sulcus and the orbital frontal cortex of the macaque monkey. *Exp. Neurol.* 57, 849–872.

Bruce, C., Desimone, R., and Gross, C. G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *J. Neurophysiol.* 46, 369–384.

Bushara, K. O., Grafman, J., and Hallett, M. (2001). Neural correlates of auditory-visual stimulus onset asynchrony detection. *J. Neurosci.* 21, 300–304.

Calvert, G. A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cereb. Cortex* 11, 1110–1123.

Calvert, G. A., and Campbell, R. (2003). Reading speech from still and moving faces: the neural substrates of visible speech. *J. Cogn. Neurosci.* 15, 57–70.

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.

Calvert, G. A., Hansen, P. C., Iversen, S. D., and Brammer, M. J. (2001). Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. *Neuroimage* 14, 427–438.

Cusick, C. G., Seltzer, B., Cola, M., and Griggs, E. (1995). Chemoarchitectonics and corticocortical terminations within the superior temporal sulcus of the rhesus monkey: evidence for subdivisions of superior temporal polysensory cortex. *J. Comp. Neurol.* 360, 513–535.

Dennett, D. (1991). *Consciousness Explained*. London: Penguin Press.

Desimone, R., and Gross, C. G. (1979). Visual areas in the temporal cortex of the macaque. *Brain Res.* 178, 363–380.

Dhamala, M., Assisi, C. G., Jirsa, V. K., Steinberg, F. L., and Kelso, J. A. (2007). Multisensory integration for timing engages different brain networks. *Neuroimage* 34, 764–773.

Dixon, N. F., and Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception* 9, 719–721.

Dosenbach, N. U., Visscher, K. M., Palmer, E. D., Miezin, F. M., Wenger, K. K., Kang, H. C., Burgund, E. D., Grimes, A. L., Schlaggar, B. L., and Petersen, S. E. (2006). A core system for the implementation of task sets. *Neuron* 50, 799–812.

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

Fairhall, S., and Macaluso, E. (2009). Spatial attention can modulate audiovisual integration at multiple cortical and subcortical sites. *Eur. J. Neurosci.* 29, 1247–1257.

Friston, K. J., Buechel, C., Fink, G. R., Morris, J., Rolls, E., and Dolan, R. J. (1997). Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage* 6, 218–229.

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.

Gil-da-Costa, R., Braun, A., Lopes, M., Hauser, M. D., Carson, R. E., Herscovitch, P., and Martin, A. (2004). Toward an evolutionary perspective on conceptual representation: species-specific calls activate visual and affective processing systems in the macaque. *Proc. Natl. Acad. Sci. U.S.A.* 101, 17516–17521.

Heekeren, H. R., Marrett, S., and Ungerleider, L. G. (2008). The neural systems that mediate human perceptual decision making. *Nat. Rev. Neurosci.* 9, 467–479.

Hikosaka, K., Iwai, E., Saito, H., and Tanaka, K. (1988). Polysensory properties of neurons in the anterior bank of the caudal superior temporal sulcus of the macaque monkey. *J. Neurophysiol.* 60, 1615–1637.

Kanowski, M., Rieger, J. W., Noesselt, T., Tempelmann, C., and Hinrichs, H. (2007). Endoscopic eye tracking system for fMRI. *J. Neurosci. Methods* 160, 10–15.

Kelly, S. D. (2005). "The puzzle of temporal experience," in *Cognition and the Brain: The Philosophy and Neuroscience movement,* ed A. B. K. Akins (Cambridge: Cambridge University Press), 208–240.

Kleinschmidt, A., Buchel, C., Zeki, S., and Frackowiak, R. S. (1998). Human brain activity during spontaneously reversing perception of ambiguous figures. *Proc. Biol. Sci.* 265, 2427–2433.

Köhler, W. (1947). *Gestalt Psychology: An Introduction to New Concepts in Modern Psychology.* New York, NY: Liveright Publication Corporation.

Macaluso, E., George, N., Dolan, R., Spence, C., and Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech: a PET study. *Neuroimage* 21, 725–732.

Marchant, J. L., Ruff, C. C., and Driver, J. (2012). Audiovisual synchrony enhances BOLD responses in a brain network including multisensory STS while also enhancing target-detection performance for both modalities. *Hum. Brain Mapp.* 33, 1212–1224.

McGrath, M., and Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *J. Acoust. Soc. Am.* 77, 678–685.

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.

Noesselt, T., Rieger, J. W., Schoenfeld, M. A., Kanowski, M., Hinrichs, H., Heinze, H. J., and Driver, J. (2007). Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *J. Neurosci.* 27, 11431–11441.

Noppeney, U., Ostwald, D., and Werner, S. (2010). Perceptual decisions formed by accumulation of audiovisual evidence in prefrontal cortex. *J. Neurosci.* 30, 7434–7446.

Ochiai, T., Grimault, S., Scavarda, D., Roch, G., Hori, T., Riviere, D., Mangin, J. F., and Regis, J. (2004). Sulcal pattern and morphology of the superior temporal sulcus. *Neuroimage* 22, 706–719.

Padberg, J., Seltzer, B., and Cusick, C. G. (2003). Architectonics and cortical connections of the upper bank of the superior temporal sulcus in the rhesus monkey: an analysis in the tangential plane. *J. Comp. Neurol.* 467, 418–434.

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.

Seltzer, B., and Pandya, D. N. (1991). Post-rolandic cortical projections of the superior temporal sulcus in the rhesus monkey. *J. Comp. Neurol.* 312, 625–640.

Slutsky, D. A., and Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect. *Neuroreport* 12, 7–10.

Spence, C., and Squire, S. (2003). Multisensory integration: maintaining the perception of synchrony. *Curr. Biol.* 13, R519–R521.

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

Stevenson, R. A., Altieri, N. A., Kim, S., Pisoni, D. B., and James, T. W. (2010). Neural processing of asynchronous audiovisual speech perception. *Neuroimage* 49, 33–38.

Szycik, G. R., Tausche, P., and Munte, T. F. (2008). A novel approach

to study audiovisual integration in speech perception: localizer fMRI and sparse sampling. *Brain Res.* 1220, 142–149.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B., and Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science* 311, 670–674.

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.

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.

Vatakis, A., and Spence, C. (2006a). Audiovisual synchrony perception for music, speech, and object actions. *Brain Res.* 1111, 134–142.

Vatakis, A., and Spence, C. (2006b). Audiovisual synchrony perception for speech and music assessed using a temporal order judgment task. *Neurosci. Lett.* 393, 40–44.

Vroomen, J., and Keetels, M. (2010). Perception of intersensory synchrony: a tutorial review. *Atten. Percept. Psychophys.* 72, 871–884.

Wright, T. M., Pelphrey, K. A., Allison, T., McKeown, M. J., and McCarthy, G. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cereb. Cortex* 13, 1034–1043.

Yeterian, E. H, Pandya, D. N, Tomaiuolo, F, and Petrides, M. (2012). The cortical connectivity of the prefrontal cortex in themonkey brain. *Cortex* 48, 58–81.

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: 06 June 2012; paper pending published: 20 June 2012; accepted: 07 August 2012; published online: 28 August 2012.

Citation: Noesselt T, Bergmann D, Heinze H-J, Münte T and Spence C (2012) Coding of multisensory temporal patterns in human superior temporal sulcus. Front. Integr. Neurosci. ***6****:64. doi: 10.3389/fnint.2012.00064*

frontiers in
INTEGRATIVE NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 20 November 2012
doi: 10.3389/fnint.2012.00105

Learning from vision-to-touch is different than learning from touch-to-vision

Dagmar A. Wismeijer*, Karl R. Gegenfurtner and Knut Drewing*

Allgemeine Psychologie, Justus-Liebig Universität Gießen, Gießen, Germany

Edited by:
Zhuanghua Shi, Ludwig-Maximilians-Universität München, Germany

Reviewed by:
Koichi Sameshima, Universidade de São Paulo, Brazil
Patrizia Fattori, University of Bologna, Italy

****Correspondence:***
Dagmar A. Wismeijer and Knut Drewing, Allgemeine Psychologie, Justus-Liebig Universität Gießen, Gießen, Germany.
e-mail: d.a.wismeijer@gmail.com; knut.drewing@psychol.uni-giessen.de

We studied whether vision can teach touch to the same extent as touch seems to teach vision. In a 2 × 2 between-participants learning study, we artificially correlated visual gloss cues with haptic compliance cues. In two "natural" tasks, we tested whether visual gloss estimations have an influence on haptic estimations of softness and vice versa. In two "novel" tasks, in which participants were either asked to haptically judge glossiness or to visually judge softness, we investigated how perceptual estimates transfer from one sense to the other. Our results showed that vision does not teach touch as efficient as touch seems to teach vision.

Keywords: arbitrary association, cue-interaction, gloss, learning, multi-sensory, softness, touch, vision

1. INTRODUCTION

Every day we obtain information about our environment using different sensory modalities. To create a unified percept of our environment, information from different senses needs to be combined. We can learn which information about our environment is most likely to belong together by analyzing statistical correlations, and by interacting with the environment. For example, we can learn that larger objects are heavier than smaller ones—given that they have the same visual appearance—by lifting them. Even more so, if we can get estimates of an environmental property from different sensory modalities, we can calibrate the senses with respect to each other and the world. Two important research questions within the field of multi-sensory perception are how such learning happens, that is how information from one sense is transferred to the other and how information from different senses is combined.

Studies on multi-sensory perception provide evidence for the theory of reliability-based cue integration, meaning that information from different senses are weighted according to their reliability before being combined, see Ernst and Bülthoff (2004) for a review.

Some of these studies focused on space perception, e.g., perceived surface orientation and depth, and showed that touch can teach vision (Ernst et al., 2000; Atkins et al., 2001; Ho et al., 2009; van Beers et al., 2011). Ernst et al. (2000), for instance, have shown that haptic feedback correlated with one of two conflicting visual slant cues entices observers to give more weight to that cue when judging slant. Evidence that vision can teach touch, is much thinner and mainly comes from developmental studies (Gori et al., 2008, 2011). From adult studies on stimulus properties such as size—which is thought to be sensed more accurately by the haptic system than the visual system (Gori et al., 2008)—and weight, there is evidence that haptic estimates are influenced by visual cues (Hillis et al., 2002; Flanagan et al., 2008; Buckingham et al., 2009; Walker et al., 2010), i.e., vision can at least influence touch. However, none of these studies have shown that vision can teach touch, in the sense that learning leads to a change in weights given to single cues, as has been shown previously for touch-to-vision learning. Thus, it is unknown whether such transfer of information can occur to the same extent from vision to touch. In this study, we investigated whether vision can teach touch to the same extent as touch seems to teach vision.

A draw-back of stimulus properties used in previous studies, e.g., surface orientation, depth, and size, is that they are naturally sensed by both the haptic and visual system. In addition, one system might be able to do so more accurately, leading to unevenly distributed weighting of visual and haptic cues, i.e., capture by one or the other sense (Ernst and Bülthoff, 2004). The fact that one sense might dominate the other, would make it difficult to study information transfer in both directions. It would thus be ideal to study multi-sensory perception of environmental properties that are naturally sensed by either of the two sensory systems, but not by both.

In such a case, one cannot disturb the natural relation between e.g., the haptic and visual counterparts of a particular property as is usually done in multi-sensory perception studies. But, one could create an artificial correlation between two unrelated sources of information. That one can learn such arbitrary statistical correlations in a relatively short time, has been studied by Ernst (2007). He showed that mandatory perceptual association between brightness and stiffness of an object occurred after exposure to an artificial statistical correlation between the two.

Using artificial correlations, we studied the influence of a non-natural cue on a perceptual estimate, and we did so in both directions: from-vision-to-touch and from-touch-to-vision. To this end, we chose two material properties that are naturally

uncorrelated to one another and can only be directly sensed by one modality. As a "haptic material property" we used compliance, which can only be directly sensed by the haptic system (because only the haptic system has direct access to force information) and any visual corrugate has to be learned (Drewing et al., 2008). As a "visual material property" we chose gloss, which has no haptic corrugate.

Real haptic objects with varying compressibility gave participants a haptic cue to the compliance of the stimulus. As a visual cue to gloss, we varied the amount of light that was reflected by a virtual cylindrical object and the size of the highlights on it. Participants were given one stimulus defined by both cues and another defined by one or the other cue and were instructed to discriminate between the two stimuli regarding gloss or softness. For instance, a participant was given one stimulus which was defined by both a compliance and a gloss cue (standard stimulus) and one which was only defined by a compliance cue (comparison stimulus). Now, we could ask the participant to judge one of two material properties: gloss or compliance. The "natural task" would consist of judging "which one feels softer," in which case the participant could compare the two haptic compliance cues, even before any statistical correlation between the haptic and visual cues was learned. In contrast, in the "novel task," judging "which one feels less glossy," the participant should, initially, hardly be able to make a reliable judgment, unless he/she relied on a pre-existing association between felt softness and seen gloss (and transferred his/her perceived softness into some estimate of felt glossiness).

We used a 2 × 2 between-participants design [2 senses or judgment modalities, (haptic or visual) × 2 judged dimensions (gloss or softness)], with a two alternative forced choice task (2-AFC task). In a 2-AFC task, the participant is forced to make a discriminative decision between two stimuli. Such a task makes it possible to measure the point of subjective equality (PSE)—which assesses the stimulus parameter value at which the two stimuli are perceived to be identical—and the just noticeable difference (JND)—which assesses the discrimination threshold. The judgment modality, or sense used to judge a particular stimulus property (dimension), not only refers to how participants had to judge ("feel" or "look"), but also reflects which cues were available. For the haptic sense, there was always a haptic cue for two stimuli (and a visual cue for just one stimulus), whereas for the visual sense, there was a visual cue for two stimuli (and a haptic cue for just one stimulus). We thus had two novel-task conditions ("which one looks softer" and "which one feels less glossy") and two natural task conditions ("which one feels softer" and "which one looks less glossy").

Since we could not be a 100% sure that participants did not have some pre-existing association between the two cues, we tested that in an initial (control) session in which we made sure that there was no overall correlation (Pearson's product-moment correlation coefficient was 0) between the two cues. Then in a second session (on another day), participants were first subjected to a short (56 trials) training in which the overall correlation (Pearson's product-moment correlation coefficient) between the two cues was 0.94, followed by training trials interleaved with test trials (the overall correlation was 0.85 in this part).

To make comparisons between experimental conditions easier, we refer in the rest of the manuscript to the cues we used as main and associated cues. The **main cue** always refers to that cue which is naturally sensed by the **judgment modality**, i.e., the haptic compliance cue for haptic judgments ("which stimulus feels . . .") and the visual gloss cue for the visual judgments ("which stimulus looks . . ."). The **associated cue** then refers to the other cue of the standard stimulus, i.e., the visual gloss cue for the haptic judgments and the haptic compliance cue for the visual judgments. In addition, the **judged dimension** relates to the material property participants had to judge: less glossy or softer.

We measured learning of the arbitrary association by changes in the estimation of the PSE and JND as follows: we introduced small discrepancies between the main and associated cue values in our test stimuli.

We predicted for the two natural tasks (haptic softness judgments and visual gloss judgments) that learning the arbitrary association would lead to estimations derived from reliability-based weighting of cues. This means that, before learning, the PSE should only depend on the main cue, because gloss and softness are not (naturally) related to one another. Learning the arbitrary association should lead to a shift in the PSE in the direction of the associated cue value, thus giving some weight to the associated cue in the judgment. In the novel task, the participant should have learned how perceived softness can be transferred into an estimate of gloss before comparing the two stimuli and thus be able to make a reliable judgment of felt gloss. After learning, cues should be combined similarly as in the natural task, leading to a similar shift in the PSE in the direction of the associated cue.

Learning the arbitrary association should lead to a decrease in the JND in all tasks. We predicted that the largest changes would occur in the novel tasks, since we did not expect participants to perform the task very well in the initial session. In addition, we predicted that after learning, the JND should be same for all tasks, given that we tried to have similar JND values for visual and haptic cues. Otherwise, the biggest difference in JND should arise between judgment modalities, but not judged dimensions.

With the experimental paradigm as sketched, we investigated how participants in the novel task learned to transfer perceptual estimates from one sense to another. And, by interchanging both the single cue stimuli (gloss vs. compliance cue only) and the dimension to be judged, we were able to compare learning between touch-to-vision and vision-to-touch transfers. In addition, the natural tasks allowed us to study whether learning arbitrary correlations influenced the integration of the associated cue into a combined percept similarly for vision-to-touch and touch-to-vision.

2. MATERIALS AND METHODS

2.1. PARTICIPANTS

All 29 participants (three male, mean age 24 years with a standard deviation of 4 years, one left-handed participant, and by accident, one participant more in the haptic soft condition) had normal or corrected-to-normal vision, a stereoacuity of at least 60 arcsec (Randot Stereo Fly, graded circle test) and no sensory or motor

deficits. They were either paid for their participation (8 Eur/h) or given credits as part of their psychology curriculum.

2.2. APPARATUS

Figure 1 shows a sketch of the experimental setup. Participants sat in front of the setup resting their heads on both a chin- and headrest to minimize head movements. To track the position of the hand, the index finger of participant's preferred hand (participants could be either left or right handed) was attached to a force-feedback device (Phantom Premium 1.5, 1000 Hz, spatial resolution: 0.03 mm), by gluing a reusable plastic finger nail with a small magnet onto the finger (see **Figure 2B**). The participants viewed the visual scene through a mirror while wearing shutter glasses (NVIDIA, 3D vision kit). The visual scene was displayed on a Samsung Syncmaster 2233RZ (120 Hz) and generated on a DELL Precision 380. Because a mirror rotates polarized light (emitted by any LCD) by 90°, we had to rotate the screen by the same amount to realign the polarizing filters of the screen and those of the shutter glasses. Due to the fact that brightness was dependent on the viewing angle along the vertical axis of the LCD screen, we corrected the whole visual scene using shaders (OpenGL/GLSL: Woo et al., 1997; Rost, 2006).

2.3. STIMULI

A stimulus consisted of either a haptic compliance cue or a visual gloss cue or a combination of both (standard stimulus). The haptic compliance cue was obtained by pressing onto custom made silicone cylindrical objects, for an example see **Figure 2A**.

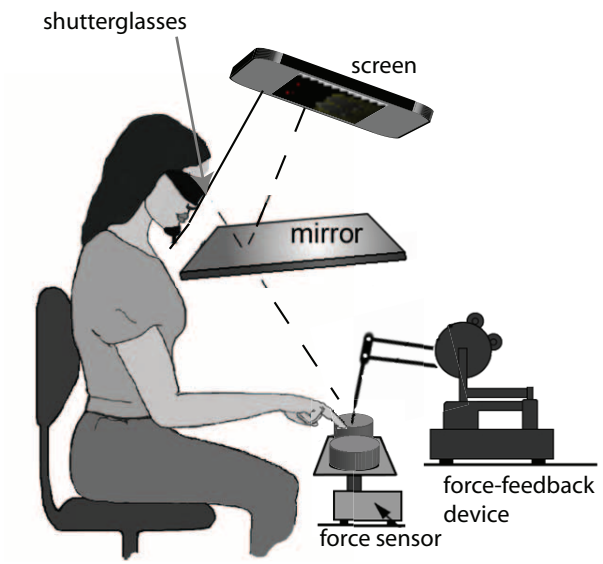


FIGURE 1 | The experimental setup. Participants were seated in front of the virtual reality setup, with their heads resting on a chin- and headrest for stabilization purposes. Participants wore shutter glasses and viewed the visual 3D scene via a silver-coated mirror. The real haptic objects were placed in front of the participant on the force sensor. We used a force-feedback device to track the location of the participant's index finger. Participants responses were registered by tracking (with the force-feedback device) which virtual decision button was touched.

They had compliance values of: 0.124, 0.136, 0.148, 0.159, 0.177, 0.189, and 0.198 mm/N, identical to the ones used by Kaim and Drewing (2011). We measured the compliance of each object using a cylindrical probe of 1 cm^2. The difference between consecutive compliance values was equal to approximately half a JND value (and see Kaim and Drewing, 2011). Instead of referring to the compliance values, we used a JND scale with one unit approximately equal to 0.5 JND (−3 . . . 3, with 0 equal to a compliance value of 0.159).

The visual gloss cue was conveyed by a virtual 3D cylindrical NURB (non-uniform rational basis spline) surface that was generated with a custom made C(++) program using OpenGL, see **Figure 2C**. NURB refers to a mathematical technique using polynomials to describe smooth surfaces, which is available in OpenGL (Woo et al., 1997). Using this mathematical technique, we were able to generate virtual cylindrical objects that had a similar non-flat top surface as our real objects, which were slightly convex. The virtual cylinders had the same dimensions (height and diameter) as the real objects. Differences in gloss appearance were established by co-varying two OpenGL defined object parameters: the shininess component, which regulates the size and brightness of specular highlights on a surface, and the specular component, which defines how a material reflects specular light. In order to generate more than one highlight on the virtual cylinders, we used several light sources to illuminate the scene, which were spread symmetrically around the vertical meridian. In a separate pilot experiment (four subjects, magnitude estimation of identical surfaces but with sparser lighting), we determined that perceived glossiness depended linearly on the specular component. In addition, there was an interaction effect between the shininess component and the specular component on perceived gloss, which became non-linear for extreme values of these two

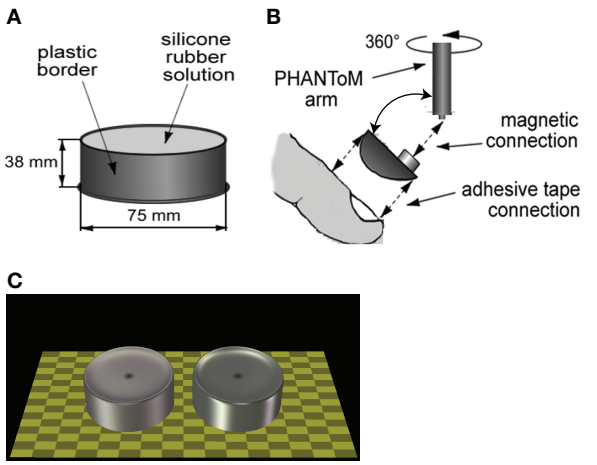


FIGURE 2 | Stimuli and setup detail. (A) Sketch of a haptic object. **(B)** Attachment of finger to force-feedback device. **(C)** Sketch of a visual scene showing the two extreme cases of our gloss axis. Note that in reality the visual correlate would only be visible when the participant pressed on the object (or in its presumable location for visual-only comparisons). Also note that the shader has been turned off to generate these images and that the actual scene was dimmer than shown here.

components. We stayed within the perceptually linear range and varied the shininess component between 30 and 90 (total range [0, 128]), and the specular component was varied between 0.25 and 0.55 (total range [0, 1]). We then chose a set of parameter values for which the difference between consecutive gloss values was approximately half a JND. We verified this assumption in another pilot experiment (almost identical to the study reported here but with longer visual stimulus exposure times, 10 subjects, 2-AFC task), in which the mean JND was 2.4 with a standard deviation of 2.66. In the current study we found a mean JND of 3.3 with a standard deviation of 2.0. Instead of referring to the gloss parameters, we used a JND scale, with one unit being approximately 0.33 JND (in the current study, range between $-3 \ldots 3$ and with 0 equal to a specular component of 0.4 and a shininess component of 60).

The virtual cylinders were positioned such that they completely coincided with the haptic objects in real space. These visual correlates were only visible when observers pressed onto the haptic object, otherwise a ring was shown to identify the location of the object. In cases without the haptic object, participants had to place their finger in the empty space where the comparison stimulus was, to make the visual cue visible.

2.4. TASK AND PROCEDURE

2.4.1. Task

Before starting the experiment, participants were told that they had to judge which stimulus felt or looked as being softer, or less glossy, respectively. In addition, we told them which type of stimuli would be present in the experiment. For instance, a participant in the haptic soft condition was told that he/she would be able to (and had to) press onto two haptic stimuli, but that only one stimulus would also be shown on the screen. After that, he/she was told to judge which one of the stimuli *felt softer* and then touch the virtual decision button ("fuehlt sich weicher an") above the softer stimulus. We additionally instructed participants to press in the middle of the haptic objects and not to slide their finger across them [a natural movement made by participants to judge surface roughness (Lederman and Klatzky, 1993)].

They were then given four initial trials, in which we used the stimuli that had the most extreme gloss and/or compliance values. After that the experiment began.

Before each trial, participants were asked to position the finger attached to the force-feedback device in the left-bottom corner of the virtual environment, which was visualized by the word "WARTEN" (wait). This gave the experimenter the opportunity to place the haptic object(s) for the next trial in the designated area(s), without touching the observer. After placement of the stimuli, the next trial started. Participants would then move their index finger (either to the left or the right stimulus) and press on the haptic object—or in the space where it should have been—and thus trigger the visual cue to be shown. They then could either touch the same object again, or move to the other object. After touching each stimulus location, they indicated which stimulus was "less glossy" or "softer" by pressing the corresponding virtual decision button. After that, the wait sign would reappear. Participants were only allowed to press each haptic object twice. Touching a third time set off a loud beep without the visual object being shown.

Participants were allowed, even encouraged, to have breaks whenever they wanted and the experimenter asked at least every hour, whether or not the participant wanted a break.

2.4.2. Procedure

We used a balanced 2×2 between-participants design and a 2-AFC task with seven participants per experimental condition. The complete experiment consisted of four experimental conditions, which are visualized in **Figure 3**. In each condition, an observer was asked to make a 2-AFC judgment based on a particular stimulus property (judged dimension), either glossiness or softness, using either vision or haptics as judgment modality.

We used a balanced design (each pair of stimuli was shown twice with stimuli changing sides) and trials were randomized across participants.

In all conditions, we presented a standard stimulus, defined by both a haptic and a visual cue. The type of comparison stimulus we provided was based on the judgment modality: it was either defined by a haptic compliance cue for haptic judgments or a visual gloss cue for visual judgments. Thus, when making a haptic judgment of softness ("which one felt as being softer"), observers were given a standard stimulus and a haptic-only comparison stimulus, whereas when making a visual judgment of softness ("which one looks as being softer"), they were given a standard stimulus and a visual-only comparison stimulus.

Because compliance and gloss are unrelated cues, we defined the compliance and gloss values in JND-related units ($-3 \ldots 3$). The haptic cue with the lowest compliance (hardest) and the most glossy visual cue were assigned a value of -3 and the haptic cue with highest compliance (softest) and respectively the least glossy visual cue, were assigned a value of 3. This assignment was in agreement with our imposed artificial correlation (Pearson's) between increasing gloss and decreasing compliance (increasing hardness).

For each experimental condition, we ran two separate sessions, on different days. In the "initial" session, we tested for any pre-existing associations between haptic compliance and visual gloss cues and in the second session learning and testing

judged dimension \ judgment modality	visual	haptic
gloss	visual judgment of gloss (VG)	haptic judgment of gloss (HG)
softness	visual judgment of softness (VS)	haptic judgment of softness (HS)

FIGURE 3 | The four experimental conditions. In our 2×2 between-participants design, participants judged on one of two dimensions, gloss or softness, using one of two judgment modalities, visual or haptic. This resulted in two natural judgment task: visual judgment of gloss and haptic judgment of softness, and two novel judgment tasks: a haptic judgment of gloss and a visual judgment of softness (both in light blue).

were combined. The set of standard stimuli that we used to test for pre-existing associations was defined by the following main cue-associated cue combinations ([main cue, associated cue]): $[-1, -1]$, $[-1, 0]$, $[-1, 1]$, $[0, -1]$, $[0, 1]$, $[1, -1]$, $[1, 0]$, $[1, 1]$. To ensure that there was no overall Pearson correlation between the two cues, we used two additional main cue-associated cue combinations: $[0, -2]$, $[0, 2]$. All of these combinations were compared to all seven comparison stimuli and were repeated six times (three repetitions and presentation on each side). In addition, we added some "noise" trials in which the following standard stimuli were only compared to the three nearest (neighbors of main cue) comparisons: $[3, -1]$, $[3, 1]$, $[-3, -1]$, $[-3, 1]$. The total session consisted of 492 ($8 \times 7 \times 3 \times 2 + 2 \times 7 \times 3 \times 2 + 4 \times 3 \times 3 \times 2$) trials. Participants did not receive any feedback on their performance during and after this session. On average, participants needed around 2 h to complete this session.

The second session consisted of a training or learning part, and a second part in which learning and testing trials were interleaved. In the training part, we used the following main cue-associated cue combination as standard stimuli ([main cue, associated cue]): $[-1, -1]$, $[-1, -2]$, $[1, 1]$, $[1, 2]$. These combinations were compared to all seven comparison stimuli and repeated twice (balanced design, total of $4 \times 7 \times 2 \times 2$) 56, trials). The overall Pearson's product-moment correlation coefficient in this part was 0.94. In the second part, the following standard stimuli were compared to the seven comparison stimuli a total of six times (including balancing): the test set: $[-1, -1]$, $[-1, 0]$, $[-1, 1]$, $[0, -1]$, $[0, 1]$, $[1, -1]$, $[1, 0]$, $[1, 1]$ and a "correlated cue set": $[-3, -3]$, $[-2, -2]$, $[0, 0]$, $[2, 2]$, $[3, 3]$. An additional set was repeated four times: $[-3, -2]$, $[-2, -3]$, $[-2, -1]$, $[-1, -2]$, $[2, 1]$, $[1, 2]$, $[3, 2]$, $[2, 3]$. This part consisted of 770 trials ($13 \times 7 \times 3 \times 2 + 8 \times 7 \times 2 \times 2$). The overall Pearson's product-moment correlation coefficient in this part was 0.85. Participants did not receive any feedback on their performance during either part. Note that learning and testing took part on the same day within this session. Participants needed 3 h to complete this session.

2.5. ANALYSIS

From the collected (judgment) data, we calculated the proportion of trials in which the comparison was perceived as softer/less glossy than the standard per softness/glossiness value of the comparison. We then fitted cumulative Gaussian distributions to these proportional values per phase of the experiment (initial or learning) under the assumption that the JND was equal for each standard stimulus (per experimental phase). To this end, we simultaneously fitted eight cumulative Gaussian distributions to the data collected with the eight standard stimuli using eight biases (PSE) and a single standard deviation (JND) as free parameters in a least-square error fit (for an example of these fits see **Figure 4**). The PSE is defined as the softness/glossiness of the comparison stimulus at which discrimination performance is random (here a performance of 0.5). The 84%-discrimination threshold (JND) is defined as the difference between the PSE and the softness/gloss of the comparison when it is judged softer/matter than the standard 84% of the time. After fitting, we selected data based on the fitted JND parameter. If the JND of both the initial and the learning phase deviated more than two standard deviations from the average JND (calculated per phase), a participant was removed, because it meant that the participant displayed random behavior in both phases. With this criterion a total of three participants were removed from further analysis, resulting in one

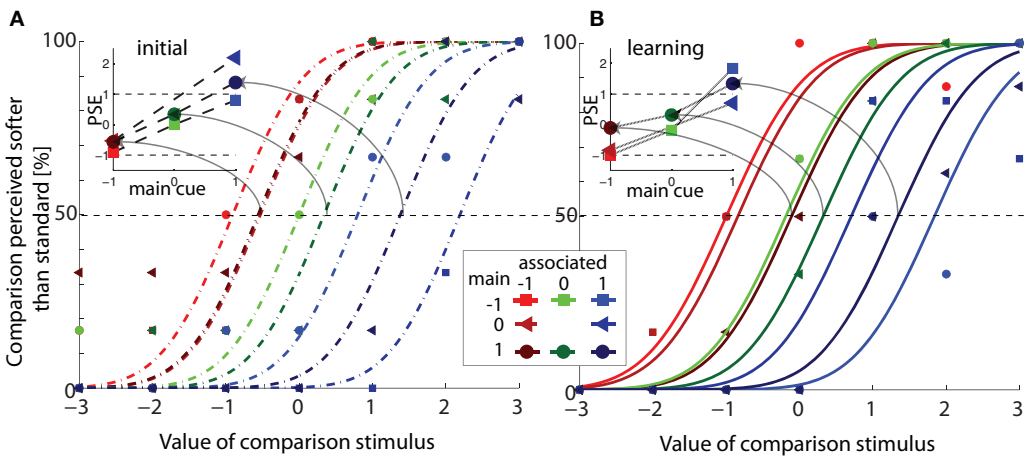


FIGURE 4 | Example of psychometric curve fits. (A) The proportion of trials in which the comparison was perceived as softer than the standard is plotted against the softness of the comparison, for one participant in the haptic soft condition in the initial session [data and fits of the learning session are shown in panel **(B)**]. Series of data points represent the percentage of judgments that the comparison was judged as being softer than the standard separated by the different standard stimuli (different symbols and colors). Lines represent a set of cumulative Gaussian functions that was fitted to the data points. Functions share the standard deviation (as defined in the fit), but differ in mean (or PSE). In the inset, PSE values are plotted against the main cue value. PSEs of data with the same associated cue value are connected by (dotted) lines. Note that contrary to our predictions, PSEs of same main cue value did not overlap, but depended on the associated cue. For this participant, the PSE values were greater than 1 (a boundary defined by a reliability-weighted cue combination model). **(B)** Fitted psychometric curves (and actual data) of the same participant in the learning session. In the inset, PSE values are plotted against the main cue value, with PSEs of data the same associated cue value connected by lines.

participant less in the haptic gloss, visual soft and visual gloss conditions.

To test for differences in learning between the four experimental groups, we used a MANOVA (multi-variate mixed-design general linear model, SPSS) on the estimated PSE and JND values. Even if the data do not perfectly comply to the assumptions of normally distributed data and homoscedasticity, these are robust test, meaning that deviations from these assumptions generally do lead to acceptable test results. Depending on the parameter to be tested, we used different within-participant and between-participant variables of which the details can be found in the corresponding results section. We used a significance criterion of $p < 0.05$. Where appropriate, we tested on individual groups using one- or two-sided Student's t-tests in order to clarify statistically significant differences between groups and/or sessions.

3. RESULTS

An example of the data we fitted under the assumption that the JND was equal for each standard stimulus (per experimental phase), can be seen in **Figure 4**, where we plotted the proportion of trials in which the comparison was perceived as softer than the standard against the softness of the comparison, for one participant in the haptic soft condition. Data and fits of the initial phase are given in panel A, and those of the learning phase in panel B. Series of data points represent the percentage of judgments that the comparison is softer than the standard separated by the different standard stimuli (different symbols and colors). Lines represent a set of cumulative Gaussian functions that was fitted to the data points. Functions share the standard deviation (as defined in the fit), but differ in mean (or PSE).

In each inset, we plotted the PSE against the main cue for each associated cue. We predicted that in the initial phase, the PSE should correspond to the main cue value and not depend on the associated cue value, which would yield overlapping curves in the inset of panel A. Whereas after learning, the PSE should depend on both cues, if cue integration did occur, and the curves should no longer overlap (inset panel B). As predicted, the PSE values of this participant did depend on the value of the main cue, the PSE increased with increasing main cue value, see both insets. Contrary to our predictions, the PSE (of this participant) also depended on the associated cue value in both experimental phases (the curves for different associated cues did not overlap). Note that reliability-weighted cue combination should yield slopes ≤ 1, because in this theory the sum of weights should always equal 1 and the slope (in the current figure) defines the weight given to the main cue. Thus, for this participant an other type of interaction between main and associated cues is observed.

In both novel-task conditions, haptic gloss and visual soft, there were four participants (one in the visual soft condition and three in the haptic gloss condition) of whom the data fitted better to a cumulative Gaussian distribution that was tilted in the direction opposite to the one defined by the correlation between the two cues in the learning phase. The data of these mirror-model participants suggested that they, at least in the initial phase, associated an increase in compliance (softness) with an increase in gloss (whereas in the training trials we correlated an increase in compliance with a decrease in gloss).

3.1. ESTIMATED POINTS OF SUBJECTIVE EQUALITY

Following the same style as that of the insets in **Figure 4**, we plotted the average PSE across participants with standard errors in **Figure 5**.

Individual PSE values entered a MANOVA with main cue (−1 vs. 0 vs. 1), associated cue (−1 vs. 1) and learning as within-participant variables, and as between-participant variables judged dimension and judgment modality. This design enabled us to separate effects of either cue. In order to get a balanced design, we did not use conditions with an associated cue value of 0. A control analysis including these stimuli yielded the same conclusions. In case of mirror-model participants, the main cues were mirrored before they entered the analysis in order to keep the direction of gloss and softness judgments the same across participants (i.e., higher/positive values meaning less glossy or softer, respectively). The associated cue was not inverted.

The overall (ANOVA) analysis showed that participants' judgments systematically depended on, and increased with, the value of the main cue, $F_{(2,\ 44)} = 143$, $p \leq 0.001$. As predicted, the relation between main cue and judgment was modified by learning, $F_{(2,\ 44)} = 3.5$, $p = 0.04$, whereby the effect of learning was modified by the judged dimension and the judgment modality (interaction: main × learn × judged dimension: $F_{(2,\ 44)} = 6.0$, $p = 0.004$; interaction main × learn × judgment modality: $F_{(2,\ 44)} = 4.7$, $p = 0.014$; interaction main × learn × judged dimension × judgment modality: $F_{(2,\ 44)} = 3.5$, $p = 0.038$). There were no other reliable effects. Even though direct effects of the associated cue—as they were predicted from the weighted averaging scheme—failed to reach significance, the manifold of interaction effects with the main cue reject the possibility that participants based their judgments solely on the main cue. In that case, the participants judgments should have depended on the main cues

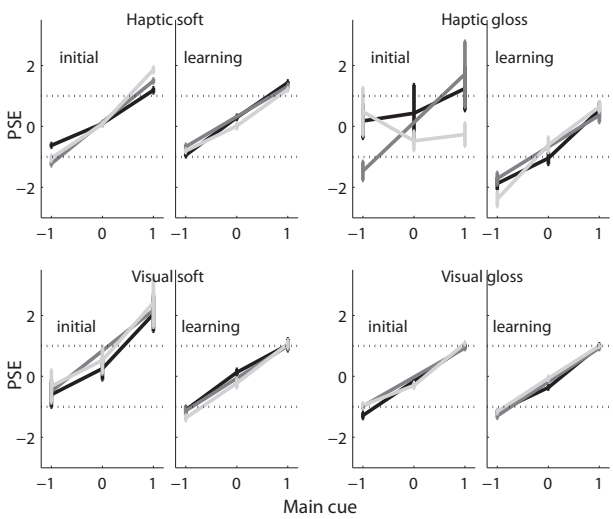


FIGURE 5 | PSEs vs. main cue. The average PSE (and standard error) across participants are plotted against the main cue value, per experimental condition and experimental session. Values belonging to the same associated cue are connected by lines, with −1 in black, 0 in gray, and 1 in light gray.

value irrespective of learning, dimension or modality (i.e., a main cue of −1, 0, or 1 should have resulted in a judgment of −1, 0, or 1 in each and every condition). Put in other words, the results suggest influences of the double-cue situation on the judgments that depended on the interaction between learning, dimension and modality condition.

In order to clarify these differences between the four tasks, we conducted four additional analyses (MANOVA) separated by judged dimension and judgment modality (within-participant variables learning, main cue and associated cue). To make the influence of the double-cue situation better visible, we conducted these analyses on PSE data from which the to-be-expected effect of a main-cue alone strategy was eliminated (simply by subtracting the main cue value from each single PSE).

In the visual gloss condition, this analysis did not yield any significant effect. This result is consistent with the view that participants judged visual gloss on the basis of the visual main cue alone, and did so similarly before and after learning the correlation with the haptic cue. The same lack of effect was observed for the visual soft condition, even if numerical effects (see **Figure 5**) seem to suggest otherwise. Thus, the PSE analyses suggest that participants based their visual judgments of both gloss and softness solely on the (main) gloss cue.

In contrast, in the haptic soft condition, participants deviated from the main-cue alone strategy: a remaining effect of the main cue, $F_{(2,\ 14)} = 6.8$, $p = 0.009$, indicated an "over-weighting" of the main cue, i.e., judgments were spread wider apart than the main cue values. The interaction main cue × associated cue × learning, $F_{(2,\ 14)} = 6.9$, $p = 0.008$, indicated a more complicated modification of this effect. Additional separate analyses (MANOVA) for each learning condition (variables: main and associated cue) tracked down these deviations to the initial phase (i.e., there was no reliable effect after learning the correlation). Here, we observed a significant interaction between the main cue and the associated cue, $F_{(2,\ 14)} = 6.6$, $p = 0.014$ (corrected according to Huynh and Feldt, 1976), indicating that over-weighting occurred in particular with one of the two associated cues (the glossier visual cue, value of 1) and before the correlation between the two cues had been learned. Taken together, these results revealed unexpected non-linear influences of the associated cue (i.e., interaction, over-weighting), before learning the correlation, which vanished with learning. Finally, in the haptic gloss condition, we observed an interaction between the main cue and learning, $F_{(2,\ 10)} = 5.6$, $p = 0.042$, which—by separate analyses for the two learning conditions—could be due to an under-weighting of the main cue in the initial phase (trend: $F_{(2,\ 10)} = 3.64$, $p = 0.081$) combined with its over-weighting after learning (trend: $F_{(2,\ 10)} = 3.6$, $p = 0.067$).

Taken together, indirect influences of the associated cue on the interpretation of the main cue were observed in the two haptic conditions, but not in the two visual conditions. The observed interactions between associated and main cues were unlike predicted, of a non-linear nature including over- and under-weighting of the main cue, and statistical interactions.

Originally, we expected only linear influences of the main and the associated cue, which should have changed by learning the correlation. However, the PSE analyses revealed that learning, with regard to the overall strategy and in contrast to judgment precision (see JND results in section 3.2), resulted in the elimination of non-linear influences. In the section 3.3, we conducted an additional sensitive analysis which supports these findings.

3.2. ESTIMATED JUST NOTICEABLE DIFFERENCES

In **Figure 6**, the mean JND value across participants with standard error are shown for each experimental phase and condition. JND values were entered in a ANOVA (mixed-design, SPSS), with learning as a within-participants variable, and judgment modality and judged dimension as between-participants variables. There was a significant main effect of learning on JNDs, $F_{(1,\ 22)} = 6$, $p = 0.025$, which can be seen in the general trend of decreasing JND with learning in all experimental conditions. This effect was modified by judged dimension and judgment modality (trend: learning × judged dimension × judgment modality $F_{(1,\ 22)} = 4$, $p = 0.07$). In addition, we found a between-participants effect of judged dimension, $F_{(1,\ 22)} = 5$, $p = 0.03$.

We predicted (see introduction) that, if there was any influence of learning, the JND values should show the largest decrease in the novel tasks. Indeed, there were trends for learning in individual conditions of the novel tasks (haptic gloss $p = 0.094$, visual soft $p = 0.12$ and all novel tasks combined $p = 0.054$, 1-tailed Student's t-tests), but not for natural tasks (individual and combined conditions $p \geq 0.2$, 1-tailed Student's t-tests). In addition, we predicted that if there would be any interaction effects with learning, these should be due to judgment modality, because the JND values could have differed between the two senses (or the two types of cues). However, there was a significant effect of judged dimension on learning (after learning: haptic gloss vs. haptic soft: $p = 0.029$, visual gloss vs. visual soft: $p = 0.014$, 1-tailed Student's t-test), and none of judgment modality (after learning: haptic gloss vs. visual gloss: $p = 0.49$, visual soft vs. haptic soft: $p = 0.37$, 1-tailed Student's t-tests). JND for gloss judgments were higher than for softness judgments both before (trend

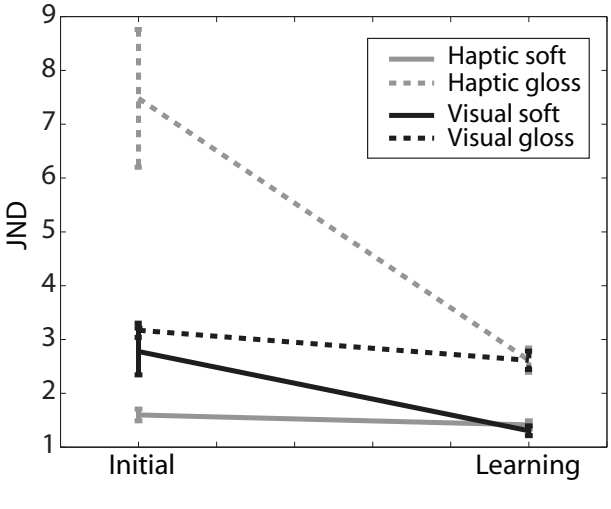


FIGURE 6 | JND. Average JND (and standard error) across participants of initial and learning sessions for each experimental condition.

$p = 0.073$, 2-tailed Student's t-test) and after learning ($p = 0.003$, 2-tailed Student's t-test), whereas different judgment modalities were indistinguishable both before ($p = 0.54$, 2-tailed Student's t-test) and after learning ($p = 0.95$, 2-tailed Student's t-test).

Overall, these effects show that, in agreement with our prediction, JND changed with learning in the novel tasks, but not in the natural tasks. And, against our prediction, the JND after learning was independent of the judgment modality, or which cues were given to the participant, but depended on the judged dimension. Thus while given the same cues, the JND values of participants judging softness were lower than of those judging glossiness.

3.3. RESIDUAL ANALYSIS

As we have seen in the fitted PSE values previously, both the main and associated cue were used to make soft and gloss judgments, but they were not combined in a way congruent with a standard cue integration hypothesis (see **Figure 5**). Therefore, investigating changes in weight given to each cue, would not completely describe effects of learning. We therefore sought of other ways to capture learning effects. We hypothesized that by learning the arbitrary correlation, the unpredicted interaction effects should decrease. In addition, learning the association should lead to an increased linear dependence on the main and associated cues, that is cue combination instead of interaction. We decided to fit a model to the PSE values that depended linearly on the main and associated cue values. The residuals of such a fit then contain any non-linear and interaction effects of the main and associated cues on the PSE values. We used the following linear cue combination model to fit the PSE values and slightly adapted it by letting go of the constraint that the sum of weights assigned to the cues should equal 1 (as in a standard cue integration model):

$$\text{PSE} = w_m \times \text{main cue} + w_a \times \text{associated cue} \qquad (1)$$

where w_m defines the weight assigned to the main cue, w_a defines the weight assigned to the associated cue. Both these weights were constrained between [0, 1] and we used an additional constraint that the sum of w_m and w_a should always be less than or equal to 1. These two fit parameters (w_m, w_a) thus give an indication of how much the PSE values linearly depended on the main and associated cue values.

The residuals of these fits now contained all the unpredicted interaction effects, such as scaling and biasing of the main cue and other interaction effects between the two cues, and they included the unpredicted variance. To separate the interaction effects from the unexplained variance, we fitted the following model to the residuals:

$$f_{\text{res}} = a_1 \times \text{main cue} + a_2 \times \text{associated cue} + a_3 \times \text{main cue} \times \text{associated cue} + a_4, \qquad (2)$$

without any constraints on the parameters. The residuals of this fit then describe the total (or final) unexplained variance.

We now had two different measures that could show an effect of learning the association: (1) the total variance explained by the learnt-linear model (Equation 1) and (2) the part of the residuals that could be fitted by our Interaction-model [because they were due to interaction effects, Equation (2)]. Learning the arbitrary association should first of all decrease the size of the residuals due to interaction effects, because the association defined a linear non-interactive correlation between the two cues. For the same reason, it should lead to an increase in the total variance explained by the learnt-linear model.

The results of these fits are shown in **Figure 7**, with in **Figure 7A**, the mean variance explained by the learnt-linear model (with standard error) per experimental phase and condition and in **Figure 7B**, the average amount (with standard error) of the residuals that was due to interaction effects (i.e., was fitted by our Interaction-model).

We then subjected the variance explained by the learnt-linear model and the part of the residuals that were fitted by our Interaction-model to a MANOVA (multivariate mixed-design general linear model, SPSS), with learning as a within-participants variable and judgment modality and judged dimension as between-participants variables. Because these two quantities are not independent from each other, they were tested simultaneously. There was a significant main effect of learning, $F_{(2,\ 21)} = 6.1$, $p = 0.008$. This effect was modified by judged dimension (interaction learning × judged dimension, $F_{(2,\ 21)} = 7.1$, $p = 0.004$), there was only a trend for the interaction with judgment modality (interaction learning × judgment modality,

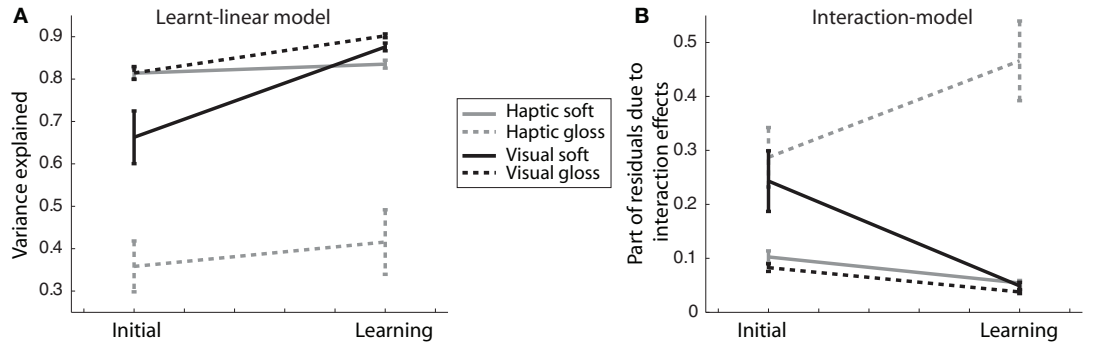


FIGURE 7 | Variance explained. (A) Variance explained by learnt-linear model fitted to the PSEs (Equation 1). The means and standard deviations across participants are shown. **(B)** Average amount (with standard error) across participants of the residuals due to interaction effects.

$F_{(2, 21)} = 3.0$, $p = 0.072$) and a significant modification by judgment modality and judged dimension simultaneously, $F_{(2, 21)} = 4.3$, $p = 0.028$. In addition, there were significant differences between groups: a between-participants interaction effect (judgment modality × judged dimension, $F_{(2, 21)} = 4.9$, $p = 0.018$) and an effect of judged dimension, $F_{(2, 21)} = 3.5$, $p = 0.05$. We further investigated the effects of learning using 1-tailed Student's *t*-tests and the interaction effects with judgment modalities and judged dimensions using 2-tailed Student's *t*-tests.

There were at least trends for an effect of learning in both visual tasks. For the visual gloss condition, both the variance explained by the learnt-linear model ($p = 0.023$, 1-tailed) and that explained by the Interaction-model ($p = 0.020$, 1-tailed) changed with learning in the predicted directions. In the visual soft condition, there was only a trend for learning for both the variance explained by the learnt-linear model ($p = 0.098$) and that explained by the Interaction-model ($p = 0.095$).

For both haptic tasks, however, we found only a trend for learning in the Interaction-model for haptic soft ($p = 0.091$) and no significant effects of learning within the learnt-linear model ($p = 0.32$) and no effects of learning in either model for the haptic gloss condition (learnt-linear: $p = 0.41$, Interaction: $p = 0.22$).

In addition, learning had a differential effect on the two novel tasks. Whereas they did not differ before learning (learnt-linear: $p = 0.18$, Interaction: $p = 0.8$, 2-tailed Student's *t*-tests), they did so afterward (learnt-linear: $p = 0.034$, Interaction: $p = 0.044$, 2-tailed Student's *t*-tests). Both before and after learning, the variance explained by the learnt-linear model was lower for the haptic gloss condition compared to those of the two natural tasks (initial: vs. haptic soft: $p = 0.005$, vs. visual gloss: $p = 0.013$, learning: vs. haptic soft: $p = 0.024$, vs. visual gloss: $p = 0.026$, 2-tailed Student's *t*-tests) and higher for the Interaction-model only after learning (vs. haptic soft: $p = 0.021$, vs. visual gloss: $p = 0.040$, 2-tailed Student's *t*-tests). The visual soft condition, however, was not different from either natural task both before and after learning.

Taken together, these results support the previous findings of the PSE analyses, namely that learning—in agreement with the correlation—occurred in visual judgment tasks (there were trends in both visual tasks), but not in either haptic judgment task. Moreover, they showed participants performed better in the visual novel task (visual soft judgment) than in the haptic novel task—reaching the same high-level performance as in the natural visual task.

4. DISCUSSION

In this study, we investigated whether learning from touch-to-vision is similar to learning from vision-to-touch. To this end, we introduced an artificial correlation between visual gloss and haptic compliance cues and investigated how learning this association influenced two natural judgments (visually judging gloss, haptically judging softness) and two novel judgments (visually judging softness and haptically judging gloss).

Our analyses of PSEs revealed unexpected non-linear interactions, whereas our sensitive analyses of explained variance revealed that learning (including de-learning of non-linear interactions) occurred, in particular, in visual tasks. In addition, the analyses of explained variance showed that "performance" after learning was better in the visual novel task (similar to "performance" in the natural visual task), than in the haptic novel task. Taken together, these results suggest that vision does not educate touch as efficiently as touch seems to educate vision.

To our knowledge, our study is the first to show that learning between the senses depends on its direction. In many previous studies, it has been shown that touch teaches vision, however, the reverse whether vision can teach touch had not been investigated thoroughly in adults, so far.

Our study is in agreement with the results from Ernst (2007) that humans have ability to learn from cue-associations and that previously unrelated cues can be recruited for a judgment task, if they are positively correlated [see also Haijiang et al. (2006) for cue recruitment in binocular rivalry tasks]. Our result that touch teaches vision is also in agreement with previous studies investigating e.g., surface orientation and depth (Ernst et al., 2000; Atkins et al., 2001; Ho et al., 2009; van Beers et al., 2011). However, we did not find any sign of reliability-weighted cue combination, as reported previously, but found cue-interaction instead. Although the difference may have occurred due to lack of learning, van Beers et al. (2011) have shown that learning to integrate haptic cues for surface slant estimation can occur quickly (with 55 trials) and within the same day. However, these time scales may apply to studies where the properties to be estimated can be sensed by both senses and not to our study in which the material properties to be estimated could only be estimated by one or the other sense.

We found that the JND depended on the judged dimension (gloss vs. soft), irrespective of the judgment modality—and thus irrespective of which cues were available for making the judgment; an effect that became clearer after learning. This may mean that participants were using different cues, or using the cues differently, in the novel tasks compared to the natural tasks. Thus, these results showed that after learning, at least in the novel tasks, estimation of the stimulus property was similar for the two senses; i.e., was the same for touch-to-vision and vision-to-touch.

In this work, we studied a kind of unconstrained basic learning. Although the cues were correlated in the learning session, participants did not receive feedback on their performance and this kind of learning could mimic learning in infants. We propose that, with our study, we tapped into the neural learning process before cue integration might occur. As a precursor to maximum-likelihood-estimated integration [see Ernst et al. (2000) for a short review], the cues both influenced the judgment, but are not yet linearly integrated. We hypothesize that given enough time, neural mechanisms related to cue integration would come to play even for seemingly arbitrary cues. However, in the case of inter-modal cue-integration, such integration is often not compulsory (Hillis et al., 2002; Gori et al., 2011) and might therefore be overshadowed by single cue effects.

Taken together, our data revealed differences in learning from-touch-to-vision and from-vision-to-touch. Learning from-touch-to-vision did occur, but not the other way around.

REFERENCES

Atkins, J. E., Fiser, J., and Jacobs, R. A. (2001). Experience-dependent visual cue integration based on consistencies between visual and haptic percepts. *Vision Res.* 41, 449–461.

Buckingham, G., Cant, J. S., and Goodale, M. A. (2009). Living in a material world: how visual cues to material properties affect the way that we lift objects and perceive their weight. *J. Neurophysiol.* 102, 3111–3118.

Drewing, K., Wiecki, T. V., and Ernst, M. O. (2008). Material properties determine how force and position signals combine in haptic shape perception. *Acta Psychol.* 128, 264–273.

Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *J. Vis.* 7, 7.1–7.14.

Ernst, M. O., Banks, M. S., and Bülthoff, H. H. (2000). Touch can change visual slant perception. *Nat. Neurosci.* 3, 69–73.

Ernst, M. O., and Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends Cogn. Sci. (Regul. Ed.)* 8, 162–169.

Flanagan, J. R., Bittner, J. P., and Johansson, R. S. (2008). Experience can change distinct size-weight priors engaged in lifting objects and judging their weights. *Curr. Biol.* 18, 1742–1747.

Gori, M., Delviva, M., Sandini, G., and Burr, D. (2008). Young children do not integrate visual and haptic form information. *Curr. Biol.* 18, 694–698.

Gori, M., Mazzilli, G., Sandini, G., and Burr, D. (2011). Cross-sensory facilitation reveals neural interactions between visual and tactile motion in humans. *Front. Psychology* 2:55. doi: 10.3389/fpsyg.2011.00055

Haijiang, Q., Saunders, J. A., Stone, R. W., and Backus, B. T. (2006). Demonstration of cue recruitment: change in visual appearance by means of pavlovian conditioning. *Proc. Natl. Acad. Sci. U.S.A.* 103, 483–488.

Hillis, J. M., Ernst, M. O., Banks, M. S., and Landy, M. S. (2002). Combining sensory information: mandatory fusion within, but not between, senses. *Science* 298, 1627–1630.

Ho, Y.-X., Serwe, S., Trommershäuser, J., Maloney, L. T., and Landy, M. S. (2009). The role of visuohaptic experience in visually perceived depth. *J. Neurophysiol.* 101, 2789–2801.

Huynh, H., and Feldt, L. S. (1976). Estimation of the box correction for degrees of freedom from sample data in randomized block and split-plot designs. *J. Ed. Behav. Stat.* 1, 69–82.

Kaim, L., and Drewing, K. (2011). Exploratory strategies in haptic softness discrimination are tuned to achieve high levels of task performance. *IEEE Trans. Haptics* 4, 242–252.

Lederman, S. J., and Klatzky, R. L. (1993). Extracting object properties through haptic exploration. *Acta psychol.* 84, 29–40.

Rost, R. J. (2006). *OpenGL Shading Lanuage*. 2nd Edn. Boston, MA: Pearson Education Inc.

van Beers, R. J., van Mierlo, C. M., Smeets, J. B. J., and Brenner, E. (2011). Reweighting visual cues by touch. *J. Vis.* 11, 1–16.

Walker, P., Francis, B. J., and Walker, L. (2010). The brightness-weight illusion. *Exp. Psychol. (formerly Z. Exp. Psychol.)* 57, 462–469.

Woo, M., Neider, J., Davis, T., and OpenGL Architecture Review Board (1997). *OpenGL Programming Guide: The Official Guide to Learning OpenGL, Version 1.1*. 2nd Edn. Boston, MA: Addison-Wesley Publishing Company.

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: 23 May 2012; paper pending published: 19 June 2012; accepted: 24 October 2012; published online: 20 November 2012.

*Citation: Wismeijer DA, Gegenfurtner KR and Drewing K (2012) Learning from vision-to-touch is different than learning from touch-to-vision. Front. Integr. Neurosci. **6**:105. doi: 10.3389/fnint.2012.00105*

frontiers in
INTEGRATIVE NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 13 February 2013
doi: 10.3389/fnint.2013.00005

Combined diffusion-weighted and functional magnetic resonance imaging reveals a temporal-occipital network involved in auditory-visual object processing

***Anton L. Beer*[1,2]*, *Tina Plank*[1], *Georg Meyer*[3] *and Mark W. Greenlee*[1,2]**

[1] *Institut für Psychologie, Universität Regensburg, Regensburg, Germany*
[2] *Experimental and Clinical Neurosciences Programme, Universität Regensburg, Regensburg, Germany*
[3] *Department of Experimental Psychology, University of Liverpool, Liverpool, UK*

Edited by:
Hermann J. Mueller, University of Munich, Germany

Reviewed by:
Toemme Noesselt, Otto-von-Guericke University, Germany
Andre J. Szameitat, Ludwig Maximilians University, Germany

***Correspondence:**
Anton L. Beer, Institut für Psychologie, Universität Regensburg, Universitätsstr. 31, 93053 Regensburg, Germany.
e-mail: anton.beer@psychologie.uni-regensburg.de

Functional magnetic resonance imaging (MRI) showed that the superior temporal and occipital cortex are involved in multisensory integration. Probabilistic fiber tracking based on diffusion-weighted MRI suggests that multisensory processing is supported by white matter connections between auditory cortex and the temporal and occipital lobe. Here, we present a combined functional MRI and probabilistic fiber tracking study that reveals multisensory processing mechanisms that remained undetected by either technique alone. Ten healthy participants passively observed visually presented lip or body movements, heard speech or body action sounds, or were exposed to a combination of both. Bimodal stimulation engaged a temporal-occipital brain network including the multisensory superior temporal sulcus (msSTS), the lateral superior temporal gyrus (lSTG), and the extrastriate body area (EBA). A region-of-interest (ROI) analysis showed multisensory interactions (e.g., subadditive responses to bimodal compared to unimodal stimuli) in the msSTS, the lSTG, and the EBA region. Moreover, sounds elicited responses in the medial occipital cortex. Probabilistic tracking revealed white matter tracts between the auditory cortex and the medial occipital cortex, the inferior occipital cortex (IOC), and the superior temporal sulcus (STS). However, STS terminations of auditory cortex tracts showed limited overlap with the msSTS region. Instead, msSTS was connected to primary sensory regions via intermediate nodes in the temporal and occipital cortex. Similarly, the lSTG and EBA regions showed limited direct white matter connections but instead were connected via intermediate nodes. Our results suggest that multisensory processing in the STS is mediated by separate brain areas that form a distinct network in the lateral temporal and inferior occipital cortex.

Keywords: multisensory processing, auditory cortex, superior temporal sulcus, extrastriate body area, fMRI, DWI, structural connectivity, fiber tractography

INTRODUCTION

Identifying objects or actions in our environment usually relies on multiple sources of sensory cues such as sounds and images. Over the last decade several functional magnetic resonance imaging (fMRI) studies have associated the superior temporal cortex (STC) with multisensory object and action processing. For instance, reliable blood-oxygenation-level-dependent (BOLD) responses to auditory and visual stimuli as well as enhanced BOLD responses to bimodal stimuli were observed in a posterior part of the superior temporal sulcus (STS) (Beauchamp et al., 2004b, 2008; Hein et al., 2007; Noesselt et al., 2007). The authors suggested that the multisensory STS (msSTS) region found in humans likely reflects a homologue of the polysensory STS region observed in macaques. As brain imaging techniques such as fMRI, electroencephalography (EEG), or magnetoencephalography (MEG) detect the activity of large neural ensembles, overlapping (or enhanced) responses may also result from separate but interspersed neural populations rather than reflecting multisensory integration. Therefore, several researchers examined violations from linear summation or race model violations in brain or behavioral responses to bimodal stimuli as such violations suggest multisensory interactions (Schröger and Widmann, 1998; Laurienti et al., 2005; Stein et al., 2009; but see also Gondan and Röder, 2006; Proctor and Meyer, 2011; Szameitat et al., 2011). For instance, (degraded) bimodal auditory-visual stimuli elicited larger BOLD responses in msSTS than predicted by the sum of the BOLD responses to corresponding unimodal stimuli (Calvert et al., 2000; Werner and Noppeney, 2010b). However, such "superadditive" responses are not always observed (Hocking and Price, 2008; Meyer et al., 2011) and likely require degraded or noise stimuli (Laurienti et al., 2005; Angelaki et al., 2009). Other studies adopting salient (non-degraded) bimodal stimuli observed "subadditive" EEG/MEG responses (Schröger and Widmann, 1998) with a source in the STS (Raij et al., 2000; Cappe et al., 2010).

Other researchers compared responses to synchronous (simultaneously presented) and asynchronous (presented with a temporal offset) auditory-visual stimulus pairs (Calvert et al.,

2000; Miller and D'Esposito, 2005; Noesselt et al., 2007, 2012; Stevenson et al., 2011). They found that synchronous stimulus pairs (that were perceived as fused) elicited stronger BOLD signals in the STS than asynchronous pairs. These findings suggest that the STS merges multimodal signals based on temporal synchrony. Other studies examined semantic congruency between pairs of object sounds and visual stimuli (Hocking and Price, 2008). For instance, semantically incongruent auditory-visual stimulus pairs elicited more pronounced MEG responses in the STS compared to that evoked by congruent pairs (Raij et al., 2000). Similarly, fMRI adaptation research found that incongruent pairs of syllable sounds and lip movies that elicited the well-known McGurk illusion (McGurk and MacDonald, 1976) were associated with more adaptation in the STS than auditory-visual pairs that failed to elicit the McGurk illusion (Benoit et al., 2010). Likewise, video clips of point-light lip or body movements elicited weaker BOLD signals in the posterior STS when paired with congruent speech sounds or body action sounds, respectively, than when paired with incongruent sounds (Meyer et al., 2011). In this latter study, a one-back task was adopted which required observers to memorize a representation of the multimodal stimuli. Moreover, the BOLD difference between congruent and incongruent sound-video pairs was only observed with stimuli of real objects and actions but not with noise stimuli. This suggests that the STS contributes to a supramodal representation of objects and actions based on converging input of auditory and visual signals.

Many studies imply that multisensory processing relies on a single region within the posterior STS. However, recent progress in fMRI research (Beauchamp et al., 2004a; Van Atteveldt et al., 2010) and cell recordings (Dahl et al., 2009) suggests that multisensory processing in the STC relies on a network of spatially distinct regions and that the STC shows a more patchy organization than previously thought. For instance, processing of multimodal synchrony seems to involve at least two distinct subparts of the STS (Stevenson et al., 2011; Noesselt et al., 2012). Congruent compared to incongruent auditory-visual motion stimuli elicited more pronounced BOLD responses in the superior temporal gyrus (STG)—rather than the STS (Baumann and Greenlee, 2007). Likewise, spatially-semantic congruent sound-picture pairs elicited more activity in the STG compared to that evoked by incongruent sound-picture pairs (Plank et al., 2012). These lateral STG regions likely correspond to lateral belt and parabelt regions of the auditory cortex as described in macaques (Petkov et al., 2006) rather than the msSTS. Regions relevant for multisensory object processing were also observed outside the STS/STG complex. Several studies showing multisensory responses in the STS also reported multisensory activity in the inferior occipito-temporal cortex, anterior insula, and ventrolateral frontal cortex (Calvert et al., 2000; Beauchamp et al., 2004b; Hein et al., 2007; Meyer et al., 2011; Nath and Beauchamp, 2012). This diversity of findings suggests that the notion of a unitary STS region related to multisensory object processing needs to be reconsidered. It is likely that multisensory object processing relies on separate but inter-connected brain areas within the STC, the inferior occipito-temporal cortex and the frontal lobe. If this network notion is true, then understanding the connections between the nodes of this network becomes crucial in understanding multisensory processing.

Diffusion weighted imaging (DWI), first described in 1985 (Le Bihan and Breton, 1985) and sometimes also referred to as diffusion tensor imaging (Basser et al., 1994), is a non-invasive MRI technique that is sensitive to the diffusion of molecules (primarily water) in the brain. Molecular diffusion is primarily caused by thermal activity and is restricted by cell membranes. Brain regions containing coherent cell structures (e.g., axons of white matter) show a higher degree of anisotropic diffusion than other brain parts (e.g., somas and dendrites in gray matter). Voxel-wise measures of diffusion parameters such as the fractional anisotropy (FA) or diffusion vectors (tensors) derived from DWI allow inferences about the white matter structure. About one decade ago, tractographic approaches emerged that infer the path of least diffusion hindrance (tracks) across the white matter based on the diffusion parameters of an assembly of voxels. These white matter tracks are likely formed by axonal fiber bundles (tracts). Therefore, DWI-based tractography allows inferences about the white matter architecture of healthy humans or patients *in vivo* (Conturo et al., 1999; Jones et al., 1999; Mori et al., 1999; Lee et al., 2005). For instance, we recently reported evidence for white matter tracts between human auditory and visual cortex (Beer et al., 2011b). Combining tracking approaches based on DWI with conventional fMRI may resolve ambiguities in brain connectivity research. For instance, concurrent functional activity or resting-state connectivity between multiple brain areas do not necessarily require a direct (monosynaptic) anatomical connection (Damoiseaux and Greicius, 2009). Instead functional connectivity may result from indirect (polysynaptic) white matter connections. Moreover, structural connectivity studies have shown that brain areas, which cannot be distinguished otherwise, may be classified by their "connectivity fingerprints" (Behrens and Sporns, 2012).

The goal of this study was to examine the structural connections of the brain network involved in auditory-visual processing by means of white matter tracking. Therefore, probabilistic fiber tracking based on DWI was performed between auditory cortex and several brain areas involved in auditory-visual processing. We were primarily interested in the connectivity profile of the msSTS and related brain areas involved in processing biological sounds and visual motion. Brain areas involved in multisensory processing of speech and body actions were localized by whole-brain fMRI. The stimuli were adapted from a previous study that showed robust activation patterns in multisensory processing areas (Meyer et al., 2011). In order to control for confounds by behavioral responses, stimuli were task-irrelevant for the observer. Observers' attention was controlled by a simple detection task. Multisensory interactions were examined by a region-of-interest (ROI) analysis.

MATERIALS AND METHODS

PARTICIPANTS

The study comprised ten healthy volunteers (including one author, 7 females, all but one right-handed). All participants reported normal or corrected-to-normal vision and no hearing impairments. Their mean age was 27 years (range from 23 to 40).

All participants gave written informed consent prior to the study. The procedure was approved by the ethical board of the University of Regensburg.

AUDITORY-VISUAL TASK

Unisensory visual and auditory as well as multisensory (auditory-visual) brain areas were identified by fMRI while participants passively perceived biological motion stimuli (**Figure 1A**; Supplementary Movies 1 and 2). Task-irrelevant stimulus presentation was chosen in order to reduce the possibility that brain activity related to multisensory processing (e.g., in frontal cortex) was confounded by activity elicited by behavioral responses (see also Hein et al., 2007). Attention was controlled by asking observers to perform a simple detection task (see below). The stimuli were adopted from previous work in which they elicited significant activation in multisensory brain areas of the STC and other parts of the cortex (Meyer et al., 2011). In particular, visual stimuli consisted of videos with point-light displays of speech (lip) (VS) or body (VB) movements. Auditory stimuli consisted of sounds corresponding to the speech (AS) (sounds generated by the lip movements) or the body (AB) movements (sounds generated by the action). Speech stimuli represented nine distinct vowel-consonant-vowel syllables such as "aba" or "igi" spoken by a male speaker. Body stimuli represented nine distinct actions such as a person walking, jumping, cycling, rowing, sawing, etc. Each stimulus was presented for 1.2 s.

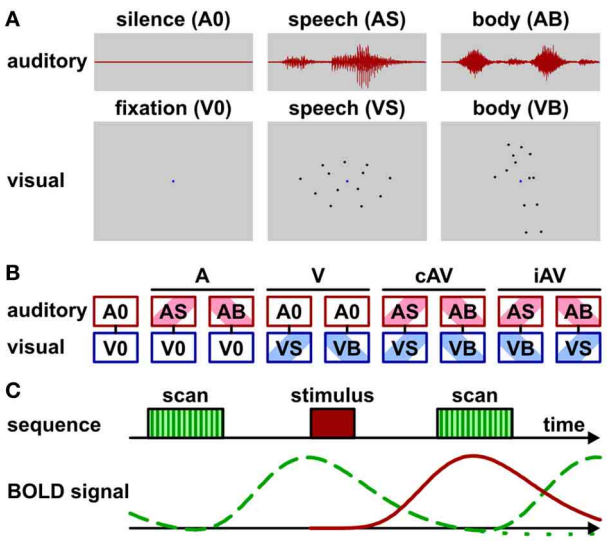


FIGURE 1 | Stimuli, conditions, and scan protocol. (A) Schematic depiction of stimulus types. Auditory stimuli (1200 ms duration) were speech sounds (AS) such as "aba" or sounds generated by body movements (AB) such as "sawing" noise. No sound was presented during baseline conditions (A0). Visual stimuli consisted of video clips (1200 ms duration) with point-light (13 dots) lip (speech) (VS) or body movements (VB). A blue dot in the display center served as fixation marker. Only the fixation dot was shown during baseline condition (V0). See also Supplementary Movies 1 and 2. **(B)** Audio-visual stimulus combinations: baseline (fixation without sound, A0V0), auditory only conditions (ASV0 or ABV0), visual only conditions (A0VS or A0VB), semantically congruent (cAV) conditions (ASVS or ABVB), semantically incongruent (iAV) conditions (ASVB or ABVS). **(C)** Trial sequence and schematic BOLD signals for sparse imaging. MR scan (acquisition) time (TA) was 2070 ms with a repetition time (TR) of 8000 ms. Stimuli were presented 4500 ms after scan onset during acquisition-free periods. Note that the stimulus-induced BOLD signal (red solid line) reaches its maximum during the acquisition time. The BOLD signal induced by the scanner noise (green dashed line) drops to about baseline at the acquisition time. Its post-stimulus undershoot (green dotted line) is masked by the peak of the subsequent BOLD response.

Each video clip showed 13 black moving dots on a gray background with a frame rate of 30 Hz. In addition, a blue dot in the center of the display served as a fixation point. The display size was normalized so that the mean dot deviation from the screen center was 2.7 degrees of visual angle (with a dot size of approximately 0.15 degrees). Participants viewed the video clips via a back-mirror (mounted within the head-coil) on a translucent screen positioned about 70 cm distant to the eye. Participants had to view the video clips while fixating the fixation point and press a button with their index finger whenever they detected a red dot. Accordingly, in some videos one dot was colored red for 300 ms during an interval between 300 and 900 ms. Sounds were presented via MR compatible headphones (MR confon, Magdeburg, Germany). Sound onset and offset was synchronized with the onset and offset of the video clips. All sounds were matched in root mean square power and presented with a sound pressure level of approximately 65 dB. Participants were asked to listen to the sounds and press a button whenever they detected a beep sound. Accordingly, in some trials a target beep sound (500 Hz, 130 ms, about 70 dB) occurred between 300 and 900 ms while the body action or speech sound was presented.

Visual and auditory stimuli were presented in nine conditions (**Figure 1B**) that were grouped into five main conditions: purely visual presentation of speech or body movements (VS or VB), purely auditory presentation of speech or body sounds (AS or AB), congruent auditory-visual presentation of speech or body stimuli (ASVS or ABVB), incongruent auditory-visual speech or body stimuli (ASVB or ABVS), or a silent baseline condition requiring participants to view the fixation point on an otherwise blank screen (A0V0). During congruent trials, each video clip was presented together with its matching sound. During incongruent trials, each clip was combined with a sound of the other category (e.g., speech sound with body action video).

In order to monitor attention, participants were asked to respond to a visual (red dot) or auditory (beep) target occurring in 20% of the trials (10% for each target). Targets were balanced across all stimulus conditions. They occurred in a pseudo-random order and never occurred on the same trial. The participants pressed a button on a response box upon target detection.

Stimuli were presented with an inter-trial-interval (ITI) of 8000 ms. This relatively large ITI was required, because a sparse-imaging MRI acquisition protocol with a repetition time (TR) of 8000 ms and an acquisition time (TA) of 2070 ms was adopted (**Figure 1C**). Sparse-imaging successfully reduces the influence of the scanner noise on the BOLD response (see Edmister et al., 1999; Engelien et al., 2002). Stimuli started 4.5 s after the onset (2.43 s after the offset) of the scanner acquisition phase. This timing was chosen for two reasons: First, in order to assure adequate perception, stimuli were presented during the silent phase

of the functional measurement stream. Second, based on previous experience the BOLD response reaches its maximum about 3–5 s after stimulus onset. A TR of 8 s and a stimulus onset at 4.5 s assured that the MR acquisition following the stimulus best captured the stimulus-induced BOLD signal with little interference by the BOLD signal induced by the scanner noise. Note that this timing ignores the "post-stimulus undershoot"—a temporary decline below baseline following the main peak—of the scanner noise BOLD response. As this post-stimulus undershoot is substantially smaller in magnitude than the main peak, equal across measurements, and masked by the main peak of subsequent measurements, it is usually ignored in auditory fMRI research (e.g., Petkov et al., 2006; Benoit et al., 2010). Each run of the multisensory task lasted about 12 min and consisted of 90 trials. At least three runs were conducted for each participant. Trials were pseudo-randomized in each run to avoid carry-over effects. Trial number per condition was balanced within each run. Stimulus pairings (e.g., targets combined with the different stimuli: "aba," "igi,". . .) were balanced across runs.

DATA ACQUISITION

All MRI data was acquired by a 3T head-only Allegra scanner (Siemens, Erlangen, Germany) using a one-channel whole-head coil while participants laid supine (head first) in the scanner bore. Head motion during the scans was constrained by cushions. For each participant, one high-resolution structural run, a series of functional runs, and three diffusion-weighted runs were acquired. The structural images were acquired with a magnetization-prepared rapid acquisition gradient echo (MPRAGE) sequence (TR = 2250 ms, echo time = 2.6 ms, inversion time = 900 ms, flip angle = 9°). The parameters were adapted from the Alzheimer's disease Neuroimaging project (Laboratory for Neuro Imaging, UCLA, Los Angeles, CA). The 160 sagittal slices covered the whole brain (voxel size = $1 \times 1 \times 1\,mm^3$, field of view = $256 \times 256\,mm^2$). Functional runs were acquired with a T2* weighted echoplanar sparse-imaging sequence (TR = 8000 ms, TA = 2070 ms, echo time = 30 ms, flip angle = 90°) with 36 axial slices (voxel size = $3 \times 3 \times 3\,mm^3$, field of view = $192 \times 192\,mm^2$, no inter-slice gap, interleaved acquisition). DWI runs were acquired with a single-shot pulsed gradient spin-echo sequence with echoplanar readout (TR = 7200 ms, echo time = 95 ms, flip angle = 90°). Diffusion was examined along 30 isotropically distributed orientations (Jones et al., 2002) and weighted by a b-value of 1000 s/mm². Five volumes without diffusion weighting (b-value of zero) were interspersed into the diffusion sequence every six volumes. The 54 axial slices covered the whole brain (voxel size = $2.5 \times 2.5 \times 2.5\,mm^3$, field of view = $240 \times 240\,mm^2$).

CORTICAL RECONSTRUCTION

Cortical reconstruction was automatically performed with Freesurfer version 4.1 (Martinos Center for Biomedical Imaging, Charlestown, MA) as described elsewhere (Beer et al., 2009). In brief, non-brain tissue was removed (Segonne et al., 2004), images were intensity corrected and normalized (Sled et al., 1998), subcortical volumetric structures were segmented (Fischl et al., 2002, 2004a), and the gray-white matter boundary was tessellated and topologic inaccuracies automatically corrected (Fischl et al., 2001; Segonne et al., 2007). Then, the cortical surface was deformed (Dale et al., 1999), inflated (Fischl et al., 1999a), registered to a spherical atlas that preserves individual folding patterns to match the cortical geometry across subjects (Fischl et al., 1999b), and automatically parcellated into units based on gyral and sulcal structures (Fischl et al., 2004b; Desikan et al., 2006).

WHOLE-BRAIN ANALYSIS OF fMRI

The fMRI data was analyzed with the FSFAST tools of Freesurfer. Pre-processing included motion correction (to the first volume of each session), intensity normalization (Cox and Jesmanowicz, 1999), and spatial smoothing with a three-dimensional Gaussian kernel of 8 mm (full-width at half-maximum). The first volume of each session was automatically co-registered to the structural volume. All co-registrations were verified by blink comparison and manually corrected if necessary.

In order to define brain regions relevant for auditory-visual processing, we performed a general linear model (GLM) whole-brain group analysis. The design matrix of the GLM contained separate predictors for all nine conditions (see **Figure 1B**) and a second order polynomial to model MR signal drift artifacts. Note that because of the sparse-imaging protocol (TR = 8000 ms, TA = 2070 ms) only one acquisition following stimulus presentation was modeled by a box-car predictor. In order to maximize statistical power and to detect all relevant brain regions, target trials (10% beep or 10% red dot) were included in the whole-brain analysis. A control analysis excluding these trials (not reported here) showed similar results. Note that target trials were excluded from the functional ROI analysis (see below). Group statistical parametric maps were calculated by a random-effects analysis. The analysis was restricted to the cortical surfaces and inter-subject normalization was performed by spherical (rather than volumetric) registration to the surface of the MNI standard brain (see above). Group significance maps were thresholded to $p = 0.01$. Additionally only clusters of contiguous vertices exceeding this threshold and spanning at least 120 mm² (approximately 10 functional voxels along the cortical surface) were considered.

Our primary motivation for the group analysis was to identify sensory-specific and putative multisensory regions of interest. Therefore, our analysis focused on five major contrasts: brain areas engaged in combined auditory and visual processing ([ASVS + ABVB + ASVB + ABVS]/4 vs. A0V0), brain areas engaged in auditory processing ([ASV0 + ABV0]/2 vs. A0V0), brain areas associated with phonological processing (ASV0 vs. ABV0), brain areas engaged in visual processing ([A0VS + A0VB]/2 vs. A0V0), and brain areas associated with processing body movements (A0VB vs. A0VS). ROIs were defined based on the group-average cortical significance maps (thresholded to $p = 0.001$) of these five main contrasts. Unisensory contrasts were used to define brain areas that were assumed to be modality-specific. These included the auditory cortex, the visual cortex, and predominantly visual areas within the parietal or frontal cortex. Moreover, a phonological processing region in the lateral superior temporal gyrus (lSTG) (Turkeltaub and Coslett, 2010; Woods et al., 2011) was defined by comparing auditory speech and body sounds (ASV0

vs. ABV0). Similarly, an extrastriate body area (EBA) (Peelen and Downing, 2007; Taylor and Downing, 2011) was defined by comparing visual body and lip movements (A0VB vs. A0VS). Brain areas assumed to be involved in multisensory processing were defined by comparing bimodal and unimodal contrasts. In particular, msSTS was assumed to be a brain region within the STS that responded to bimodal as well as to unimodal stimuli. However, bimodal contrasts had a higher statistical power than unimodal contrasts (as there were twice as many conditions). Consequently, at a given threshold bimodal contrasts showed slightly larger activity maps in putative multisensory regions than unimodal contrasts. Therefore, the borders of putative msSTS were marked based on bimodal contrasts. Unimodal responses within msSTS were verified by unimodal activity maps with a reduced threshold ($p = 0.05$). Implications of this relatively liberal definition criterion on our main findings are considered in the discussion (see below). Group labels (ROIs based on the cortical reconstruction) were reverse-mapped to individual cortical surfaces based on the spherical registrations. Both functional and structural landmarks (gyri and sulci) were used to segregate neighboring labels (ROIs).

DWI-BASED TRACTOGRAPHY

Pre-processing

DWI pre-processing and fiber tracking was conducted with the FDT toolbox of FSL (Centre for Functional Magnetic Resonance Imaging of the Brain, University of Oxford, Oxford, UK) (Smith et al., 2004; Woolrich et al., 2009). First, all diffusion-weighted runs were concatenated and corrected for head motion and image distortions due to eddy currents. Then, distributions on diffusion parameters were estimated for each voxel by means of Markov Chain Monte Carlo sampling (Behrens et al., 2003). Two fibers were estimated for each voxel unless prevented by automatic relevance detection (Behrens et al., 2007) in order to model complex fiber architectures. Finally, the first b-zero weighted image of the DWI series was automatically co-registered to the high-resolution T1-weighted anatomical image using Freesurfer tools. Each co-registration was inspected by "blink comparison" and manually corrected if necessary.

Fiber tracking

The fiber tracking procedure essentially followed the protocol as described elsewhere (Beer et al., 2011b). In brief, fibers were tracked by repetitively sampling from the distributions on voxel-wise principal diffusion directions. Each time a streamline through these local samples was calculated. Connectivity distributions were built by sampling many streamlines. We computed 5000 trajectories per seed voxel (resulting in 5000 × number of seed voxels tracks) with 2000 steps per sample (step length was 0.5 mm). Streamline trajectories were terminated when the angle between two steps fell below 60° (curvature threshold) or when the trajectory turned back on itself (loop criterion). Furthermore, tracking was constrained to the ipsilateral cortex. No FA threshold was applied. Path distributions were corrected for the length of the pathways. All trajectories were seeded with labels (ROIs) derived from the functional analysis (see above). These included the Heschl's region (H), the planum temporale (PT), msSTS, lSTG (sensitive to phonological sounds), and the EBA. Moreover, for further analysis several other labels (ROIs) based on the results of the functional or tractographic whole-brain group analysis were used as seeds (see "Results"). All labels (and seeds) were defined in structural space. Note that tracking was performed in diffusion space whereas tracking results were transformed back into structural space (same as seed space) using the co-registration matrices (see above).

The tracking analysis was limited to the ipsilateral cortical terminations of the tracks. This procedure proved to be most informative in previous studies (Beer et al., 2011b). For each voxel along the cortical surface, track probabilities were calculated by dividing the number of tracks reaching this voxel by the overall number of tracks. In order to reject (reduce) false positive tracks, probability maps were thresholded to $p = 5 \times 10^{-4}$. This threshold was adopted from our previous study (Beer et al., 2011b). Individual thresholded track termination maps were projected to the reconstructed surface of the standard brain by spherical surface-based (rather than volumetric) normalization (Fischl et al., 1999b) and aggregated. Finally, group aggregated track termination maps were thresholded (only track terminations found in at least 3 subjects were considered). Moreover, only track terminations that formed clusters spanning at least 120 mm² on the cortical surface were considered.

Previous research has shown that tracks seeded in the auditory cortex terminated in two distinct regions of the STS: anterior (aSTS) and posterior (pSTS). As we were interested in their role in multisensory processing, two ROIs (labels) were defined based on the group aggregate track termination maps for tracks seeded in H and PT, respectively. Moreover, additional ROIs (labels) in the inferior occipital cortex (IOC) were defined based on the results of the whole-brain track termination analysis (described in the "Results" section).

ROI ANALYSIS OF fMRI

Pre-processing for the ROI analysis was identical to the whole-brain analysis except that no spatial smoothing was applied. The mean MR signal change for each ROI was extracted by the GLM similar to the whole-brain analysis. In particular, the mean MR signal was estimated for each of the eight stimulus conditions and the baseline condition (A0V0). In addition, target trials (10% beep or 10% red dot) were modeled by separate predictors in order to exclude possible response-related activity. BOLD signals were calculated by subtracting the baseline signal (A0V0) from the MR signal of all eight stimulus conditions (e.g., [ASVS] - [A0V0]). Moreover, all BOLD signals were normalized to percent signal change relative to the ROI-specific global mean (constant predictor of the GLM). Our analysis focused on two aspects: (1) BOLD signals in response to auditory, visual, and bimodal stimuli regardless of stimulus type (speech or body movements) and (2) BOLD signal differences between speech (S) and body (B) stimuli reflecting feature-specific activity. For the first analysis, BOLD signals to speech and body (S, B) stimuli were pooled. One-sample *t*-tests were performed on BOLD signals for unimodal auditory and visual stimulus conditions in order to classify ROIs as being primary auditory, visual, or multimodal. In order to detect non-linear multisensory interaction effects (superadditive

or subadditive) BOLD signals to combined auditory-visual stimuli were compared to the sum of BOLD signals for unimodal stimulus conditions (e.g., [ASVS + ABVB]/2 vs. [ASV0 + ABV0]/2 + [A0VS + A0VB]/2). In order to detect multisensory congruency effects, BOLD signals to incongruent bimodal stimuli (iAV) were compared to BOLD signals of congruent (cAV) bimodal stimuli (e.g., [ASVB + ABVS]/2 vs. [ASVS + ABVB]/2). Note that all BOLD signals reflect differences to the baseline (A0V0) condition normalized to percent signal change (see above). For the second analysis, BOLD signals to body movements (B) were subtracted from BOLD signals to speech stimuli (S). The same comparisons were performed as for the feature-unspecific analysis. Note that for incongruent bimodal stimuli (ASVB, ABVS) the sign of the S-B difference of visual stimuli (B-S) is opposite of auditory stimuli (S-B). Therefore, responses to bimodal stimuli were compared to the sum of unimodal responses showing the same sign ([ASVS-ABVB] vs. [ASV0-ABV0] + [A0VS-A0VB] for congruent bimodal stimuli; [ASVB-ABVS] vs. [ASV0-ABV0] + [A0VB-A0VS] for incongruent stimuli).

STRUCTURAL CONNECTIVITY BETWEEN ROIs

White matter connections between ROIs as defined by the whole-brain analysis were estimated by additional trackings. Here, each ROI served as seed for separate trackings. Moreover, each ROI served as target area. In order to derive a measure of pair-wise connectivity strength, we counted the number of tracks emitting from the seed ROI and reaching the target ROI. Note that the number of tracks emitting from the seed is proportionate to the number of seed voxels. Moreover, the number of tracks terminating in the target ROI increases with the number of target voxels. In order to compensate for this dependency on ROI size, a normalized track connectivity index (TCI) was calculated by dividing the number of tracks by the product of the number of voxels in the seed and target ROI, respectively. All ROIs served as both seed and target ROI. The resulting two connectivity indices were averaged. Moreover, connectivity indices from left and right hemispheres were averaged. Finally, TCI values from the whole group were averaged.

RESULTS

BEHAVIORAL RESPONSES

All participants successfully detected most of the target stimuli (>95% hits, <1% false alarms). Moreover, no significant differences in response times were observed between visual and auditory targets suggesting that both modalities were attended equally well.

WHOLE-BRAIN ANALYSIS OF fMRI

In order to identify unisensory and putative multisensory brain regions involved in auditory and visual processing, a whole-brain analysis of the functional data was performed (see **Figure 2**). First, the response to all bimodal stimuli ([ASVS + ABVB + ASVB + ABVS]/4) was contrasted with the response to the control condition A0V0 (silence, blank screen). The BOLD response for this contrast revealed a large network of brain areas primarily in the temporal and occipital cortex extending to dorsal parietal and posterior frontal areas (**Figure 2A**). Comparing responses to unimodal auditory stimuli ([ASV0 + ABV0]/2) with the baseline condition (A0V0) showed that activity in the STC primarily reflected brain areas relevant for auditory processing (**Figure 2B**). Activity maps for both contrasts overlapped with the Heschl's region and the planum temporale—representing the core and caudal belt of auditory cortex, respectively (Petkov et al., 2006; Da Costa et al., 2011). The activity map in the auditory cortex further extended to lateral parts of the STG, which likely correspond to lateral belt and parabelt regions of the auditory cortex. The contrast comparing responses to speech sounds with sounds generated by body movements (AS vs. AB) (**Figure 2C**) revealed a relatively distinct region in the lateral STG associated with phonological processing (Turkeltaub and Coslett, 2010; Woods et al., 2011).

The contrast comparing responses to unimodal visual stimuli ([A0VS + A0VB]/2) with responses to the baseline condition (A0V0) revealed a network of brain areas primarily in the occipital and dorsal parietal cortex (**Figure 2D**). Two small clusters of activity were found in the anterior part of the calcarine sulcus (CaS) and the occipital pole (OcPo), respectively. These two clusters fell within the primary visual cortex (V1) as verified by the automatic parcellation of Freesurfer and likely reflect peripheral and central representations of the visual field. Larger clusters were found in the lateral occipital cortex stretching anterior to the posterior part of the middle temporal gyrus—a region overlapping with the motion-sensitive MT complex (MT+) (Zihl et al., 1983; Tootell et al., 1995). Activities in the parietal cortex were limited to ventral and dorsal parts of the intraparietal sulcus (IPS). A relatively large cluster of activity was found in the inferior and lateral parts of the occipito-temporal cortex. The posterior part of this cluster overlapped with track terminations (see below). The anterior part of this cluster was primarily found in the fusiform gyrus (FG). Furthermore, a small cluster of activity was found in the parietal (PIC) and the anterior insular cortex (AIC). Moreover, several adjacent clusters of activity were observed in ventral (vlFC) and dorsal (dlFC) parts of the posterior lateral frontal cortex. The contrast comparing responses to body movements with speech (lip) movements (VB vs. VS) revealed two relatively distinct regions in the lateral occipital cortex and FG presumably reflecting the EBA and the fusiform body area (Peelen and Downing, 2007; Taylor and Downing, 2011).

Although both unimodal auditory and unimodal visual contrasts showed distinct activity patterns throughout most of the cortex, several brain regions were activated by unimodal auditory, visual, and bimodal contrasts. This putative multisensory brain network included a region in the posterior part of the STS that most likely corresponds to the msSTS area (Beauchamp et al., 2004b, 2008). Note that the borders of msSTS were based on the bimodal contrast and verified by unimodal contrasts with a lower threshold (see "Materials and Methods" section for details). In addition, regions in the parieto-occipital sulcus (POS) and the CaS were activated by visual, auditory, and bimodal stimuli. Finally, the ventrolateral frontal cortex was responsive to both unimodal and bimodal stimuli. Note that most functional activities were similar in the two hemispheres.

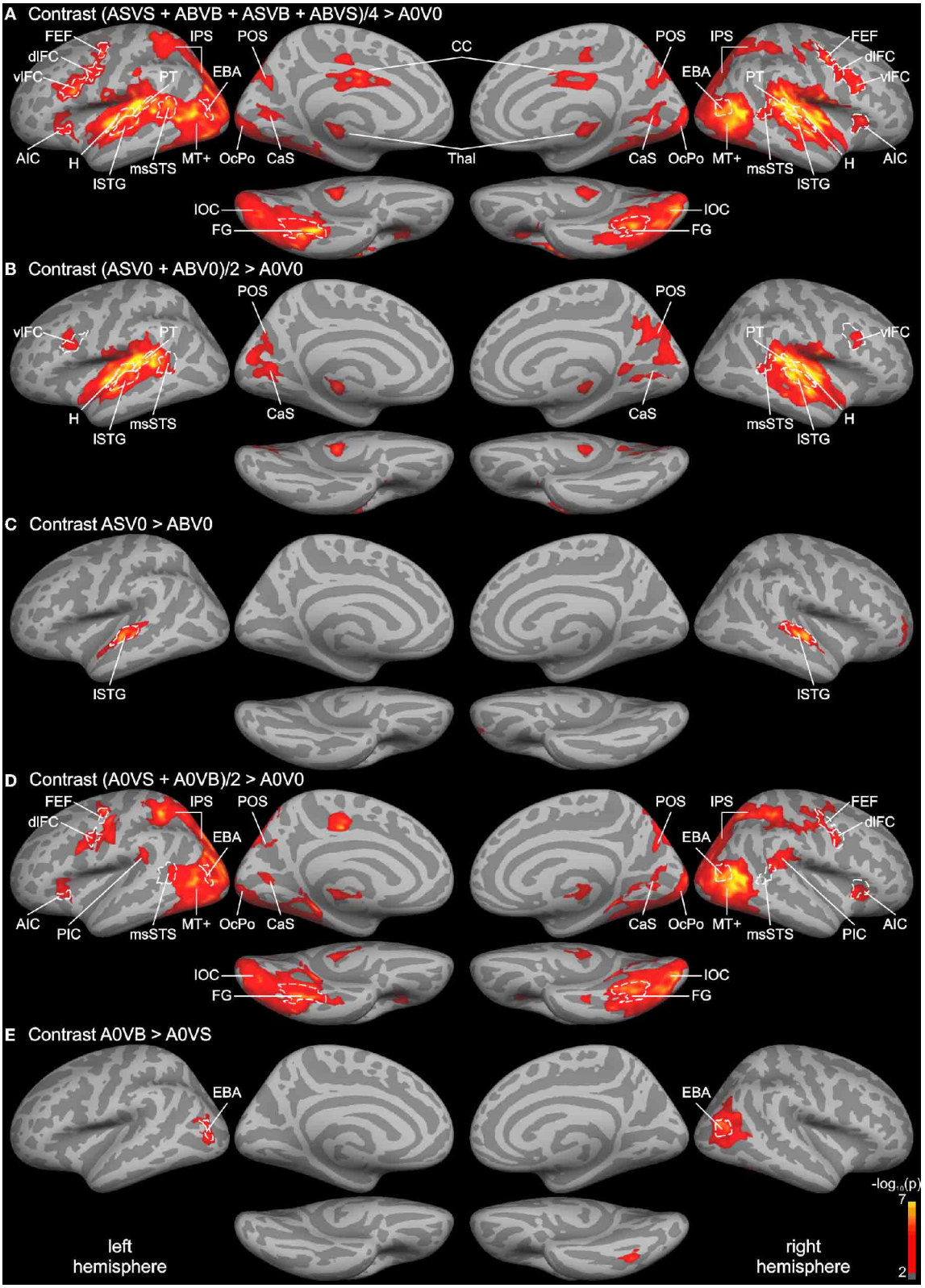


FIGURE 2 | Whole-brain group analysis statistical parametric maps overlaid on cortical surfaces of the MNI standard brain. Both left and right hemispheres are shown from a lateral, medial, and inferior view. Five relevant contrasts are shown: **(A)** Contrasting BOLD responses to all bimodal stimuli with BOLD responses to control stimuli (A0V0) showed brain areas relevant for auditory, visual, and multisensory processing. **(B)** Brain areas primarily involved in auditory processing were identified by contrasting

(Continued)

FIGURE 2 | Continued
responses to unimodal auditory stimuli with the control. **(C)** Brain areas specific to phonological processing were identified by contrasting auditory phonological (lip) sounds (AS) with body sounds (AB). **(D)** Brain areas primarily involved in visual motion processing were identified by contrasting responses to unimodal visual stimuli with the control condition. **(E)** Brain areas specific to body processing were revealed by contrasting responses to visual body movements (VB) with visual lip movements (VS). All contrasts were thresholded to $p = 0.01$ (red) and color-coded (yellow: $p = 10^{-7}$). An additional cluster threshold of 120 mm² was applied. Regions of interests (ROIs) were defined as cortical labels (marked in white) based on functional (threshold $p = 0.001$) and structural (gyri and sulci) criteria (see text). Labels (ROIs) are indicated as dashed white lines. AIC, anterior insular cortex; CaS, calcarine sulcus; CC, cingulate cortex; CenS, central sulcus; dlFC and vlFC, dorsal and ventral parts of posterior lateral frontal cortex; EBA, extrastriate body area; FEF, frontal eye field; FG, fusiform gyrus; H, Heschl's region; IOC, inferior occipital cortex; IPS, intraparietal sulcus; lSTG, lateral superior temporal gyrus; msSTS, multisensory superior temporal sulcus; MT+, motion-sensitive middle temporal area plus satellites; OcPo, occipital pole; PIC, parieto-insular cortex; POS, parieto-occipital sulcus; PT, planum temporale.

WHOLE-BRAIN ANALYSIS OF TRACK TERMINATIONS

Probabilistic tracking was performed in the same participants to determine the white matter connectivity across cortical brain regions involved in auditory and visual processing of objects. The tracking algorithm was seeded in several auditory, visual, and multimodal cortical ROIs (labels) as determined by functional and structural (gyral and sulcal structure) criteria (see above). **Figure 3** shows the cortical track terminations of the whole group for several seed regions. Only track terminations that exceeded the track frequency threshold in at least three individual brains and which exceeded the cluster threshold are displayed. Tracks seeded in the Heschl's region primarily terminated in several distinct cortical regions of the temporal and occipital cortex (**Figure 3A**). Track terminations in the temporal cortex were seen in adjacent auditory cortex including the planum temporale and the lSTG. Projections were also observed in an anterior division of the STS (aSTS). Furthermore, several foci of H track terminations were observed in the IOC (see details below). H track terminations were also found in several areas of the medial occipital lobe (see **Figure 4** for an enlarged view): the OcPo, a region in the anterior CaS, and dorsal POS. Other track terminations were found at the cortical border to the corpus callosum (Cal) and thalamus (Thal)—most likely reflecting inter-hemispheric and thalamic fiber connections—and the anterior insula.

Combined functional and structural criteria were adopted for the PT seed. Only the part of the planum temporale that was functionally active during the auditory or bimodal task was included. Tracks seeded in PT terminated in two distinct regions of the STS: anterior (aSTS) and posterior (pSTS) divisions (see **Figure 5** for an enlarged view). PT tracks further projected to the IOC and a distinct region in the central sulcus (CenS). Tracks also reached the hemisphere border to the corpus callosum. Little or no PT track terminations were observed in the medial occipital lobe.

Our primary interest was in the connectivity profile of the msSTS region. Therefore, tracks were seeded in the part of the STS that was active during the multisensory localizer task (see above). Our primary interest was to evaluate the white matter connectivity of this region with auditory and visual cortex. As illustrated in **Figure 3C**, little or no connections between the msSTS region and the Heschl's region of the auditory cortex or early visual cortex (e.g., medial occipital cortex) were observed. Moreover, no track terminations were observed in the IOC. Instead terminations of the msSTS region primarily terminated in the planum temporale of the auditory cortex, the lateral STG, other parts of the STS, the middle temporal gyrus, inferior parietal cortex, and the CenS.

Our functional analysis revealed two regions that were either specific to phonological processing (lSTG) or visual body movements (EBA). Both regions showed signs of multisensory interactions (see below). Therefore, we were interested in the connectivity profile of these two regions. Tracks seeded in the phonological processing area (lSTG) showed wide-spread terminations in the auditory cortex including the H and PT region and the STS (**Figure 3D**). However, STS terminations were primarily observed in the aSTS region with limited connections to the msSTS region. Substantial track terminations were observed in the IOC. Tracks seeded in the EBA (**Figure 3E**) primarily terminated in adjacent regions. In addition, white matter connections were observed with the aSTS region in both hemispheres in some brains.

In order to compare functional activity maps with track termination maps enlarged views of the cortical surfaces are shown in **Figures 4–6**. As shown in **Figure 4**, H track terminations in the medial occipital cortex (CaS, OcPo, POS) corresponded quite well with the activity clusters as determined by the functional localizer. However, the connectivity profile of tracks terminating in the STS (**Figure 5**) was more complex than expected. One of the most relevant findings was that H and PT track terminations (aSTS and pSTS) showed little or no overlap with the functional msSTS region. In order to further investigate the connectivity profile of the STS, we performed additional tracking using the aSTS and pSTS regions as seeds. The results of this tracking revealed that the aSTS region showed relatively strong connections to the auditory cortex (H and PT), the phonological processing area (lSTG) as well as to the msSTS region. By contrast, the pSTS region was primarily connected to posterior parts of the auditory cortex (PT) and the aSTS region. These findings suggest that the msSTS region showed no direct white matter connections with the primary auditory cortex (H) but instead is connected to the auditory cortex via intermediate brain areas such as PT and aSTS.

Several previous studies have shown that activity in the STS is accompanied by activity in the inferior or lateral occipital cortex (Beauchamp et al., 2004b; Meyer et al., 2011). **Figure 6** compares the activity maps of our functional localizer with the connectivity profiles of tracks terminating in the IOC. Comparing terminations of tracks from different seeds suggests that the IOC can be sub-divided into three main areas based on their connectivity profile. At the most posterior end—overlapping with the collateral transverse sulcus (CTS)—terminations were primarily found from H tracks. Adjacent to the CTS—overlapping with posterior parts of the lateral occipito-temporal gyrus or FG (lOTG)—terminations were observed from H tracks, lSTG

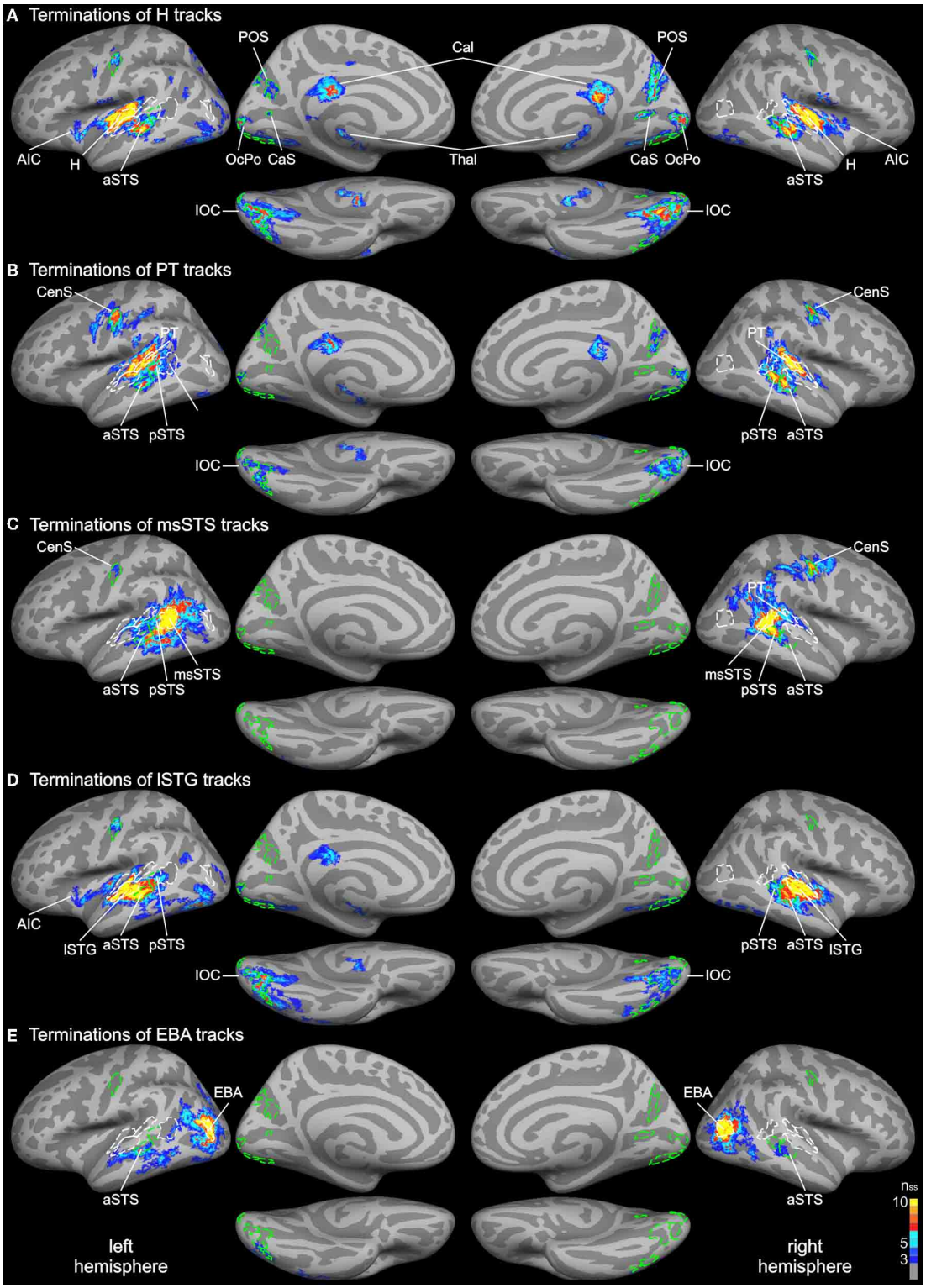


FIGURE 3 | Group average track termination maps overlaid on cortical surfaces of the MNI standard brain. Termination maps were thresholded to $n_{ss} = 3$. Color scale: dark blue, terminations found in three hemispheres (threshold); yellow, terminations found in all hemispheres. The different panels show separate termination maps for tracks seeded in **(A)** Heschl's region (H), **(B)** planum temporale (PT), **(C)** multisensory STS (msSTS), **(D)** lateral STG (lSTG) sensitive to phonological sounds, **(E)** extrastriate body area (EBA). Labels (ROIs) based on track terminations are indicated as dashed green lines. Functional labels are shown in white. AIC, anterior insular cortex; aSTS and pSTS, anterior and posterior region of superior temporal sulcus; CaS, calcarine sulcus; CenS, central sulcus; IOC, inferior occipital cortex; OcPo, occipital pole; POS, parieto-occipital sulcus.

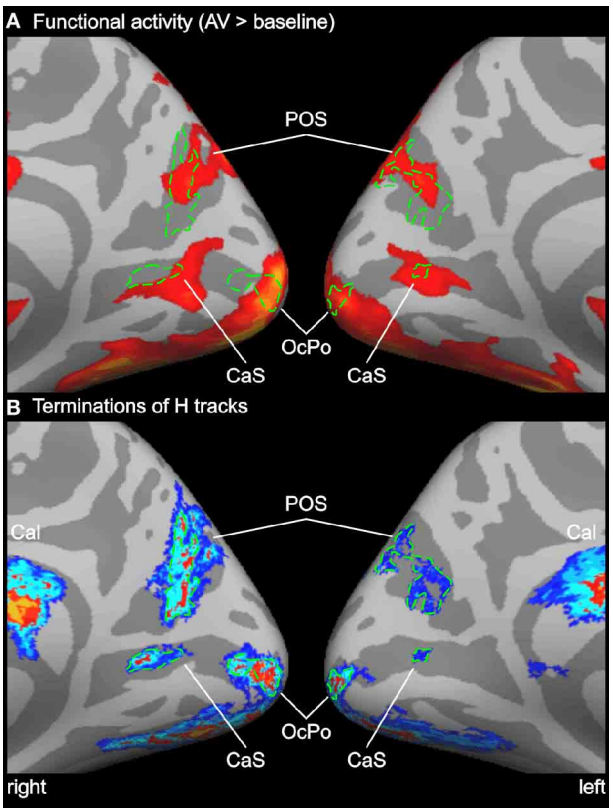


FIGURE 4 | Enlarged view of functional activity and track terminations in medial occipital cortex. (A) Functional activity contrasting BOLD responses to all bimodal stimuli with BOLD responses to control stimuli (A0V0). **(B)** Termination maps for tracks seeded in the Heschl's region (H). Labels (ROIs) based on track terminations are shown in green. See also **Figures 2** and **3** for abbreviations.

tracks, and aSTS tracks. More lateral—overlapping with the lateral occipito-temporal sulcus (lOTS)—track terminations were primarily observed from lSTG tracks and aSTS tracks. Note that no direct white matter connections were observed between the STC and anterior parts of the lateral occipito-temporal/FG that showed activity in the fMRI analysis (but see ROI connectivity analysis below).

ROI ANALYSIS OF fMRI

In order to further quantify the contribution of each ROI on multisensory processing, we performed a ROI analysis on the functional data. We were primarily interested in determining the extent to which the region was primarily auditory, visual, or multisensory. Therefore, we compared the mean activity of each ROI during unimodal auditory (ASV0, ABV0) or visual stimulation (A0VS, A0VB) with the baseline condition (A0V0) (see **Table 1** and **Figures 5I–M**). Furthermore, we were interested in whether the response to bimodal stimuli was larger (superadditive) or smaller (subadditive) than the sum of unimodal responses. Therefore, the sum of unimodal responses was subtracted from the response of congruent (ASVS, ABVB) and incongruent (ASVB, ABVS) stimuli, respectively. Finally, responses to congruent and incongruent were compared. Repeated-measures ANOVAs conducted separately for each comparison showed that BOLD responses significantly differed across ROIs. **Table 1** shows the results of the ROI analysis averaged across speech and body movement conditions. As expected ROIs in the auditory cortex (H, PT) showed significant BOLD responses to auditory but not visual stimuli. Furthermore, unimodal auditory stimuli elicited significant BOLD responses in the lateral STG and the STS. Interestingly, auditory stimuli also elicited BOLD responses in several brain areas that were assumed to be primarily visual: CTS, lOTG, POS, CaS, and OcPo. Unimodal visual responses were observed in the medial occipital cortex (OcPo, POS) and the inferior occipito-temporal cortex (CTS, lOTS, lOTG, FG). Furthermore, visual responses were observed in the EBA region and posterior parts of the STS (pSTS, msSTS). No significant unimodal visual responses were observed in the auditory cortex.

Subadditive responses to bimodal stimuli (irrespective of stimulus type) were primarily observed in the posterior part of the STS (pSTS, msSTS). Moreover, subadditive responses to congruent bimodal stimuli were observed in the inferior occipito-temporal cortex (FG, lOTG, CTS), parts of the occipital cortex (POS, OcPo), and the planum temporale. Significant BOLD response differences between congruent and incongruent bimodal stimuli were observed in the planum temporale, EBA, IOC (lOTS, CTS), the OcPo, and anterior insula.

As our whole-brain analysis revealed cortical regions that responded preferably to speech sounds (S) or visual body movements (B), an additional ROI analysis was performed on the differences between S and B stimulus types (S-B) (see **Table 2** and **Figures 5J–M**). Note that positive differences reflect stronger BOLD responses to speech (S) stimuli and negative differences reflect stronger responses to body (B) stimuli. Further note that responses to bimodal stimuli were compared with the sum of its corresponding unimodal differences (see "Methods and Materials" section). Repeated-measures ANOVAs conducted separately for each comparison showed that BOLD responses significantly differed across ROIs. As expected from the results of the whole-brain analysis, the lSTG region showed stronger responses to unimodal phonological sounds (ASV0) than to body sounds (ABV0). By contrast, weaker responses to speech sounds were observed in the OcPo. Similarly, the EBA region and the anterior part of the FG showed stronger responses to body movements (A0VB) than to lip (speech) movements (A0VS). In addition, unimodal visual lip (speech) movements (A0VS) elicited stronger responses than body movements (A0VB) in the phonological STG region (lSTG), subparts of the STS (aSTS and pSTS but not msSTS), the inferior (CTS) and posterior (OcPo) occipital cortex.

Congruent bimodal stimuli elicited subadditive difference responses (ASVS-ABVB) in several brain regions including lSTG, pSTS (but not msSTS), and the EBA. By contrast, incongruent bimodal stimuli (ASVB-ABVS) elicited superadditive responses compared to their unimodal counter-parts in most of these regions including the OcPo. Accordingly, a significant difference in the bimodal response pattern of congruent and incongruent bimodal stimuli was observed in lSTG, aSTS, pSTS, EBA, and vlFC, suggesting that congruent and incongruent bimodal stimuli were processed differently in these areas.

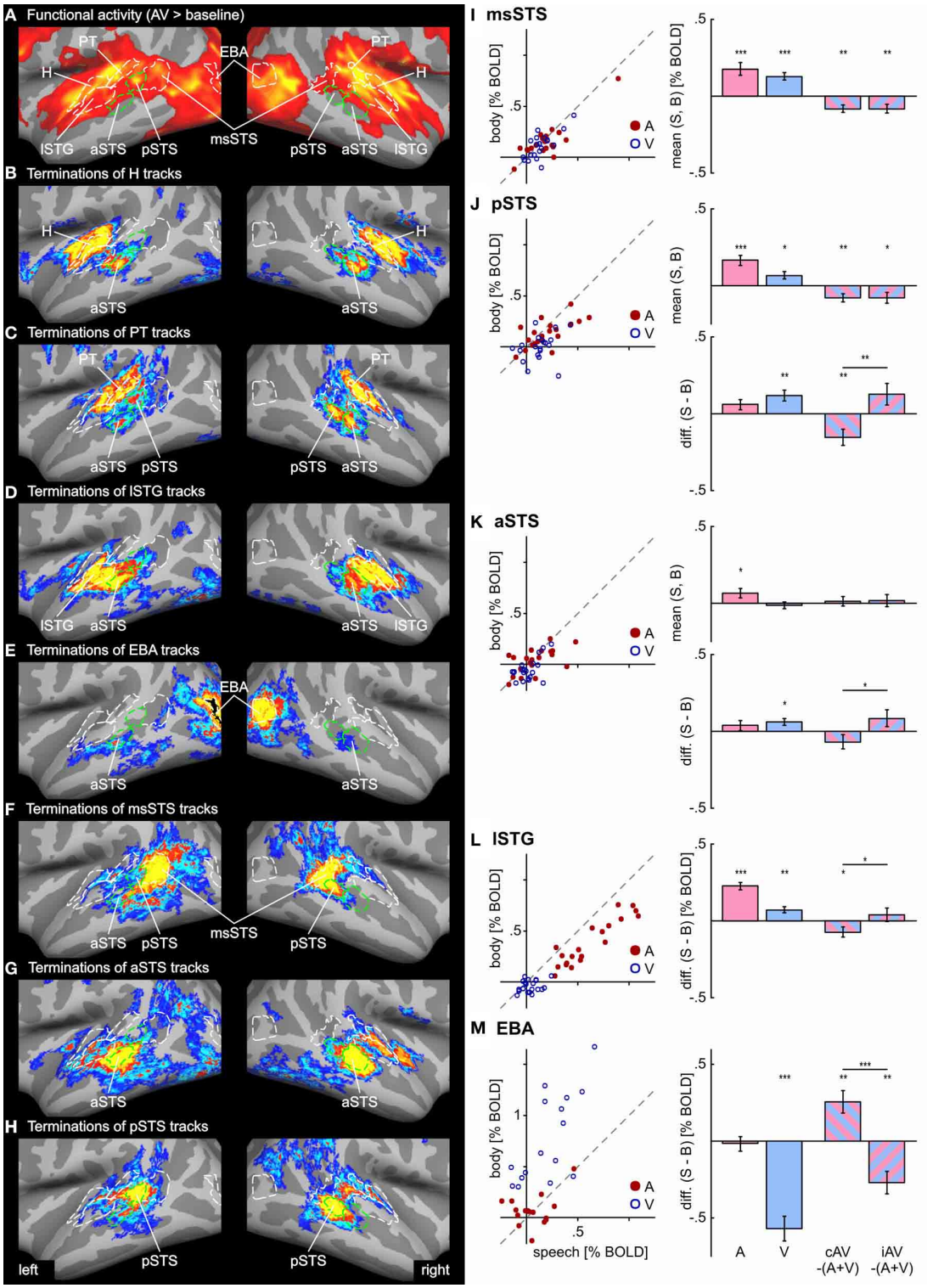


FIGURE 5 | Functional activity and track terminations in temporal cortex. Enlarged views of functional activity and track termination maps are presented in panels **(A–H)**. Labels (ROIs) based on track terminations are shown in green, labels (ROIs) based on functional activity in white. Panels **(I–M)** present the results of the functional ROI analysis for temporal cortex regions (see **Tables 1** and **2** for other ROIs). BOLD responses to unimodal (auditory or visual) conditions (relative to baseline) are shown separate for all hemispheres ($n = 20$) in scatter plots. Deviations from the main diagonal indicate specificity for stimulus type (speech or body). Bar graphs depict the group mean BOLD responses to unimodal conditions as well as differences between bimodal responses to the sum of unimodal responses (AV − [A + V]) separate for congruent (cAV) and incongruent (iAV) stimulus pairs (***$p < 0.001$; **$p < 0.01$; *$p < 0.05$). Separate graphs are shown for the mean responses across stimulus type and the response difference between speech and body (S-B) stimuli. See also **Figures 2** and **3** for abbreviations.

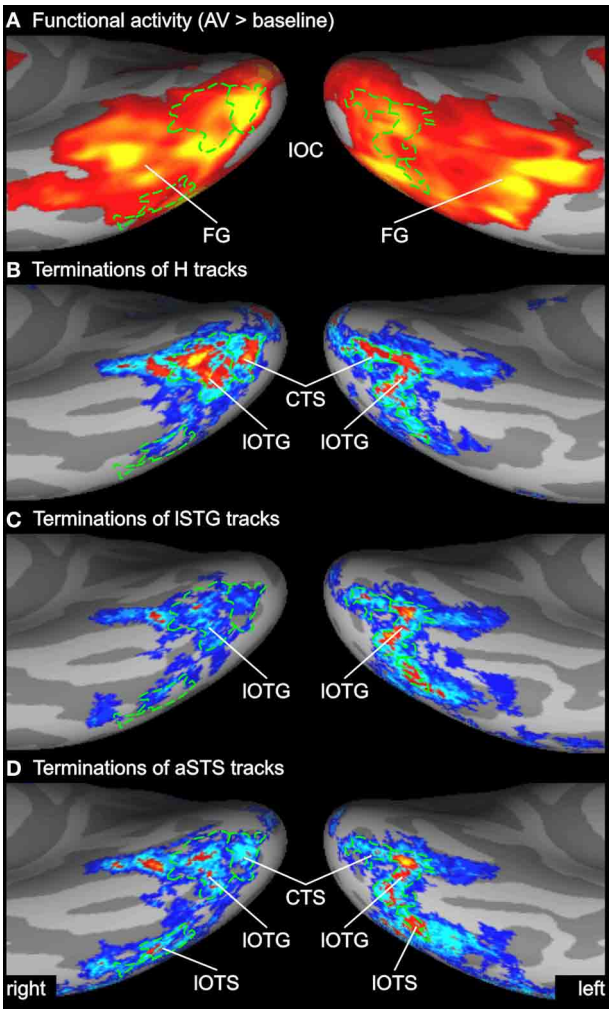


FIGURE 6 | Enlarged view of functional activity and track terminations in inferior occipito-temporal cortex. (A) Functional activity contrasting BOLD responses to all bimodal stimuli with BOLD responses to control stimuli (A0V0). **(B–D)** Termination maps for tracks seeded in the Heschl's region (H), lateral superior temporal gyrus (lSTG), and anterior superior temporal sulcus (aSTS), respectively. Labels (ROIs) based on track terminations are shown in green. CTS, collateral transverse sulcus; lOTG and lOTS, lateral occipito-temporal gyrus and sulcus. See also **Figures 2** and **3** for other abbreviations.

STRUCTURAL CONNECTIVITY BETWEEN ROIs

Our ROI analysis on functional data and our whole-brain tracking results indicated that several distinct brain areas were involved in the processing of auditory and visual stimuli. Therefore, we performed pair-wise probabilistic tracking in order to quantify the strengths of white matter connections using the ROIs as seed and targets. Track probabilities were normalized by the size of the seed and target regions, respectively. The group average normalized ROI-to-ROI track probabilities (pooled across both hemispheres) are shown in **Table 3**. This analysis primarily confirmed the results on the whole-brain track terminations reported above. That is, H tracks primarily terminated in the lSTG, aSTS (but not msSTS), IOC, and medial occipital cortex. PT tracks primarily terminated in the lSTG, aSTS, pSTS, CenS, and dlFC. Tracks seeded in the phonological lSTG area primarily terminated in the aSTS and pSTS region as well as the lateral inferior occipito-temporal cortex. The aSTS region showed a connectivity profile distinct from the more posterior pSTS and msSTS areas. Whereas the aSTS region showed strong projections to the IOC, no such projections were observed in more posterior regions of the STS (pSTS, msSTS). The pSTS region showed relatively strong connectivity with area PT. The msSTS region was primarily connected with the aSTS region. The EBA was primarily connected with aSTS and sub-regions of the IOC (lOTS).

The ROI-to-ROI connectivity analysis further revealed the connectivity profile of inferior occipital regions. These regions were primarily connected with the H part rather than the PT part of the auditory cortex. Furthermore, anterior regions of the IOC (lOTG, lOTS) were connected with the lSTG and the aSTS, but little connectivity was observed with the msSTS. Moreover, parts of the IOC (lOTG) were connected to the medial occipital cortex (CaS).

The ventrolateral frontal cortex as well as the anterior insular region observed in fMRI showed little direct connectivity with the STC. Instead these two regions were connected via intermediate nodes such as the dorsolateral frontal cortex or the inferior occipito-temporal cortex (e.g., FG, lOTS).

DISCUSSION

Our whole-brain analysis revealed a large network of brain areas responding to auditory, visual, or multimodal stimuli. Processing of unimodal auditory stimuli (**Figures 2B,C**) primarily involved the STC, distinct parts of the medial occipital cortex, and the ventrolateral frontal cortex. Activity clusters in the temporal lobe included the Heschl's region and the planum temporale, which likely correspond to the core and caudal belt of auditory cortex, respectively (Petkov et al., 2006; Da Costa et al., 2011). Consistent with previous research (Turkeltaub and Coslett, 2010; Woods et al., 2011) a relatively distinct region in the lateral anterior STG sensitive to phonological (speech) processing was observed in both hemispheres.

Processing of unimodal visual stimuli involved a network of brain areas primarily in the occipital and dorsal parietal cortex (**Figures 2D,E**). This network largely corresponds to brain networks related to visual motion processing as described elsewhere (Kovacs et al., 2008; Beer et al., 2009). It included activity in the posterior part of the middle temporal gyrus—a region known as the motion-sensitive MT complex (MT+) (Zihl et al., 1983; Tootell et al., 1995). The visual network also involved distinct regions in the medial occipital cortex (CaS and POS) and inferior and lateral parts of the occipito-temporal cortex. Two distinct regions in the lateral occipital cortex and FG, respectively, were sensitive to visual body movements. Similar regions were described before as EBA and fusiform body area (Peelen and Downing, 2007; Taylor and Downing, 2011). Moreover, several adjacent clusters of activity were found in the frontal cortex.

Although auditory and visual processing was associated with distinct brain areas throughout most of the cortex, several cortical sites were activated by both auditory and visual stimuli. In particular, we observed a region in the posterior part of the

Table 1 | Mean BOLD responses per ROI.

Label	Unimodal				Bimodal					
	A		V		cAV−(A+V)		iAV−(A+V)		iAV−cAV	
	Mean	(SE)	Mean	(SE)	Mean	(SE)	Mean	(SE)	Mean	(SE)
H	**1.06*****	(0.06)	0.02	(0.01)	−0.05	(0.03)	−0.04	(0.03)	0.01	(0.01)
PT	**0.73*****	(0.06)	0.03	(0.02)	**−0.06***	(0.03)	−0.01	(0.04)	**0.05***	(0.02)
lSTG	**0.51*****	(0.05)	0.00	(0.01)	−0.01	(0.03)	−0.01	(0.03)	0.00	(0.01)
aSTS	**0.06***	(0.03)	−0.02	(0.02)	0.02	(0.04)	0.01	(0.03)	−0.01	(0.02)
pSTS	**0.16*****	(0.03)	**0.07***	(0.02)	**−0.08***	(0.03)	**−0.08****	(0.03)	0.00	(0.02)
msSTS	**0.17*****	(0.04)	**0.13*****	(0.03)	**−0.08****	(0.03)	**−0.09****	(0.02)	0.00	(0.01)
EBA	0.07	(0.03)	**0.49*****	(0.08)	**−0.13***	(0.04)	−0.04	(0.03)	**0.09***	(0.04)
lOTG	**0.07***	(0.03)	**0.38*****	(0.05)	**−0.07***	(0.03)	−0.03	(0.02)	0.04	(0.02)
lOTS	0.01	(0.03)	**0.18*****	(0.03)	−0.04	(0.04)	0.02	(0.03)	**0.05***	(0.02)
CTS	**0.11****	(0.03)	**0.47*****	(0.05)	**−0.09***	(0.04)	−0.03	(0.03)	**0.06***	(0.03)
FG	0.06	(0.05)	**0.31*****	(0.04)	**−0.12***	(0.05)	−0.05	(0.04)	0.07	(0.03)
POS	**0.10****	(0.03)	**0.10****	(0.03)	**−0.10***	(0.04)	−0.05	(0.03)	0.06	(0.03)
CaS	**0.15****	(0.04)	0.08	(0.05)	−0.07	(0.03)	−0.04	(0.04)	0.03	(0.04)
OcPo	**0.11****	(0.04)	**0.24*****	(0.05)	**−0.10****	(0.03)	−0.03	(0.04)	**0.07***	(0.03)
AIC	0.03	(0.02)	**0.09****	(0.02)	−0.04	(0.02)	0.02	(0.03)	**0.06***	(0.02)
CenS	0.06	(0.03)	**0.06****	(0.02)	**−0.09***	(0.04)	−0.05	(0.04)	0.04	(0.02)
dlFC	−0.02	(0.03)	**0.14*****	(0.03)	0.01	(0.04)	−0.01	(0.03)	−0.01	(0.03)
vlFC	0.07	(0.04)	**0.11****	(0.02)	−0.05	(0.02)	−0.01	(0.04)	0.04	(0.03)

*Note: The table lists the mean BOLD responses (percent signal change) relative to baseline (A0V0) averaged across speech (S) and body (B) conditions for each ROI (label). Standard errors (SE) are shown in parenthesis. Significant differences are highlighted in bold (***$p < 0.001$; **$p < 0.01$; *$p < 0.05$, uncorrected). A, unimodal auditory; V, unimodal visual; cAV, congruent bimodal; iAV, incongruent bimodal. ROIs were defined as labels on cortical surfaces (see **Figures 2**, **3**, **6**). AIC, anterior insular cortex; aSTS and pSTS, anterior and posterior region of superior temporal sulcus; CaS, calcarine sulcus; CenS, central sulcus; CTS, collateral transverse sulcus; dlFC and vlFC, dorsal and ventral parts of the posterior lateral frontal cortex; EBA, extrastriate body area; FG, fusiform gyrus; H, Heschl's region; lOTG and lOTS, lateral occipito-temporal gyrus and sulcus; lSTG, lateral superior temporal gyrus; msSTS, multisensory superior temporal sulcus; OcPo, occipital pole; POS, parieto-occipital sulcus; PT, planum temporale.*

STS that was activated by auditory, visual, and bimodal stimuli (**Figures 2**, **5**). Moreover, auditory and visual activity was also observed in the medial occipital cortex (POS, CaS) and the ventrolateral frontal cortex. We subsequently discuss the activity and connectivity profiles of sub-parts of the auditory-visual brain network in detail.

MEDIAL OCCIPITAL CORTEX

The medial occipital cortex showed two clusters in the anterior part of the CaS and the OcPo that were activated by auditory-visual stimuli. These two regions likely reflect peripheral and central representations of the primary visual cortex (V1), respectively (e.g., Beer et al., 2009). In addition, auditory-visual stimuli elicited activity in dorsal parts of the POS. Although these regions are considered modality-specific visual areas, they also showed BOLD responses to purely unimodal auditory and subadditive responses to bimodal stimulation (see **Table 1** and **Figure 2**). This finding of sound-induced activity and crossmodal response modulation in visual cortex is in accordance with a number of EEG/MEG (McDonald et al., 2003; Raij et al., 2010), positron emission tomography (Weeks et al., 2000; Gougoux et al., 2005), and fMRI (Röder et al., 2002) studies. Traditionally, this crossmodal recruitment of visual cortex was attributed to feedback signals from multisensory association cortex (McDonald et al., 2003). However, several lines of research suggest that there are even direct connections between primary sensory cortices (Foxe and Schroeder, 2005). For instance, sounds presented prior to a (peripheral) visual target facilitated visual perception only when sounds and visual stimuli were spatially aligned but not when they were misaligned by as little as 6 degrees of visual angle (Beer et al., 2011a). The sharp spatial tuning of crossmodal facilitation suggests that it relies on brain structures with constrained receptive fields. Similarly, sounds facilitate visual perceptual learning only at visual field locations that were aligned with the sound source (Beer and Watanabe, 2009). MEG combined with source analysis revealed that sounds elicited responses in primary visual cortex at latencies (53 ms) that seem to be too early to be mediated by association cortex (Raij et al., 2010). Functional connectivity MRI showed that BOLD signals between early sensory cortices are highly correlated whereas limited correlation was observed in other brain regions (Eckert et al., 2008; Lewis and Noppeney, 2010; Werner and Noppeney, 2010a). Anatomical tracer studies reported direct axonal connections between auditory and early visual cortex in non-human primates (Falchier et al., 2002; Rockland and Ojima, 2003; Clavagnier et al., 2004; Bizley et al., 2007). Recently, we reported direct white matter tracts between the Heschl's region and the medial occipital cortex in humans (Beer et al., 2011b). The present tracking results with seeds in

Table 2 | Difference (speech minus body) BOLD responses per ROI.

Label	Unimodal				Bimodal					
	A_{S-B}		V_{S-B}		$cA_{S-B}V_{S-B}-(A_{S-B}+V_{S-B})$		$iA_{S-B}V_{B-S}-(A_{S-B}-V_{S-B})$		$(iAV-[A-V]_i)-(cAV-[A+V]_c)$	
	Mean	***(SE)***	**Mean**	***(SE)***	**Mean**	***(SE)***	**Mean**	***(SE)***	**Mean**	***(SE)***
H	0.03	*(0.04)*	0.03	*(0.02)*	−0.01	*(0.04)*	0.00	*(0.07)*	0.01	*(0.06)*
PT	0.03	*(0.04)*	0.02	*(0.03)*	−0.03	*(0.05)*	−0.07	*(0.06)*	−0.04	*(0.05)*
lSTG	**0.22*****	*(0.02)*	**0.07****	*(0.02)*	**−0.08***	*(0.03)*	0.04	*(0.04)*	**0.11***	*(0.05)*
aSTS	0.04	*(0.03)*	**0.06***	*(0.02)*	−0.07	*(0.05)*	0.08	*(0.05)*	**0.15***	*(0.07)*
pSTS	0.06	*(0.03)*	**0.12****	*(0.04)*	**−0.16****	*(0.05)*	0.12	*(0.07)*	**0.28****	*(0.09)*
msSTS	0.03	*(0.02)*	0.02	*(0.02)*	−0.09	*(0.05)*	0.01	*(0.05)*	0.10	*(0.07)*
EBA	−0.02	*(0.05)*	**−0.57*****	*(0.08)*	**0.25****	*(0.07)*	**−0.27****	*(0.07)*	**−0.53*****	*(0.11)*
lOTG	−0.02	*(0.03)*	0.07	*(0.04)*	0.00	*(0.06)*	0.06	*(0.05)*	0.06	*(0.06)*
lOTS	−0.07	*(0.05)*	−0.03	*(0.03)*	0.01	*(0.05)*	0.04	*(0.06)*	0.03	*(0.07)*
CTS	−0.04	*(0.04)*	**0.11***	*(0.04)*	0.02	*(0.08)*	0.14	*(0.07)*	0.12	*(0.08)*
FG	0.05	*(0.06)*	**−0.07***	*(0.03)*	−0.05	*(0.07)*	−0.05	*(0.11)*	−0.01	*(0.09)*
POS	0.01	*(0.03)*	−0.03	*(0.03)*	−0.02	*(0.05)*	−0.04	*(0.05)*	−0.01	*(0.05)*
CaS	−0.02	*(0.05)*	−0.02	*(0.07)*	0.03	*(0.12)*	−0.01	*(0.07)*	−0.04	*(0.15)*
OcPo	**−0.10*****	*(0.03)*	**0.16****	*(0.05)*	0.06	*(0.06)*	**0.13***	*(0.05)*	0.07	*(0.08)*
AIC	0.02	*(0.03)*	−0.03	*(0.03)*	−0.05	*(0.05)*	−0.04	*(0.05)*	0.01	*(0.06)*
CenS	−0.01	*(0.03)*	−0.02	*(0.03)*	0.02	*(0.04)*	0.02	*(0.06)*	0.00	*(0.07)*
dlFC	**0.09***	*(0.04)*	0.05	*(0.06)*	−0.16	*(0.08)*	0.05	*(0.10)*	0.22	*(0.14)*
vlFC	0.03	*(0.04)*	0.05	*(0.03)*	**−0.21*****	*(0.04)*	0.03	*(0.08)*	**0.24***	*(0.09)*

*Note: The table lists the differences in BOLD responses (percent signal change) to speech (S) minus body (B) stimuli for each ROI (label). Standard errors (SE) are shown in parenthesis. Significant differences are highlighted in bold (***$p < 0.001$; **$p < 0.01$; *$p < 0.05$, uncorrected). A_{S-B}, unimodal auditory; V_{S-B}, unimodal visual; $cA_{S-B}V_{S-B}$, congruent bimodal; $iA_{S-B}V_{B-S}$, incongruent bimodal. See also* ***Table 1*** *for abbreviations.*

Table 3 | ROI-to-ROI white-matter connectivity.

Label	PT	lSTG	aSTS	pSTS	msSTS	EBA	lOTG	lOTS	CTS	FG	POS	CaS	OcPo	AIC	CenS	dlFC	vlFC
H	**3.3**	**12.7**	**10.3**	**2.4**	1.1	0.4	**22.1**	**13.2**	**13.3**	0.6	**2.3**	1.3	**5.6**	<0.1	1.8	**2.3**	0.2
PT		**7.4**	**7.4**	**16.4**	**3.1**	0.2	**2.4**	**2.3**	**3.6**	<0.1	0.8	0.1	1.9	<0.1	**21.5**	**11.6**	0.2
lSTG			**40.4**	**5.9**	**2.4**	0.7	**8.0**	**17.7**	**4.4**	0.5	0.4	<0.1	1.7	<0.1	1.6	**3.2**	<0.1
aSTS				**6.5**	**8.0**	**2.7**	**24.7**	**35.3**	**8.1**	0.3	0.7	0.1	**5.5**	<0.1	**8.7**	**4.1**	<0.1
pSTS					**5.5**	0.2	0.7	0.2	0.5	<0.1	<0.1	<0.1	0.1	<0.1	**7.5**	0.6	<0.1
msSTS						**2.0**	0.2	0.2	0.3	<0.1	<0.1	<0.1	0.3	<0.1	<1.1	0.1	<0.1
EBA							0.4	**2.4**	<0.1	<0.1	0.1	<0.1	<0.1	<0.1	0.3	<0.1	<0.1
lOTG								**16.0**	**5.3**	0.7	0.4	**5.4**	1.5	**2.3**	0.1	<0.1	<0.1
lOTS									0.1	1.4	0.1	<0.1	0.2	**4.4**	<0.1	<0.1	<0.1
CTS										<0.1	1.0	**11.6**	**9.4**	0.4	<0.1	<0.1	<0.1
FG											<0.1	<0.1	<0.1	**3.2**	<0.1	<0.1	0.1
POS												**19.3**	0.9	0.1	0.1	<0.1	<0.1
CaS													**9.5**	<0.1	<0.1	<0.1	<0.1
OcPo														0.1	0.1	<0.1	<0.1
AIC															<0.1	<0.1	0.1
CenS																**14.0**	0.1
dlFC																	**7.2**

Note: The table lists group average ROI-to-ROI white-matter track connectivity indices (TCI) reflecting normalized track probabilities. The number of tracks connecting ROIs were divided by the product of voxels of each ROI. TCI-values are given in 10^{-8}. Large values reflect high probabilities of tracks between the voxels of each ROI. Values greater than 2 are highlighted in bold. See also ***Table 1*** *for abbreviations.*

the auditory cortex (H or PT) essentially replicated our previous finding on an independent sample. White matter tracks seeded in the Heschl's region terminated in anterior parts of the CaS, the OcPo, and the dorsal part of the POS. In addition, the present findings showed (whole-brain and ROI analysis) for the first time that these track termination areas also correspond to sub-parts of the visual cortex that were recruited by auditory processing.

TEMPORAL CORTEX

Consistent with previous findings (Calvert et al., 2000; Beauchamp et al., 2004b, 2008; Baumann and Greenlee, 2007; Hein et al., 2007; Noesselt et al., 2007; Van Atteveldt et al., 2010; Werner and Noppeney, 2010b; Meyer et al., 2011; Nath and Beauchamp, 2012; Plank et al., 2012) our functional MRI results showed that several brain regions of the temporal cortex were involved in auditory and visual object and action processing. Unisensory auditory processing primarily recruited superior parts of the temporal cortex whereas unisensory visual processing primarily involved inferior parts. These networks overlapped at the posterior part of the STS. It likely corresponds to the msSTS region as described in previous studies (Beauchamp et al., 2004b, 2008). No overlap of unimodal activity maps was observed in other STS regions. As indicated in the introduction, brain imaging techniques such as fMRI, EEG, or MEG detect responses of large neural ensembles and overlapping unimodal activity maps may simply result from separate but interspersed neural populations (Laurienti et al., 2005). However, multisensory integration may be inferred, if the brain responses to bimodal stimuli are not a linear summation of brain responses to unimodal stimuli ($AV \neq A + V$). Therefore, we performed a ROI analysis comparing BOLD signals to bimodal (AV) with the sum of unimodal responses ($A + V$). This ROI analysis showed subadditive responses within our putative msSTS region. Note, however, that limitations to the criterion of non-linearity ($AV - [A + V]$) as stated elsewhere (Gondan and Röder, 2006; Proctor and Meyer, 2011; Szameitat et al., 2011) must be considered. That is, responses to two trials are subtracted from responses to one trial. Accordingly, it may be argued that subadditivity merely results from double subtraction of BOLD responses that are common to all trials. However, we believe that this argument may not (or only partially) account for multisensory interaction effects in our study for several reasons: Common BOLD responses may reflect cognitive processes associated with the task (e.g., alertness or motor responses, etc.). However, these task-dependent responses were minimized in our paradigm as we used passive viewing/listening and the biological stimuli were task-irrelevant. Additionally, common BOLD responses may result from task-independent activity that is observed even in the resting brain (Gusnard and Raichle, 2001). Although it is disputed whether this resting state activity reflects a task-independent default brain state or just another task-specific activity (Morcom and Fletcher, 2007), it should be considered when examining non-linear interactions. Note that our paradigm contained baseline trials that were similar to the stimulus trials on most aspects (e.g., timing, task, etc.) except that they did not contain the target stimuli. Therefore, these baseline trials likely elicited task-dependent and task-independent BOLD responses that are common to all trials. Further note that in our study the MR signal for this baseline (A0V0) was subtracted from the MR signal of each stimulus condition (e.g., ASVS). Therefore, common BOLD responses were likely eliminated by this comparison prior to testing for interaction effects. Furthermore, unspecific BOLD responses should affect all or most brain areas in a similar manner. However, we found no subadditive responses in several low-level and high-level brain areas such as primary auditory cortex (H), some parts of visual cortex (CaS), and frontal cortex (see **Tables 1**, **2**). Furthermore, several brain areas (e.g., aSTS) also showed differential responses to congruent and incongruent bimodal stimuli. Congruency effects cannot be explained by double subtraction of common BOLD response components. Similarly, subadditive responses were also found to be selective for stimulus type (speech vs. body). These comparisons ($A_{S-B}V_{S-B} - [A_{S-B} + V_{S-B}]$) implicitly controlled for common response components. Another concern is whether multisensory interactions reflect perceptual or post-perceptual processing (McDonald et al., 2000). Our task (passive viewing) discouraged participants to adopt post-perceptual (e.g., decision, response, etc.) processes on the biological stimuli. However, observers may have engaged in such processes implicitly. If so, multisensory processing of biological stimuli should have interfered with the main (unimodal) detection task. No interference effects were observed.

Our finding of subadditive responses within the msSTS region is consistent with a number of EEG/MEG studies (Raij et al., 2000; Cappe et al., 2010) and electrophysiology studies in animals (Barraclough et al., 2005; Dahl et al., 2009) that observed similar subadditive responses in the STS. However, the nature of subadditive responses is still debated (Laurienti et al., 2005; Stein et al., 2009; Cappe et al., 2010). As subadditive responses in the superior colliculus were usually observed when auditory and visual stimuli were slightly mis-aligned, some researchers suggested that it reflects integration at the inhibitory surround receptive field of multisensory neurons (Stein and Stanford, 2008). However, subadditive responses in our and other studies (e.g., Barraclough et al., 2005) were observed even with spatially aligned auditory-visual stimuli. Another possibility is that crossmodal signals sharpen the tuning curve of object-encoding neurons (Raij et al., 2000). Alternatively, subadditive multisensory interactions may reflect converging auditory and visual input to the same neural (multimodal) representation of an object. As with salient stimuli—but not with degraded ones (Werner and Noppeney, 2010b)—either modality is sufficient to activate this representation, the response to bimodal stimuli reflects the maximum of unimodal responses (rather than their sum). Accordingly, subadditivity might reflect adaptation or saturation of a bimodal neural population (Weigelt et al., 2008) and its associated hemodynamic response (Toyoda et al., 2008).

Consistent with previous research (Meyer et al., 2011), most of the brain areas in the temporal cortex were recruited by both speech and body stimuli. However, a sub-region of that network (lSTG) showed additional selectivity for phonological sounds

(compared to body action sounds). A similar region responsive to sub-lexical speech sounds was observed before (Turkeltaub and Coslett, 2010). It likely corresponds to the lateral belt or parabelt of the auditory cortex (Woods et al., 2011). Although the lSTG is primarily auditory, it was also activated by visual stimuli and showed subadditive responses to congruent bimodal stimuli. Consistent with our finding, intracranial recordings from the lateral belt of rhesus monkeys showed multisensory modulation of facial and vocal signals (Ghazanfar et al., 2005). Furthermore, a region in the lateral occipital cortex showed stronger responses to visual body movements compared to lip movements. This region most likely corresponds to the EBA as described before (Peelen and Downing, 2007; Cziraki et al., 2010; Taylor and Downing, 2011). Our ROI analysis showed that the BOLD response in the EBA was modulated by concurrent auditory stimuli. An enhanced response (superadditive) was observed with incongruent bimodal stimuli and a reduced response (subadditive) was observed for congruent bimodal stimuli. Previous research has shown that sounds can affect visual processing of biological motion (Baart and Vroomen, 2010). However, to our knowledge this is the first demonstration of superadditive response enhancement by concurrent auditory-visual stimuli in the EBA. Interestingly, sounds modulated the response to visual stimuli although sounds alone did not elicit responses in the EBA. However, as shown by previous animal physiology, even subthreshold auditory connections can substantially influence visual processing (Clemo et al., 2008).

The primary motivation for our study was to examine the structural connectivity of the multisensory integration regions in the STC with the auditory cortex and other relevant brain areas. Our previous study demonstrated white matter tracts between auditory cortex and two distinct regions within the STS (aSTS, pSTS). The tracking results of the present study essentially replicated these previous findings by showing that tracks seeded in the Heschl's region and the planum temporale terminated in an anterior (aSTS) and posterior (pSTS) part of the STS. We were interested in the relationship of these structurally-defined regions with the msSTS region that was observed with functional MRI (Beauchamp et al., 2004b, 2008). We expected that the functionally-defined msSTS region overlapped with the STS regions that were connected with the auditory cortex via white matter tracks. Contrary to this hypothesis we found only limited overlap suggesting that msSTS is not directly connected with the core (H) of the auditory cortex. Additional tracking revealed that the msSTS also showed little direct connectivity with early visual brain areas including IOC. Instead msSTS seems to be primarily connected to other STS regions such as aSTS. Note that the borders of our msSTS region were defined by relatively liberal criteria. Therefore, it is unlikely that our msSTS region was too small to show sufficient overlap with terminations from auditory cortex tracks. Our tracking results further showed that areas lSTG and EBA—both regions selective for stimulus type (S vs. B) and modulated by sensory signals from its non-preferred modality—showed no direct white matter connections. Instead these two regions seemed to be connected via intermediate nodes such as aSTS or the IOC (e.g., lOTS).

Our finding of no direct connections between msSTS and auditory cortex seem to be inconsistent with previous connectivity research. For instance, functional connectivity based on fMRI suggested direct connections between primary auditory or visual cortex with msSTS (Noesselt et al., 2007; Werner and Noppeney, 2010a). However, functional connectivity does not necessarily require direct (monosynaptic) anatomical connections but instead may be mediated by polysynaptic connections (Damoiseaux and Greicius, 2009). Our results are partially consistent with tracer studies in animals. For instance, retrograde tracers injected into the STS of rhesus monkeys revealed that separate parts of the STS receive afferents from segregated areas of the STG (Seltzer and Pandya, 1994). Similarly, we found at least two regions within the STS (aSTS and pSTS) that were connected to separate regions of the STG and auditory cortex. However, direct axonal connections were also observed between polysensory STS and V1 in monkeys (Falchier et al., 2002). By contrast, we did not observe corresponding white matter tracks in humans. Tracer studies in rodents also observed sparse axonal projections from auditory cortex to other parts of the brain that were not detected in our study (Budinger and Scheich, 2009). Some of these differences might reflect species-specific characteristics. Alternatively, these differences may result from methodological limitations of fiber tracking (see below). For instance, DWI-based fiber tracking tends to neglect small fibers. In addition, small fiber tracts may have been obscured by our clustering procedure that we adopted in order to minimize false positives.

Our tracking results suggest that multisensory integration in the STS is not mediated by a single brain area but instead by a cascade of inter-connected brain areas located in the lateral temporal cortex (and IOC). Our ROI analysis of functional MRI data further suggests that aSTS, pSTS, and msSTS differ by the pattern of multisensory processing. Whereas activity in the aSTS region was sensitive to stimulus type (speech vs. body), no such sensitivity was observed in msSTS. Moreover, some regions (e.g., aSTS, lSTG) were predominantly involved in auditory processing, but auditory responses were modulated by visual signals. Both our connectivity findings and our functional results suggest that the STC is subdivided into several distinct regions and is best conceived as a multisensory network or complex rather than as a single region. This notion of a multisensory STC network may also help us to understand conventional fMRI findings that are difficult to accommodate with the notion of a single msSTS region. For instance, several brain imaging studies in humans found multiple distinct brain areas within the STS that may be segregated based on their multisensory integration patterns (Beauchamp et al., 2004a; Van Atteveldt et al., 2010; Werner and Noppeney, 2010b; Stevenson et al., 2011; Noesselt et al., 2012). Recent electrophysiological recordings from the STS of rhesus monkeys revealed separate patches within the STS that differ by the type of multisensory interactions (e.g., superadditive vs. subadditive) (Dahl et al., 2009). Our results elaborate these reports by showing that STS patches related to multisensory processing may also be characterized by distinct "connectivity fingerprints" (Behrens and Sporns, 2012).

INFERIOR OCCIPITO-TEMPORAL CORTEX

Our whole-brain analysis showed that visual speech and body motion also activated a relatively large cluster in the inferior and lateral occipito-temporal cortex. Our tracking results showed that the posterior part overlapped with track terminations from auditory cortex or the STC (H, lSTG, aSTS) (**Figure 6**). These track termination patterns further suggest subdivisions within this area: Tracks seeded in the Heschl's region primarily terminated in the collateral-transverse sulcus (CTS), tracks seeded in lSTG primarily terminated in the lOTG, and tracks seeded in aSTS terminated in the lOTS. No track terminations were observed in anterior parts of the inferior occipito-temporal cortex (primarily overlapping with the FG). Although all of these regions were primarily visual, two regions (CTS, lOTG) also showed BOLD responses to auditory stimuli. Moreover, bimodal stimuli elicited subadditive responses in the IOC. Similar multisensory responses in the IOC were observed in previous studies (Beauchamp et al., 2004b; Hocking and Price, 2008). Tracer studies in ferrets showed axonal connections between the auditory core and area 20 (corresponding to ventral/inferior occipital cortex) (Bizley et al., 2007). Moreover, axonal connections were observed between ventral preoccipital regions and the STS in rhesus monkeys (Yeterian and Pandya, 2010). Given that our tracking failed to find direct connections between STS areas and early visual cortex, it is possible that the IOC provides the major visual input to the STC regions.

FRONTAL CORTEX

Several studies examining multisensory integration observed multimodal responses in the anterior insula and ventrolateral frontal cortex - in addition to STS activity (Calvert et al., 2000; Beauchamp et al., 2004b; Taylor et al., 2006; Hein et al., 2007; Meyer et al., 2011; Nath and Beauchamp, 2012). For instance, familiar incongruent stimuli (animal sounds paired with animal pictures)—but not pairs of unfamiliar artificial stimuli—elicited larger responses in the inferior frontal cortex compared to that evoked by congruent stimulus pairs (Hein et al., 2007). Electrophysiological recordings in non-human primates showed that the ventrolateral frontal cortex contains a relatively large proportion of bimodal neurons that are responsive to faces and animal vocalizations (Romanski, 2007). We also observed a brain area in the ventrolateral frontal cortex that was activated by bimodal stimuli and that showed subadditive responses to congruent (but not to incongruent) bimodal speech stimuli (**Figure 2**, **Table 1**). Unfortunately, relatively little is known about the conditions leading to this vlFC activation. Some researchers suggested that it reflects cognitive demands associated with the multisensory task such as task difficulty and memory retrieval (Beauchamp et al., 2004b). Alternatively, it might reflect rehearsal processes or motor-related activity (Meyer et al., 2011; Wuerger et al., 2011). However, our results and the study by Hein and colleagues (Hein et al., 2007) showed vlFC modulation even with task-irrelevant multimodal stimuli. Therefore, its primary role could be multisensory binding of meaningful or communication-related (speech) signals (Taylor et al., 2006). Tracer studies in monkeys observed axonal projections from the anterior STG/STS to the ventrolateral frontal cortex and from posterior STG/STS to the dorsolateral frontal cortex (Romanski et al., 1999). Our tracking results showed connections between the STC and dorsal parts of the frontal cortex. However, we observed no detectable direct tracks between posterior parts of the STG (H or PT) or other parts of our STC network with the vlFC. Instead, vlFC connections to the multisensory STC network seem to be indirect (e.g., via dlFC).

LIMITATIONS

Although combining functional MRI and fiber tracking based on diffusion-weighted MRI provides relevant information that goes beyond the information provided by either method alone, both methods suffer from limitations that should be considered (Ramnani et al., 2004; Wakana et al., 2007; Damoiseaux and Greicius, 2009; Beer et al., 2011b). For instance, fiber tracks as determined by diffusion-weighted tensor imaging are best interpreted as white matter paths of least diffusion hindrance. Therefore, white matter tracks may result from axonal bundles (tracts) following these paths, but may also result from other sources. Moreover, tracks as determined by DWI bear no directional information as implied by the term "seed". Seed points in tractography do not specify the origin of the fibers (e.g., cell soma) but instead an anchor for the tracking algorithm. Seed points may equally well be the targets of fibers (e.g., synapses). Similarly, connectivity cascades (e.g., from auditory cortex to msSTS via aSTS) revealed from white matter tracks must be interpreted with caution. Tracks converging at a region may result from axonal fiber bundles targeting neural populations that are separate or only polysynaptically linked (rather than monosynaptically). Furthermore, the large volume effect and crossing fiber architectures may produce inaccurate tracking results. In order to address these problems, a probabilistic (rather than deterministic) tracking algorithm was applied. This algorithm describes each voxel by its probability of being connected to the seed region. Track probabilities are a descriptive measure and the validity of tracking inferences depends on the choice of an adequate track probability threshold (Morris et al., 2008). In light of these limitations, tracking results must be reproduced across subjects (Glasser and Rilling, 2008), across studies (Wakana et al., 2007) and—if possible—across fiber tracing methods (Catani et al., 2003) in order to gain credibility. In our study tracking was performed on a group of ten brains. Consistent estimates were found across brains for most tracks. Moreover, the H and PT track estimates essentially replicated our previous findings on an independent sample (Beer et al., 2011b).

CONCLUSION

In summary, functional MRI revealed a network of brain regions primarily within the temporal and occipital cortex involved in multisensory object and action processing including the msSTS region, a speech-selective lSTG region, and the EBA. Probabilistic tracking revealed white matter tracks between the auditory and the medial occipital cortex. However, brain areas involved in multisensory processing within the superior temporal and lateral occipital cortex showed little direct connectivity with primary sensory cortices. Instead these brain areas were connected to early sensory cortices via intermediate nodes of the STS and IOC. Our findings

suggest that combining structural connectivity and functional imaging reveals mechanisms related to multisensory integration that remain undetected by either technique alone.

ACKNOWLEDGMENTS

This work was supported in part by the Elitenetzwerk Bayern and the Deutsche Forschungsgemeinschaft (FOR 1075). We thank E.-R. Symeonidou and F. Weiß for assistance in conducting the experiment.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Integrative_Neuroscience/10.3389/fnint.2013.00005/abstract

REFERENCES

Angelaki, D. E., Gu, Y., and Deangelis, G. C. (2009). Multisensory integration: psychophysics, neurophysiology, and computation. *Curr. Opin. Neurobiol.* 19, 452–458.

Baart, M., and Vroomen, J. (2010). Do you see what you are hearing? Cross-modal effects of speech sounds on lipreading. *Neurosci. Lett.* 471, 100–103.

Barraclough, N. E., Xiao, D., Baker, C. I., Oram, M. W., and Perrett, D. I. (2005). Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *J. Cogn. Neurosci.* 17, 377–391.

Basser, P. J., Mattiello, J., and Lebihan, D. (1994). Estimation of the effective self-diffusion tensor from the NMR spin echo. *J. Magn. Reson. B* 103, 247–254.

Baumann, O., and Greenlee, M. W. (2007). Neural correlates of coherent audiovisual motion perception. *Cereb. Cortex* 17, 1433–1443.

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., and Martin, A. (2004a). Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nat. Neurosci.* 7, 1190–1192.

Beauchamp, M. S., Lee, K. E., Argall, B. D., and Martin, A. (2004b). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron* 41, 809–823.

Beauchamp, M. S., Yasar, N. E., Frye, R. E., and Ro, T. (2008). Touch, sound and vision in human superior temporal sulcus. *Neuroimage* 41, 1011–1020.

Beer, A. L., Batson, M. A., and Watanabe, T. (2011a). Multisensory perceptual learning reshapes both fast and slow mechanisms of crossmodal processing. *Cogn. Affect. Behav. Neurosci.* 11, 1–12.

Beer, A. L., Plank, T., and Greenlee, M. W. (2011b). Diffusion tensor imaging shows white matter tracts between human auditory and visual cortex. *Exp. Brain Res.* 213, 299–308.

Beer, A. L., and Watanabe, T. (2009). Specificity of auditory-guided visual perceptual learning suggests crossmodal plasticity in early visual cortex. *Exp. Brain Res.* 198, 353–361.

Beer, A. L., Watanabe, T., Ni, R., Sasaki, Y., and Andersen, G. J. (2009). 3D surface perception from motion involves a temporal-parietal network. *Eur. J. Neurosci.* 30, 703–713.

Behrens, T. E., Berg, H. J., Jbabdi, S., Rushworth, M. F., and Woolrich, M. W. (2007). Probabilistic diffusion tractography with multiple fibre orientations: what can we gain? *Neuroimage* 34, 144–155.

Behrens, T. E., and Sporns, O. (2012). Human connectomics. *Curr. Opin. Neurobiol.* 22, 144–153.

Behrens, T. E., Woolrich, M. W., Jenkinson, M., Johansen-Berg, H., Nunes, R. G., Clare, S., et al. (2003). Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magn. Reson. Med.* 50, 1077–1088.

Benoit, M. M., Raij, T., Lin, F. H., Jaaskelainen, I. P., and Stufflebeam, S. (2010). Primary and multisensory cortical activity is correlated with audiovisual percepts. *Hum. Brain Mapp.* 31, 526–538.

Bizley, J. K., Nodal, F. R., Bajo, V. M., Nelken, I., and King, A. J. (2007). Physiological and anatomical evidence for multisensory interactions in auditory cortex. *Cereb. Cortex* 17, 2172–2189.

Budinger, E., and Scheich, H. (2009). Anatomical connections suitable for the direct processing of neuronal information of different modalities via the rodent primary auditory cortex. *Hear. Res.* 258, 16–27.

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.

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.

Catani, M., Jones, D. K., Donato, R., and Ffytche, D. H. (2003). Occipito-temporal connections in the human brain. *Brain* 126, 2093–2107.

Clavagnier, S., Falchier, A., and Kennedy, H. (2004). Long-distance feedback projections to area V1: implications for multisensory integration, spatial awareness, and visual consciousness. *Cogn. Affect. Behav. Neurosci.* 4, 117–126.

Clemo, H. R., Sharma, G. K., Allman, B. L., and Meredith, M. A. (2008). Auditory projections to extrastriate visual cortex: connectional basis for multisensory processing in "unimodal" visual neurons. *Exp. Brain Res.* 191, 37–47.

Conturo, T. E., Lori, N. F., Cull, T. S., Akbudak, E., Snyder, A. Z., Shimony, J. S., et al. (1999). Tracking neuronal fiber pathways in the living human brain. *Proc. Natl. Acad. Sci. U.S.A.* 96, 10422–10427.

Cox, R. W., and Jesmanowicz, A. (1999). Real-time 3D image registration for functional MRI. *Magn. Reson. Med.* 42, 1014–1018.

Cziraki, C., Greenlee, M. W., and Kovacs, G. (2010). Neural correlates of high-level adaptation-related aftereffects. *J. Neurophysiol.* 103, 1410–1417.

Da Costa, S., Van Der Zwaag, W., Marques, J. P., Frackowiak, R. S., Clarke, S., and Saenz, M. (2011). Human primary auditory cortex follows the shape of Heschl's gyrus. *J. Neurosci.* 31, 14067–14075.

Dahl, C. D., Logothetis, N. K., and Kayser, C. (2009). Spatial organization of multisensory responses in temporal association cortex. *J. Neurosci.* 29, 11924–11932.

Dale, A. M., Fischl, B., and Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage* 9, 179–194.

Damoiseaux, J. S., and Greicius, M. D. (2009). Greater than the sum of its parts: a review of studies combining structural connectivity and resting-state functional connectivity. *Brain Struct. Funct.* 213, 525–533.

Desikan, R. S., Segonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* 31, 968–980.

Eckert, M. A., Kamdar, N. V., Chang, C. E., Beckmann, C. F., Greicius, M. D., and Menon, V. (2008). A cross-modal system linking primary auditory and visual cortices: evidence from intrinsic fMRI connectivity analysis. *Hum. Brain Mapp.* 29, 848–857.

Edmister, W. B., Talavage, T. M., Ledden, P. J., and Weisskoff, R. M. (1999). Improved auditory cortex imaging using clustered volume acquisitions. *Hum. Brain Mapp.* 7, 89–97.

Engelien, A., Yang, Y., Engelien, W., Zonana, J., Stern, E., and Silbersweig, D. A. (2002). Physiological mapping of human auditory cortices with a silent event-related fMRI technique. *Neuroimage* 16, 944–953.

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

Fischl, B., Liu, A., and Dale, A. M. (2001). Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE Trans. Med. Imaging* 20, 70–80.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* 33, 341–355.

Fischl, B., Salat, D. H., Van Der Kouwe, A. J., Makris, N., Segonne, F., Quinn, B. T., et al. (2004a). Sequence-independent segmentation of magnetic resonance images. *Neuroimage* 23(Suppl. 1), S69–S84.

Fischl, B., Van Der Kouwe, A., Destrieux, C., Halgren, E., Segonne, F., Salat, D. H., et al. (2004b). Automatically parcellating the human cerebral cortex. *Cereb. Cortex* 14, 11–22.

Fischl, B., Sereno, M. I., and Dale, A. M. (1999a). Cortical surface-based analysis. II: inflation, flattening, and a surface-based coordinate system. *Neuroimage* 9, 195–207.

Fischl, B., Sereno, M. I., Tootell, R. B., and Dale, A. M. (1999b).

High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum. Brain Mapp.* 8, 272–284.

Foxe, J. J., and Schroeder, C. E. (2005). The case for feedforward multisensory convergence during early cortical processing. *Neuroreport* 16, 419–423.

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.

Glasser, M. F., and Rilling, J. K. (2008). DTI tractography of the human brain's language pathways. *Cereb. Cortex* 18, 2471–2482.

Gondan, M., and Röder, B. (2006). A new method for detecting interactions between the senses in event-related potentials. *Brain Res.* 1073–1074, 389–397.

Gougoux, F., Zatorre, R. J., Lassonde, M., Voss, P., and Lepore, F. (2005). A functional neuroimaging study of sound localization: visual cortex activity predicts performance in early-blind individuals. *PLoS Biol.* 3:e27. 10.1371/journal.pbio.0030027

Gusnard, D. A., and Raichle, M. E. (2001). Searching for a baseline: functional imaging and the resting human brain. *Nat. Rev. Neurosci.* 2, 685–694.

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.

Hocking, J., and Price, C. J. (2008). The role of the posterior superior temporal sulcus in audiovisual processing. *Cereb. Cortex* 18, 2439–2449.

Jones, D. K., Simmons, A., Williams, S. C., and Horsfield, M. A. (1999). Non-invasive assessment of axonal fiber connectivity in the human brain via diffusion tensor MRI. *Magn. Reson. Med.* 42, 37–41.

Jones, D. K., Williams, S. C., Gasston, D., Horsfield, M. A., Simmons, A., and Howard, R. (2002). Isotropic resolution diffusion tensor imaging with whole brain acquisition in a clinically acceptable time. *Hum. Brain Mapp.* 15, 216–230.

Kovacs, G., Raabe, M., and Greenlee, M. W. (2008). Neural correlates of visually induced self-motion illusion in depth. *Cereb. Cortex* 18, 1779–1787.

Laurienti, P. J., Perrault, T. J., Stanford, T. R., Wallace, M. T., and Stein, B. E. (2005). On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. *Exp. Brain Res.* 166, 289–297.

Le Bihan, D., and Breton, E. (1985). Imagerie de diffusion *in vivo* par résonance magnétique nucléaire. *CR Acad. Sci. Paris* 201, 1109–1112.

Lee, S. K., Kim, D. I., Kim, J., Kim, D. J., Kim, H. D., Kim, D. S., et al. (2005). Diffusion-tensor MR imaging and fiber tractography: a new method of describing aberrant fiber connections in developmental CNS anomalies. *Radiographics* 25, 53–65. discussion: 66–68.

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.

McDonald, J. J., Teder-Sälejärvi, W. A., Di Russo, F., and Hillyard, S. A. (2003). Neural substrates of perceptual enhancement by cross-modal spatial attention. *Cogn. Affect. Behav. Neurosci.* 15, 10–19.

McDonald, J. J., Teder-Sälejärvi, W. A., and Hillyard, S. A. (2000). Involuntary orienting to sound improves visual perception. *Nature* 407, 906–908.

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748.

Meyer, G. F., Wuerger, S., and Greenlee, M. (2011). Interactions between auditory and visual semantic stimulus classes: evidence for common processing networks for speech and body actions. *J. Cogn. Neurosci.* 23, 2291–2308.

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.

Morcom, A. M., and Fletcher, P. C. (2007). Does the brain have a baseline? Why we should be resisting a rest. *Neuroimage* 37, 1073–1082.

Mori, S., Crain, B. J., Chacko, V. P., and Van Zijl, P. C. (1999). Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging. *Ann. Neurol.* 45, 265–269.

Morris, D. M., Embleton, K. V., and Parker, G. J. (2008). Probabilistic fibre tracking: differentiation of connections from chance events. *Neuroimage* 42, 1329–1339.

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.

Noesselt, T., Bergmann, D., Heinze, H.-J., Münte, T., and Spence, C. (2012). Coding of multisensory temporal patterns in human superior temporal sulcus. *Front. Integr. Neurosci.* 6:64. doi: 10.3389/fnint.2012.00064

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.

Peelen, M. V., and Downing, P. E. (2007). The neural basis of visual body perception. *Nat. Rev. Neurosci.* 8, 636–648.

Petkov, C. I., Kayser, C., Augath, M., and Logothetis, N. K. (2006). Functional imaging reveals numerous fields in the monkey auditory cortex. *PLoS Biol.* 4:e215. doi: 10.1371/journal.pbio.0040215

Plank, T., Rosengarth, K., Song, W., Ellermeier, W., and Greenlee, M. W. (2012). Neural correlates of audio-visual object recognition: effects of implicit spatial congruency. *Hum. Brain Mapp.* 33, 797–811.

Proctor, B. J., and Meyer, G. F. (2011). Electrophysiological correlates of facial configuration and audio-visual congruency: evidence that face processing is a visual rather than a multisensory task. *Exp. Brain Res.* 213, 203–211.

Raij, T., Ahveninen, J., Lin, F. H., Witzel, T., Jaaskelainen, 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.

Raij, T., Uutela, K., and Hari, R. (2000). Audiovisual integration of letters in the human brain. *Neuron* 28, 617–625.

Ramnani, N., Behrens, T. E. J., Penny, W., and Matthews, P. M. (2004). New approaches for exploring anatomical and functional connectivity in the human brain. *Biol. Psychiatry* 56, 613–619.

Rockland, K. S., and Ojima, H. (2003). Multisensory convergence in calcarine visual areas in macaque monkey. *Int. J. Psychophysiol.* 50, 19–26.

Röder, B., Stock, O., Bien, S., Neville, H., and Rösler, F. (2002). Speech processing activates visual cortex in congenitally blind humans. *Eur. J. Neurosci.* 16, 930–936.

Romanski, L. M. (2007). Representation and integration of auditory and visual stimuli in the primate ventral lateral prefrontal cortex. *Cereb. Cortex* 17(Suppl. 1), i61–i69.

Romanski, L. M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P. S., and Rauschecker, J. P. (1999). Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nat. Neurosci.* 2, 1131–1136.

Schröger, E., and Widmann, A. (1998). Speeded responses to audiovisual signal changes result from bimodal integration. *Psychophysiology* 35, 755–759.

Segonne, F., Dale, A. M., Busa, E., Glessner, M., Salat, D., Hahn, H. K., et al. (2004). A hybrid approach to the skull stripping problem in MRI. *Neuroimage* 22, 1060–1075.

Segonne, F., Pacheco, J., and Fischl, B. (2007). Geometrically accurate topology-correction of cortical surfaces using nonseparating loops. *IEEE Trans. Med. Imaging* 26, 518–529.

Seltzer, B., and Pandya, D. N. (1994). Parietal, temporal, and occipital projections to cortex of the superior temporal sulcus in the rhesus monkey: a retrograde tracer study. *J. Comp. Neurol.* 343, 445–463.

Sled, J. G., Zijdenbos, A. P., and Evans, A. C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans. Med. Imaging* 17, 87–97.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* 23(Suppl. 1), S208–S219.

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.

Stein, B. E., Stanford, T. R., Ramachandran, R., Perrault, T. J. Jr., and Rowland, B. A. (2009). Challenges in quantifying multisensory integration: alternative criteria, models, and inverse effectiveness. *Exp. Brain Res.* 198, 113–126.

Stevenson, R. A., Vanderklok, R. M., Pisoni, D. B., and James, T. W. (2011). Discrete neural substrates underlie complementary audiovisual speech integration processes. *Neuroimage* 55, 1339–1345.

Szameitat, A. J., Schubert, T., and Müller, H. J. (2011). How to test for dual-task-specific effects in brain imaging studies—an evaluation of potential analysis methods. *Neuroimage* 54, 1765–1773.

Taylor, J. C., and Downing, P. E. (2011). Division of labor between lateral

and ventral extrastriate representations of faces, bodies, and objects. *J. Cogn. Neurosci.* 23, 4122–4137.

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.

Tootell, R. B. H., Reppas, J. B., Kwong, K. K., Malach, R., Born, R. T., Brady, T. J., et al. (1995). Functional analysis of human MT and related visual cortical areas using magnetic resonance imaging. *J. Neurosci.* 15, 3215–3230.

Toyoda, H., Kashikura, K., Okada, T., Nakashita, S., Honda, M., Yonekura, Y., et al. (2008). Source of nonlinearity of the BOLD response revealed by simultaneous fMRI and NIRS. *Neuroimage* 39, 997–1013.

Turkeltaub, P. E., and Coslett, H. B. (2010). Localization of sublexical speech perception components. *Brain Lang.* 114, 1–15.

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

Wakana, S., Caprihan, A., Panzenboeck, M. M., Fallon, J. H., Perry, M., Gollub, R. L., et al. (2007). Reproducibility of quantitative tractography methods applied to cerebral white matter. *Neuroimage* 36, 630–644.

Weeks, R., Horwitz, B., Aziz-Sultan, A., Tian, B., Wessinger, C. M., Cohen, L. G., et al. (2000). A positron emission tomographic study of auditory localization in the congenitally blind. *J. Neurosci.* 20, 2664–2672.

Weigelt, S., Muckli, L., and Kohler, A. (2008). Functional magnetic resonance adaptation in visual neuroscience. *Rev. Neurosci.* 19, 363–380.

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.

Werner, S., and Noppeney, U. (2010b). Superadditive responses in superior temporal sulcus predict audiovisual benefits in object categorization. *Cereb. Cortex* 20, 1829–1842.

Woods, D. L., Herron, T. J., Cate, A. D., Kang, X., and Yund, E. W. (2011). Phonological processing in human auditory cortical fields. *Front. Hum. Neurosci.* 5:42. doi: 10.3389/fnhum.2011.00042

Woolrich, M. W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., et al. (2009). Bayesian analysis of neuroimaging data in FSL. *Neuroimage* 45, S173–S186.

Wuerger, S. M., Parkes, L., Lewis, P. A., Crocker-Buque, A., Rutschmann, R., and Meyer, G. F. (2011). Premotor cortex is sensitive to auditory-visual congruence for biological motion. *J. Cogn. Neurosci.* 24, 575–587.

Yeterian, E. H., and Pandya, D. N. (2010). Fiber pathways and cortical connections of preoccipital areas in rhesus monkeys. *J. Comp. Neurol.* 518, 3725–3751.

Zihl, J., von Cramon, D., and Mai, N. (1983). Selective disturbance of movement vision after bilateral brain damage. *Brain* 106(Pt 2), 313–340.

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: 31 May 2012; accepted: 25 January 2013; published online: 13 February 2013.

Citation: Beer AL, Plank T, Meyer G and Greenlee MW (2013) Combined diffusion-weighted and functional magnetic resonance imaging reveals a temporal-occipital network involved in auditory-visual object processing. Front. Integr. Neurosci. 7:5. doi: 10.3389/fnint.2013.00005

frontiers in
INTEGRATIVE NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 25 August 2012
doi: 10.3389/fnint.2012.00063

Temporal-order judgment of visual and auditory stimuli: modulations in situations with and without stimulus discrimination

Elisabeth Hendrich[1]*, Tilo Strobach[1,2], Martin Buss[3], Hermann J. Müller[1] and Torsten Schubert[2]

[1] *Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany*
[2] *Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany*
[3] *Institute of Automatic Control Engineering, Technische Universität München, Munich, Germany*

Edited by:
Zhuanghua Shi, Ludwig-Maximilians-Universität München, Germany

Reviewed by:
Toemme Noesselt, Otto-von-Guericke-University, Germany
Antonio Pereira, Federal University of Rio Grande do Norte, Brazil

***Correspondence:**
Elisabeth Hendrich, Department of Psychology, Ludwig-Maximilians-Universität München, Leopoldstraße 13, D-80802 Munich, Germany.
e-mail: elisabeth.hendrich@psy.lmu.de

Temporal-order judgment (TOJ) tasks are an important paradigm to investigate processing times of information in different modalities. There are a lot of studies on how temporal order decisions can be influenced by stimuli characteristics. However, so far it has not been investigated whether the addition of a choice reaction time (RT) task has an influence on TOJ. Moreover, it is not known when during processing the decision about the temporal order of two stimuli is made. We investigated the first of these two questions by comparing a regular TOJ task with a dual task (DT). In both tasks, we manipulated different processing stages to investigate whether the manipulations have an influence on TOJ and to determine thereby the time of processing at which the decision about temporal order is made. The results show that the addition of a choice RT task does have an influence on the TOJ, but the influence seems to be linked to the kind of manipulation of the processing stages that is used. The results of the manipulations indicate that the temporal order decision in the DT paradigm is made after perceptual processing of the stimuli.

Keywords: dual task, choice RT task, temporal order judgments, comparison, time of temporal order decision

INTRODUCTION

To form an adequate representation of the environment, we often have to integrate visual and auditory information into a multisensory representation (Stein and Meredith, 1994; King, 2005; Spence, 2007). An important factor influencing this integration is the different processing duration of visual and auditory perceptual processing. Physically, an auditory signal coming from a certain source is slower in reaching the observer than a corresponding visual signal (Sugita and Suzuki, 2003). This is somehow compensated for by faster sound transduction than light transduction and by faster neural processing of auditory information in the human neural system (King, 2005).

Researchers have developed a number of paradigms that are directed at investigating the principles of information processing in different information modalities (i.e., perceptual latencies). Temporal-order judgment (TOJ) is one of these paradigms and it is commonly used for comparing perceptual latencies in different information modalities (e.g., Spence et al., 2001; Miller and Schwarz, 2006; Cardoso-Leite et al., 2007; Shi et al., 2008; Boenke et al., 2009). In a typical TOJ task, two stimuli are presented with varying stimulus onset asynchronies (SOAs) and participants are asked to indicate the temporal order of the two stimuli. The point in time at which the two stimuli are rated as presented at the same time, is called the Point of Subjective Simultaneity (PSS). Several factors have been identified to have an effect on the perception of temporal order. One of them is the modality of the stimuli (e.g., Hirsh and Sherrick, 1961; Rutschmann and Link, 1964; Roufs, 1974; Jaśkowski et al., 1990; Spence et al., 2001). Many authors consent that an auditory stimulus has to be presented after a visual stimulus to be perceived as simultaneous in a TOJ task (Hirsh and Sherrick, 1961; Dinnerstein and Zlotogura, 1968; Jaśkowski et al., 1990; Zampini et al., 2003; Keetels and Vroomen, 2005; Van Eijk et al., 2008; Boenke et al., 2009; but see, e.g., Rutschmann and Link, 1964). This effect is ascribed to the faster sound transduction in the human ear than light transduction in the human eye (King, 2005; Arrighi et al., 2006). Additionally neural transmission times are shorter in the auditory system than in the visual system (King, 2005). Therefore, the onset of the auditory stimulus has to be delayed compared to the onset of a visual stimulus if both ought to be perceived as simultaneous. Times of reported auditory delay vary from 5 ms (Hirsh and Sherrick, 1961) to 71 ms (Dinnerstein and Zlotogura, 1968). However, there are some studies that report the opposite effect of visual delays (e.g., Rutschmann and Link, 1964). According to Boenke et al. (2009), this might be explained by higher intensity of the visual stimuli and/or lower intensities of the auditory stimuli. Stimulus intensity is therefore another factor that seems to influence TOJ (Neumann, 1982; Boenke et al., 2009).

However, what has, to our knowledge, not been studied so far, is whether the particular processing requirements that are associated with the auditory and the visual stimulus affect the TOJs in addition to temporal delay and intensity characteristics. A number of studies have advocated for a close relation between perception and action planning (e.g., Deubel and Schneider,

1996; Deubel et al., 1998; Witt et al., 2005; Zwickel et al., 2007; Witt and Proffitt, 2008), pointing to mutual dependencies between processes involved in the early perception of sensory information and in processes planning actions with the perceived sensory information. Witt et al. (2005), for instance, found that people perceive a target that is just beyond arm's reach as closer when they intend to reach it with a tool compared to when they plan to reach it without the tool. Deubel and colleagues also showed the close connection between intended actions and perceptual processing. For instance, Deubel and Schneider (1996) showed that the planning of an action has an early influence on perceptual processes. In a dual-task (DT) paradigm, they asked participants to plan a saccade to a specific location. Additionally, the participants had to discriminate between the symbols "E" and "mirror-E," either at the target location of the saccade or at an adjacent location. Stimulus discrimination performance was best, if the two tasks, i.e., planning a saccade and discriminating, involved the same stimulus at one location and dropped if they involved different stimuli at different locations. Deubel et al. (1998) showed similar findings for the planning of a manual reaching task. If action planning requirements influence early perceptual processes, e.g., by influencing the allocation of attentional resources to the processed stimuli, then this may have an additional effect on the processing speed of the perceptual stimuli. It is well known that attention facilitates the detection (e.g., Posner, 1980) and the identification of visual stimuli (e.g., Desimone and Duncan, 1995) and this may influence subsequent judgments about stimulus features, e.g., the temporal order of their presentation.

A way to investigate this research question is to add the requirement to carry out a discriminative response on the processed visual and auditory stimuli in the context of a TOJ task. This can be done by administering different types of visual and auditory stimuli, which have to be discriminated in order to perform a choice reaction in addition to the TOJ. The question of TOJ under the condition of an additional choice requirement is of relevance because visual and auditory information often not only have to be noticed but also require an appropriate reaction from the observer in a laboratory context as well as in a real world environment. In this situation it is of special interest, whether such an enforced choice reaction will or will not have an effect on the order with which the auditory and the visual signal are perceived. From a theoretical perspective, an answer on this question may also provide valuable insights into our understanding of the processing architectures of TOJ tasks and DT situations with variable task orders. It should be mentioned at this point, that simultaneity judgments (SJ) are an alternative method to investigate time characteristics of stimulus processing and this method can show results different from TOJ results (see Van Eijk et al., 2008). For SJs, participants are asked to indicate whether two stimuli are present at the same time or not. We choose TOJ instead of SJs, because the determination of temporal order of stimuli is more complex and therefore more sensitive to detect possible differences between different conditions within TOJ (e.g., high or low stimulus contrast) and between TOJ and DT. Compared to that, indicating whether two stimuli are presented at the same time or not, is relatively easy. Therefore, we considered TOJs to be the more appropriate paradigm to investigate timing of stimulus processing in contrast to SJs.

Several authors conducted experiments on DT with variable intervals between stimuli (SOA) and unpredictable task order, which necessitated an additional judgment of the temporal order of the two stimuli (e.g., De Jong, 1995; Luria and Meiran, 2003; Sigman and Dehaene, 2006; Szameitat et al., 2006). In these DT situations participants are presented with stimuli in different modalities and they have to perform a choice reaction on the stimuli. Most often participants are required to perform the two tasks in the order of the stimulus presentation. While this task basically requires first TOJ about the presentation of stimuli in different modalities, it subsequently requires that different visual and auditory stimuli need to be distinguished, related to a pre-specified response category and a subsequent motor response needs to be selected and executed (Umiltà et al., 1992; De Jong, 1995; Sigman and Dehaene, 2006).

Surprisingly, while TOJs play a role in research on DT situations with unpredictable SOA, only one study, i.e., De Jong (1995), compared the response order results of DTs to the response order in typical TOJ tasks. In particular, De Jong realized this comparison when he used TOJ as a control condition for a DT experiment with varying task order. In the DT task, an auditory and a visual choice RT task were presented and participants were either asked to respond in the order the two stimuli were presented (forced-order) or they received no specific instruction regarding response order (free-order). To certify that participants were able to judge the order of the two stimuli correctly, De Jong added a control condition in which the participants were exclusively asked to judge the order of the stimulus presentation without conducting discriminative choice reactions. The results of this study showed that, in a substantial number of trials, the participants responded to the stimuli in the opposite order of their presentation ("response reversals") in both instruction-conditions. In the control condition, in which participants judged only the order of the stimulus presentation, the number of response reversals was much lower than in the experimental condition. A reason for the higher number of response reversals in the experimental condition (DT) compared to the number of response reversals in the control condition (TOJ-type) could be the additional requirement to make speeded responses to the stimuli.

However, De Jong (1995) discarded this idea of an impact of the additional speeded responses on temporal order decision as unlikely and did not pursue it any further. This decision was based on findings from a study of Sternberg et al. (1971), who combined a TOJ task with an RT task. According to De Jong, the results reported by Sternberg et al. suggested that "such interfering effects (caused by an additional RT task) were probably quite minimal" (p. 15). However, the tasks used by Sternberg et al. were quite different from the ones De Jong had used. Sternberg et al. had presented an auditory and a cutaneous stimulus. A cue before one of the two stimuli told the participants that they had to pull a lever as quickly as possible in reaction to the cued stimulus and after that judge the order of the two stimuli by pressing a corresponding button. Thus, participants had to do first a reaction-time task to one of the stimuli and then, in a second step, judge the order of

the stimuli. In the study of De Jong, however, participants had to do two choice reaction tasks plus a TOJ and respond to the stimuli in the order of their presentation.

In our opinion, the two studies, i.e., De Jong (1995) and Sternberg et al. (1971), focused on two task conditions, which exposed highly different task demands for the participants and their results are not comparable with each other. Therefore, we think that the question whether an additional choice-RT task on the presented stimuli has an influence on TOJs has not yet been addressed sufficiently and, aim to shed new light on it. Based on De Jong's results (De Jong, 1995), we expect that an additional identification task has an influence on TOJs. The study of De Jong, however, did not provide any details about the specific processing architecture of TOJ with and without selecting responses following stimulus identifications.

A second aim of this study was to investigate, when, precisely, the judgment of temporal order takes place during task processing in pure TOJ tasks and in DT situations with unpredictable task order. As already noted, task processing in a RT task is assumed to consist of several consecutive processes or stages (Sternberg, 1969; Sanders, 1980, 1990): perception, response selection, and response execution. At some point in time during the information processing chains of two tasks the temporal order of the two stimuli has to be decided. First empirical hints about the location of the order decision in DT situations come from Sigman and Dehaene (2006). These authors investigated whether the prolongation of certain processing stages has an influence on the processing order of the two component tasks in a DT situation. To analyze this question, they used a DT paradigm with unpredictable stimulus order, in which participants were asked to respond to an auditory and a visual stimulus in the order of their presentation. Sigman and Dehaene found that the prolongation of the perception stage of the visual task (if it was presented first) led to an increase in response reversals, while the prolongation of the response-selection stage of the visual task did not. To explain their results, they assumed that task order is decided after the perceptual processes of the first presented stimulus, in a definite executive control processing stage (see also Lien et al., 2003; Liepelt et al., 2011; Strobach et al., in press for alternative ideas on such control processes).

To investigate the precise location of the task order decision in the TOJ paradigm and in the DT paradigm, we manipulated the length of the first two task processing stages (perception and response-selection stage) in the DT paradigm and the first stage, the perception stage, in the TOJ paradigm. We tried to do so in a more direct way, in comparison to Sigman and Dehaene (2006), and compared the effects on response order in DT with effects in TOJ where feasible, as there is no response-selection stage in a TOJ task which can be manipulated. For the particular case of task order decision in DT situations, the manipulations could lead to three different outcomes: if the TOJ occurs at the very beginning of task processing, then both manipulations, the manipulation of the perception stage and the manipulation of the response-selection stage (resulting in manipulations of the stage latencies), should fail to show an effect on response order. This is because manipulations of perceptual and response-selection latencies do not affect the outcome of a TOJ process located before these manipulated stages. If the order judgment happens after perception, but before the response-selection stage, then only the manipulation of the perception stage should have an influence on response order. This hypothesis results from the assumption that manipulations of perceptual stage latencies but not latencies of response selections have an effect on the outcome of a TOJ process located between the former and the latter manipulated stage. If the judgment about the temporal order of the two stimuli occurs after the response-selection stage, then both manipulations have an effect on response order. This is because we assume that manipulations of perceptual and response-selection latencies affect the outcome of a TOJ process located after these manipulated stages.

For the particular case of order decision in the TOJ paradigm, the manipulation of the perception stage is decisive: if this manipulation provides an effect on the order decision, then the decision is located after the perception stage.

In sum, the first aim of this study was to investigate whether TOJ is principally affected by the requirement to perform a discriminative response on the perceived stimuli in addition to the TOJ. Furthermore, we investigated the point in time at which the decision about the temporal order of the two stimuli is made in the condition with additional identification of the stimulus. For that purpose, we manipulated the length of the perception stage (Experiment 1) and of the response-selection stage (Experiment 2) of one of the two tasks.

EXPERIMENT 1: MANIPULATION OF THE PERCEPTION STAGE

In Experiment 1, we compared "pure" TOJ in a TOJ paradigm with order judgments in a DT paradigm requiring the additional identification of the stimuli plus a response selection in the component tasks. We presented an auditory and a visual task in varying order. For the visual task, one of three numbers was shown randomly. For the auditory task, one of three different tones was presented. Usually, in the TOJ paradigm participants have to press one of two keys indicating whether stimulus A or stimulus B was presented first. In order to minimize differences between the two tasks, participants in the TOJ-task condition judged the order of the two stimuli by pressing two buttons in the corresponding order of stimulus presentation. In the DT condition, participants had to identify the stimuli and press two corresponding buttons (1 out of 3 for the visual task and 1 out of 3 for the auditory task) in the order of the stimulus presentation. The length of the perception stage of the visual task was manipulated by weak or strong contrast of the presented numbers and their background.

METHOD

Participants

Eighteen participants (12 female) took part in the experiment. The participants were all students of the LMU Munich, who received course credit or payment (8 Euro/hour) for their participation. The average age was 25.0 years (SD = 3.0 years). All subjects were right-handed and reported normal or corrected-to-normal vision and hearing.

Apparatus and stimuli

The participants were tested individually in a sound-attenuated, darkened room. They sat in front of a CRT monitor (85 Hz) at a distance of about 60 cm and wore headphones. Responses were given on the QWERT-keyboard. The experimental code was written in Presentation (Version 14.4 02.24.10) and run on a Dell Optiplex GX620 with Windows XP Professional.

Three digits were presented as the visual stimuli: "2," "5," or "9." The numbers were either presented with strong contrast (white font color, 55 cd/m^2, against dark-gray background) or with weak contrast (gray font color, 0.09 cd/m^2, against dark-gray background). Each number subtended a visual angle of 1° × 1.5° (1 cm × 1.5 cm). Dark gray background (0.11 cd/m^2) was used instead of black background to minimize visual after-effects. The auditory stimuli were three sine-wave tones with frequencies of 250, 500, and 1000 Hz and a volume of 58 decibel. They were presented via headphones. Both types of stimuli, visual, and auditory, were presented for 200 ms each.

Design and procedure

The experimental design formed a 2 × 2× 7 factorial model with task condition (TOJ vs. DT), contrast of the visual stimulus (weak vs. strong contrast) and SOA (−400 ms, −120 ms, −60 ms, 0 ms, +60 ms, +120 ms, +400 ms) as within-subjects factors.

In the TOJ-task condition, participants completed 20 practice trials before starting with the main experiment. Each trial started with the presentation of a fixation point in the centre of the screen for 500 ms. The fixation point was followed by a blank screen for 600 ms, then the first stimulus presentation (i.e., either visual or auditory) and, after a variable SOA, the second presentation (i.e., auditory or visual, respectively). Participants then responded by pressing the "c"-key to the auditory stimulus and the ","-key to the visual stimulus in the order of the perceived stimulus presentation order. Each trial had a constant length of 4,500 ms.

In the DT condition, participants completed two practice blocks for the single tasks (15 trials each) and one practice block with both tasks (20 trials). The procedure of trials with both tasks was identical to the TOJ-task condition. In contrast to the TOJ-task condition, however, participants responded by pressing "y," "x," or "c"-key for low, middle, and high tone (i.e., auditory task) and ","; "."; or "-"-key for 2, 5, and 9 (i.e., visual task) in the DT-part, respectively. Feedback on the correctness of the responses was given in the practice blocks.

All possible combinations of SOA (±400 ms, ±120 ms, ±60 ms, 0 ms), visual stimuli (2, 5, 9), and auditory stimuli (250, 500, 1000 Hz) resulted in 63 different trial types. The order of these trial types varied randomly for each participant in each block. For each task condition (TOJ and DT), 3 blocks with strong-contrast and 3 blocks with weak-contrast visual stimuli were presented in alternating order; therefore, 378 trials were presented for each task condition. Half of the participants started with a strong-contrast block, the other half with a weak-contrast block. Response order and PSS were measured as dependent variables.

The complete experimental session lasted approximately 90 min. It consisted of two parts, TOJ and DT, presented in this order for every participant. Participants had the opportunity to have a short break between the two parts.

RESULTS

Temporal-order judgment

To compare the TOJs between the TOJ-task condition and the DT condition, we calculated the percentage of trials in which the tone was reported as first for each task condition, contrast condition, and SOA. The percentage of trials in which the tone was reported as first was submitted to an analysis of variance (ANOVA) including the within-subjects factors task condition (TOJ vs. DT), contrast (strong vs. weak contrast), and SOA (±400 ms, ±120 ms, ±60 ms, 0 ms).

As illustrated in **Figure 1**, participants tried to follow the instruction to respond in the order of presentation, which is reflected in a significant effect of SOA, $F_{(6,\ 102)} = 145.119$, $p < 0.01$. As can be seen in **Figure 1**, participants had a higher number of tone-first responses under conditions with positive SOA, i.e., conditions in which the tone was presented before the visual stimulus compared to conditions with negative SOA, i.e., conditions in which the tone was presented second.

We also found a significant difference between the task conditions, $F_{(1,\ 17)} = 6.029$, $p < 0.05$. Participants responded to the auditory task first significantly more often in the DT condition ($m = 55.8\%$) than in the TOJ-task condition ($m = 51.1\%$). The ANOVA also revealed a significant interaction of task condition and SOA, $F_{(6,\ 102)} = 2.276$, $p < 0.05$. Further t-tests revealed significant differences between the two task conditions at SOAs −120 ms [$t_{(17)} = -2.184$, $p < 0.05$], −60 ms [$t_{(17)} = -2.130$, $p < 0.05$], and 0 ms [$t_{(17)} = -2.613$, $p < 0.05$]. In the DT condition, participants reported the tone task as first more often than in the TOJ-task condition.

The manipulation of the contrast led to a significant effect, $F_{(1,\ 17)} = 22.962$, $p < 0.01$. Participants reported the tone as first more often in the condition with weak contrast ($m = 57.8\%$) than in the strong-contrast condition ($m = 49.1\%$). There was also a significant interaction of contrast and SOA,

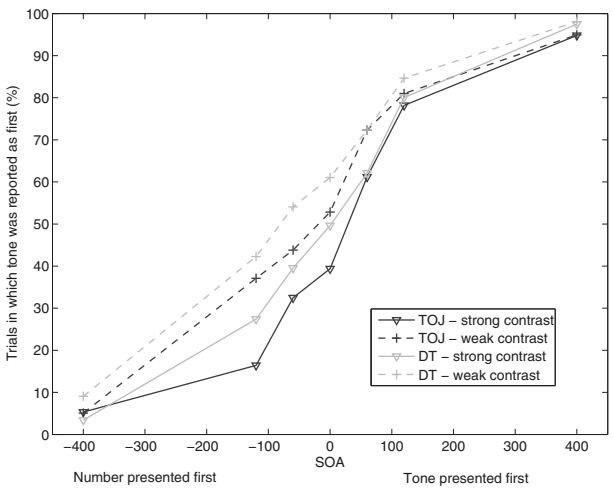


FIGURE 1 | Proportion of trials in which the participants reported the tone task as first in the conditions of temporal-order judgment (TOJ) and dual task (DT) in Experiment 1, each with two contrast conditions (strong and weak contrast).

$F_{(6,\ 102)} = 7.052$, $p < 0.01$. The percentage of trials in which the participants reported the tone task as first was higher for weak contrast at the SOA levels −400 ms [$t_{(17)} = -2.536$, $p < 0.05$], −120 ms [$t_{(17)} = -4.367$, $p < 0.01$], −60 ms [$t_{(17)} = -4.473$, $p < 0.01$], 0 ms [$t_{(17)} = -3.226$, $p < 0.01$], and 60 ms [$t_{(17)} = -3.578$, $p < 0.01$]. There were no further interactions.

Point of subjective simultaneity (PSS)

The PSS denotes the SOA, at which the participants report the tone as first in 50% of the trials. It was calculated by fitting logistic regression functions to the data of each participant. For each condition, the PSS was calculated by estimating the 50% performance point on the fitted logistic function (Treutwein and Strasburger, 1999). As illustrated in **Figure 2**, for the TOJ-task condition, the mean PSSs were 26.3 ms for the strong-contrast condition and −33.0 ms for the weak-contrast condition. The results indicate that, in the strong-contrast condition, the number digit had to be presented 26 ms after the tone to be perceived as simultaneous. In the weak-contrast condition, the number had to be presented 33 ms before the tone task to be perceived as simultaneous. In the DT condition, the mean PSSs were −4.8 ms for the strong contrast condition and −71.5 ms for the weak contrast condition. In both contrast conditions, the number had to be presented before the tone to be perceived as simultaneous. PSSs were submitted to a 2 × 2 ANOVA with task condition (TOJ vs. DT) and contrast (strong vs. weak contrast) as within-subjects-factors. The ANOVA revealed a significant effect of task condition, $F_{(1,\ 17)} = 4.451$, $p < 0.05$. The PSS was significantly more negative in the DT condition ($m = -38.2$ ms) than in the TOJ-task condition ($m = -3.4$ ms). The factor contrast did also show a significant effect, $F_{(1,\ 17)} = 19.878$, $p < 0.01$. PSS values were significantly more negative in the condition with weak contrast ($m = -52.2$ ms) than in the strong-contrast condition ($m = 10.7$ ms). The interaction of task condition and contrast was not significant, $F_{(1,\ 17)} < 1$.

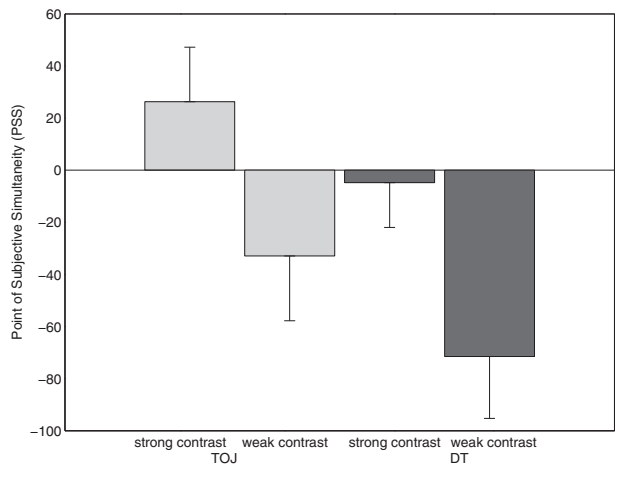


FIGURE 2 | Points of subjective simultaneity (PSS) for the two task conditions temporal-order judgment (TOJ) and dual task (DT) in Experiment 1, each measured in two contrast conditions, strong and weak contrast.

DISCUSSION

In a TOJ-task, participants have to indicate the order of two stimuli which are presented with varying SOAs. In Experiment 1 we investigated whether the requirement to identify and discriminate the stimuli plus subsequent response selection processes in an auditory-visual DT have an effect on the TOJ. Additionally, the perception stage of the visual task was manipulated to localize the processing stage at which the decision about the temporal order is made. In fact, the present contrast manipulation aims at testing whether the judgment is made before or after perceptual processes associated with the contrast manipulation.

The results of Experiment 1 show that TOJ, as measured by the response order, is influenced by the kind of task requirement the participants have to complete. When participants have to judge the temporal order of two stimuli and, additionally, have to identify the stimuli, then they report the auditory stimulus significantly more often as first compared to when they do not have to identify the specific auditory stimulus. This effect was especially pronounced for trials in which the visual stimulus was presented first. A possible reason for this observation might be that the perceptual stage of the visual task was manipulated in Experiment 1. Participants might have considered the visual task to be the more difficult one and therefore might have preferred to do the easier auditory task first. We will return to this effect in Experiment 2 and in the General discussion.

Furthermore, we found an effect of task condition on the PSS, which fits to the results of the TOJ data.

For the issue of the localization of the point in time at which the temporal order decision is made, the results of the manipulation of the visual perception stage are important. The results showed that the contrast manipulation of the visual stimulus had an effect in both task conditions, the TOJ- and DT condition. In the condition with weak contrast, the visual task had to be presented earlier than in the condition with strong contrast to be perceived as simultaneous. Importantly, concerning the point in time at which the TOJ occurs in the DT- and TOJ-task condition, we assume that it must happen after the perception stage, because the manipulation of the perception stage influences the TOJs in both the TOJ- and the DT condition.

This effect was especially pronounced for trials in which the visual task was presented first. These results are in line with Boenke et al. (2009), who argued that relative stimulus intensity influences the PSS in TOJ-tasks. More specifically, they claimed that studies which found that the visual stimulus has to be presented before the auditory stimulus to be perceived as simultaneous used auditory stimuli of higher intensity and/or visual stimuli of lower intensity than studies who found the opposite. This indicates that higher intensity of the visual stimulus leads to a shift of the PSS. The results of the TOJ-task condition in Experiment 1 support this idea. In the condition with strong contrast of the visual stimulus, the PSS is positive, which means that the auditory stimulus has to be presented before the visual stimulus to be perceived as simultaneous. In the condition with low visual intensity, however, the PSS is negative; thus, there is a requirement to present this visual before the auditory stimulus to generate a percept of simultaneous stimulus presentation.

In the DT condition, we also found a shift of the PSS from strong to weak contrast. The PSS in the condition with low stimulus intensity was more negative than the PSS in the condition with high stimulus intensity. This means, in the condition with low intensity the visual stimulus has to be presented even longer before the auditory stimulus than in the high intensity condition to be perceived as simultaneous.

EXPERIMENT 2: MANIPULATION OF THE RESPONSE-SELECTION STAGE

The results of Experiment 1 showed that the manipulation of the perception stage has an influence on the judgment of temporal order. Thus, the judgment must be localized after the perception stage because otherwise we should not have found an effect of the duration of the visual stimulus on judgment order. In Experiment 2, we aimed to further specify the temporal location of the order judgments within the particular task processing architecture of the current task situations. While the processing chain in TOJ tasks is mostly restricted to the perception of the stimuli, the comparison of their presentation times, and the programming of motor responses, the component tasks in the DT situation involve an additional response-selection stage each. In Experiment 2, we aimed to assess whether the TOJ in the DT situation is located before the response-selection stage or not. Note that the findings of Experiment 1 leave open that question because they localized the order judgment only non-specifically as later than the perception stage. For that purpose, in Experiment 2 we manipulated the duration of the response-selection stage in the visual task of the DT situation by manipulating the stimulus-response compatibility.

If judging the temporal order occurs after the response-selection stage, then the manipulation of its duration should have a notable effect on the order judgments. In case judging the order occurs before that stage, it should not have an effect on order judgments. We manipulated the stimulus-response compatibility of the visual task by administering a compatible condition, in which the numbers (2, 5, 9) presented in the visual task were mapped to keys of right hand motor responses according to numerical magnitude. In the incompatible condition, numbers were mapped in a non-standard way to the response keys of the right hand, the 5 to the leftmost key, the 9 to the middle key, and the 2 to the rightmost key. This manipulation should affect merely the duration of the response-selection stage in the visual task of the DT condition (Sanders, 1980; McCann and Johnston, 1992).

METHOD

Participants

In Experiment 2, 19 (17 female) participants took part, who had not participated before. The participants were again students of the LMU, who received course credit or payment (8 Euro/hour) for their participation. The average age was 23.9 years (SD = 3.6 years). All subjects except for one were right-handed and all reported normal or corrected-to-normal vision and hearing.

Apparatus, stimuli, design, and procedure

These characteristics of Experiment 2 were identical to Experiment 1 with the following exceptions. In Experiment 2, the stimulus-response mapping of the visual task was manipulated. The mapping could be either compatible or incompatible. In the compatible-mapping condition, the participants responded to the three numbers 2, 5, and 9 by pressing the ","; "."; and "-"-key, respectively. In the incompatible-mapping condition, the numbers were mapped to the same keys, but in a different, non-standard order: 5 was mapped on the ","-key, 9 on the "."-key, and 2 on the "-"-key.

The design formed an incomplete factorial model. As there was only one response key for all three different stimuli in the TOJ-task condition, the stimulus-response mapping could only be manipulated for the DT condition. For the TOJ-task condition, participants completed three blocks of TOJs after the completion of 20 practice trials. For the DT condition, three blocks with compatible stimulus-response mapping and three blocks with incompatible stimulus-response mapping of the visual stimuli were presented in alternating order. Half of the participants started with a block with compatible mapping, the other half with an incompatible-mapping block. Before the six experimental DT-blocks, participants completed three practice blocks for the single tasks (20 trials each for auditory task, visual task with compatible stimulus-response mapping and visual task with incompatible stimulus-response mapping) and two practice block with both tasks with the two different stimulus-response mappings (30 trials each). The complete experimental session lasted approximately 75 min and consisted of TOJ and DT, which were conducted in this order by every participant.

RESULTS

Temporal-order judgments

We calculated the percentage of trials in which the participants reported the auditory stimulus as first for every task condition and every SOA and present them in **Figure 3**. The data were submitted to a repeated-measures ANOVA with task condition

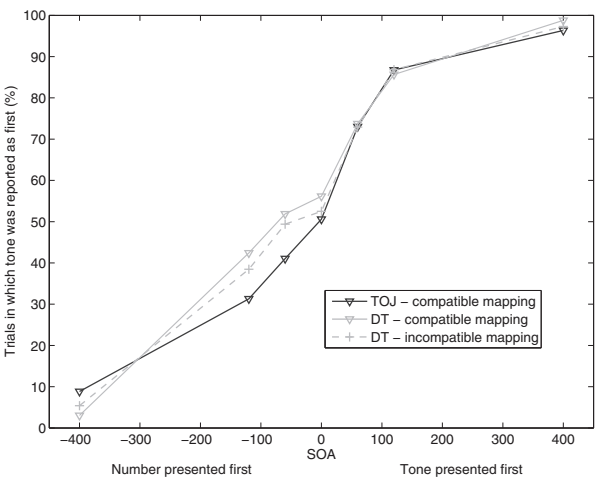


FIGURE 3 | Proportion of trials in which the auditory task was reported as first in the conditions of temporal-order judgment (TOJ) and dual task (DT) in Experiment 2. Two DT conditions were presented: DT with compatible stimulus-response mapping and DT with incompatible stimulus-response mapping.

(TOJ, DT compatible, and DT incompatible) and SOA as within-subjects factors.

The factor SOA was significant, indicating that the percentage of trials in which the auditory task was responded to first, increased with increasing SOA, $F_{(6,\ 108)} = 104.992$, $p < 0.01$. Thus, similar to Experiment 1, participants generally tried to follow the instructions to judge the order of the stimuli because as can be seen in **Figure 3**, the proportion of trials in which the auditory stimulus was reported as first, was higher at positive SOAs compared to negative SOAs.

We did not find a significant effect of the factor task condition, $F_{(1.195,\ 21.510)} < 1$, (Greenhouse-Geisser corrected) on order judgment. Because that factor reflects the difficulty manipulation in the DT condition, the lacking effect of task condition reflects the fact that response order did not differ significantly between the parts TOJ, DT compatible, and DT incompatible. The manipulation of the stimulus-response mapping in the DT condition had no effect on the TOJs consistent with the assumption that the temporal-order decision occurred before the response-selection stage in the DT condition. Also, the interaction between SOA and task condition was not significant, $F_{(12,\ 216)} = 1.543$, $p = 0.19$. However, visual inspection of the data suggested a difference between the task conditions for the negative SOAs (trials in which the number was presented first). Indeed, if only TOJ and DT compatible data were included in the analysis, we found a significant interaction of task condition and SOA, $F_{(6,\ 108)} = 2.343$, $p < 0.05$. One-tailed t-tests showed significant differences between the task conditions for the SOAs −400 ms, −60 ms, and 400 ms. At SOA −400, the proportion of trials in which the participants reported the auditory task as first was higher in the TOJ-task condition than in the DT condition, $t_{(18)} = -1.767$, $p < 0.05$. At SOA −60, $t_{(18)} = 1.868$, $p < 0.05$, and SOA 400, $t_{(18)} = 2.117$, $p < 0.05$, the proportion of trials in which the participants reported the auditory task as first, was higher in the DT condition than in the TOJ-task condition. The results indicate a similar pattern as in Experiment 1, where the proportion of trials in which the tone was reported as first was significantly higher in the DT condition than in the TOJ condition, especially for negative SOAs, i.e., trials in which the number was presented first.

Point of subjective simultaneity (PSS)

Again, we calculated the PSSs by submitting the data of each subject to separate logistic regression analyses for each task and mapping condition. Then we calculated the mean PSS of all participants by estimating the 50% point of the logistic function, at which the participants report the auditory task and the visual task as first equally often (see **Figure 4**).

For the TOJ-task condition, the mean PSS amounted to −36 ms, which indicates that the tone had to be presented 36 ms after the number stimulus to be perceived as simultaneous. In the DT condition, the mean PSS amounted to −60 ms for both the conditions with compatible and incompatible response mappings. We submitted the PSS-values to a repeated-measures ANOVA with task condition as within-subjects factor (TOJ, DT compatible, and DT incompatible). This ANOVA revealed no effect of task condition, $F_{(2,\ 38)} < 1$, suggesting that the three different conditions, including the two DT conditions with different response selection difficulty, did not differ with respect to PSS.

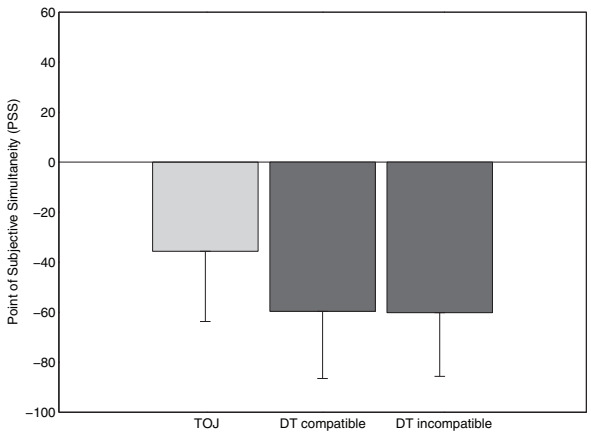


FIGURE 4 | Points of subjective simultaneity (PSS) for the temporal-order judgment task (TOJ) and the two dual-task (DT) conditions (compatible and incompatible stimulus-response mapping) in Experiment 2.

DISCUSSION

The results of Experiment 2 show that the TOJ in the DT condition was not influenced by the manipulation of the response-selection stage. Neither was the PSS shifted by the manipulation when comparing across all conditions. These results indicate that the point in time at which the TOJ is made in the DT condition, was located before the response-selection stage, as the manipulation of this stage had no effect on TOJ.

An issue which needs additional consideration is the effect of the task condition on the overall order performance: in Experiment 2, the ANOVA did not reveal a main factor of task condition, which, on first glance, might be puzzling because in Experiment 1 we found that order judgments differed between the TOJ and the DT conditions. Thus, while we could not find an overall effect of task condition (including TOJ, DT compatible, and DT incompatible as levels) in Experiment 2, the visual inspection of the data suggested an effect of task condition for those trials in which the visual stimulus was presented first. This visual impression was confirmed when we included only the data of the TOJ and the DT compatible condition in the analysis. In that particular case, we obtained a significant interaction between task condition and SOA, which indicates a similar response pattern as in Experiment 1 thus replicating those findings. We will come back to this issue in the "General Discussion."

In sum, the current findings suggest that the order judgments do not differ between the DT compatible and the DT incompatible conditions in Experiment 2, suggesting that the order decision must have taken place before the response-selection stage in the component tasks of the DT condition.

GENERAL DISCUSSION

In the present experiments, one aim was to determine the point in processing when the decision about the temporal order of two presented stimuli is made. The other aim was to investigate

whether the additional insertion of a response selection requirement for the processing of the stimuli in a TOJ context has an influence on TOJs.

In order to address the first aim, we manipulated the perception stage of the TOJ task and the DT in Experiment 1 and the response-selection stage of the DT in Experiment 2. As the manipulation of the perception stage had an effect on TOJs in both task conditions, we conclude that in both conditions the decision about the temporal order occurs after the perceptual stage. Experiment 2 showed that the manipulation of the response-selection stage did not have an influence on TOJ in the DT condition. For the DT condition, we therefore conclude that the decision about the temporal order of the two stimuli is made after perceptual processes and before the response-selection processes start. This finding corresponds with the findings of Sigman and Dehaene (2006), who also found that a manipulation of the perception stage does have an influence on response order, while a manipulation of the response-selection stage does not. The authors hypothesized that an executive process determines the order of the two stimuli which is located after the perceptual processes of the first task.

Concerning the question whether the additional insertion of a choice reaction has an influence on TOJs, the current data suggest that this is indeed the case. We found a difference between the TOJs in the TOJ-task and the DT conditions in Experiment 1. In the DT condition, the auditory task was reported as first more often than in the TOJ-task condition, especially in trials in which the visual task was presented first. In Experiment 2, we found a similar result, albeit not as clear as in Experiment 1. The auditory task was reported as first more often in the DT condition than in the TOJ-task condition. This effect appeared especially in those trials in which the visual task was presented first. Although the data in Experiment 2 did not show a general effect of task condition, the interaction between task condition and SOA indicates a similar pattern as in Experiment 1. The results of both experiments suggest that the additional requirement to discriminate between different stimuli in the DT condition has an influence on temporal order decisions. This effect might be the result of differences in attention allocation between the two task conditions, which result in differences in perception speed (e.g., Posner, 1980; Desimone and Duncan, 1995). Most importantly, we can thereby show, that in addition to the already known factors that have an influence on the perception of temporal order (e.g., stimulus modality, see e.g., Hirsh and Sherrick, 1961; Rutschmann and Link, 1964; Roufs, 1974; Jaśkowski et al., 1990; Spence et al., 2001), also specific task requirements influence temporal order decisions.

What are the relations between the current findings and findings of earlier saccadic studies (e.g., Deubel and Schneider, 1996)? As reported in the introduction, Deubel and Schneider (1996) showed in a DT study, that the planning of a saccade to a target stimulus improves perceptual processing of said visual stimulus in a concurrent discrimination task by allocating attention to it. Discrimination performance and thereby perception of the visual stimulus in this study is best when the saccade and the discrimination task involve the same location compared to when they involve two stimuli at different locations. Our results are also consistent with the assumption that visual stimulus processing (in relation to auditory stimulus processing) is potentially improved under particular conditions. That is, the visual task is reported as first more often in the TOJ-task condition requiring no response selections when contrasted to the DT condition that requires these selection processes in two tasks.

Why is the effect of task condition especially prominent in the trials in which the visual task was presented first? In Experiment 1, the perception stage of the visual task was manipulated in both task conditions, first in the TOJ-task condition followed by the manipulation in the DT condition. It could be that further task changes, like the addition of a discrimination requirement in the DT condition, had a stronger effect on the manipulated visual task. Compared to the TOJ condition, the processing of the visual task might have been slowed by this additional processing requirement because the stimulus processing rate is slowed per se or additional information may be processed when participants are instructed for stimulus discrimination. What might also play a role in explaining the prominent effect of task condition in trials in which the visual task was presented first is the usually faster processing of auditory stimuli compared to visual stimuli, which was particularly shown in TOJ studies (e.g., Hirsh and Sherrick, 1961; Dinnerstein and Zlotogura, 1968; Jaśkowski et al., 1990; Zampini et al., 2003; Keetels and Vroomen, 2005; Van Eijk et al., 2008; Boenke et al., 2009; but see e.g., Rutschmann and Link, 1964, for the opposite effect).

In Experiment 2, there was no manipulation of the TOJ-task condition, but still an interaction of task and SOA was found, when the TOJ-task condition was compared with the DT condition with compatible mapping. The same interaction was found in Experiment 1. In Experiment 2, it was also the visual task that was manipulated in the DT condition. Therefore, the same possible explanations as just mentioned for Experiment 1 might apply here: the additional processing stages necessary for DT compared to TOJ-tasks might have a greater effect on the visual task for the aforementioned reasons (see above).

A recent study by McDonald et al. (2005) found that an attended object is reported as being presented earlier than a simultaneously presented unattended object in a TOJ task. Attention was modulated by cueing one of the two visual stimuli with a sound. The authors recorded event-related brain potentials (ERPs) during the task and found that the attended stimulus was not processed faster than the unattended (which would have been indicated by latency shifts in early ERPs). Instead, attention had an effect on the amplitude of the ERPs: the attended stimulus showed a higher amplitude than the unattended one. The authors suggest that these attention-induced enhancements in signal strength of the cued stimulus are then "interpreted as a timing difference by a later comparator mechanism" (McDonald et al., 2005, p. 1201). In regard to our study, this could mean that the manipulation of the visual task led to a shift in attention to the auditory task, either because of an alerting quality of the auditory signal or because of differences in difficulty between auditory and visual task. This attention shift could have led to a strengthening of the auditory signal, which would have been interpreted as a difference in presentation time. In order to investigate this assumption, further experiments have to be done.

Summing up, our study gives new evidence for the time in processing at which the decision about the temporal order of two stimuli is made, both in a TOJ-task and a DT. Also, we could show that the additional task requirement to discriminate the stimuli has an influence on the TOJs of a visual and an auditory stimulus.

ACKNOWLEDGMENTS

This research was supported by the excellence cluster Cognition for Technical Systems (CoTeSys). We would like to thank Zhuanghua Shi and Heng Zou for valuable discussions as well as Christina Reimer, Tobias Baumann, and Dejan Draschkow for their assistance with data collection.

REFERENCES

Arrighi, R., Alais, D., and Burr, D. (2006). Perceptual synchrony of audiovisual streams for natural and artificial motion sequences. *J. Vis.* 6, 260–268.

Boenke, L. T., Deliano, M., and Ohl, F. W. (2009). Stimulus duration influences perceived simultaneity in audiovisual temporal-order judgment. *Exp. Brain Res.* 198, 233–244.

Cardoso-Leite, P., Gorea, A., and Mamassian, P. (2007). Temporal order judgment and simple reaction times: evidence for a common processing system. *J. Vis.* 7, 1–14.

De Jong, R. (1995). The role of preparation in overlapping-task performance. *Q. J. Exp. Psychol. A* 48, 2.

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* 18, 193–222.

Deubel, H., and Schneider, W. X. (1996). Saccade target selection and object recognition: evidence for a common attentional mechanism. *Vision Res.* 36, 1827–1837.

Deubel, H., Schneider, W. X., and Paprotta, I. (1998). Selective dorsal and ventral processing: evidence for a common attentional mechanism in reaching and perception. *Visual Cogn.* 5, 81–107.

Dinnerstein, A. J., and Zlotogura, P. (1968). Intermodal perception of temporal order and motor skills: effects of age. *Percept. Mot. Skills* 26, 987–1000.

Hirsh, I. J., and Sherrick, C. E. Jr. (1961). Perceived order in different sense modalities. *J. Exp. Psychol.* 62, 423–432.

Jaśkowski, P., Jaroszyk, F., and Hojan-Jezierska, D. (1990). Temporal-order judgments and reaction time for stimuli of different modalities. *Psychol. Res.* 52, 35–38.

Keetels, M., and Vroomen, J. (2005). The role of spatial disparity and hemifields in audio-visual temporal order judgments. *Exp. Brain Res.* 167, 635–640.

King, A. J. (2005). Multisensory integration: strategies for synchronization. *Curr. Biol.* 15, 339–341.

Lien, M.-C., Schweickert, R., and Proctor, R. W. (2003). Task switching and response correspondence in the psychological refractory period paradigm. *J. Exp. Psychol. Hum. Percept. Perform.* 29, 692–712.

Liepelt, R., Strobach, T., Frensch, P., and Schubert, T. (2011). Improved intertask coordination after extensive dual-task practice. *Q. J. Exp. Psychol.* 64, 1251–1272.

Luria, R., and Meiran, N. (2003). Online order control in the psychological refractory period paradigm. *J. Exp. Psychol. Hum. Percept. Perform.* 29, 556–574.

McCann, R. S., and Johnston, J. C. (1992). Locus of the single-channel bottleneck in dual-task interference. *J. Exp. Psychol. Hum. Percept. Perform.* 18, 471.

McDonald, J. J., Teder-Sälejärvi, W. A., Di Russo, F., and Hillyard, S. A. (2005). Neural basis of auditory-induced shifts in visual time-order perception. *Nat. Neurosci.* 8, 1197–1202.

Miller, J., and Schwarz, W. (2006). Dissociations between reaction times and temporal order judgments: a diffusion model approach. *J. Exp. Psychol. Hum. Percept. Perform.* 32, 394–412.

Neumann, O. (1982). "Experimente zum Fehrer-Raab-Effekt und das 'Wetterwart'-Modell der visuellen Maskierung [Experiments on the Fehrer-Raab effect and the 'Weather-Station' model of visual masking]", in *Report, Department of Psychology at the Ruhr-University Bochum, Cognitive Psychology Unit* (Bochum).

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

Roufs, J. A. (1974). Dynamic properties of vision. V. Perception lag and reaction time in relation to flicker and flash thresholds. *Vision Res.* 14, 853–869.

Rutschmann, J., and Link, R. (1964). Perception of temporal order of stimuli differing in sense mode and simple reaction time. *Percept. Mot. Skills* 18, 345–352.

Sanders, A. F. (1980). Stage analysis of reaction processes. *Tutor. Mot. Behav.* 1, 331–354.

Sanders, A. F. (1990). Issues and trends in the debate on discrete vs. continuous processing of information. *Acta Psychol.* 74, 123–167.

Shi, Z., Hirche, S., Schneider, W., and Müller, H. (2008). "Influence of visuomotor action on visual-haptic simultaneous perception: a psychophysical study," in *Symposium on Haptic Interfaces for Virtual Environments and Teleoperator Systems* (Reno, NV).

Sigman, M., and Dehaene, S. (2006). Dynamics of the central bottleneck: dual-task and task uncertainty. *PLoS Biol.* 4:e220. doi: 10.1371/journal.pbio.0040220

Spence, C. (2007). Audiovisual multisensory integration. *Acoust. Sci. Technol.* 28, 61–70.

Spence, C., Shore, D. I., and Klein, R. M. (2001). Multisensory prior entry. *J. Exp. Psychol. Gen.* 130, 799–832.

Stein, B. E., and Meredith, M. A. (1994). *The Merging of the Senses. Cognitive Neuroscience Series*, 2nd Edn. Cambridge, MA: The MIT Press.

Sternberg, S. (1969). The discovery of processing stages: extensions of Donders' method. *Acta Psychol.* 30, 276–315.

Sternberg, S., Knoll, R. L., and Gates, B. A. (1971). *Prior entry reexamined: the effect of attentional bias on order perception*. Bell Laboratories Technical Memorandum, MM 71-1221–1217.

Strobach, T., Frensch, P. A., Soutschek, A., and Schubert, T. (in press). Investigation on the improvement and transfer of dual-task coordination skills. *Psychol. Res.*

Sugita, Y., and Suzuki, Y. (2003). Audiovisual perception: implicit estimation of sound-arrival time. *Nature* 421, 911.

Szameitat, A. J., Lepsien, J., Cramon, D. Y., Sterr, A., and Schubert, T. (2006). Task-order coordination in dual-task performance and the lateral prefrontal cortex: an event-related fMRI study. *Psychol. Res.* 70, 541–552.

Treutwein, B., and Strasburger, H. (1999). Fitting the psychometric function. *Percept. Psychophys.* 61, 87–106.

Umiltà, C., Nicoletti, R., Simion, F., Tagliabue, M. E., and Bagnara, S. (1992). The cost of a strategy. *Eur. J. Cogn. Psychol.* 4, 21–40.

Van Eijk, R. L. J., Kohlrausch, A., Juola, J. F., and Van De Par, S. (2008). Audiovisual synchrony and temporal order judgments: effects of experimental method and stimulus type. *Percept. Psychophys.* 70, 955–968.

Witt, J. K., and Proffitt, D. R. (2008). Action-specific influences on distance perception: a role for motor simulation. *J. Exp. Psychol. Hum. Percept. Perform.* 34, 1479–1492.

Witt, J. K., Proffitt, D. R., and Epstein, W. (2005). Tool use affects perceived distance, but only when you intend to use it. *J. Exp. Psychol. Hum. Percept. Perform.* 31, 880–888.

Zampini, M., Shore, D. I., and Spence, C. (2003). Audiovisual temporal order judgments. *Exp. Brain Res.* 152, 198–210.

Zwickel, J., Grosjean, M., and Prinz, W. (2007). Seeing while moving: measuring the online influence of action on perception. *Q. J. Exp. Psychol.* 60, 1063–1071.

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: 31 May 2012; accepted: 07 August 2012; published online: 25 August 2012.

*Citation: Hendrich E, Strobach T, Buss M, Müller HJ and Schubert T (2012) Temporal-order judgment of visual and auditory stimuli: modulations in situations with and without stimulus discrimination. Front. Integr. Neurosci. **6**:63. doi: 10.3389/fnint.2012.00063*

frontiers in
INTEGRATIVE NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 16 October 2012
doi: 10.3389/fnint.2012.00095

Duration reproduction with sensory feedback delay: differential involvement of perception and action time

***Stephanie Ganzenmüller*[1,2]*, *Zhuanghua Shi*[1] *and Hermann J. Müller*[1,3]**

[1] *Department Psychology, General and Experimental Psychology, LMU Munich, Germany*
[2] *Graduate School of Systemic Neuroscience, LMU Munich, Germany*
[3] *Department of Psychological Sciences, Birkbeck College (University of London), London, UK*

Edited by:
Micah M. Murray, University Hospital Center and University of Lausanne, Switzerland

Reviewed by:
Melissa J. Allman, Michigan State University, USA
Guido M. Cicchini, Consiglio Nazionale delle Ricerche, Italy

***Correspondence:**
Stephanie Ganzenmüller, Department Psychology, General and Experimental Psychology, Ludwig-Maximilians-Universität München, Leopoldstr. 13, 80802 Munich, Germany.
e-mail: s_ganzenmueller@psy.lmu.de

Previous research has shown that voluntary action can attract subsequent, delayed feedback events toward the action, and adaptation to the sensorimotor delay can even reverse motor-sensory temporal order judgments. However, whether and how sensorimotor delay affects duration reproduction is still unclear. To investigate this, we injected an onset- or offset-delay to the sensory feedback signal from a duration reproduction task. We compared duration reproductions within (visual, auditory) modality and across audiovisual modalities with feedback signal onset- and offset-delay manipulations. We found that the reproduced duration was lengthened in both visual and auditory feedback signal onset-delay conditions. The lengthening effect was evident immediately, on the first trial with the onset-delay. However, when the onset of the feedback signal was prior to the action, the lengthening effect was diminished. In contrast, a shortening effect was found with feedback signal offset-delay, though the effect was weaker and manifested only in the auditory offset-delay condition. These findings indicate that participants tend to mix the onset of action and the feedback signal more when the feedback is delayed, and they heavily rely on motor-stop signals for the duration reproduction. Furthermore, auditory duration was overestimated compared to visual duration in crossmodal feedback conditions, and the overestimation of auditory duration (or the underestimation of visual duration) was independent of the delay manipulation.

Keywords: action, audition, time perception, time reproduction, vision

INTRODUCTION

Accurate timing is essential for our everyday activities, like dancing, playing music, or catching a moving object. In order to accomplish precise timing in a complex environment, our brain has to frequently update its internal representation of multiple sensory inputs. Precisely inferring the timing and duration of events as well as correctly judging temporal order in the sub-second range can be challenging, since neural representations of time may be confounded by noise and delay perturbation in sensory pathways. For example, the neural transmission time can vary across different sensory modalities (King and Palmer, 1985; Regan, 1989), and physical transmission distances (Campbell et al., 1981; Shadmehr et al., 2010), as well as stimulus intensities (Purpura et al., 1990). Continuous changes of the body and the environment provide a further challenge for accurate action timing (Shadmehr et al., 2010). However, in daily life, accurate sensorimotor temporal coordination remains possible, indicating that our brain is able to calibrate and compensate for temporal inconsistencies among different sensory inputs as well as delays in the sensorimotor loop.

Indeed, research has demonstrated that the brain can dynamically realign the perceived timing of multisensory or sensorimotor events. For example, Fujisaki et al. (2004) have shown adaptive changes in synchrony perception between vision and audition: after exposure to a fixed audiovisual asynchrony, the point of subjective simultaneity (PSS, a measure of point in time at which observers perceive maximum simultaneity) of an audiovisual event was shifted toward the previous "lagging" modality. Other work has revealed similar temporal recalibration mechanisms across other modalities (Vroomen et al., 2004; Navarra et al., 2005; Hanson et al., 2008; Harrar and Harris, 2008; Takahashi et al., 2008; Di Luca et al., 2009). Temporal recalibration has also been found between an action and its sensory feedback. The first study that demonstrated compensation for temporal delays in the visuomotor feedback loop confronted participants with a visual-motor lag (delayed visual feedback while controlling the horizontal movement of a small airplane as it moved down the screen through an obstacle field) (Cunningham et al., 2001). Participants' performance improved after some time of practice. Interestingly, when the lag was removed after the adaptation, the adapted behavior persisted and participants, suffering from the adaptation, often made movements too early, leading to more crashes. In another study, Stetson et al. (2006) demonstrated that following brief exposure to delayed visual feedback of a voluntary action the subjective temporal order of a motor-sensory event might even be reversed when the delay was removed. This effect was attributed to dynamical shifts of the appearance of the visual stimulus with respect to the perceived timing of the key press,

in order to maintain appropriate causality perception. This proposal goes along with earlier findings that a delayed sensory effect is perceived as having appeared slightly earlier in time if it follows a voluntary action (Eagleman and Holcombe, 2002; Haggard et al., 2002)—a phenomenon referred to as "intentional binding." Studies have also demonstrated that intentional binding attracts a voluntary action toward its sensory effect, so that the action is perceived as having occurred slightly later in time and the interval between the action and its sensory feedback as shorter than the actual interval (Haggard et al., 2002; Engbert et al., 2007, 2008). Wearden et al. (2009) proposed that the shortening effect is driven by a transient slowdown of an internal clock after a voluntary action, and this shortening effect might be reinforced by everyday experience which leads us to assume sensorimotor synchrony between the start of a motor action and its sensory consequence (Heron et al., 2009). However, whether sensorimotor temporal calibration is due to timing changes in the motor system or in the perceptual system is still under debate. Some researchers have suggested that sensorimotor temporal calibration is induced mainly by a temporal shift in the motor system (Sugano et al., 2010), whereas others have attributed sensorimotor temporal calibration to pure perceptual learning (Kennedy et al., 2009).

Alternatively, sensorimotor temporal (re-)calibration has been taken to only reflect modification of predictive feed-forward actions, reducing the errors between the internal prediction and the external feedback (Miall and Jackson, 2006; Shadmehr et al., 2010). Such error correction mechanisms have been used for explaining sensorimotor synchronization, as for instance in the frequently used paradigm of finger tapping to an external pacing source (metronome). When the changes of the pacing source are detectable and regular, participants are able to reduce their sensorimotor asynchronies by predicting upcoming changes. When temporal changes are unpredictable, the time to the next motor response is automatically adjusted in proportion to the asynchrony in the previous sensorimotor event (Repp, 2005).

However, it is important to note that most of the aforementioned studies focused on sensorimotor calibration of a point in time. By contrast, the effects of delayed feedback on the voluntary duration reproduction are as yet little understood. Unlike a point in time, subjective duration can be distorted in many ways, such as by a saccadic eye movement shortly before or after the to-be-estimated event (Morrone et al., 2005), a voluntary action immediately prior to the critical event (Park et al., 2003), the emotional state of the observer (Angrilli et al., 1997; Shi et al., 2012), stimulus properties (such as intensity) (Eagleman, 2008), or pharmacological agents (such as cocaine or methamphetamine) (Meck, 1996) (see review Buhusi and Meck, 2005). Perceived durations in different modalities can also differ. For example, sounds are often perceived as longer than light flashes of the same physical duration (Walker and Scott, 1981; Wearden et al., 1998). Furthermore, there is evidence that the auditory system dominates the visual system, causing the durations of visual stimuli, presented simultaneously with an auditory stimuli, to be perceived as longer than they physically are (Walker and Scott, 1981; van Wassenhove et al., 2008; Burr et al., 2009; Chen and Yeh, 2009; Shi et al., 2010a; Klink et al., 2011). In addition, not only the use of different signal modalities during a timing task, but also the encoding of multiple signal durations, can lead to distortions in temporal memory—an effect recently termed as "memory-mixing" (Gu and Meck, 2011). Such high variability in subjective timing is quite surprising considering how important accurate timing is for our actions.

The purpose of the present study was to investigate how asynchronous-feedback signals would influence motor timing. We adopted an action-based duration reproduction paradigm combined with feedback onset- and, respectively, offset-delay manipulations. That is, participants had to reproduce auditory or visual durations and received (auditory or visual) feedback signals[1]. The feedback could either be synchronized or delayed with participants' button presses (onsets or offsets), and could be delivered in the same or different modality. We specifically asked participants to focus on the reproduction of the standard duration and not pay attention to the feedback. There are two sources of temporal information available for duration reproduction: motor timing (i.e., the duration of the button press) and the feedback timing. If participants only rely on the motor timing for their ongoing reproduction, reproduction errors would be expected to be the same or similar across all trials, no matter whether the feedback is synchronous or delayed. If participants get influenced by the feedback signal during their reproduction, despite the instruction, different reproduction errors for synchronized versus delayed feedback would be predicted. Furthermore, we examined influences of action-effect causal relationship on the duration reproduction, by presenting the feedback signal randomly near the onset or offset of participants' action.

GENERAL METHODS

SUBJECTS

Sixty nine naive volunteers (53 females, mean age 27.6) participated in each experiment for payment (Experiments 1–4: 14 participants, Experiment 5: 13 participants). All participants had normal or corrected-to-normal vision; none of them reported any history of somatosensory disorders. They gave written informed consent before the experiments.

STIMULI AND APPARATUS

All experiments were conducted in a dimly lit cabin (0.21 cd/m^2). Auditory tones (400 Hz and 600 Hz, 64 dB) and LED lights (84 cd/m^2 blue and 67 cd/m^2 red) were presented as stimuli. Stimulus presentation and data acquisition were controlled by a National Instrument PXI system, ensuring highly accurate timing (<1 ms). The experimental programs were developed using MatLab and the Psychophysics Toolbox (Brainard, 1997). The auditory stimuli were delivered to participants via headphones (Pro-luxe XL-300); the LED stimuli (two LEDs, blue and red)

[1]In this study we refer to the second stimulus—that is presented during the reproduction—as a "feedback signal" to highlight the causal relationship between the action and sensory effect. The terms "feedback signal" and "feedback" are used interchangeably in the text.

were positioned 2 cm apart horizontally. The response button was placed on the table in-between the participant and the LEDs. Reproduction times were measured using the response button, which participants pressed with their right-hand index finger.

PROCEDURE

We adopted and modified an action-based duration reproduction task with feedback, as introduced by Bueti and Walsh (2010). Each trial started with a standard duration, either 800 or 1200 ms in length, in the form of an auditory tone (Experiments 1 and 4) or an LED light (Experiments 2 and 3). Following the presentation of the standard duration, participants were asked to reproduce the duration as accurately as possible by button press, with reproduction duration demarcated by the onset and offset of the press action. Pressing the button also induced a feedback signal (a tone in Experiments 1 and 3, an LED light in Experiments 2 and 4) whose onset or offset could deviate from the onset or offset of the button press (see **Figure 1** and next paragraph). Subjects were told that feedback signal could be either dependent or independent of their button press. They were specifically instructed to reproduce the standard duration as accurately as possible by pressing down the button, regardless of the feedback signals (see the detail instruction in the "Appendix"). To distinguish and counter-balance the standard and feedback stimuli, half of the participants received high tones (or red lights) as standard stimuli and low tones (or blue lights) as the ± feedback stimuli, and vice versa for the other half.

For the first four experiments, there were three different temporal manipulations of feedback signals: synchronous-feedback, onset-delay feedback, and offset-delay feedback. In the synchronous-feedback condition, the onset and offset of the feedback occurred synchronously with the onset of the button press and the release of the button. In the onset-delay condition, the onset of the feedback signal was delayed by 200 ms following the onset of the button press, while feedback offset occurred synchronously with the release of the button. In the offset-delay condition, the feedback signal started synchronously with the button press, but the feedback offset occurred only 200 ms after the release of the button. These three conditions were varied block-wise, with 10 trials per block. Both the onset- and offset-delay blocks were preceded and followed by a synchronous-feedback block. The order of the onset- and offset-delay blocks was randomized.

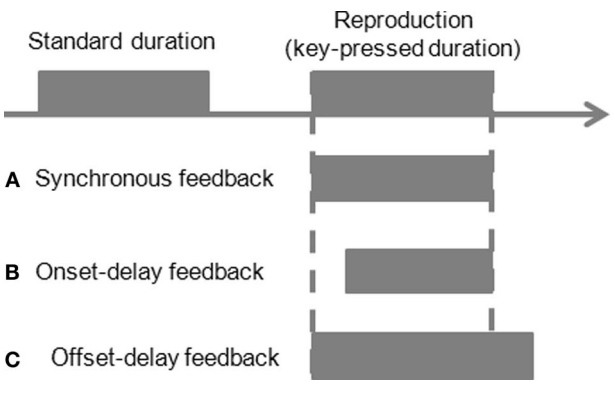


FIGURE 1 | Schematic illustration of the experimental design. A standard duration reproduction paradigm with manipulation of feedback delays during reproduction. An auditory or visual stimulus is presented first as a standard duration. Participants reproduce the standard by pressing a button. Another auditory or visual stimulus is fed back to participants based on the action. The feedback signal could be synchronous to the key press (**A** synchronous-feedback condition), or be delayed 200 ms at the onset of the feedback but simultaneously stops at button release (**B** onset-delay feedback condition), or starts synchronously with the button press but stops 200 ms after the button release (**C** offset-delay feedback condition).

In Experiment five, we used the same block-design as in previous experiments, but randomized the onset and offset of the feedback signal relative to the button press. To do this, for each synchronous-feedback block we measured the mean reproduction durations for 800 and 1200 ms, and the mean response onset asynchrony. During the onset-manipulation blocks, the feedback signal started independently of the button press, with random jittering ±200, ±100, or 0 ms around the mean response onset asynchrony measured in the preceding synchronous block. The feedback signal stopped when the button was released. During the offset-manipulation blocks, the feedback signal started synchronously with the button press, but stopped automatically with a duration randomly jittering ±200, ±100, or 0 ms around the mean reproduction duration (either 800 or 1200 ms corresponding to the duration in the current trial) measured in the preceding synchronous block. The random jittering was used in order to ensure that participants would not be able to predict the onset or offset of the manipulated feedback signal, thus we could obtain about half of all trials with feedback prior to participants' actions. We further increased the number of the trials to 20 for the onset- and offset-manipulation blocks to ensure enough trials with the feedback before participants' action. The task instruction was kept the same as during the previous four experiments.

Note that the standard and feedback stimuli were kept within the same modality in Experiments 1, 2, and 5, but presented in separate modalities in Experiments 3 and 4 (see **Table 1**).

In the first four experiments, there were 10 repetitions for the onset- and offset-delay blocks and 20 repetitions for the synchronous-feedback signal blocks. Participants took a short break after every eight blocks. In Experiment 5, there were eight repetitions for the onset- and offset-manipulation blocks (each consisting of 20 trials) and 16 repetitions for the synchronous-feedback signal blocks (each consisting of 10 trials). Here, participants took a short break after four blocks

Table 1 | Modalities of the standard and feedback stimuli.

Experiment	Standard	Feedback
1	Auditory	Auditory
2	Visual	Visual
3	Visual	Auditory
4	Auditory	Visual
5	Auditory	Auditory

(= 60 trials). In addition, there were two practice blocks with the synchronous-feedback signal condition run prior to the formal experiment.

DATA ANALYSIS

Mean measures and standard deviations of time reproduction have been shown to vary linearly with standard durations, so that after normalization the same form of distribution of relative time and constant timing sensitivity can be found (Gibbon et al., 1984). In line with this, reproduction errors (i.e., the difference between the reproduced duration and the standard duration) in the present study exhibited differences between the two standard durations (800 and 1200 ms), that is, the amount of over-/underestimation (in ms) is proportional to the respective standard duration. To take this into account, we calculated reproduction errors and then normalized them by the corresponding physical duration. Normalized reproduction errors of zero indicate perfect reproduction, positive values an overestimation, and negative values an underestimation of the standard duration. In order to examine dynamic influences of the onset- and offset-delay manipulation, we selected four trials from the synchronous block prior to and the synchronous block after the delay manipulation. The first four trials served as baseline and the last four trials for analyzing after-effects of the delay manipulation. Henceforth, we refer to the former four synchronous-feedback trials as baseline phase, the latter four synchronous-feedback trials as post phase, and the 10 trials from the (intervening) delay block as delay phase. We omitted the middle two trials in the synchronous-feedback block to separate the post and baseline phases. Repeated-measures analyses of variance (ANOVAs) of the normalized reproduction errors in the three different phases (baseline phase, delay phase, and post phase) were run separately for the onset- and offset-delay conditions. Bonferroni-corrected *t*-tests for multiple comparisons were carried out for a-posteriori comparisons to assess differences in reproduction errors.

For Experiment 5, we focused on analyzing linear correlations between the onset- and offset-manipulations and normalized reproduction errors. Thus, linear regression and correlation analyses were applied. We realigned the onsets of the feedback relative to the onsets of the actual response, and compared the differential influences between the feedback before and after participants' action. For the offset-manipulation condition, we used an alternative approach: we calculated the offset jitters relative to the standard durations and analyzed the general relationship between the offset jitters and the reproduction errors. We did not align the offsets relative to the responses, since the mean feedback duration was close to the mean reproduction time, which would inevitably lead to pseudo negative correlation between the relative offset and the reproduced duration. Such correlation could not reflect the influence of the offset-manipulation. In both cases, we normalized feedback jitters with their correspondent standard durations, such that the feedback jitter has the same unit as the normalized reproduction error.

RESULTS

GENERAL REPRODUCTION RESULTS

We analyzed reproduction times for the synchronous-feedback condition for all five experiments, comparing reproduction performance after the short (800 ms) and long (1200 ms) standards. Reproduced durations in milliseconds are presented in **Figure 2**. We found a significant difference between the reproduced times of the short and long standard stimuli (all $p < 0.01$) across all five experiments, suggesting participants were actually able to perform the task.

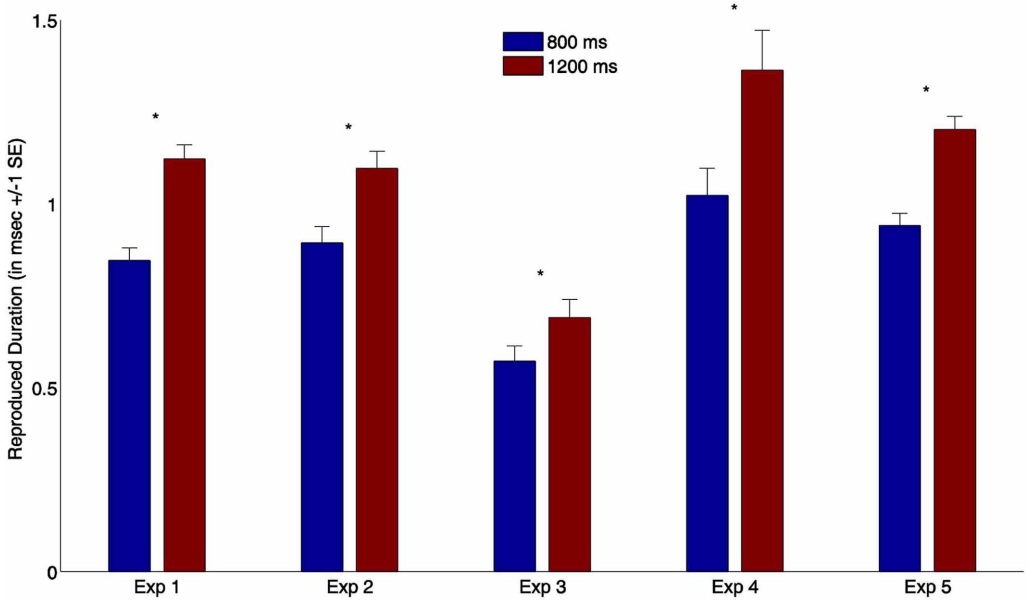


FIGURE 2 | Mean reproduction times (and associated standard errors) from all synchronous trials for all five experiments. Blue bars depict produced durations after short standard stimuli (800 ms); red bars indicate reproduction times after long standard stimuli (1200 ms) (* indicates $p < 0.01$).

EFFECTS OF UNIMODAL FEEDBACK ONSET- AND OFFSET-MANIPULATION ON THE DURATION REPRODUCTION

Normalized reproduction errors, and associated standard errors, for the first four experiments and all conditions are presented in **Table 2**. **Figure 3** shows the normalized reproduction errors for the onset- and offset-delay manipulation for the unimodal auditory and visual feedback.

In the onset-delay conditions (**Figure 3**, up-panels), normalized reproduction errors were significantly influenced by the delay manipulation, [$F_{(2,\ 26)} = 246.78$; $p < 0.01$], and [$F_{(2,\ 26)} = 43.30$, $p < 0.01$] for the auditory and visual conditions respectively. The overestimation during the onset-delay phase for both auditory and visual conditions proved to be significantly larger compared to the baseline ($p < 0.01$) and the post phase ($p < 0.01$) (**Figure 3**, low-panels). Normalized reproduction errors in the post phase (overestimation) were raised reliably relative to the baseline ($p < 0.01$) for the auditory condition, but not for the visual condition ($p = 0.16$). Interestingly, the overestimation on the onset-delay phase was 21% for the auditory and 19% for the visual, which are statistically not different from the onset-delay manipulation (all $p > 0.1$). Furthermore, the overestimation started with the first trial of the delay manipulation (condition) and stopped as soon as the delay was removed (**Figure 3**, up-panels). Paired t-tests showed no significant difference in the overestimation between the first versus the remaining trials in both delay and post phase, (all $p > 0.1$).

In contrast to the onset-delay manipulation (which made participants overestimate the standard durations), the offset-delay manipulation (**Figure 3**, mid-panels) showed different patterns for the auditory and visual conditions. In the auditory condition (**Figure 3**, middle left panel), the offset-delay led participants to significantly underestimate the standard durations during the offset-delay phase, [$F_{(2,\ 26)} = 13.73$; $p < 0.01$]. This effect derived mainly from a significantly negative increase in normalized reproduction errors during the delay phase versus the baseline ($p < 0.01$). Normalized errors were also negatively increased in the post phase compared to the baseline ($p < 0.01$). However, there was no reliable difference between the delay and post phases ($p = 0.99$). Paired t-tests showed that the underestimation started only from the second trial with delay manipulation, as there was no effect in the first trial of the delay phase (significant difference between the first and the remaining trials, [$t_{(13)} = 9.30$, $p < 0.01$]). Also, underestimation only stopped on the second trial of the post phase, with reproduction errors on the first trial still differing significantly from the errors on the other trials, [$t_{(13)} = -5.26$, $p < 0.01$]. In contrast to the auditory condition, manipulation of the visual offset-delay feedback had no significant influence on normalized reproduction, [$F_{(2,\ 26)} = 1.60$, $p = 0.22$] (baseline vs. delay: $p = 1.00$; delay vs. post phase: $p = 0.36$; baseline vs. post phase: $p = 0.45$).

EFFECTS OF CROSSMODAL FEEDBACK ONSET- AND OFFSET-MANIPULATION ON DURATION REPRODUCTION

Overall, there was strong underestimation of the visual standard with synchronous auditory feedback signal (hereafter we refer to as the visual-auditory experiment), and strong overestimation of the auditory standard with visual feedback signal (hereafter the auditory-visual experiment), all $p < 0.01$. Trial-wise normalized reproduction errors for the onset- and offset-delay manipulations are depicted in **Figure 4**.

For the onset-delay conditions (**Figure 4**, up-panels), the normalized reproduction errors were significantly modulated by onset-delays for the visual-auditory experiment, $F_{(2,\ 26)} = 185.41$, $p < 0.01$, and the auditory-visual experiment, $F_{(2,\ 26)} = 39.06$, $p < 0.01$. The underestimation (in the visual-auditory experiment, **Figure 4A**) and the overestimation (in the auditory-visual experiment, **Figure 4B**) in the onset-delay phase, were significantly different from the correspondent baseline and the post phase (all $p < 0.01$), while there were no differences between the baseline and post phase (all $p > 0.1$). Interestingly, the reproduced duration during the onset-delay phase compared to the baseline was increased 21% for the visual-auditory experiment and 16% for the auditory-visual experiment. Both are comparable to the overestimation observed in Experiment 1 and 2 (21 and 19% respectively). Further pair-wise sequential-trial analysis showed that the manipulation effect of the onset-delay in the visual-auditory experiment started on the first trial of delay manipulation ($p = 0.78$) and stopped as soon as the delay was removed ($p = 0.28$). However, in the auditory-visual experiment, participants needed one trial to adjust their behavior to the onset-delay, as evidenced by significantly different normalized reproduction errors in the first trial compared to the remaining trials of the delay phase, $t_{(13)} = -2.57$, $p < 0.05$. However, the effect ceased as soon as the delay was removed ($p = 0.59$).

For the visual-auditory experiment, a general, significant underestimation was also found in the offset-delay condition, $F_{(2,\ 26)} = 8.15$, $p < 0.01$ (**Figure 4A**, mid-panel). Relative to the baseline, the normalized reproduction error (underestimation) was negatively increased in the offset-delay phase ($p < 0.05$) and

Table 2 | Normalized reproduction errors (± standard errors) in percentage by onset- and offset-delay manipulation and different phases in Experiments 1–4.

	Onset-delay manipulation			Offset-delay manipulation		
	Baseline phase	Delay phase	Post phase	Baseline phase	Delay phase	Post phase
Experiment 1	−0.55 ± 2.5	21.73 ± 2.0	3.89 ± 2.8	1.18 ± 2.9	−4.57 ± 1.8	−3.69 ± 2.9
Experiment 2	−0.28 ± 4.4	19.09 ± 2.7	3.51 ± 4.4	0.95 ± 4.3	−1.72 ± 2.8	4.91± 4.2
Experiment 3	−33.88 ± 3.1	−12.16 ± 2.2	−31.48 ± 3.4	−33.06 ± 3.4	−37.93 ± 2.1	−38.19 ± 3.6
Experiment 4	21.01 ± 4.6	37.21 ± 3.3	24.39 ± 5.3	22.55 ± 4.9	23.35 ± 3.4	25.47 ± 5.6

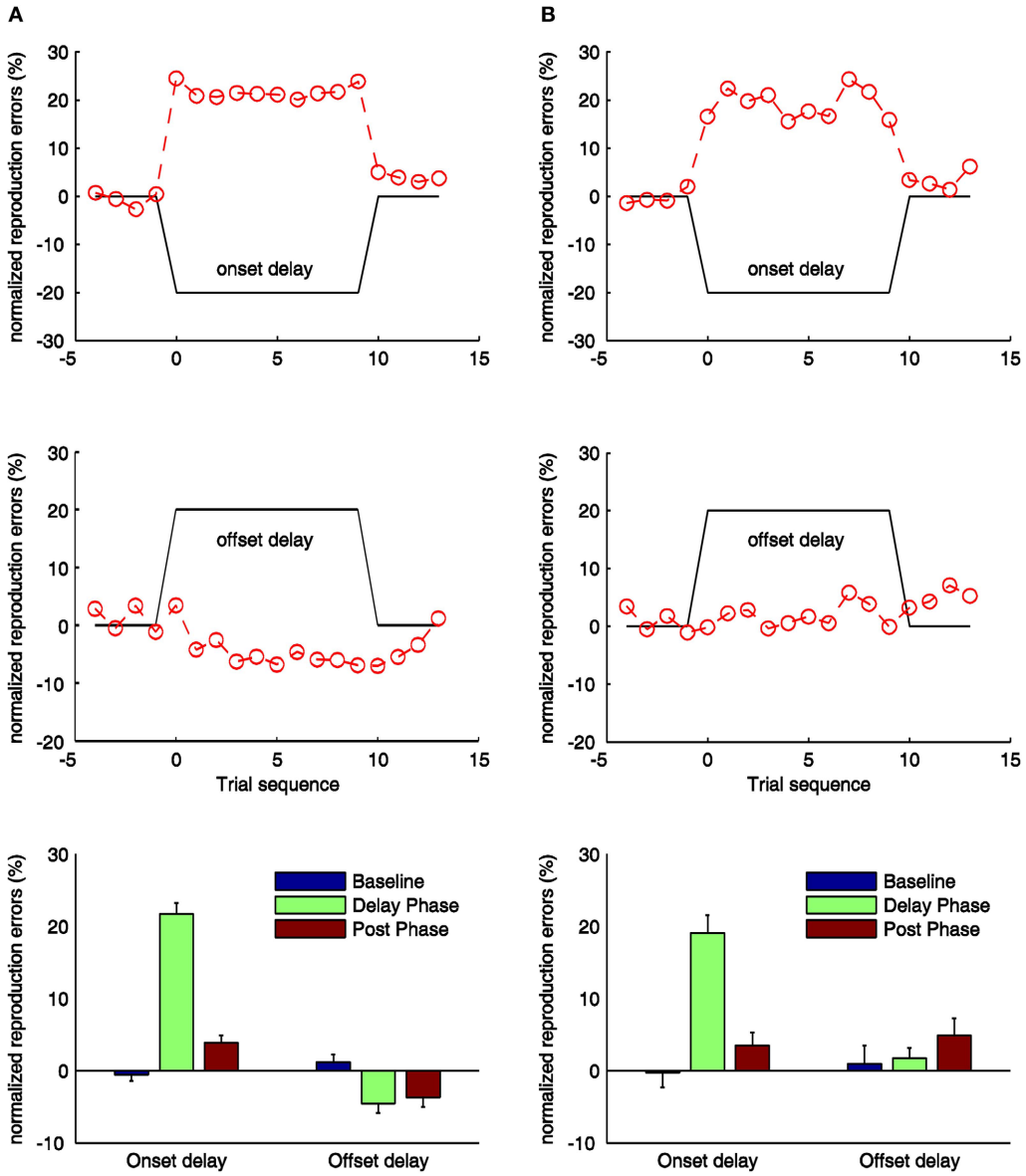


FIGURE 3 | Normalized reproduction errors [(subjective duration—physical duration)/physical duration] for the onset- and offset-delay condition of Experiment 1 (A) and Experiment 2 (B). In the upper and middle panels trial-wise dynamic changes of normalized reproduction are shown. Four trials from the synchronous block before the delay manipulation (baseline phase), delay block (delay phase), and four trials after the delay manipulation (post phase) are displayed. The black lines indicate the physical delay. The red dashed curves and circles depict mean normalized reproduction errors as a function of trial sequence and the onset-delay (up-panel) or offset-delay (middle panel). In the low-panels mean normalized reproduction errors (and associated standard errors) are plotted against baseline, delay and post phase for the onset- and offset-delay conditions.

in the post phase ($p < 0.05$); there was no difference between the latter two phases ($p = 1.00$). The increased underestimation due to the offset-delay manipulation is again comparable to the results of Experiment 1. Sequential-trial analysis revealed both the first and the second trial to differ significantly from the remaining trials in the delay phase [first: $t_{(13)} = 2.58$, $p < 0.05$; second: $t_{(13)} = 5.03$, $p < 0.01$]. In the post phase, normalized reproduction errors did not change over trials ($p > 0.1$). Trial-wise comparisons of delay- and post-phase reproduction errors yielded no significant differences (all $p > 0.1$). Thus, participants either needed more than four trials to readjust their reproduction performance to the synchronous-feedback, or normalized reproduction errors were too variable within trials. However, for the auditory-visual experiment, the offset-delay manipulation did not influence the reproduction performance, $F_{(2,\ 26)} = 0.95$, $p = 0.40$. None of the phases differed from any other (all $p > 0.1$). This result is similar to that obtained in Experiment 2.

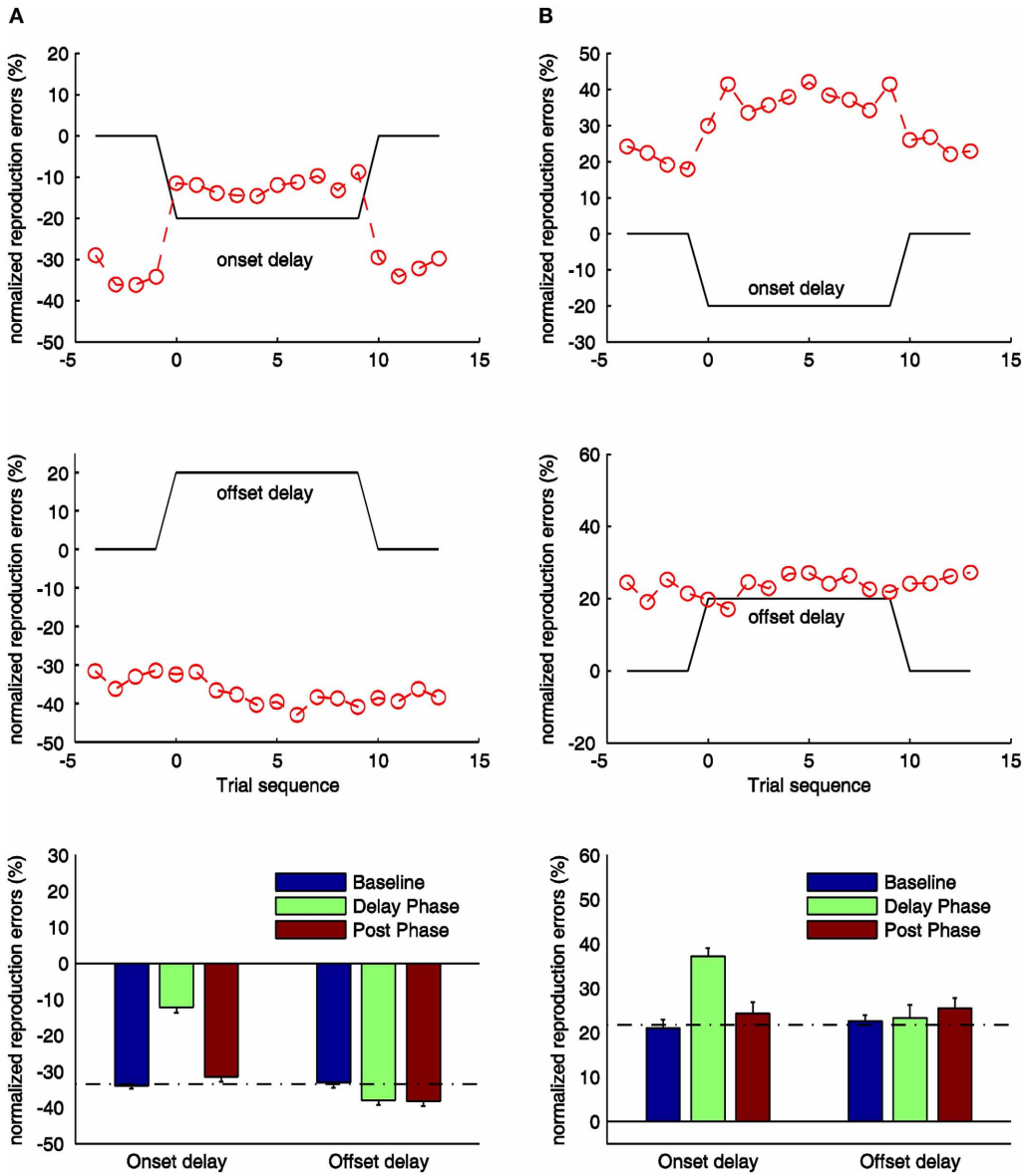


FIGURE 4 | Normalized reproduction errors for the onset- and offset-delay condition of Experiment 3 (A) and Experiment 4 (B). In the upper and middle panels trial-wise dynamic changes of normalized reproduction are shown. Four trials from the synchronous block before the delay manipulation (baseline phase), delay block (delay phase), and four trials after the delay manipulation (post phase) are displayed. The black lines indicate the physical delay. The red dashed curves and circles depict mean normalized reproduction errors as a function of trial sequence and the onset-delay (up-panel) or offset-delay (middle panel). In the low-panels mean normalized reproduction errors (and associated standard errors) are plotted against baseline, delay, and post phase for the onset- and offset-delay conditions. The dashed line indicates the mean normalized reproduction error in the baseline condition.

EFFECTS OF RANDOM ONSET- AND OFFSET-MANIPULATION ON THE DURATION REPRODUCTION

Figure 5 illustrates relationships between the reproduction error and the relative feedback onset (left panel) and offset (right panel) for a typical participant. For the onset-manipulation condition, there was a significant correlation between positive feedback delays and reproduction errors (correlation coefficient: 0.41, linear slope: 0.89, all $p < 0.05$). The steep slope indicates an about 89% compensation for the delayed onset in the duration reproduction, which was similar to the finding in Experiment 1. However, such correlation was broken down when the feedback was presented before participants' actions. There was no correlation [mean: 0.1, $t_{(12)} = 0.81$, $p = 0.43$] for those "preceded" feedback trials, and the mean slope (0.17) did not significantly differ from zero, $t_{(12)} = 0.90$, $p = 0.39$. For the offset-manipulation condition, the correlation between reproduction errors and random offsets was mildly related, mean correlation coefficient 0.31, $t_{(12)} = 6.53$, $p < 0.05$. The mean slope (0.3) was significant higher than zero, $t_{(12)} = 8.31$, $p < 0.05$, though it was significantly lower than the mean slope of the "delayed" onset

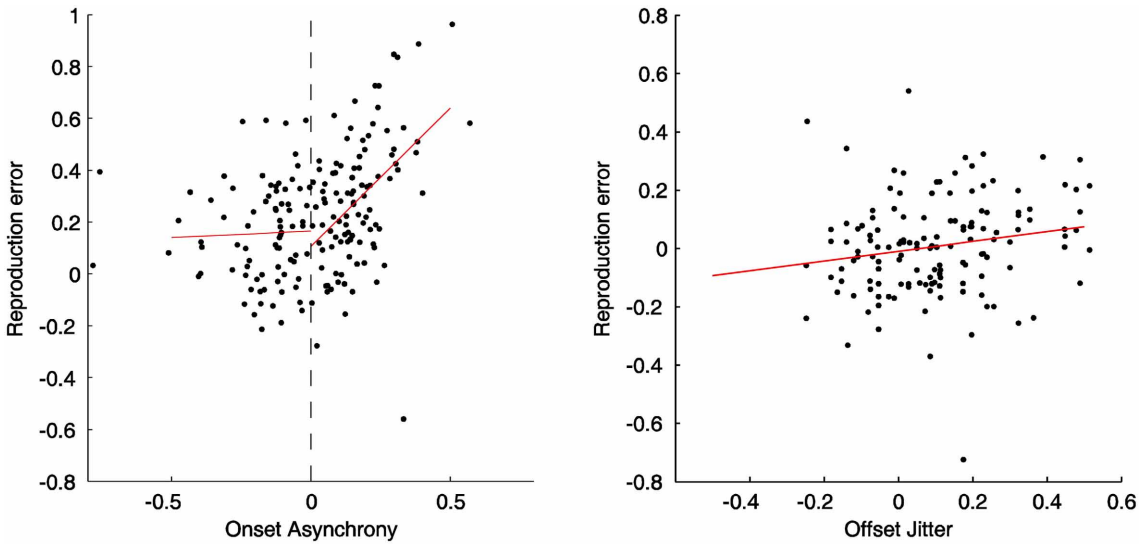


FIGURE 5 | Normalized reproduction errors and linear regression lines (red) for the onset (left side) and offset (right side) manipulation conditions from a typical dataset. In the onset-manipulation condition, the fitted slope for the feedback signal started before the action onset (0.06) is not significant different from zero, while the slope for the delayed feedback (1.07) is significant higher than zero. In the offset-manipulation condition the slope is 0.30, significantly higher than zero.

condition, $t_{(12)} = 3.83$, $p < 0.05$. The mild offset modulation confirmed the findings in Experiments 1 and 3.

DISCUSSION

The results of the present study illustrate how the onset- and offset-manipulation of the feedback signal influences the duration reproduction. In all experiments, we found an increase in duration reproduction for conditions with positive onset-delay feedback manipulation. The lengthening of the reproduced duration could almost compensate the onset-delay (about 90% for the auditory feedback and 75–90% for the visual feedback). The subjective lengthening started immediately with the first trial (or second in Experiment 4), and ended with the last trial of the delay phase. Despite our explicit instruction for reproducing the standard duration regardless the feedback signal, the reproduced duration was still heavily influenced by the onset of the delayed feedback. However, such influence was broken down when the feedback signal was presented before participant's button press.

The results suggest that the action-effect causal relationship may play a critical role in the duration reproduction. Through prior experience, we have learnt that the effect of an action is not always immediate (Pesavento and Schlag, 2006). For example, the response of a tap on the computer keyboard becomes visible as a letter on the screen only after a delay of some 20–50 ms, and the response of a remote control might even be slower (Rank et al., 2010; Shi et al., 2010a,b; Sugano et al., 2010). The action-effect causal relationship may lead to bind and recalibrate motor-sensory timing (Cunningham et al., 2001; Stetson et al., 2006), to attract a voluntary action toward its sensory effect (Haggard et al., 2002; Engbert et al., 2007, 2008), and to shift attention toward to the sensory feedback (Buehner and Humphreys, 2009). Such causal binding may well relate to the memory-mixing model (Gu and Meck, 2011). Due to limited capacity of working memory and the cause-effect relationship, motor timing, and *caused*-feedback timing may share the same representation, which pulls both onsets closer. Other studies have also shown similar binding and regression effects in the reproduction task (Teghtsoonian and Teghtsoonian, 1978; Lejeune and Wearden, 2009; Jazayeri and Shadlen, 2010). For example, participants are able to use temporal context (such as mean duration) to reduce variability of their performance by sacrificing accuracy during a reproduction task (Lejeune and Wearden, 2009; Jazayeri and Shadlen, 2010). However, when the causal relationship is violated (i.e., the feedback was prior to the action in Experiment 5), linkage between two events—the action and sensory feedback—becomes weak, which leads to less memory interference between the two representations. The causal binding and memory-mixing could also explain the quick adjustment to the onset-delay, since the binding and immediate adjustment of the reproduction can take place in the same trial.

In contrast to the effects of introducing feedback onset-delays, offset-delay manipulation appears to modulate duration reproduction in a modality-dependent manner, though with comparatively small effects. Duration reproduction for the auditory offset-feedback delay (Experiments 1, 3, and 5) was shortened by only some 25–30% of the delay manipulation, while there was no shortening effect for the visual offset-delay manipulation. The latter was probably due to sluggish visuomotor timing (Jäncke et al., 2000; Repp, 2005). With the auditory offset-delay manipulation, the shortening effect became manifested not on the first trial with a delay, but only on the second or third trial. Similarly, the shortening effect diminished more gradually after the removal of the delay (after one trial in Experiment 1 and probably more than four trials in Experiment 3). This dynamic

adaptation is comparable to previously observed adaptive changes in synchrony perception (Fujisaki et al., 2004; Vroomen et al., 2004). Also, the amount of adaptation (25% of the auditory offset-delay manipulation) resembles previously reported shifts in PSEs for point-in-time calibration [e.g., 10% for multisensory adaptation (Fujisaki et al., 2004; Di Luca et al., 2009), and 29% for sensorimotor adaptation (Sugano et al., 2010)]. The partial compensation has been attributed to the fact that the brain takes into account a long history of "veridical" sensory inputs throughout lifetime, as compared to only a short adaptation phase during typical psychophysical experiments (Fujisaki et al., 2004). Similar in our study, the asynchrony between the end of an action and the end of the *auditory* feedback may be used as an error signal (Shadmehr et al., 2010) for sensorimotor adaptation to partially adjust future actions. As suggested by the memory-mixing account (Gu and Meck, 2011), participants may use the representation of previous experienced offset-delay for predicting a potential delay on a given offset-manipulation trial.

Mild partial compensation also suggests that participants trust their own stop signal more than the delayed offset signal. This may relate to the switch of the internal clock model (Gibbon, 1977; Gibbon et al., 1984), which consists of a pacemaker emitting pulses at a certain rate and a mode switch that can open and close to permit an accumulator to collect emitted pulses. When the switch closes, the number of pulses in the accumulator is compared against a reference time from memory. Larger amounts of accumulated pulses mean longer estimated durations. Recent striatal beat-frequency (SFB) model provides a neurobiological plausible model of interval timing and switch (Matell and Meck, 2004), which suggests timing is based on the coincidental activation of medium spiny neurons in the basal ganglia by cortical neural oscillators. At trial onset the synchronization of cortical oscillators is triggered by the dopaminergic burst, and at expected offset a burst is reflected on cortico-striatal transmission (see review Buhusi and Meck, 2005). It has been shown that neurons in the motor cortex increase their synchrony when animals are trained to expect an action (Riehle et al., 1997). The synchronization triggered by the expected stop-action might be considered as the more reliable switch-off signal than the offset of the external sensory feedback, leading to the offset-delay interval being largely neglected and to less memory-mixing than during the onset condition. This could also explain the findings in Experiment 5, where the feedback offset was random and unreliable.

In Experiments 3 and 4, in which the standard duration and the feedback signal were presented in different modalities, we observed a strong distortion of perceived durations: visual standard durations were strongly underestimated by presentation of auditory feedback signals during the reproduction, and this finding was mirrored by a strong overestimation of auditory standard durations when the feedback signal was a visual stimulus. The over- and underestimations across the audiovisual modalities are analog to previous findings. For example, Wearden et al. (1998) have provided evidence that the auditory pacemaker ticks faster than the visual pacemaker, as a result of which auditory durations are perceived as longer than physically equivalent visual durations. However, it remains an open question whether the observed audiovisual effects are mainly caused by the crossmodal memory-mixing. Nevertheless, recall that the overestimation (underestimation) was additive to the effects of delay manipulation, which suggests that the crossmodal standard-feedback signals comparison (i.e., presenting a standard stimulus in one modality and providing a feedback signal stimulus in another modality) is operating mainly on the perceptual level, relatively independent of sensorimotor adjustments.

CONCLUSION

In summary, the present study investigated the effects of feedback signal delay manipulation on active duration reproduction. When the onset of sensory feedback signals was delayed, reproduced durations lengthened immediately to compensate for the feedback signal delays in large proportion. The feedback before action onset was neglected. However, when the offset of sensory feedback signals was delayed, reproduced durations only shortened by about 25–30% of the delay with auditory feedback signals, while there was no compensation for visual feedback signals. These results suggest that active duration reproduction is heavily mixed with the delayed feedback onset and mildly influenced by the feedback offset. The results can be explained with causal binding and the memory-mixing accounts. Moreover, the observed under- and overestimation due to crossmodal manipulation of the standard and feedback signal stimuli is additive to the sensorimotor delay adaptation.

ACKNOWLEDGMENTS

This research was conducted at the LMU in Munich and was supported in part by German Research Council (DFG) project grant SH166 to Zhuanghua Shi and Hermann J. Müller. The authors express their gratitude to Kristian Hristov for his assistance in carrying out the experimental work. We thank the reviewers for insightful suggestions.

REFERENCES

Angrilli, A., Cherubini, P., Pavese, A., and Manfredini, S. (1997). The influence of affective factors on time perception. *Percept. Psychophy.* 59, 972–982.

Brainard, D. H. (1997). The psychophysics toolbox. *Spat. Vis.* 10, 433–436.

Buehner, M. J., and Humphreys, G. R. (2009). Causal binding of actions to their effects. *Psychol. Sci.* 20, 1221–1228.

Bueti, D., and Walsh, V. (2010). Memory for time distinguishes between perception and action. *Perception* 39, 81–90.

Buhusi, C. V., and Meck, W. H. (2005). What makes us tick? Functional and neural mechanisms of interval timing. *Nat. Rev. Neurosci.* 6, 755–765.

Burr, D., Banks, M. S., and Morrone, M. C. (2009). Auditory dominance over vision in the perception of interval duration. *Exp. Brain Res.* 198, 49–57.

Campbell, W. W., Ward, L. C., and Swift, T. R. (1981). Nerve conduction velocity varies inversely with height. *Muscle Nerve* 4, 520–523.

Chen, K.-M., and Yeh, S.-L. (2009). Asymmetric cross-modal effects in time perception. *Acta Psychol.* 130, 225–234.

Cunningham, D. W., Billock, V. A., and Tsou, B. H. (2001). Sensorimotor adaptation to violations of temporal contiguity. *Psychol. Sci.* 12, 532.

Di Luca, M., Machulla, T.-K., and Ernst, M. O. (2009). Recalibration of multisensory simultaneity: cross-modal transfer coincides with a change in perceptual latency. *J. Vis.* 9, 1–16.

Eagleman, D. M. (2008). Human time perception and its illusions. *Curr. Opin. Neurobiol.* 18, 131–136.

Eagleman, D. M., and Holcombe, A. O. (2002). Causality and the perception of time. *Trends Cogn. Sci.* 6, 323–325.

Engbert, K., Wohlschläger, A., and Haggard, P. (2008). Who is causing what? The sense of agency is relational and efferent-triggered. *Cognition* 107, 693–704.

Engbert, K., Wohlschläger, A., Thomas, R., and Haggard, P. (2007). Agency, subjective time, and other minds. *J. Exp. Psychol. Hum. Percept. Perform.* 33, 1261–1268.

Fujisaki, W., Shimojo, S., Kashino, M., and Nishida, S. (2004). Recalibration of audiovisual simultaneity. *Nat. Neurosci.* 7, 773–778.

Gibbon, J. (1977). Scalar expectancy theory and Weber's law in animal timing. *Psychol. Rev.* 84, 279–325.

Gibbon, J., Church, R. M., and Meck, W. H. (1984). Scalar timing in memory. *Ann. N.Y. Acad. Sci.* 423, 52–77.

Gu, B.-M., and Meck, W. (2011). "New perspectives on Vierordt's law: memory-mixing in ordinal temporal comparison tasks," in *Multidisciplinary Aspects of Time and Time Perception, Vol. 6789,* eds A. Vatakis, A. Esposito, M. Giagkou, F. Cummins, and G. Papadelis (Berlin, Heidelberg: Springer), 67–78.

Haggard, P., Clark, S., and Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nat. Neurosci.* 5, 382.

Hanson, J. V. M., Heron, J., and Whitaker, D. (2008). Recalibration of perceived time across sensory modalities. *Exp. Brain Res.* 185, 347–352.

Harrar, V., and Harris, L. R. (2008). The effect of exposure to asynchronous audio, visual, and tactile stimulus combinations on the perception of simultaneity. *Exp. Brain Res.* 186, 517–524.

Heron, J., Hanson, J. V. M., and Whitaker, D. (2009). Effect before cause: supramodal recalibration of sensorimotor timing. *PLoS ONE* 4:e7681. doi: 10.1371/journal.pone.0007681

Jäncke, L., Loose, R., Lutz, K., Specht, K., and Shah, N. J. (2000). Cortical activations during paced finger-tapping applying visual and auditory pacing stimuli. *Brain Res. Cogn. Brain Res.* 10, 51–66.

Jazayeri, M., and Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nat. Neurosci.* 13, 1020–1026.

Kennedy, J. S., Buehner, M. J., and Rushton, S. K. (2009). Adaptation to sensory-motor temporal misalignment: instrumental or perceptual learning? *Q. J. Exp. Psychol.* 62, 453–469.

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.

Klink, P. C., Montijn, J. S., and van Wezel, R. J. A. (2011). Crossmodal duration perception involves perceptual grouping, temporal ventriloquism, and variable internal clock rates. *Attent. Percept. Psychophys.* 73, 219–236.

Lejeune, H., and Wearden, J. (2009). Vierordt's the experimental study of the time sense (1868) and its legacy. *Eur. J. Cogn. Psychol.* 21, 941–960.

Matell, M. S., and Meck, W. H. (2004). Cortico-striatal circuits and interval timing: coincidence detection of oscillatory processes. *Cogn. Brain Res.* 21, 139–170.

Meck, W. H. (1996). Neuropharmacology of timing and time perception. *Brain Res. Cogn. Brain Res.* 3, 227–242.

Miall, R. C., and Jackson, J. K. (2006). Adaptation to visual feedback delays in manual tracking: evidence against the Smith Predictor model of human visually guided action. *Exp. Brain Res.* 172, 77–84.

Morrone, M. C., Ross, J., and Burr, D. (2005). Saccadic eye movements cause compression of time as well as space. *Nat. Neurosci.* 8, 950–954.

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., and Spence, C. (2005). Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Cogn. Brain Res.* 25, 499–507.

Park, J., Schlag-Rey, M., and Schlag, J. (2003). Voluntary action expands perceived duration of its sensory consequence. *Exp. Brain Res.* 149, 527–529.

Pesavento, M. J., and Schlag, J. (2006). Transfer of learned perception of sensorimotor simultaneity. *Exp. Brain Res.* 174, 435–442.

Purpura, K., Tranchina, D., Kaplan, E., and Shapley, R. M. (1990). Light adaptation in the primate retina: analysis of changes in gain and dynamics of monkey retinal ganglion cells. *Vis. Neurosci.* 4, 75–93.

Rank, M., Shi, Z., Müller, H. J., and Hirche, S. (2010). Perception of delay in haptic telepresence systems. *Presence Teleoper. Virtual Environ.* 19, 389–399.

Regan, D. (1989). *Human Brain Electrophysiology: Evoked Potentials and Evoked Magnetic Fields in Science and Medicine*. New York, NY: Elsevier.

Repp, B. H. (2005). Sensorimotor synchronization: a review of the tapping literature. *Psychon. Bull. Rev.* 12, 969–992.

Riehle, A., Grun, S., Diesmann, M., and Aertsen, A. (1997). Spike synchronization and rate modulation differentially involved in motor cortical function. *Science* 278, 1950–1953.

Shadmehr, R., Smith, M. A., and Krakauer, J. W. (2010). Error correction, sensory prediction, and adaptation in motor control. *Annu. Rev. Neurosci.* 33, 89–108.

Shi, Z., Chen, L., and Müller, H. (2010a). Auditory temporal modulation of the visual Ternus effect: the influence of time interval. *Exp. Brain Res.* 203, 723–735.

Shi, Z., Zou, H., Rank, M., Chen, L., Hirche, S., and Müller, H. J. (2010b). Effects of packet loss and latency on the temporal discrimination of visual-haptic events. *IEEE Trans. Haptics* 3, 28–36.

Shi, Z., Jia, L., and Mueller, H. J. (2012). Modulation of tactile duration judgments by emotional pictures. *Front. Integr. Neurosci.* 6:24. doi: 10.3389/fnint.2012.00024

Stetson, C., Cui, X., Montague, P. R., and Eagleman, D. M. (2006). Motor-sensory recalibration leads to an illusory reversal of action and sensation. *Neuron* 51, 651–659.

Sugano, Y., Keetels, M., and Vroomen, J. (2010). Adaptation to motor-visual and motor-auditory temporal lags transfer across modalities. *Exp. Brain Res.* 201, 393–399.

Takahashi, K., Saiki, J., and Watanabe, K. (2008). Realignment of temporal simultaneity between vision and touch. *Neuroreport* 19, 319–322.

Teghtsoonian, R., and Teghtsoonian, M. (1978). Range and regression effects in magnitude scaling. *Percept. Psychophys.* 24, 305–314.

van Wassenhove, V., Buonomano, D. V., Shimojo, S., and Shams, L. (2008). Distortions of subjective time perception within and across senses. *PLoS ONE* 3:e1437. doi: 10.1371/journal.pone.0001437

Vroomen, J., Keetels, M., de Gelder, B., and Bertelson, P. (2004). Recalibration of temporal order perception by exposure to audio-visual asynchrony. *Cogn. Brain Res.* 22, 32–35.

Walker, J. T., and Scott, K. J. (1981). Auditory–visual conflicts in the perceived duration of lights, tones, and gaps. *J. Exp. Psychol. Hum. Percept. Perform.* 7, 1327–1339.

Wearden, J. H., Edwards, H., Fakhri, M., and Percival, A. (1998). Why 'sounds are judged longer than lights': application of a model of the internal clock in humans. *Q. J. Exp. Psychol. B* 51B, 97–120.

Wenke, D., and Haggard, P. (2009). How voluntary actions modulate time perception. *Exp. Brain Res.* 196, 311–318.

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: 14 May 2012; accepted: 27 September 2012; published online: 16 October 2012.

*Citation: Ganzenmüller S, Shi Z and Müller HJ (2012) Duration reproduction with sensory feedback delay: differential involvement of perception and action time. Front. Integr. Neurosci. **6**:95. doi: **10.3389/fnint.2012.00095***

APPENDIX

INSTRUCTION

In this experiment your task is to reproduce the duration of a tone by pressing a button. For each trial, you will first hear a tone for a certain duration. Please try to memorize the temporal information as accurately as possible! As soon as the tone stops, you are asked to press the button in front of you for as long as you heard the tone before. It is important for the experiment that you reproduce the duration of the first tone as accurately as possible! While you press the button, another tone will be presented. This tone could be either dependent or independent of your button press. Therefore, please try to reproduce the duration of the first tone, regardless of the second tone! There will be a practice block in the beginning for familiarization with the task. After the practice block the actual experiment will be started automatically. There will be 10 blocks for the whole experiment, which lasts about 45 min.

frontiers in
INTEGRATIVE NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 26 September 2012
doi: 10.3389/fnint.2012.00083

The hand-reversal illusion revisited

Sang W. Hong[1]*, Linda Xu[2], Min-Suk Kang[3] and Frank Tong[3]

[1] *Department of Psychology, Florida Atlantic University, Boca Raton, FL, USA*
[2] *Vanderbilt Center for Science Outreach, Vanderbilt University, Nashville, TN, USA*
[3] *Department of Psychology, Vanderbilt University, Nashville, TN, USA*

Edited by:
Zhuanghua Shi, Ludwig-Maximilians-Universität München, Germany

Reviewed by:
Hao Zhang, Duke University, USA
Zhuanghua Shi, Ludwig-Maximilians-Universität München, Germany

***Correspondence:**
Sang W. Hong, Department of Psychology, Florida Atlantic University, BS 209, 777 Glades Rd., Boca Raton, FL 33431, USA.
e-mail: shong6@fau.edu

The hand-reversal illusion is a visuomotor illusion that is commonly seen in children's play. When participants attempt to lift a designated finger while their hands are cross-folded, they are likely to erroneously lift the matched finger of the other hand; however, such errors are rare when subjects close their eyes. Based on the fact that the illusion disappears without visual input, researchers previously concluded that the illusion depends upon visual and proprioceptive conflict (Van Riper, 1935). Here, we re-evaluated this visual-proprioceptive conflict hypothesis by obtaining reaction time measurements because, in the original study, subjects might have relied on a strategy of responding more slowly to minimize making errors. We found that the impairment due to cross-folding one's hand persisted in the absence of the visual input, as evidenced by delayed response times (RTs). Further, we found that such impairment occurred when the fingers of only one hand were tested, indicating that the impairment was not due to left-right confusions of the hands during tactile identification or response selection. Based on these results, we suggest that the illusion is not solely due to the conflict between visual and proprioceptive information. Instead, we propose that the unusual configuration itself that involves a reversal of the left and right hands in external space also contributes to the impaired motor response.

Keywords: hand-reversal illusion, visuo-tactile-motor interaction, multisensory perception, proprioception, remapping

INTRODUCTION

One of the most important goals of sensory processing is to guide action. For example, the execution of a goal-directed movement such as grasping or pointing requires the subject to determine the location, size, and shape of the target object through sensory processing (for review, Goodale and Servos, 1996). Coordination of vision and proprioception is crucial for goal-directed hand movements (Rossetti et al., 1995; van Beers et al., 1999). Specifically, it has been shown that integration of both visual and proprioceptive information improves spatial localization performance (van Beers et al., 1999). When visual and proprioceptive information about hand position is in conflict, as can be induced by placing a wedge prism in front of the subject's eyes, the subject perceives the hand position somewhere between the vision-based and the proprioception-based location, slightly closer to the vision-based position (Pick et al., 1969; Warren, 1980; Touzalin-Chretien et al., 2010).

The hand-reversal illusion, originally called the "Japanese Illusion" (Burnett, 1904; Klein and Schilder, 1929; Van Riper, 1935) has been suggested to provide a compelling example of the importance of multisensory integration in making simple hand movements, such as lifting a finger. After folding the two hands naturally, as shown in the right panel of **Figure 1A**, a participant can easily lift the index finger of the left hand upon instruction. In contrast, when the two hands are cross-folded, as shown in the left panel of **Figure 1A**, the participant often lifts the right-hand index finger when visually instructed to lift the index finger of the left hand. This type of error in motor behavior was believed to occur because of conflict between visual and proprioceptive information: all right-hand fingers appear to belong to the left-hand and vice versa, even though one knows that the positions of the two hands have been reversed and folded based on proprioceptive information. Consistent with this hypothesized conflict between vision and proprioception, it was reported that errors were virtually eliminated if conflicting visual information was prevented by blindfolding the participant, and instruction was given solely by touching the designated finger (Burnett, 1904; Van Riper, 1935). The importance of vision in body representation can be also found in the "rubber-hand illusion" (Botvinick and Cohen, 1998; Ehrsson et al., 2004) and in mirror therapy for "phantom limb" pain (Ramachandran and Rogers-Ramachandran, 1996). For example, in clinical trials of mirror therapy, patients with traumatic amputations reported vivid kinesthetic and somatic sensations in the missing hand when looking at the mirror image of the intact hand when instructed to perform coordinated bimanual movements.

There could be, however, an alternative explanation for the hand-reversal illusion. It is possible that the hand-reversal illusion is simply due to the greater confusability of finger representations of the two hands in the cross-folded configuration, rather than the conflict between visual and proprioceptive information. It has been shown that unnatural configurations of fingers and hands,

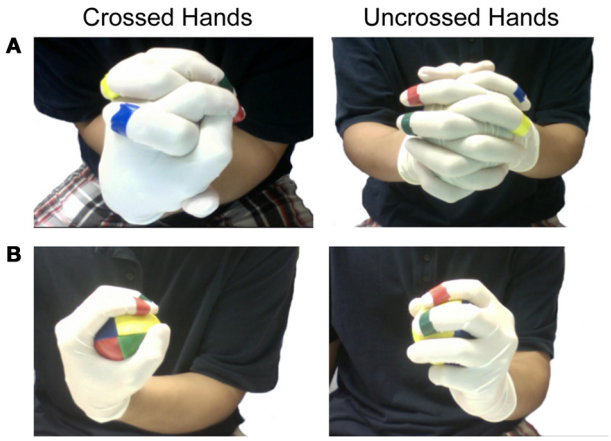


FIGURE 1 | Hand and finger postures. (A) Example of the hand and finger posture for the Hand-Reversal Illusion (left) and a naturally folded hand posture (right). **(B)** Hand and finger postures with the ball in Experiment 3.

such as interweaving fingers (Zampini et al., 2005; Haggard et al., 2006; Riemer et al., 2010; Overvliet et al., 2011) and crossing hands (Benedetti, 1985; Yamamoto and Kitazawa, 2001; Heed et al., 2012), can affect tactile localization on the fingers and hands. The effect of the configuration of fingers and hands on localization and identification of touch has been observed both with (Haggard et al., 2006) and without (Riemer et al., 2010; Overvliet et al., 2011) visual information, indicating that visual information may not be the main source of confusion. Instead, these studies suggest that the effect of unnatural body configuration on localization of touch is due to the conflict between the somatotopic body coordinate and the external spatial coordinate. For example, when two hands are crossed-folded (see **Figure 1A**, left), the right hand belongs to the right side of the body in terms of somatotopic coordinates but is located on the left side in terms of external spatial coordinates. According to a somato-perceptual information processing model (Longo et al., 2010), the spatial location of the finger within the body schema should be first identified, which is achieved based on somatotopic organization, and then transformed to an external spatial representation to execute a finger movement. This remapping of the representation of the body part to the external spatial representation is important for performing goal-directed movements such as reaching and pointing (Sarlegna et al., 2009), since such actions require localization of both the target object and the hands in the external three-dimensional space.

In the current study, we were specifically interested in which of three factors might cause the hand-reversal illusion. One possibility, as was originally proposed by Van Riper (1935), is that potentially confusing visual information in the crossed-hands position leads to slower response times (RTs) and more errors. Another possibility is that unnatural hand configuration itself (switch between left and right body part) impairs a person's ability to localize the touched digit or to make the relevant motor action in response to that touch due to greater confusability between left and right hands. Lastly, the impairment may be due to the mismatch between the somatotopic body representation and the external spatial representation (Longo et al., 2010), as suggested by previous research on tactile localization (Haggard et al., 2006; Riemer et al., 2010; Overvliet et al., 2011). We conducted three experiments to distinguish between these possible accounts.

In our experiments, we re-evaluated the illusion by obtaining reaction time measurements because, in the original study, subjects might have relied on a strategy of responding more slowly to minimize making errors. If potentially confusing visual information is not the main cause of the illusion, we should be able to observe evidence of the illusion that is slower responses with cross-folded hands, even when conflicting visual input is eliminated. We used only tactile cues (tapping the designated finger) to directly compare the results from different visual conditions (i.e., with vs. without input). In the second experiment, we examined whether RT delays in the crossed-hands configuration was attributable to left-right confusions during response selection, by testing fingers from only one hand. Moving the finger that was touched requires localization of the tactile input, response selection of the finger to move, and execution of the movement. RT delays might occur at any of these processing stages. By testing only a single hand in Experiment 2, we minimized the potential for left-right confusion between the hands at the stage of both identification and response selection. If delays in RT are still observed in the crossed-hands position, such a result would indicate that the impairment is unlikely to reflect confusion at these stages. In the third experiment, we determined whether RT delays in the crossed-hands configuration might simply be due to the unnatural posture of the hands and fingers, by testing only one hand. Note that both hands were used for cross-folding but fingers from only one hand were tested in the second experiment. If confusability of the left and right hands is the primary cause of RT delays in the crossed-hands configuration, then no impairment should occur when only a single hand makes a similar unnatural configuration.

MATERIALS AND METHODS

PARTICIPANTS

Twenty healthy adult volunteers with normal or corrected-to-normal vision participated in the experiments. Each participant took part in two of the three experiments. All participants provided informed consent to participate in the study, which was approved by the Vanderbilt University Institutional Review Board.

PROCEDURE

For the experiment, the participant sat between two desks that each supported a MacBook Pro 13" computer used for video recording (PhotoBooth software) the participant's hands from both sides. The participant wore latex gloves, and each of the index and middle fingers were marked with a unique color band for experimental coding (**Figure 1**). This experiment focused on the index and middle fingers exclusively, as it proved more difficult to move the third or fourth digits independently, especially when the hands were positioned in the reversed configuration.

An experimenter stood in front of the participant and waited for the instruction generated by a separate MacBook Pro 15" computer, which indicated the specific finger to be tested on each trial. When the experiment began, an assistant started video recording the participant's arms, hands, and fingers; the camera viewpoint was adjusted so that other body parts remained out of view, including the face. The recording frequency was set to 30 frames per second. The assistant also started a computer program that showed the experimenter which finger to tap on each trial, by presenting pictures of the hands and the designated colors. An auditory beep occurred 3 s after the picture was displayed to the experimenter, cuing the experimenter to use a hard plastic pen tip to touch the relevant finger between the first joint from the fingertip and the second joint. The beep also served as a temporal cue to prepare the participant, and was presented in every experimental condition. The participant's task was to lift the tapped finger as quickly as possible without making errors.

RT for lifting a finger on a given trial was measured by counting video frames. Each frame was calculated as 33 ms with a 30 frame/s recording frequency. The starting point for counting frames was defined as time when the pen tip first touched the finger, and the end point was defined as the first frame that showed a finger rising away from the back of the folded hands. The frame count included both the starting and end points. RT was calculated by multiplying the frame count by 33 ms. Since overall RTs differed considerably across participants, RTs were normalized by each individual's mean reaction time in the experiment, resulting in a value greater than 1 for slower responses and a value lower than 1 for faster responses.

The experiment consisted of a 2 × 2 design, with the participant's hands arranged in a normal or reversed configuration and the eyes open or closed. In the eyes-open condition, participants were asked to look steadily at their hands; in the eyes-closed condition, participants were instructed to keep their eyes closed. In the uncrossed hands condition (**Figure 1A**, right), participants folded their two hands naturally. In the reversed hands configuration (**Figure 1A**, left), they crossed the wrists, interlaced the fingers with the thumbs pointing downwards, and then turned the arms and hands in and around toward the body until the fifth fingers were closest to the body and the thumbs pointed upwards and outwards. The RT for lifting the fingers was measured eight times for each of the four fingers and the order of the 32 finger taps in each condition was completely randomized. It took about 5 min to complete one condition.

In Experiment 1, we tested the original hand-reversal illusion with two hand positions: cross-folded hands (**Figure 1A**, left), and uncrossed hands (**Figure 1A**, right). In Experiment 2, participants were instructed to make the same hand positions as in Experiment 1, but the fingers of only one hand were tested to minimize potential left-right confusion between the hands. In Experiment 3, participants grasped a ball with their dominant hand. In this position, participants mimicked the cross-folded and uncrossed hand positions (**Figure 1B**). By using only one hand, we could further remove left-right confusion, so that we were able to test whether simple conflict between the somatotopic representation of a single hand and its position in external space would be sufficient to induce the impairment. The two hand positions (cross-folded and uncrossed) were tested in combination with the two visual conditions (eyes open and closed) for all three experiments, resulting in a total of 12 different conditions. Each subject performed two (out of three) randomly chosen experiments. Within each experimental session, the order of four conditions was randomized.

RESULTS

We measured how long it took participants to move the relevant finger that was tapped under the following conditions, with eyes opened or closed and with hands folded in a normal or reversed configuration. Normalized RTs are shown in **Figure 2A**. The RT was measured 32 times for each condition, and only correct responses (lifting the finger indicated by the experimenter) were used for analysis. Incorrect responses were very

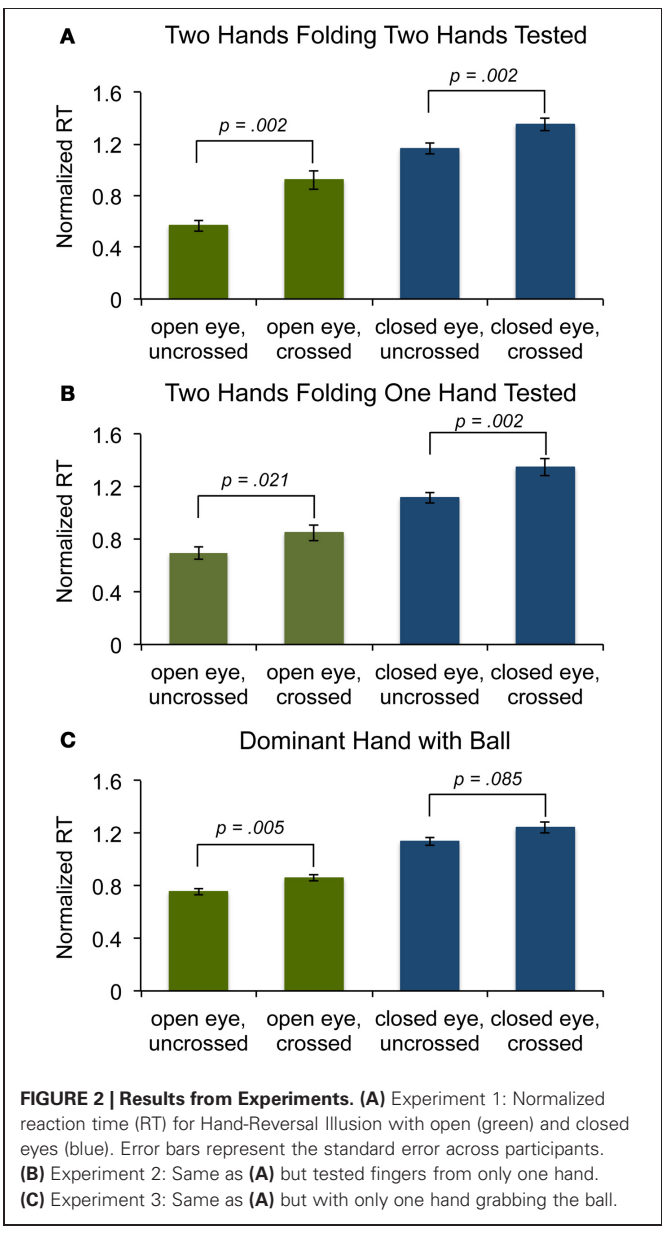


FIGURE 2 | Results from Experiments. (A) Experiment 1: Normalized reaction time (RT) for Hand-Reversal Illusion with open (green) and closed eyes (blue). Error bars represent the standard error across participants. **(B)** Experiment 2: Same as **(A)** but tested fingers from only one hand. **(C)** Experiment 3: Same as **(A)** but with only one hand grabbing the ball.

rare for all participants (one or two errors, if any, for the cross-folded hand condition). The very low frequency of errors was due to the fact that the relevant finger was directly touched in these experiments.

A Two-Way analysis of variance (ANOVA) within-subject design revealed a significant main effect of both hand position (cross-folded and uncrossed) [$F_{(1, 9)} = 34.24, p < 0.01$] and eye condition (open and closed) [$F_{(1, 9)} = 34.74, p < 0.01$]. The faster RTs in the open-eye condition might reflect an overall advantage of multi-sensory information on localization performance (Kennett et al., 2001; Forster et al., 2002). However, we did not find evidence of a significant interaction between two conditions [$F_{(1, 9)} = 3.41$, $p = 0.10$], indicating that the hand-reversal illusion occurred regardless of whether the eyes were open or closed and that the conflict visual information did not provide additional source of confusion when tactile cue was used. A planned orthogonal contrast confirmed that RTs were significantly slower in the hands-crossed configuration than in the normal configuration, both when participants had their eyes open [$F_{(1, 9)} = 19.22, p < 0.01$] and when their eyes were closed [$F_{(1, 9)} = 18.61, p < 0.01$].

The visual conflict hypothesis cannot explain these results, as we find that the hand-reversal illusion, in terms of RT, persists without visual input. Instead, this result indicates that the unnatural hand posture itself causes difficulty in localizing or responding with the designated finger to lift. This proved true even though the designated finger was directly touched, thereby providing unambiguous information about the relevant finger to move. Previous studies on tactile localization with unnatural hand configuration have found that mismatches between the body schema representation for the left and right hands, and their location in external space (left and right side from the body center) may cause difficulty in one's ability to localize a tactile stimulus.

Lifting an indicated finger involves multiple stages of processing, including identification of tactile input, response selection and execution of finger movement. RT delays could occur at any or all processing stages. To examine whether the confusion occurs at the stage of tactile identification or response selection or execution of movement, we measured RTs to test probes presented exclusively to the fingers of just one hand. With this experiment, potential left-right hand confusion at the level of tactile identification and response selection can be minimized, since participants have to use fingers from only one hand. A Two-Way within-subject ANOVA revealed that the main effect of the hand position was significant [$F_{(1, 9)} = 17.84$, $p < 0.01$], indicating that slower RTs in crossed-hands configuration was unlikely to reflect confusions at the processing stage of identification or response selection (**Figure 2B**). A planned orthogonal contrast confirmed that RTs were significantly slower in the hands-crossed configuration than in the normal configuration for both eye conditions (open: [$F_{(1, 9)} = 7.81, p < 0.05$] and closed: [$F_{(1, 9)} = 19.27, p < 0.01$]).

In the crossed hands configuration with participants' own two hands, the impairment in response latency could occur due to the confusion in the bodily representation of handedness (left-right hands). Alternatively, it remains a possibility that simple conflict between the somatotopic representation of a single hand and its position in external space would be sufficient to induce the impairment without confusion in left-right hands. To address this potential concern, we conducted a control experiment that required participants to use only a single hand. We hypothesized that the use of a single hand should further minimize the left-right confusions when participants attempted to plan the correct motor action, but conflict between somatotopic representation of the hand and its position in external space remains. Participants were instructed to make a hand posture similar to that required for Experiment 1, by holding a ball in the palm of their dominant hand (**Figure 1B**). We tested participants' dominant hand since we found no difference in reaction times between the two hands for the participants in Experiment 1. A Two-Way within-subject ANOVA revealed a significant main effect of both eye condition [$F_{(1, 9)} = 73.16$, $p < 0.001$] and hand position [$F_{(1, 9)} = 17.81$, $p < 0.01$], indicating that the impairment in finger responses still occurred when only one hand was positioned in a reversed configuration over the body midline (**Figure 2C**). A planned contrast showed that, however, RT was significantly slower for crossed hand condition with open eye [$F_{(1, 9)} = 13.48$, $p < 0.01$]. In closed eye condition, RT was slow overall for crossed hand but was not significant [$F_{(1, 9)} = 3.73, p = 0.085$].

DISCUSSION

In this study, we investigated possible explanations for the hand-reversal illusion, also known as Japanese Illusion, other than visuo-proprioceptive conflict hypothesis. When RTs, instead of error rate, were measured, we found that the impairment in finger response persisted even after conflicting visual information was eliminated. This result indicates that, contrary to the long-standing belief, the conflict between visual and proprioceptive information is not the only cause of the illusion. Instead, we propose that the hand-reversal illusion can be understood within the same framework that explains the effect of various unnatural hand configurations on tactile perception.

Previous research has shown that temporal order judgments (TOJ) are less precise for tactile stimulation delivered to the two hands when those hands are crossed and positioned in the contralateral hemifield, compared to when the two hands are normally positioned (Yamamoto and Kitazawa, 2001; Shore et al., 2002; Schicke and Röder, 2006). This hand-crossing effect has been interpreted as evidence of a conflict between the somatotopic body representation (e.g., right hand) and the representation of that body part in external space (e.g., left side of one's body), which can result in a difficulty in tactile processing. Resolution of this conflict may be required to execute a correct action (Heed et al., 2012). Also, interweaving one's fingers can disrupt the precise localization of which finger was touched, the spatial sequence of multiple touches (Haggard et al., 2006; Riemer et al., 2010; Overvliet et al., 2011) and the discrimination of tactile stimulation (Zampini et al., 2005). These previous studies are generally consistent with the present findings.

From this previous work, however, it was not clear whether a reconciliation between somatotopic representation and external spatial representation would be necessary for making a simple finger movement in response to local tactile stimulation.

The present study required participants to respond directly to local touch with a finger movement. Our experimental procedure therefore avoided requiring participants from having to make an explicit judgment regarding the location of the cued body part according to external spatial coordinates. Nevertheless, we observed a behavioral cost in RTs for the crossed-hands position.

Previous studies, however, had not tested whether such conflict would occur when only one hand is located in the contralateral side from the body center. If the mismatch between somatotopic body representation and the external spatial representation is the real cause of the crossing effect, the effect should persist when only one hand crosses over the body midline in external space. Our result support this hypothesis by showing that RTs to the tactile cue are slower even when only one hand was tested, indicating that simple conflict between the somatotopic representation of a single hand and its position in external space is sufficient to induce the impairment.

In the current study, we were able to rule out the processing stages of tactile identification and response selection as a possible locus for the confusion, but it remains to be determined whether the impairment occurs at the stage of motor planning and execution (sending the motor command to move the finger). Of potential relevance, Shore et al. (2002) found that when a cue was provided in the visual domain (i.e., an LED light on the designated finger) instead of the tactile domain (i.e., a tap on the designated finger), the effect of hand-crossing on TOJ was reduced. This result suggests that the left-right confusion occurred at the tactile localization stage rather than at the execution stage. In contrast, the hand-reversal illusion has been found to be stronger when instruction is given through visual rather than tactile cues (Burnett, 1904; Van Riper, 1935). This discrepancy may suggest that the confusion occurs at the stage of motor planning in the hand-reversal illusion. Consistently, we observed delays in RT when any possibility of confusion in the stage of identification and response selection was minimized by testing fingers from only one hand (Experiment 2). The result suggests that behavioral cost in finger movement with cross-folded hands may occur at the stage of motor planning and execution.

The hand-reversal illusion has been believed to occur due to visual-proprioceptive conflict since errors in finger lifting response are virtually abolished without visual input. Here we show that the visual-proprioceptive conflict may not be the only cause of the illusion. Although, it has been reported earlier and replicated here again that errors in lifting indicated fingers almost never occur with tactual cue, our results suggest that the cost of unnatural hand configurations, shown by delays in RT, persists. Unnatural hand configurations can induce impairment in localization of finger with visual cue (Shore et al., 2002) as well as tactile cue. Together with our results, we suggest that the unnatural hand configuration itself (reversal of between left and right body parts), which induces conflict between the somatotopic body representation and representation of the body in the external space, also contributes to the hand-reversal illusion.

ACKNOWLEDGMENTS

This research was made possible by grants from the National Eye Institute R01 EY017082, supplementary funds (R01 EY017082-03S1) provided by the American Recover and Reinvestment Act, as well as the Vanderbilt Center for Science Outreach.

REFERENCES

Azanon, E., and Soto-Faraco, S. (2007). Alleviating the 'crossed-hands' deficit by seeing uncrossed rubber hands. *Exp. Brain Res.* 182, 537–548.

Benedetti, F. (1985). Processing of tactile spatial information with crossed fingers. *J. Exp. Psychol. Hum. Percept. Perform.* 11, 517–525.

Botvinick, M., and Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature* 391, 756.

Burnett, C. T. (1904). Studies in influence of abnormal position on motor impulse. *Pscychol. Rev.* 11, 370–394.

Ehrsson, H. H., Spence, C., and Passingham, R. E. (2004). That's my hand! Activity in premotor cortex reflects feeling of ownership of a limb. *Science* 305, 875–877.

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.

Goodale, M. A., and Servos, P. (1996). "Visual control of prehension," in *Advances in Motor Learning and Control*, ed H. N. Zelanznik (Windsor, ON: Human Kinetics), 87–115.

Haggard, P., Kitadono, K., Press, C., and Taylor-Clarke, M. (2006). The brain's fingers and hands. *Exp. Brain Res.* 172, 94–102.

Heed, T., Backhaus, J., and Röder, B. (2012). Integration of hand and finger location in external spatial coordinates for tactile localization. *J. Exp. Psychol. Hum. Percept. Perform.* 38, 386–401.

Kennett, S., Taylor-Clarke, M., and Haggard, P. (2001). Noninformative vision improves the spatial resolution of touch in humans. *Curr. Biol.* 11, 1188–1191.

Klein, E., and Schilder, P. (1929). The Japanese illusion and the postural model of the body. *J. Nerv. Ment. Dis.* 70, 241–263.

Longo, M. R., Azañón, E., and Haggard, P. (2010). More than skin deep: body representation beyond primary somatosensory cortex. *Neuropsychologia* 48, 655–668.

Overvliet, K. E., Anema, H. A., Brenner, E., Dijkerman, H. C., and Smeets, J. B. J. (2011). Relative finger position influences whether you can localize tactile stimuli. *Exp. Brain Res.* 208, 245–255.

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

Rossetti, Y., Desmurget, M., and Prablanc, C. (1995). Vectorial coding of movement: vision, proprioception, or both? *J. Neurophys.* 74, 457–463.

Ramachandran, V. S., and Rogers-Ramachandran, D. (1996). Synaesthesia in phantom limbs induced with mirrors. *Proc. R. Soc. Lond. B Biol. Sci.* 263, 377–386.

Riemer, M., Trojan, J., Kleinbohl, D., and Hölzl, R. (2010). Body posture affects tactile discrimination and identification of fingers and hands. *Exp. Brain Res.* 206, 47–57.

Rossetti, Y., Desmurget, M., and Prablanc, D. (1995). Vectorial coding of movement: vision, proprioception, or both? *J. Neurophysiol.* 74, 457–463.

Sarlegna, F. R., Przybyla, A., and Sainburg, R. L. (2009). The influence of target sensory modality on motor planning may reflect errors in sensori-motor transformations. *Neuroscience* 164, 597–610.

Schicke, T., and Röder, B. (2006). Spatial remapping of touch: confusion of perceived stimulus order across hand and foot. *Proc. Natl. Acad. Sci. U.S.A.* 103, 11808–11813.

Shore, D. I., Spry, E., and Spence, C. (2002). Confusing the mind by crossing the hands. *Brain Res. Cogn. Brain Res.* 14, 153–163.

Touzalin-Chretien, P., Ehrler, S., and Dufour, A. (2010). Dominance of vision over proprioception on motor programming: evidence from ERP. *Cereb. Cortex* 20, 2007–2016.

van Beers, R. J., Sittig, A. C., and Denier van der Gon, J. J. (1999). Integration of proprioceptive and visual position-information: an experimentally supported model. *J. Neurophysiol.* 81, 1355–1364.

Van Riper, C. (1935). An experimental study of the Japanese illusion. *Am. J. Psychol.* 47, 252–263.

Warren, D. H. (1980). Response factors in intermodality localization under conflict conditions. *Percept. Psychophys.* 27, 28–32.

Yamamoto, S., and Kitazawa, S. (2001). Reversal of subjective temporal order due to arm crossing. *Nat. Neurosci.* 4, 759–765.

Zampini, M., Harris, C., and Spence, C. (2005). Effect of posture change on tactile perception: impaired direction discrimination performance with interleaved fingers. *Exp. Brain Res.* 166, 498–508.

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: 03 May 2012; accepted: 07 September 2012; published online: 26 September 2012.

*Citation: Hong SW, Xu L, Kang M-S and Tong F (2012) The hand-reversal illusion revisited. Front. Integr. Neurosci. **6**:83. doi: 10.3389/fnint.2012.00083*

frontiers in
INTEGRATIVE NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 31 May 2012
doi: 10.3389/fnint.2012.00026

Electrophysiological correlates of predictive coding of auditory location in the perception of natural audiovisual events

Jeroen J. Stekelenburg* and Jean Vroomen

Department of Cognitive Neuropsychology, Tilburg University, Tilburg, Netherlands

Edited by:
Hermann J. Mueller, University of Munich, Germany

Reviewed by:
Daniel Senkowski, Charité, University Medicine, Germany
Edmund C. Lalor, Trinity College Dublin, Ireland

****Correspondence:***
Jeroen J. Stekelenburg, Department of Medical Psychology and Neuropsychology, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, Netherlands.
e-mail: j.j.stekelenburg@uvt.nl

In many natural audiovisual events (e.g., a clap of the two hands), the visual signal precedes the sound and thus allows observers to predict *when*, *where*, and *which* sound will occur. Previous studies have reported that there are distinct neural correlates of temporal (when) versus phonetic/semantic (which) content on audiovisual integration. Here we examined the effect of visual prediction of auditory location (where) in audiovisual biological motion stimuli by varying the spatial congruency between the auditory and visual parts. Visual stimuli were presented centrally, whereas auditory stimuli were presented either centrally or at 90° azimuth. Typical sub-additive amplitude reductions (AV − V < A) were found for the auditory N1 and P2 for spatially congruent and incongruent conditions. The new finding is that this N1 suppression was greater for the spatially congruent stimuli. A very early audiovisual interaction was also found at 40–60 ms (P50) in the spatially congruent condition, while no effect of congruency was found on the suppression of the P2. This indicates that visual prediction of auditory location can be coded very early in auditory processing.

Keywords: audiovisual integration, spatial congruity, visual prediction

INTRODUCTION

Many ecological settings are multisensory in nature with a causal relationship between the involved unisensory modalities, as in the case of hearing and seeing someone speak (Winkler et al., 2009). Quite often, visual information also leads the auditory information as in the case where two objects collide, or in the case of audiovisual speech where lip movements precedes actual phonation for up to several hundredths of milliseconds (Klucharev et al., 2003; van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007). This visual anticipatory information allows observers to predict several aspects of the upcoming auditory signal, like its timing and content. Several electrophysiological markers underlying this predictive information have been found in studies aimed at tracking the time course of audiovisual speech integration. These report that the auditory-evoked N1 and P2 components of the event-related brain potential (ERP) are attenuated and speeded up when the auditory signal (monosyllabic syllables) is accompanied by concordant visual speech input (Klucharev et al., 2003; Besle et al., 2004; van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007; Arnal et al., 2009). These sub-additive interactions are not only found in speech, but also in other naturalistic and artificial non-speech events provided that the visual information precedes and predicts sound onset as in the case of a clap of the two hands (Stekelenburg and Vroomen, 2007) or a collision of two disks (Vroomen and Stekelenburg, 2010). Of equal importance, there is no N1-suppression when there is no visual anticipatory information about sound onset as in the case in a video recording of a saw that suddenly moves (Stekelenburg and Vroomen, 2007; Vroomen and Stekelenburg, 2010). The functional interpretation of the N1-suppression may be related to a reduction of signal uncertainty, dampened sensation of loudness, or lowered computational demands for auditory brain areas.

This hypothesis fits with results from the early 1970s where in motor-sensory research it was found that the auditory N1 is dampened by self-generated sounds (Schafer and Marcus, 1973; McCarthy and Donchin, 1976; Martikainen et al., 2005) if compared to sounds replayed to the participant. Similar effects of motor prediction were found for the visually evoked N1 (Gentsch and Schütz-Bosbach, 2011; Gentsch et al., 2012). This motor-induced effect on the N1 has been attributed to reduced temporal uncertainty induced by a forward model that predicts and inhibits the sensory consequences of one's own actions (Schafer and Marcus, 1973). Also, work on unimodal auditory processing indicates that the auditory N1 can be attenuated by expectations of time (Lange, 2009, 2010).

From this literature, it also appears that predictions about the informational content are processed in a later stage of auditory processing. In multisensory studies it has been found that the N1-suppression is not affected by whether the auditory and visual information are congruent or incongruent (e.g., hearing /ba/ while lipreading /fu/), but AV integration of informational content (whether phonetic or semantic) affects the auditory P2 component as it is modulated by stimulus congruency in both speech (Klucharev et al., 2003) and

non-speech stimuli (Stekelenburg and Vroomen, 2007). This suggests the existence of two functionally distinct integrative mechanisms with different time-courses (see also Arnal et al., 2009). Klucharev et al. (2003) hypothesized that the early effects at N1 reflect AV interactions in the processing of general features shared by the acoustic and visual stimulus such as coincidence in time, and—at this stage untested—spatial location.

The current study was set-up to further explore the time-course and functional significance of visual predictive coding on auditory processing. One hitherto unexplored aspect is that, besides prediction of time and content, visually anticipatory information can also predict the likely *location* of the auditory signal because the origin of a sound usually corresponds with the location of the visual signal. Here, we thus examined whether spatial congruency between auditory and visual anticipatory information affects the auditory-evoked potentials N1, P2, or other components. For spatially congruent events, the location of the auditory and visual stimulus were aligned in the center, while for the incongruent condition there was a large separation of 90° between the auditory and visual stimulus. This large separation effectively prevented a ventriloquist effect (i.e., vision capturing the apparent sound location) to occur (Colin et al., 2001), and the reported effects were, therefore, devoid of neural correlates associated with ventriloquism (Bonath et al., 2007). We expected that if predictive coding entails prediction of sound location (over and above timing), then more suppression should be found when the locations of the auditory and visual stimulus were congruent rather than incongruent because a "confirmed" prediction lowers computational demands. Alternatively, if the brain does not use visual anticipatory information about sound location, then no effect of audiovisual spatial congruency should be observed.

METHODS

PARTICIPANTS

Twenty-two (19 woman, mean age 18.5, SD 1.1) healthy participants took part in the experiment. All were students from Tilburg University who reported normal hearing and normal or corrected-to-normal vision. All of them were naive to the purpose of the study. They received course credits for their participation. Informed consent was obtained from all participants. This study was conducted in accordance with the Declaration of Helsinki and was approved by the local Ethics Committee of Tilburg University.

STIMULI AND PROCEDURE

The experiment took place in a dimly lit and sound attenuated room. Visual stimuli were presented on a 19-inch monitor positioned at eye-level, at 70 cm from the participant's head. The sounds emanated from either of two speakers. One speaker was located directly below the monitor, the other one was located on the left side of the participant, perpendicular to the left ear, at the same height and distance as the central speaker. Four audiovisual stimuli were used that in previous studies induced reliable N1-suppression (Stekelenburg and Vroomen, 2007). Two stimuli were the syllables /bi/ and /fu/ pronounced by a Dutch female speaker whose entire face was visible on the screen. The other stimuli were natural human actions, i.e., a clap of two hands and a tap of a spoon against a cup. For each stimulus category three exemplars were recorded so that there were 12 unique recordings in total. The videos were presented at a rate of 25 frames/s with an auditory sample rate of 44.1 kHz. The size of the video frames subtended 14° horizontal and 12° vertical visual angle. Sound level was measured with a sound-level meter with the microphone pointing toward the auditory source. Peak intensity was 65 dB(A) for the central and lateral speakers. The duration of the auditory sample was 306–325 ms for /bi/, 594–624 ms for /fu/, 292–305 ms for the spoon tapping on a cup, and 103–107 ms for the clapping hands. Average duration of the video was 3 s, including a 200-ms fade-in and fade-out, and a still image (200–500 ms) at the start (**Figure 1**). A blank screen of 500–1000 ms followed each trial. The inter-stimulus interval (from auditory onset) was on average 3.7 s. The time from the start of the articulatory movements until voice onset was, on average, 160 ms for /bi/ and 200 ms for /fu/. The time from the start of the movements of the arm(s) until sound onset in the non-speech stimuli was 280 ms for the clapping hands and 320 ms for the tapping spoon.

There were five experimental conditions; Ac (audio from the center, no video), Al (audio from lateral, no video), Vc (video from central, no audio), AcVc (audio and video from central), and AlVc (audio from lateral, video from center). For each condition, a total of 72 experimental trials were presented, separately for each of the four stimuli across 12 blocks, amounting to a total of 1440 trials. Trial order was randomized. To ensure that participants were looking at the video during stimulus presentation, they had to detect, by key press, the occasional occurrence of catch trials (8% on top of the total number of experimental trials). Catch trials occurred equally likely in all conditions. Catch trials contained a superimposed small white spot—either between the lips and nose for the speech stimulus, or at collision site for

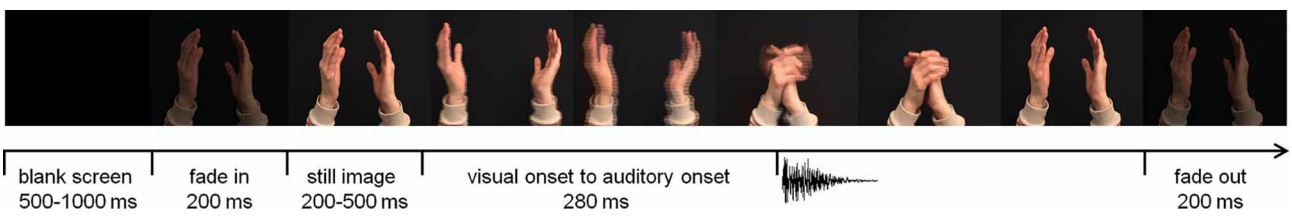


FIGURE 1 | Time-course of an audiovisual trial (hand clap). The visual onset to auditory onset differed per stimulus type (hand clap 280 ms; tapping spoon 320 ms; /bi/ 160 ms; /fu/ 200 ms).

the hands or at the site where the spoon hit the cup—for 120 ms. The appearance of the spot varied quasi-randomly within 300 ms before or after the onset of the sound. In the Ac and Al conditions the spot was presented on a dark screen at about the same position and at the same time as in the other conditions.

EEG RECORDING AND ANALYSIS

The electroencephalogram (EEG) was recorded at a sample rate of 512 Hz from 49 locations using active Ag-AgCl electrodes (BioSemi, Amsterdam, The Netherlands) mounted in an elastic cap and two mastoid electrodes. Electrodes were placed according the extended International 10–20 system. Two additional electrodes served as reference (Common Mode Sense [CMS] active electrode) and ground (Driven Right Leg [DRL] passive electrode). EEG was referenced offline to an average of left and right mastoids and band-pass filtered (0.5–30 Hz, 24 dB/octave). The raw data were segmented into epochs of 800 ms, including a 100-ms prestimulus baseline. ERPs were time-locked to the sound onset in the AV and A conditions, and to the corresponding time stamp in the V condition. After EOG (Gratton et al., 1983), epochs with an amplitude change exceeding ±120 μV at any EEG channel were rejected. ERPs of the non-catch trials were averaged per condition (Ac, Al, Vc, AcVc, and AlVc) across all stimuli. As in previous studies, multisensory interactions were examined by comparing ERPs evoked by A stimuli with the corresponding AV minus V (AV − V) ERPs (Besle et al., 2004; Stekelenburg and Vroomen, 2007; Arnal et al., 2009; Vroomen and Stekelenburg, 2010). The additive model (A = AV − V) assumes that the neural activity evoked by the AV stimuli is equal to the sum of activities of A and V if the unimodal signals are processed independently. This assumption is valid for extracellular media, and is based on the law of superposition of electric fields (Barth et al., 1995). If the bimodal response differs (supra-additive or sub-additive) from the sum of the two unimodal responses, this is attributed to the interaction between the two modalities (Giard and Peronnet, 1999; Molholm et al., 2002; Klucharev et al., 2003; Besle et al., 2004; Teder-Sälejärvi et al., 2005; Stekelenburg and Vroomen, 2007; Vroomen and Stekelenburg, 2010). Critical comparisons in the current study were between the AV interactions of which visual and auditory signals originated from the same location and AV interactions of which visual and auditory signals originated from different locations. We, therefore, calculated congruent (AcVc − Vc) and incongruent (AlVc − Vc) difference waves and compared them to the corresponding A conditions (i.e., Ac and Al, respectively). The auditory N1 and P2 had a central maximum, and analyses were, therefore, conducted at nine central electrodes surrounding Cz. The peak amplitude of N1 was scored in a window of 70–150 ms. The peak amplitude of P2 was scored in a window of 120–250 ms. To test possible differences in AV interactions of congruent and incongruent sound locations, N1 and P2 scores of the congruent and incongruent difference waves were subtracted from the corresponding A conditions; (Ac − [AcVc − Vc]) and (Al − [AlVc − Vc]). These difference scores were submitted to a repeated measures MANOVA with as within-subjects variables Congruency (AV locations congruent versus incongruent) and Electrode (FC1, FCz, FC2, C1, Cz, C2, CP1, CPz, CP2).

RESULTS

Participants detected 99% of the catch trials, indicating that they indeed watched the monitor. **Figures 2** and **3** show that for both auditory locations, AV interactions were associated with N1 and P2 suppression. As reported before (Stekelenburg and Vroomen, 2007), the video substantially reduced the amplitude of the auditory N1, [$F_{(1, 21)} = 30.54, p < 0.001$]. Most importantly, this intersensory effect was influenced by spatial Congruency, [$F_{(1, 21)} = 5.53, p < 0.05$], indicating that the N1 reduction was greater for the spatially congruent AcVc condition (a 1.7 μV reduction) than for the spatially incongruent AlVc condition (a 1.1 μV reduction). This congruency effect was not affected by Electrode ($F < 1$) (**Figure 4**).

We also tested for each sound location separately the intersensory effect on N1 amplitude with the variables Modality (A versus AV − V) and Electrode in a repeated measure MANOVA. For both congruent and incongruent conditions, N1 suppression was significant, [$F_{(1, 21)} = 29.99, p < 0.001$, $F_{(1, 21)} = 17.83, p < 0.001$], respectively. To further delineate whether the difference in N1 suppression should be attributed to differences in A-only versus AV − V, we separately tested A-only and AV − V between the two locations. The N1 of the AcVc − Vc condition was smaller than AlVc − Vc condition, [$F_{(1, 21)} = 14.66, p < 0.01$], but there was no difference in the N1 between Ac and Al, [$F_{(1, 21)} = 1.39, p = 0.25$]. This further suggests that the effect of location on N1-suppression was due to differences in AV integration, and not to differences in Ac versus Al *per se*. The same MANOVA as for the N1 amplitude on the latency scores showed that N1 latency was not affected by stimulus modality ($F < 1$), nor was there an effect of Congruency ($F < 1$) or a Congruency × Electrode interaction, [$F_{(1, 21)} = 1.02, p = 0.47$].

The same MANOVA on the P2 showed that the P2 was also reduced in amplitude (1.9 μV) and speeded up (7 ms) alike in the bimodal conditions AcVc and AlVc, [$F_{(1, 21)} = 23.03, p < 0.001$] and [$F_{(1, 21)} = 5.98, p < 0.05$], respectively. Importantly, the intersensory effects on the P2 amplitude and P2 latency were *not* affected by Congruency, [$F_{(1, 21)} = 1.07, p = 0.31$ and $F < 1$], respectively. There were also no Congruency × Electrode interactions for P2 amplitude (**Figure 4**) and P2 latency, [$F_{(1, 21)} = 1.16, p = 0.38$ and $F_{(1, 21)} = 1.33, p = 0.31$].

Figure 2 also suggests that there was an early effect of spatial congruency at the P50 component. The P50 was scored by calculating mean activity in a 40–60 ms window and showed a maximum at the fronto-central electrodes. The difference scores (Ac − [AcVc − Vc]) and (Al − [AlVc − Vc] were submitted to a repeated measures MANOVA with the within-subjects variables Congruency (AV locations congruent versus incongruent) and Electrode (F1, Fz, F2, FC1, FCz, FC2, C1, Cz, C2). There was a Congruency × Electrode interaction, [$F_{(8, 14)} = 2.90, p < 0.05$]. Simple effect test, examining the effect of congruency at each electrode, showed that this effect was localized mainly at electrode Cz ($p < 0.05$). Separate tests for the congruent and incongruent conditions showed that at Cz significant AV interactions were found for congruent presentations, [$t_{(21)} = 2.62, p < 0.05$], but not for incongruent presentations ($t < 1$).

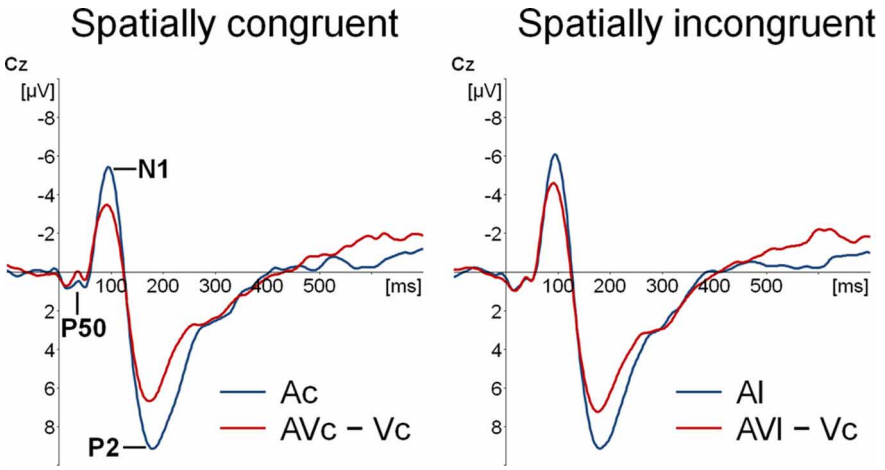


FIGURE 2 | Event-related potentials (ERPs) time-locked to auditory onset at electrode Cz. Separately for spatially congruent and incongruent audiovisual presentations the auditory-only (Ac and Al) ERP and the audiovisual minus visual-only difference wave (AVc − Vc and AVl − Vc) are displayed.

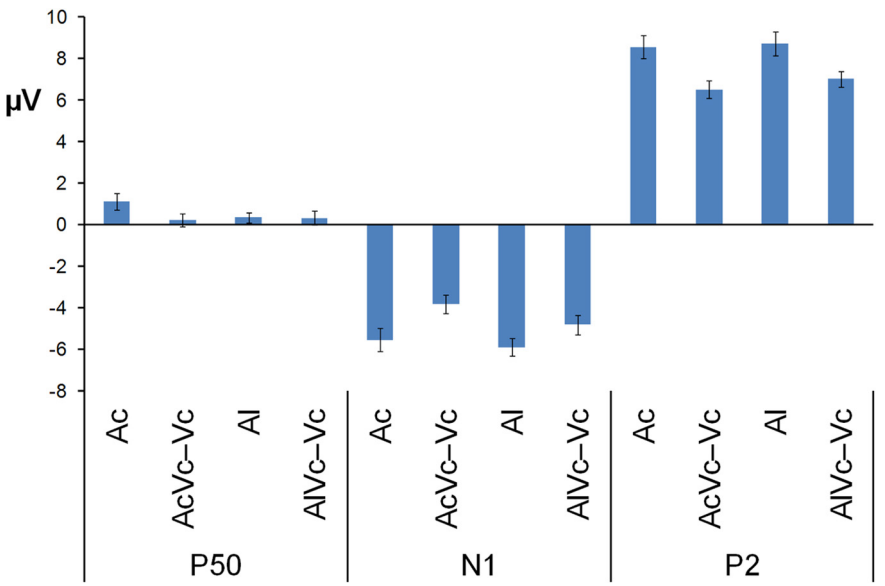


FIGURE 3 | Mean voltage in μV of P50, N1, and P2 averaged across the electrodes used in the MANOVA for auditory-only stimuli presented centrally and laterally (Ac and Al) and audiovisual difference waves for spatially congruent and incongruent AV presentations (AVc − Vc and AVl − Vc). The bars indicate one standard error of mean.

DISCUSSION

Our results support theoretical models that assume that the brain uses distinct sources of information to predict subsequent sensory inputs. More specifically, our results replicate the by now well-established finding that suppression of auditory N1 and P2 constitutes the neural consequence of an interaction of audiovisual stimuli containing anticipatory visual motion (Besle et al., 2004; van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007; Arnal et al., 2009; Vroomen and Stekelenburg, 2010). One small difference with a previous study (Stekelenburg and Vroomen, 2007) that used the same stimuli was that both N1 and P2 peaked earlier in the bimodal condition whereas in the current study latency facilitation was limited to P2. It may be inferred that the latency facilitation of N1 is less robust than the suppression of the N1 amplitude. This is further supported by a study using audiovisual speech also showing a reduction in N1 amplitude, but not in N1 latency (Besle et al., 2004). The new finding here is that the N1 suppression was greater for spatially congruent than incongruent AV stimuli. We hypothesized that a visual signal that is naturally leading the auditory signal would allow observers to predict not only the onset and content, but also the location of the sound. As demonstrated before, temporal prediction is predominantly reflected in N1-suppression because it only occurs when anticipatory visual movements reliably predict sound onset,

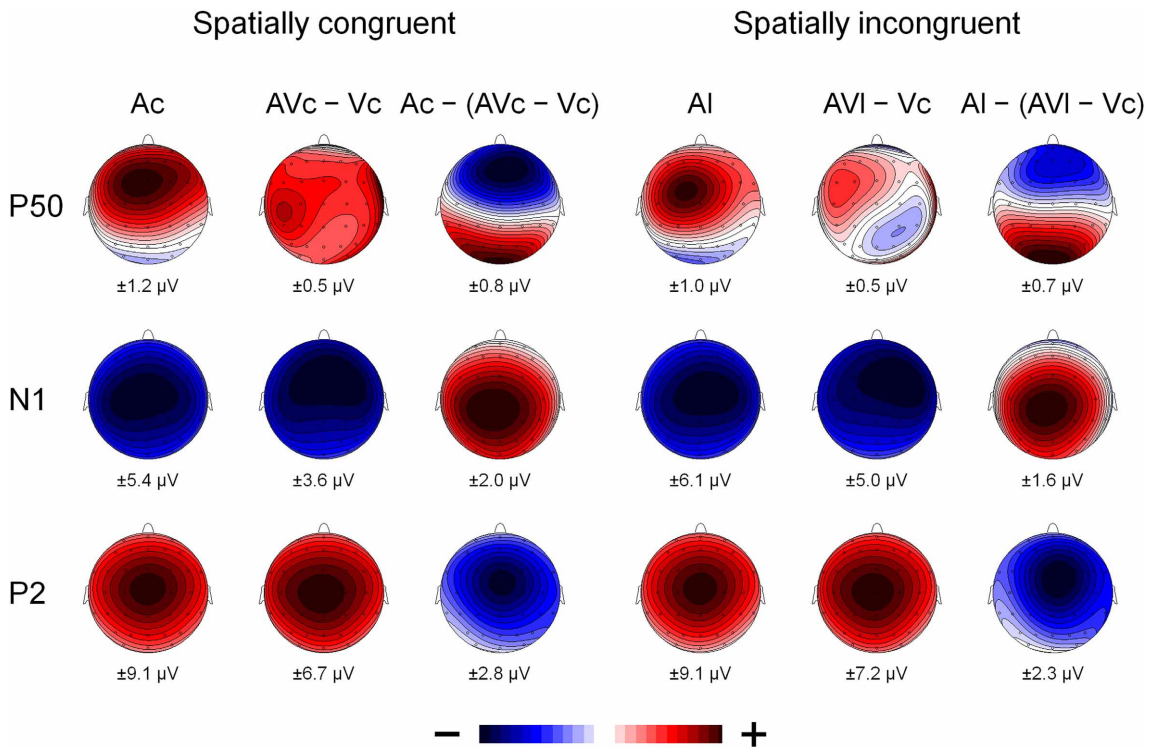


FIGURE 4 | The scalp topography of P50, N1, and P2 for the A-only ERPs (Ac and Al), the AV difference waves (AVc − Vc and AVl − Vc) and the multisensory interactions represented by the A-only ERP minus the AV difference wave (Ac − [AVc − Vc] and Al − [AVl − Vc]) for spatially congruent and incongruent presentations. The range of the voltage maps in μV is displayed below each map.

while it is abolished when vision does not predict sound onset (Stekelenburg and Vroomen, 2007; Vroomen and Stekelenburg, 2010). In a similar vein, it thus appears that the N1 is also sensitive to spatial prediction, given that the visually induced auditory N1-suppression was reduced when the auditory location did not match the predicted location. It thus seems likely that the N1 suppression reflects a process in which both the temporal onset and the location of the sound is predicted on the basis of the leading visual signal.

The spatially congruent AV stimuli also induced early integration effects at 40–60 ms at the central sites, while no such early integration was found for spatially discordant AV stimuli. Similar early AV interactions have been demonstrated in studies on AV integration with more basic artificial stimuli (Giard and Peronnet, 1999; Molholm et al., 2002; Talsma et al., 2007; Sperdin et al., 2009). This suggests that spatial congruity is a necessary condition for these early interactions.

Whereas spatial congruency affected AV interactions at N1, no such effect was found at the P2 component. This is in line with a study demonstrating that same- and different AV location pairings showed both similar and different AV interactions (Teder-Sälejärvi et al., 2005). The dissociation between N1 and P2 effects also verifies the hypothesis of a study (Klucharev et al., 2003) stating that the AV interactions at N1 reflect interactions in the processing of general features shared by the acoustic and visual stimulus, specifically spatial and temporal correspondence, while later interactions at P2 latency reflect interactions at phonetic, semantic, or associative level (Stekelenburg and Vroomen, 2007). The distinction between these two qualitatively different integration mechanisms with different underlying time-courses is supported by a MEG/fMRI study (Arnal et al., 2009). The latter study proposed that two distinct neural routes are involved in the audiovisual integration of speech. These authors conjectured that predictive visual information affects auditory perception via a fast direct visual to auditory pathway which conveys physical visual but no phonological characteristics. After the visual-to-auditory predictive mechanism a secondary feedback signal is followed via STS, which signals the error (if present) between visual prediction and auditory input. Because visual predictive information about auditory location affects early (P50, N1) potentials, it seems reasonable to maintain that within this dual route model, AV integration of location is realized via a fast direct route.

One can also ask to which extent the present results are modulated by the effects of attention on multisensory processing. The task of participants was to detect visual catch trials in the center of fixation. This implies that in the congruent condition, auditory stimuli were presented at the attended location (the center), whereas in the incongruent condition they were presented at an unattended location. Could it be, then, that differences in "spatial attention" rather than "accuracy of sensory prediction" underlie the present results of spatial congruency. Indeed, it has been argued that attention can modulate the neural correlates of multisensory integration (Talsma et al., 2010). Typically,

in these studies (Senkowski et al., 2005; Talsma and Woldorff, 2005) attention is manipulated by presenting auditory, visual, and audiovisual stimuli randomly to two lateral spatial positions and instructing participants to focus their attention at only one of these locations during a block of trials. When stimuli are presented at the attended location, multisensory (AV) stimuli elicit *larger* ERP waveforms (N1, P2) than the sum of the visual and auditory (A + V) parts alone, whereas at the unattended location, the difference between the AV and A + V is smaller. Note that this result is exactly the opposite pattern what was found here, because we obtained *smaller* ERPs, not larger, if sounds were presented at the audiovisual congruent (i.e., attended) location. In addition, the interaction of attention with multisensory integration was associated with enhanced late fronto-centrally distributed potentials (Busse et al., 2005; Talsma and Woldorff, 2005), whereas in the current study there was no hint of late congruency effects. We conjecture that the critical difference is that we used stimuli with visual predictive information that preceded sound onset, whereas these other studies used *synchronized* AV stimuli, thus without visual anticipatory information. Future studies might try to further disentangle the effects of spatial attention and sensory prediction on multisensory integration. One could, for example, envisage a study in which visual stimuli with predictive information are presented at fixation or far from fixation, while sounds are presented from audiovisual congruent or incongruent locations. On the attentional account, distance from fixation should matter, while on the sensory prediction account it is the spatial congruency between the auditory and visual information that matters.

The here reported effects of spatial congruity differ in several aspects (timing and location over the scalp) from earlier studies on spatial location. One study (Teder-Sälejärvi et al., 2005) found ERP interactions that differed according to spatial congruity which included a phase and amplitude modulation of visual-evoked activity localized to the ventral occipito-temporal cortex at 100–400 ms, and an amplitude modulation of activity localized to the superior temporal region at 260–280 ms. Another study (Gondan et al., 2005) also found effects of spatial congruity as ERPs to spatially congruent and spatially incongruent bimodal stimuli started to differ over the parietal cortex around 160 ms after stimulus onset. We conjecture, though, that a critical difference is that both studies used *synchronized* AV stimuli, thus without visual anticipatory information. Other potentially relevant differences are that we used natural rather than artificial stimuli (flashes and beeps), and we used a larger degree of separation between auditory and visual stimuli [90° in our study versus 40° in Gondan et al. (2005) and 60° in Teder-Sälejärvi et al. (2005)].

Our study is also relevant for the question as to whether N1-suppression to audiovisual presentations is evoked by factors other than visual prediction. Initial studies on visually induced suppression of auditory N1 (e.g., Besle et al., 2004; van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007) cannot rule out that visual anticipatory movement might have summoned involuntary transient attention to the visual modality, thereby depleting attentional resources of the auditory modality (Pilling, 2009). This depletion of auditory resources might then be reflected in a suppression of the auditory N1. However, if this kind of non-spatial depletion of auditory resources were the sole determinant of the N1-supression, one would expect no differential effect of spatial congruity on N1-suppression because it should be identical for both congruent and incongruent locations. The current results, therefore, refute a depletion account of N1-suppression.

In summary, we found that the auditory-evoked N1 and P2 were suppressed when accompanied by their corresponding visual signals. These sub-additive AV interactions have previously been attributed to visual prediction of auditory onset (Stekelenburg and Vroomen, 2007; Vroomen and Stekelenburg, 2010). The crucial finding here is that spatial congruity between A and V also affected AV interactions at early ERP components: for spatially incongruent pairings, no AV interactions at P50 and less N1-suppression was found than for spatially congruent pairings, whereas suppression of P2 remained unaffected by spatial congruency. This suggests that visuo-spatial and visuo-temporal information have different time-courses in AV integration: spatial prediction has earlier effects on auditory processing (P50, N1) than temporal prediction (N1, P2).

REFERENCES

Arnal, L. H., Morillon, B., Kell, C. A., and Giraud, A. L. (2009). Dual neural routing of visual facilitation in speech processing. *J. Neurosci.* 29, 13445–13453.

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.

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.

Bonath, B., Noesselt, T., Martinez, A., Mishra, J., Schwiecker, K., Heinze, H. J., and Hillyard, S. A. (2007). Neural basis of the ventriloquist illusion. *Curr. Biol.* 17, 1697–1703.

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.

Colin, C., Radeau, M., Deltenre, P., and Morais, J. (2001). Rules of intersensory integration in spatial scene analysis and speechreading. *Psychol. Belg.* 41, 131–144.

Gentsch, A., Kathmann, N., and Schütz-Bosbach, S. (2012). Reliability of sensory predictions determines the experience of self-agency. *Behav. Brain Res.* 228, 415–422.

Gentsch, A., and Schütz-Bosbach, S. (2011). I did It: unconscious expectation of sensory consequences modulates the experience of self-agency and its functional signature. *J. Cogn. Neurosci.* 23, 3817–3828.

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.

Gondan, M., Niederhaus, B., Rösler, F., and Röder, B. (2005). Multisensory processing in the redundant-target effect: a behavioral and event-related potential study. *Percept. Psychophys.* 67, 713–726.

Gratton, G., Coles, M. G., and Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalogr. Clin. Neurophysiol.* 55, 468–484.

Klucharev, V., Möttönen, R., and Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Brain Res. Cogn. Brain Res.* 18, 65–75.

Lange, K. (2009). Brain correlates of early auditory processing are attenuated by expectations for time and pitch. *Brain Cogn.* 69, 127–137.

Lange, K. (2010). Can a regular context induce temporal orienting to a target sound? *Int. J. Psychophysiol.* 78, 231–238.

Martikainen, M. H., Kaneko, K., and Hari, R. (2005). Suppressed responses to self-triggered sounds in the human auditory cortex. *Cereb. Cortex* 15, 299–302.

McCarthy, G., and Donchin, E. (1976). The effects of temporal and event uncertainty in determining the waveforms of the auditory event related potential (ERP). *Psychophysiology* 13, 581–590.

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. *Brain Res. Cogn. Brain Res.* 14, 115–128.

Pilling, M. (2009). Auditory event-related potentials (ERPs) in audiovisual speech perception. *J. Speech Lang. Hear. R.* 52, 1073–1081.

Schafer, E. W., and Marcus, M. M. (1973). Self-stimulation alters human sensory brain responses. *Science* 181, 175–177.

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.

Sperdin, H. F., Cappe, C., Foxe, J. J., and Murray, M. M. (2009). Early, low-level auditory-somatosensory multisensory interactions impact reaction time speed. *Front. Integr. Neurosci.* 3:2. doi: 10.3389/neuro.07.002.2009

Stekelenburg, J. J., and Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *J. Cogn. Neurosci.* 19, 1964–1973.

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.

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.

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.

Teder-Sälejärvi, W. A., Di Russo, F., McDonald, J. J., and Hillyard, S. A. (2005). Effects of spatial congruity on audio-visual multimodal integration. *J. Cogn. Neurosci.* 17, 1396–1409.

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.

Vroomen, J., and Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *J. Cogn. Neurosci.* 22, 1583–1596.

Winkler, I., Denham, S. L., and Nelken, I. (2009). Modeling the auditory scene: predictive regularity representations and perceptual objects. *Trends Cogn. Sci.* 13, 532–540.

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: 23 March 2012; accepted: 14 May 2012; published online: 31 May 2012.

Citation: Stekelenburg JJ 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*

frontiers in
INTEGRATIVE NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 11 September 2012
doi: 10.3389/fnint.2012.00075

The effect of task order predictability in audio-visual dual task performance: Just a central capacity limitation?

Thomas Töllner[1]*, Tilo Strobach[1,2], Torsten Schubert[2] and Hermann J. Müller[1,3]

[1] Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany
[2] Department of Psychology, Humboldt Universität Berlin, Berlin, Germany
[3] Department of Psychological Sciences, Birkbeck College London, London, UK

Edited by:
Zhuanghua Shi, Ludwig-Maximilians-Universität München, Germany

Reviewed by:
Hao Zhang, Duke University Medical Center, USA
Joseph Krummenacher, University of Fribourg, Switzerland

***Correspondence:**
Thomas Töllner, Department of Psychology, Ludwig-Maximilians-Universität München, Leopoldstrasse 13, D-80802 Munich, Germany.
e-mail: thomas.toellner@psy.lmu.de

In classic Psychological-Refractory-Period (PRP) dual-task paradigms, decreasing stimulus onset asynchronies (SOA) between the two tasks typically lead to increasing reaction times (RT) to the second task and, when task order is non-predictable, to prolonged RTs to the first task. Traditionally, both RT effects have been advocated to originate exclusively from the dynamics of a central bottleneck. By focusing on two specific electroencephalographic brain responses directly linkable to perceptual or motor processing stages, respectively, the present study aimed to provide a more detailed picture as to the origin(s) of these behavioral PRP effects. In particular, we employed 2-alternative forced-choice (2AFC) tasks requiring participants to identify the pitch of a tone (high versus low) in the auditory, and the orientation of a target object (vertical versus horizontal) in the visual, task, with task order being either predictable or non-predictable. Our findings show that task order predictability (TOP) and inter-task SOA interactively determine the speed of (visual) perceptual processes (as indexed by the PCN timing) for both the first and the second task. By contrast, motor response execution times (as indexed by the LRP timing) are influenced independently by TOP for the first, and SOA for the second, task. Overall, this set of findings complements classical as well as advanced versions of the central bottleneck model by providing electrophysiological evidence for modulations of both perceptual and motor processing dynamics that, in summation with central capacity limitations, give rise to the behavioral PRP outcome.

Keywords: attention, decision making, executive control, central bottleneck, PCN, LRP

INTRODUCTION

In classic Psychological-Refractory-Period (PRP) dual-task paradigms, the time taken to respond to the stimulus of the second task typically increases with decreasing inter-task interval (i.e., stimulus onset asynchrony, SOA), whereas there is no influence of inter-task SOA on reaction times (RT) to the stimulus of the first task (e.g., Welford, 1952; Pashler and Johnston, 1989). This well-established and extensively studied effect has traditionally been explained in terms of a sequential processing model consisting of three stages: (1) a *perceptual* stage, which selects the task-relevant stimulus (e.g., based on a spatial characteristic: left versus right positioning of the stimulus relative to the vertical midline of the display) and, if required, extracts the response-critical stimulus attribute (e.g., exact featural identity: red versus green) required for subsequent response decisions; (2) a *central* stage which decides upon the appropriate motor response (e.g., left versus right index finger press) on the basis of a pre-specified task setting (i.e., stimulus-response, S-R, and mapping); and (3) a *motor* stage, which produces and executes this response. While both perceptual and motor stages are generally assumed to operate in parallel, the commonly advocated view (e.g., Pashler, 1984; Luck, 1998; Schubert, 1999) is that the effect of inter-task SOA on RTs to the second task may originate exclusively from a processing *bottleneck* located at the central stage, in particular: central-stage processing of the second task is delayed until central processing of the first task has been completed. Accordingly, RTs to the second task depend on the onset of responses to the first task, rather than the onset of the respective (first-task) stimuli.

Recent findings (e.g., Schubert, 1996, 2008; Jiang et al., 2004; Sigman and Dehaene, 2006), however, have challenged the traditional view that responses to the *first* task are processed independently of the inter-task interval. For instance, responses to the first task are slowed down compared to when this task is executed in isolation (e.g., Jiang et al., 2004; Sigman and Dehaene, 2006), or when the sequence of the two upcoming (dual) tasks is made unpredictable (e.g., De Jong, 1995; Szameitat et al., 2002, 2006; Sigman and Dehaene, 2006). To explain this set of findings, Sigman and Dehaene (2006) introduced an "*extended central bottleneck*" *view*, according to which additional central executive processes, including task scheduling and task disengagement (see also Lien et al., 2003; Liepelt et al., 2011; Strobach et al., 2012, in press), are assumed to give rise to the increased processing times for the first task especially at short SOAs (e.g., <300 ms). Crucially, both task control processes are scheduled within the *central* system (involving the operation of executive control); thus, again assuming solely *central processes* as origin of the RT cost associated with unpredictable, relative to predictable, task orders.

On this background, the aim of the present electroencephalogram (EEG) study was twofold: First, we intended to gain deeper insights into the question of whether the SOA effect on RTs to the first task under conditions of unpredictable task order is indeed due to the dynamics of task coordination processes that operate exclusively at the *central stage*, as proposed by the "extended central bottleneck" model of Sigman and Dehaene (2006; see also Schubert, 1996); or, alternatively, whether there might also be modulations evident at the preceding *perceptual* and/or the subsequent *motor* stage that, when combined with the central processing dynamics, contribute to this RT effect. Second, we asked whether the SOA effect on RTs to the second task is, again, solely driven by central capacity limitations, as advocated by traditional central bottleneck models (e.g., Pashler, 1994), and/or whether this effect may be further influenced by the predictability of the task order. To address these questions, we employed a 2-alternative forced-choice (2AFC) audio-visual dual task, requiring participants to identify the pitch of a tone (high versus low) in the auditory, and the orientation of a laterally presented target object (vertical versus horizontal) in the visual, task, with the order of the dual tasks being either fixed (predictive task order) or random (non-predictive task order), with variable inter-task intervals (SOAs). In addition, we combined mental chronometry data with two specific electroencephalographic brain responses directly linkable to either pure perceptual or pure motor stages of the information-processing stream.

The first EEG parameter, the *Lateralized-Readiness-Potential* (LRP), is a well-known and extensively studied event-related potential (ERP) component generally agreed to reflect the activation and execution of effector-specific motor responses (e.g., Coles, 1989; Osman and Moore, 1993; Eimer, 1998). In more detail, the LRP is negativity strongest over the motor areas contralateral to the side of a uni-manual response, typically elicited in the 150 ms time window pre-response. To dissociate the LRP from overlapping motor response-unspecific ERPs, the waveforms recorded ipsilateral to the response side are subtracted from contralateral waveforms, resulting in the so-called (contralateral-minus-ipsilateral) LRP difference wave. These subtractions can be performed time-locked to either stimulus or response onset. Accordingly, the timing of the stimulus-locked LRP (sLRP) can be regarded as indexing the start of effector-specific motor activation after the completion of response selection (i.e., central) processes (e.g., Sommer et al., 2001; Töllner et al., 2011b), whereas the time demands required by response execution processes are derivable from the response-locked LRP (rLRP) onset timing (e.g., Miller, 2007).

The second parameter of interest, the *Posterior-Contralateral-Negativity* (PCN), is a similarly prominent and extensively explored electroencephalographic brain response that has been linked to the focal-attentional selection of task-relevant target objects in visual space (e.g., Luck and Hillyard, 1994; Eimer, 1996; Woodman and Luck, 1999). [Traditionally, this component has been referred to as N2-posterior-contralateral (N2pc). However, based on recent evidence (e.g., Shedden and Nordgaard, 2001) that underscores the independence of this component in terms of both timing and activation from the non-lateralized N2, we prefer the term PCN (instead of N2pc) in order to avoid misleading associations or interpretations.] Specifically, the PCN is a negative-going deflection most prominent over the visual areas contralateral to the side of an attended object, elicited—depending on a variety of *top-down* (e.g., Eimer and Kiss, 2008; Töllner et al., 2010, 2012a) and *bottom-up* (e.g., Brisson et al., 2007; Töllner et al., 2011a) factors—in the time window approximately 175–300 ms post-stimulus. As for LRP computations, it is strongly recommended to subtract the waveforms recorded ipsilateral to the stimulus side from contralateral waveforms to cancel out overlapping target selection-unspecific ERPs.

Taken together, for auditory and visual responses, the coupling of mental chronometry to the rLRP allows us to dissociate pre-motor (i.e., perceptual and central processes) and motor processes that, in combination, may contribute to the interactive RT effect of "task-order predictability" and "stimulus-onset-asynchrony" in audio-visual dual-task performance (see **Figure 1**). In addition, for responses to visual stimuli, we can further split pre-motor times into processing components related to pre-attentive, perceptual and, respectively, post-selective, perceptual plus central (i.e., stimulus-response translation) processes on the basis of PCN computations.

MATERIALS AND METHODS

PARTICIPANTS

Thirteen participants (seven female) took part in the present study. Their ages ranged from 24 to 32 (median 28) years. All had normal or corrected-to-normal vision and reported no history of neurological disorders. Observers were either paid or received course credit for participating. One observer was excluded due to excessive eye movement artifacts. The experimental procedure was approved by the ethics committee of the Department of Psychology, Ludwig-Maximilians-University Munich, in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki).

STIMULI AND STUDY DESIGN

Visual stimulation consisted of two colored shape stimuli (radius: 1.2° of visual angle) presented against a black background and positioned equidistantly (visual angle: 3.0°) from a white central fixation cross in the lower visual field. On each trial, one of the two lateralized locations contained a task-relevant target stimulus, equally likely defined in the pre-instructed color *red* (CIE 0.544, 0.403, 68) or *blue* (CIE 0.213, 0.264, 68), together with a task-irrelevant distracter stimulus at the opposite location defined in the alternative color (blue or red, respectively). Each stimulus outline contained a grating composed of three black bars (0.4° × 2.4°) separated by two gaps (0.3° × 2.4°), which were randomly oriented either vertically or horizontally. Auditory stimuli were pure sine-waves tones, of a frequency of either 350 or 900 Hz, and of 100 ms duration.

The experiment was performed in a dimly lit, sound-attenuated, and electrically shielded experimental cabin (Industrial Acoustics Company GmbH). Visual stimuli were presented on a 17″ computer screen, mounted at a viewing distance of approximately 75 cm. Auditory stimuli were presented simultaneously via two stereo loudspeakers, placed approximately 10 cm to the left and right side of the monitor,

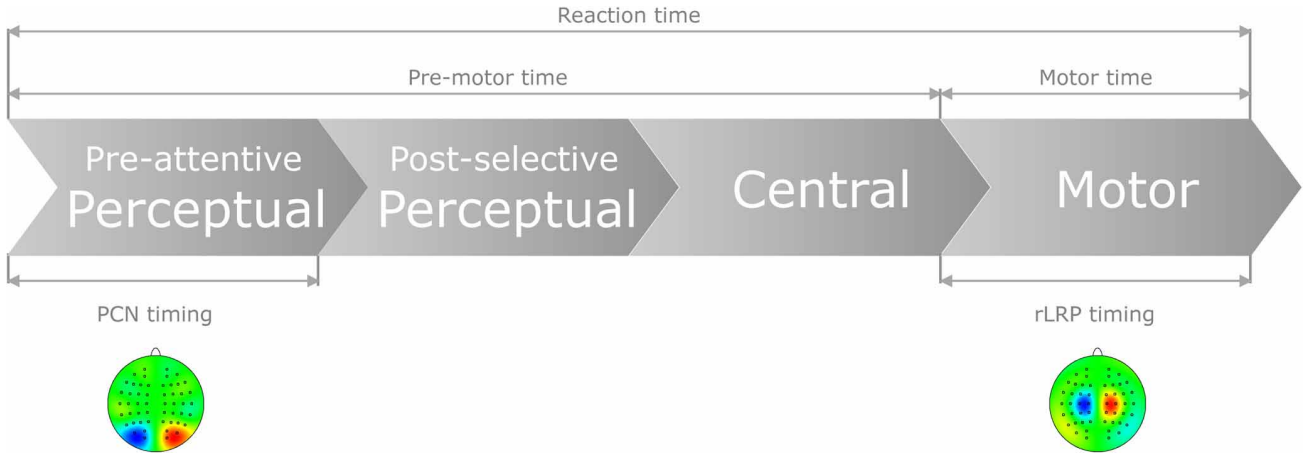


FIGURE 1 | Schematic of the present approach to control for pure perceptual and/or pure motor capacity limitations as a function of task order predictability (TOP) and stimulus onset asynchrony (SOA) in dual task. In particular, by computing the difference waves of the PCN and LRP responses, the following time demands can be electro-cortically extracted: (1) Pre-attentive perceptual processes, as determined by feature-contrast and salience coding, necessary to focally select the visual target stimulus; (2) Motor production processes, as determined by the required motor effector, necessary to activate and execute the selected response.

respectively. In order to obtain a reasonable number of trials necessary to analyze all experimental conditions of interest (see below), two separate recording sessions were conducted for each individual participant, with the second session performed within one week of, and at a similar time of day, as the first session. One experimental session consisted of 24 blocks of 48 trials each, resulting in a total of 2304 trials for each participant for both sessions. Each session was further divided into four parts of six blocks each, with two parts with predictive and two with non-predictive task-order. The sequence of these four parts was counterbalanced across subjects, but held constant across the two sessions for each individual participant. A trial started with the presentation of a white fixation cross for 500 ms, which was immediately followed by the stimuli of the first task (i.e., visual or auditory, respectively) presented for 100 ms. After a randomly chosen SOA of 150, 300, or 600 ms, the stimuli of the second task (i.e., visual or auditory, respectively) appeared for 100 ms. Trials were terminated by the participant's response(s) or after a maximum response window of 3 s for both tasks. In case of an incorrect response, the word "FEHLER" (German word for "ERROR") was centrally presented for 1 s, signaling erroneous behavior. Subsequently, a blank screen was shown during an intertrial interval of 1 s. Participants were clearly instructed to maintain central eye fixation throughout the experiment and to respond as quickly and accurately as possible, with the order of motor responses matching the order of (visual versus auditory) task occurrence.

In both tasks, there were always two stimuli concurrently presented at two lateralized locations. In the auditory task, both loudspeakers presented one-and-the-same stimulus requiring participants to identify the pitch of a tone (i.e., high versus low). The visual task, by contrast, involved the presentation of two different stimuli so as to be able to compute the PCN component (see also Brisson and Jolicoeur, 2007a,b). Prior to the start of each experiment, the task-relevant (visual) stimulus was specified by a semantic pre-cue (e.g., the word "BLUE") indicating the defining color of the target stimulus (in the example: blue) in the upcoming block of trials. Independent of the target-defining color, however, participants task was to identify the target's orientation (i.e., vertical versus horizontal). It should be noted that this difference between the two types of task, which permitted the PCN to be computed for visual stimuli, had no consequences for the second parameter of interest: the LRP, which is computed relative to the respective side of the executed motor response (see above). Further, we deliberately introduced the same task requirements for both the auditory and the visual task (i.e., stimulus identification, rather than detection or localization; see Töllner et al., 2012b, for a systematic comparison of different task settings), in order to obtain comparable response latencies (see RT analysis below). Participants responded, for example, to the auditory task with a single key press with the left hand, using the index and middle finger to indicate the high versus low pitch of the tones, respectively; and with a single key press using the right index and middle finger to indicate the target's (vertical versus horizontal) orientation in the visual task. The S–R mappings were reversed across hands and fingers after the first half of each experimental session and counterbalanced across participants. Prior to the start of the first as well as second half of each experimental session, at least one block of practice was administered to permit participants to become familiar with the required S–R mapping in each task. After each block, participants received summary performance statistics (mean error rate and RT).

EEG RECORDING AND DATA ANALYSIS

The EEG was continuously digitized from 64 Ag/AgCl active electrodes (actiCAP system, BrainProducts Munich) at 1 KHz. Electrodes were mounted on an elastic cap (Easy Cap, FMS) and placed in accord to the International 10–10 System (American

Electroencephalographic Society, 1994). The horizontal and vertical electrooculogram was monitored by means of electrodes placed at the outer canthi of the eyes, and the superior and inferior orbits, respectively. All electrophysiological signals were amplified by BrainAmp amplifiers (BrainProducts, Munich) using a 0.1–250 Hz bandpass filter, and filtered offline with a 0.5–40 Hz band-pass (Butterworth infinite-impulse-response filter, 24 dB/Oct). All electrodes were referenced to FCz and re-referenced offline to averaged mastoids. Impedances were kept below 5 kΩ.

Prior to segmenting the EEGs, the raw data was visually inspected in order to identify and manually remove non-stereotypical noise in the signals. This was followed by an infomax independent-component analysis (ICA) run to identify components representing blinks and/or horizontal eye movements, and to remove these artefacts before back-projection of the residual components. The continuous EEG was then epoched into 3.0 s segments, ranging from 1.3 s before to 1.7 s after stimulus onset. Next, a baseline correction was performed based on the 200 ms pre-stimulus interval. Only trials with correct responses in both (dual) tasks and without artifacts—defined as any signal exceeding ±60 μV, bursts of electromyographic activity (permitted maximal voltage steps/sample point of 50 μV), and activity fluctuating less than 0.5 μV within 500 ms (indicating "dead" channels)—were considered on an individual-channel basis for further analysis. The signals were then re-epoched into 0.6 s segments ranging from 200 ms before to 400 ms s after stimulus onset for the PCN analysis and, respectively, into 1.2 s segments ranging from 1 s before to 200 ms after response onset for the rLRP analysis, before the ERP waveforms were averaged.

The LRP was quantified by subtracting ERPs measured at medial central electrodes (C3/C4) ipsilateral to the response side from contralateral ERPs. The onset latencies of the LRPs were computed according to the jackknife-based scoring method (Ulrich and Miller, 2001), which defines the LRP onset as the point in time at which the LRP activation meets a specific criterion value relative to the pre-stimulus baseline. As proposed by Ulrich and Miller (2001), we used 90% of the maximum LRP activation as optimal criterion for defining rLRP onset latencies. LRP amplitudes were calculated averaging five data points before and after the maximum deflection obtained in the 250 ms pre-response time interval. The PCN was computed by subtracting ERPs measured at lateral parieto-occipital electrode sites (PO7/PO8) ipsilateral to the target's location from contralateral ERPs. The latencies of the PCNs were defined individually as the maximum negative-going deflection in the 150–350 ms post-stimulus interval. PCN amplitudes were computed averaging five data points before and after this maximum deflection.

For both the first and second (dual) task responses, differences in behavioral (RTs, error rates) as well as electrophysiological measures (rLRP onset latencies/amplitudes; PCN latencies/amplitudes) were assessed by carrying out separate two-way repeated-measure analyses of variance (ANOVAs) with the factors Task Order Predictability (TOP) (predictive, non-predictive) and SOA (150 ms, 300 ms, and 600 ms). Significant main effects and/or interactions were further examined by means of *post-hoc* comparisons (Tukey HSD).

RESULTS

RESPONSES TO THE FIRST TASK

Behavior

When the auditory task was performed first, we found RTs to be modulated interactively by TOP and SOA. As can be seen from the left panel of **Figure 2**, RTs increased monotonically with decreasing inter-task interval for non-predictive task orders (denoted by red lines), whereas there was no SOA effect for predictive task orders (denoted by blue lines). Statistically, both main effects [TOP: $F_{(1, 11)} = 45.09$, $p < 0.001$; SOA: $F_{(2, 22)} = 26.44$, $p < 0.001$] as well as their interaction [$F_{(2, 22)} = 20.732$, $p < 0.001$] were significant. *Post-hoc* analyses confirmed that, for non-predictive task orders, RTs increased from long to intermediate SOAs [$p < 0.05$] and from intermediate to short SOAs [$p < 0.001$]. By contrast, no statistical differences were evident among the various SOA levels for predictive task order conditions [all $p > 0.07$]. With regard to the error rates (depicted in **Table 1**), more incorrect responses were made when the order of the two

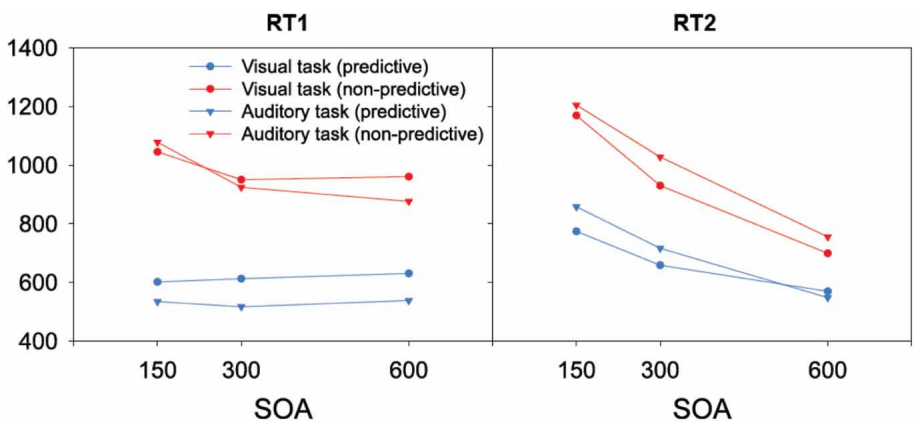


FIGURE 2 | Reaction times of the present study as a function of task order (first task, second task), task order predictability (predictive, non-predictive), and stimulus onset asynchronies (150 ms, 300 ms, 600 ms).

Table 1 | Error rates in the present study as a function of task order (first task, second task), task order predictability (predictive, non-predictive), and stimulus onset asynchrony (150 ms, 300 ms, 600 ms).

Task	SOA	First task		Second task	
		predictive	non-predictive	predictive	non-predictive
Visual	150	2.0	3.0	3.6	4.0
Task	300	1.7	3.1	4.2	3.2
	600	1.7	2.5	4.5	3.5
Auditory	150	4.6	3.5	2.4	3.7
Task	300	3.5	2.6	1.1	2.0
	600	3.5	3.0	0.9	1.9

tasks was non-predictive (2.5% vs. 1.4%) and when they were separated by short (3.0%) rather than intermediate (1.5%) and long (1.4%) inter-task intervals, yielding significant main effects of TOP [$F_{(1, 11)} = 12.24$, $p < 0.001$] and SOA [$F_{(2, 22)} = 12.00$, $p < 0.001$].

The same overall data pattern was revealed when the visual task was performed first (left panel of **Figure 2**). RTs again increased monotonically with decreasing inter-task interval for non-predictive task orders (denoted by red lines), with no SOA influences evident for predictive orders (denoted by blue lines). As for the auditory task, this was substantiated by a significant main effect of TOP [$F_{(1, 11)} = 49.01$, $p < 0.001$], which interacted significantly with SOA [$F_{(2, 22)} = 5.362$, $p < 0.05$]. In more detail, intermediate and short inter-task intervals differed significantly for non-predictive task orders [$p < 0.01$], but not for predictive orders [all $p > 0.07$]. Participants again made more errors with non-predictive relative to predictive task orders (2.9% vs. 1.8%), evidenced by a significant main effect of TOP [$F_{(1, 11)} = 7.723$, $p < 0.05$].

Posterior-contralateral-negativity

Grand average ERP waveforms elicited by visual first-task displays are shown separately for contra—and ipsilateral target stimuli with respect to the hemisphere of the recording electrode (PO7/PO8) in the top panel of **Figure 3**, while the bottom panel presents the corresponding (contralateral-minus-ipsilateral) difference waves as a function of SOA (short, intermediate, and long) and TOP (predictive, non-predictive). For all six (TOP × SOA) conditions, a solid PCN was evoked, visible as a more negative (i.e., less positive) voltage in the time range approximately 150–250 ms *post-stimulus*. To statistically corroborate that the PCN was elicited reliably for the first task, we initially performed a repeated-measures ANOVA with the single factor Period (Baseline versus PCN activation). Baseline activation values were determined—similar to the PCN amplitudes (see above)—by averaging across five sample points prior to and following the maximum negatively directed deflection in the 200 ms pre-stimulus interval. This analysis revealed the effect of Period [$F_{(1, 11)} = 13.81$, $p < 0.003$] to be significant, confirming the presence of the PCN.

As further can be seen from (the bottom panel of) **Figure 3**, the PCN was more pronounced for predictive relative to non-predictive task orders at short (−1.72 μV vs. −1.29 μV) and intermediate (−2.16 μV vs. −1.06 μV), but not long (−1.83 μV vs. −2.05 μV), inter-task intervals. In addition, for non-predictive task orders, the rise of the PCN appeared to be slightly delayed for short (221 ms) and intermediate (218 ms), as compared to long (209 ms), inter-task intervals; in contrast, this pattern was reversed for predictive task orders (short: 207 ms; intermediate: 211 ms; long: 223 ms). Both observations were substantiated by a significant main effect of TOP for PCN amplitudes [$F_{(1, 11)} = 6.74$, $p > 0.025$], as well as a significant interaction of both factors for PCN amplitudes [$F_{(2, 22)}$: 3.99, $p < 0.033$] and latencies [$F_{(2, 22)} = 6.22$, $p < 0.007$]. Subsequent *post-hoc* contrasts confirmed faster PCN elicitation with predictive relative to non-predictive task orders for short inter-task intervals, and vice versa for long intervals (all $p < 0.05$).

Lateralized-readiness-potential

The top panel of **Figure 4** presents grand average ERP waveforms elicited by both visual and auditory stimuli, separately for the recording electrodes (C3/C4) contra—and ipsilateral to the side of the respective motor response, while the bottom panel shows the corresponding (contralateral-minus-ipsilateral) difference waves as a function of inter-task SOA (short, intermediate, and long) and TOP (predictive, non-predictive). All six (TOP × SOA) conditions triggered a solid LRP, visible as a more negative (i.e., less positive) voltage most pronounced approximately in the 200 ms pre-response time window. First, we compared activation values obtained during the baseline and LRP time windows (see above) in a repeated-measure ANOVA with the factor Period (Baseline versus LRP activation). A highly significant main effect [$F_{(1, 11)} = 13.01$, $p < 0.004$] of Period corroborated that the LRP was reliably triggered. As further illustrated in **Figure 4** (bottom panel), the rise of the LRP occurred earlier (relative to response onset) and was more pronounced for predictive (158 ms, −1.32 μV) relative to non-predictive task order trials (210 ms, −1.06 μV), with no differences discernable across inter-task intervals. Statistically, these observations were confirmed for LRP onset latencies [$F_{c(1, 11)}$: 5.61, $p_c > 0.037$], but failed to reach significance level for LRP amplitudes [$F_{(1, 11)}$: 2.05, $p > 0.180$].

RESPONSES TO THE SECOND TASK

Behavior

As illustrated in the right panel of **Figure 2**, RTs to the *auditory* second task were generally increased with non-predictive (denoted by red lines) relative to predictive (denoted by blue lines) task orders [$F_{(1, 11)} = 53.40$, $p < 0.001$], and decreasing inter-task intervals [$F_{(2, 22)} = 459.74$, $p < 0.001$], with the latter replicating the classic SOA effect in PRP dual-tasks. Further, TOP and SOA interacted significantly [$F_{(2, 22)} = 19.84$, $p < 0.001$], owing to a monotonically increasing TOP effect from long to intermediate SOAs [$p < 0.001$], and intermediate to short SOAs [$p < 0.001$]. In addition, participants exhibited significantly [$F_{(2, 22)} = 6.86$, $p < 0.01$] more errors with short (4.0%) relative to intermediate (3.1%) and long (3.2%) inter-task intervals (see **Table 1**). There was no main effect of, or interaction with, TOP on errors.

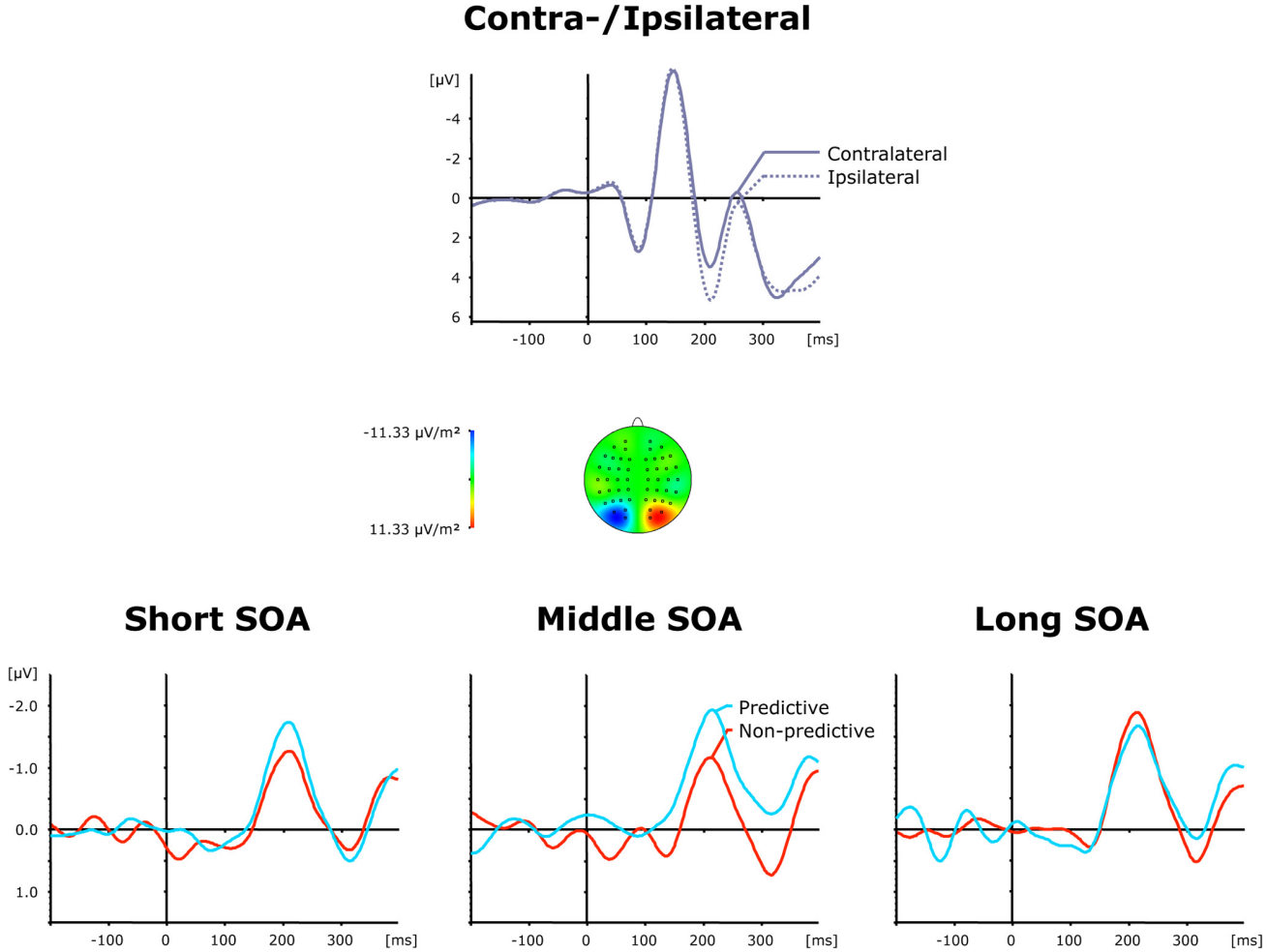


FIGURE 3 | Grand-averaged event-related brain responses time-locked to the visual stimulus of the first task at electrodes PO7/PO8. Top panel: Waveforms contra—and ipsilateral to the target location. **Central panel**: Topographical map of the PCN scalp distribution at the point in time when the difference between contra—and ipsilateral waveforms reached its maximum. These maps were computed by mirroring the contra-minus-ipsilateral difference waves to obtain symmetrical values for both hemispheres (using spherical spline interpolation). **Bottom panel**: PCN difference waves obtained by subtracting ipsilateral from contralateral activity for each of the six TOP × SOA conditions.

Similar to the response times to the first task, RTs to the *visual* second-task displays generally matched the overall pattern of the auditory RTs: both main effects [TOP: $F_{(1,\ 11)} = 36.40$, $p < 0.001$; SOA: $F_{(2,\ 22)} = 117.11$, $p < 0.001$] as well as their interaction [$F_{(2,\ 22)} = 44.43$, $p < 0.001$] were significant. As can be seen in the right panel of **Figure 2**, the TOP effect was again stronger for intermediate than for long inter-task intervals [$p < 0.001$], and even more pronounced for short relative to intermediate SOAs [$p < 0.001$]. No effects reached statistical significance in the error data.

Posterior-contralateral-negativity

Grand average ERP waveforms elicited by visual second-task displays are presented separately for contra—and ipsilateral target stimuli relative to the hemisphere of the recording electrode (PO7/PO8) in the top panel of **Figure 5**; the bottom panel shows the corresponding difference waves as a function of SOA (short, intermediate, and long) and TOP (predictive, non-predictive). In all six experimental conditions, a solid PCN was elicited, evident as a more negative (i.e., less positive) voltage in a time range similar to the PCNs evoked by the first task. An initial ANOVA with the single main term Period (Baseline versus PCN activation) yielded a highly significant main effect [$F_{(1,\ 11)} = 10.99$, $p < 0.007$], confirming PCN elicitation in response to visual second-task displays. As further shown by the bottom panel of **Figure 5**, the TOP effect on the PCN timing was dependent on the inter-task interval. The PCN was delayed for non-predictive as compared to predictive task orders at short SOAs (233 ms vs. 204 ms), but expedited at long SOAs (202 ms vs. 214 ms), without any timing difference at intermediate SOAs (210 ms vs. 210 ms)—statistically confirmed by a significant TOP × SOA interaction on PCN latencies [$F_{(2,\ 22)}$: 3.49, $p < 0.048$]. By contrast, there were no reliable differences in the associated PCN amplitudes [main effect TOP: $F < 1$, $p >$

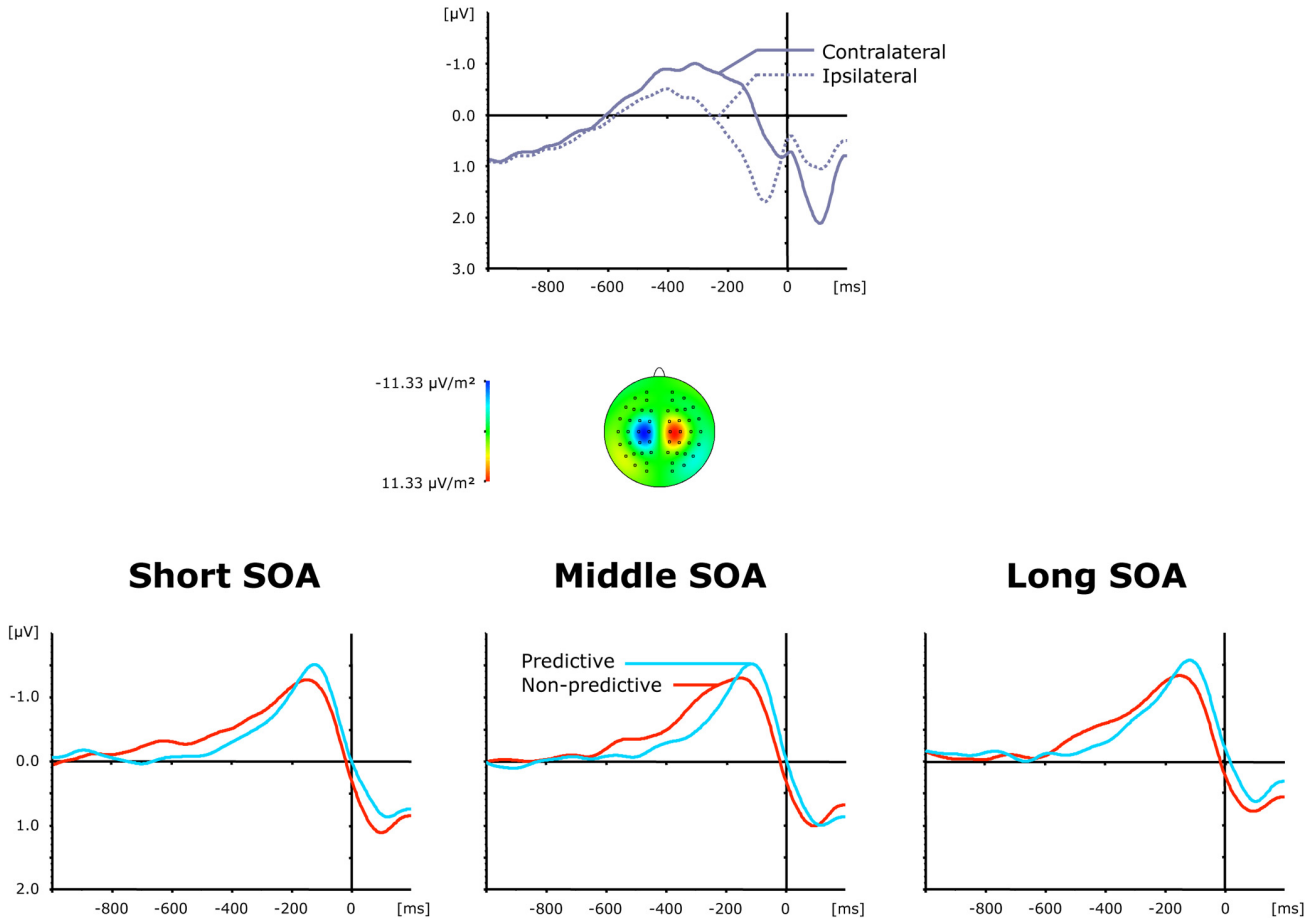


FIGURE 4 | Grand-averaged event-related brain responses time-locked to the motor response of the first task at electrodes C3/C4. Top panel: Waveforms contra—and ipsilateral to the response side. **Central panel**: Topographical map of the LRP scalp distribution at the point in time when the difference between contra—and ipsilateral waveforms reached its maximum. These maps were computed by mirroring the contra-minus-ipsilateral difference waves to obtain symmetrical values for both hemispheres (using spherical spline interpolation). **Bottom panel**: LRP difference waves obtained by subtracting ipsilateral from contralateral activity for each of the six TOP × SOA conditions.

0.416; main effect SOA: $F < 1$, $p > 0.548$; interaction: $F < 1.15$, $p > 0.338$].

Lateralized-readiness-potential

The top panel of **Figure 6** presents average ERP waveforms elicited by visual and auditory stimuli belonging to the second task, separately for the recording electrode (C3/C4) contra—and ipsilateral to the motor-response side, while the bottom panel shows the corresponding difference waves as a function of SOA (short, intermediate, and long) and TOP (predictive, non-predictive). In all six experimental conditions, a solid LRP was triggered, which can be seen as a more negative (i.e., less positive) voltage strongest in the 200 ms pre-response time range, and following the preceding contralateral positivity (i.e., representing the corresponding contralateral motor response) associated with the first task. Statistically, an initial ANOVA with the single factor Period (Baseline versus LRP activation) confirmed LRP presence [$F_{(1,\ 11)} = 11.73$, $p < 0.006$]. Furthermore, as shown in **Figure 6** (bottom panel), the rise of the LRP occurred faster (relative to response onset) for long (134 ms) relative to intermediate × (162 ms) and short (161 ms) inter-task intervals, with a slightly expedited LRP onset for predictive (126 ms) relative to non-predictive (142 ms) task order trials at long SOAs. By contrast, there were no reliable differences in LRP magnitude across the SOA conditions. Statistically, these observations were substantiated by a significant main effect of SOA for LRP onset latencies [$F_{c(2,\ 22)}$: 3.57, $p_c > 0.045$], and non-significant effects [TOP: $F < 1$, $p > 0.712$; SOA: $F < 1$, $p > 0.638$; interaction: $F < 1$, $p > 0.879$] for LRP amplitudes. Note that the TOP × SOA interaction on LRP onset latencies failed to reach significance [$F_c < 1$, $p_c > 0.794$].

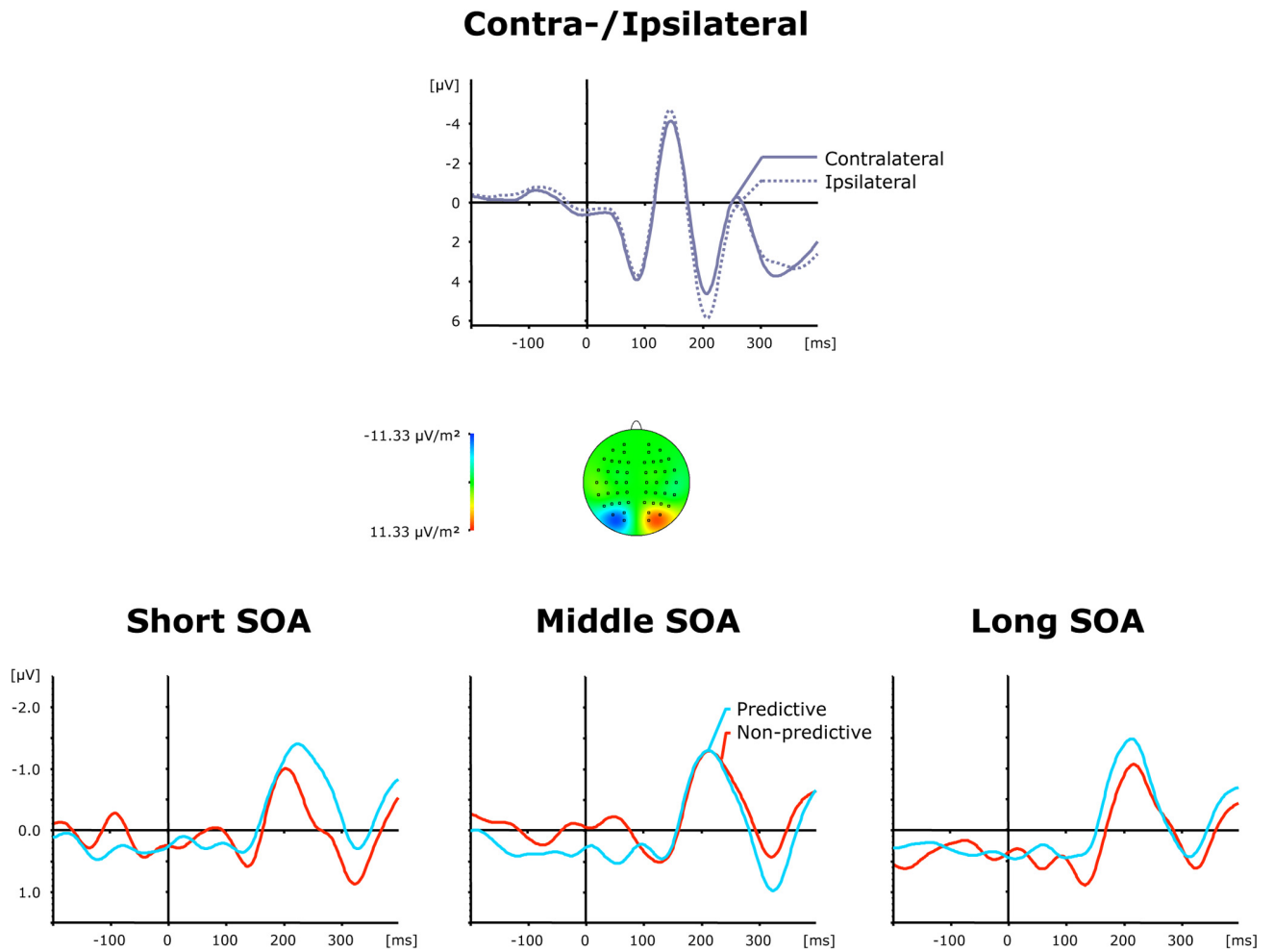


FIGURE 5 | Grand-averaged event-related brain responses time-locked to the visual stimulus of the second task at electrodes PO7/PO8. Top panel: Waveforms contra—and ipsilateral to the target location. **Central panel**: Topographical map of the PCN scalp distribution at the point in time when the difference between contra—and ipsilateral waveforms reached its maximum. These maps were computed by mirroring the contra-minus-ipsilateral difference waves to obtain symmetrical values for both hemispheres (using spherical spline interpolation). **Bottom panel**: PCN difference waves obtained by subtracting ipsilateral from contralateral activity for each of the six TOP × SOA conditions.

DISCUSSION

When investigating the processing dynamics underlying bottlenecks in PRP-type of dual-task paradigms, two behavioral findings robustly emerge: (1) for the first task, increasing RTs with decreasing inter-task SOAs when the task order is made non-predictive; and (2), independent of TOP, increasing RTs with decreasing inter-task SOAs for the second task. Both RT effects traditionally have been advocated to originate *exclusively* from capacity limitations (i.e., bottlenecks) residing at the *central* stage (i.e., central system) of the information-processing stream (e.g., Pashler, 1994; Sigman and Dehaene, 2006). Here, we provide electrophysiological evidence that challenges this classic view.

ELECTROENCEPHALOGRAPHIC MEASUREMENT OF PERCEPTUAL CAPACITY LIMITATIONS IN DUAL-TASK

Evidence for capacity limitations at a perceptual stage of processing derives from the activation pattern of the PCN, which—owing to its latency and extraction method from the ERP—has been demonstrated to reflect pure perceptual processes (e.g., Luck and Hillyard, 1994; Eimer, 1996) within the visual modality. In particular, the latency of this lateralized brain response is indicative of the time required by the human visual system to focally select the target amongst distracter items in visual space (e.g., Woodman and Luck, 1999; Töllner et al., 2011a), whereas its magnitude indicates the amount of focal-attentional resource allocation to the target location.

For the *first* task, we found TOP and inter-task SOA to interactively influence the speed and activation strength of the PCN: for predictive task orders, the PCN was elicited (on average ∼14 ms) earlier with short relative to long inter-task intervals, whereas this pattern was reversed for non-predictive task orders, with (on average ∼14 ms) shorter PCN latencies for long relative to short SOAs. Associated with this, the magnitude of the PCN was reduced for short and intermediate, but not long,

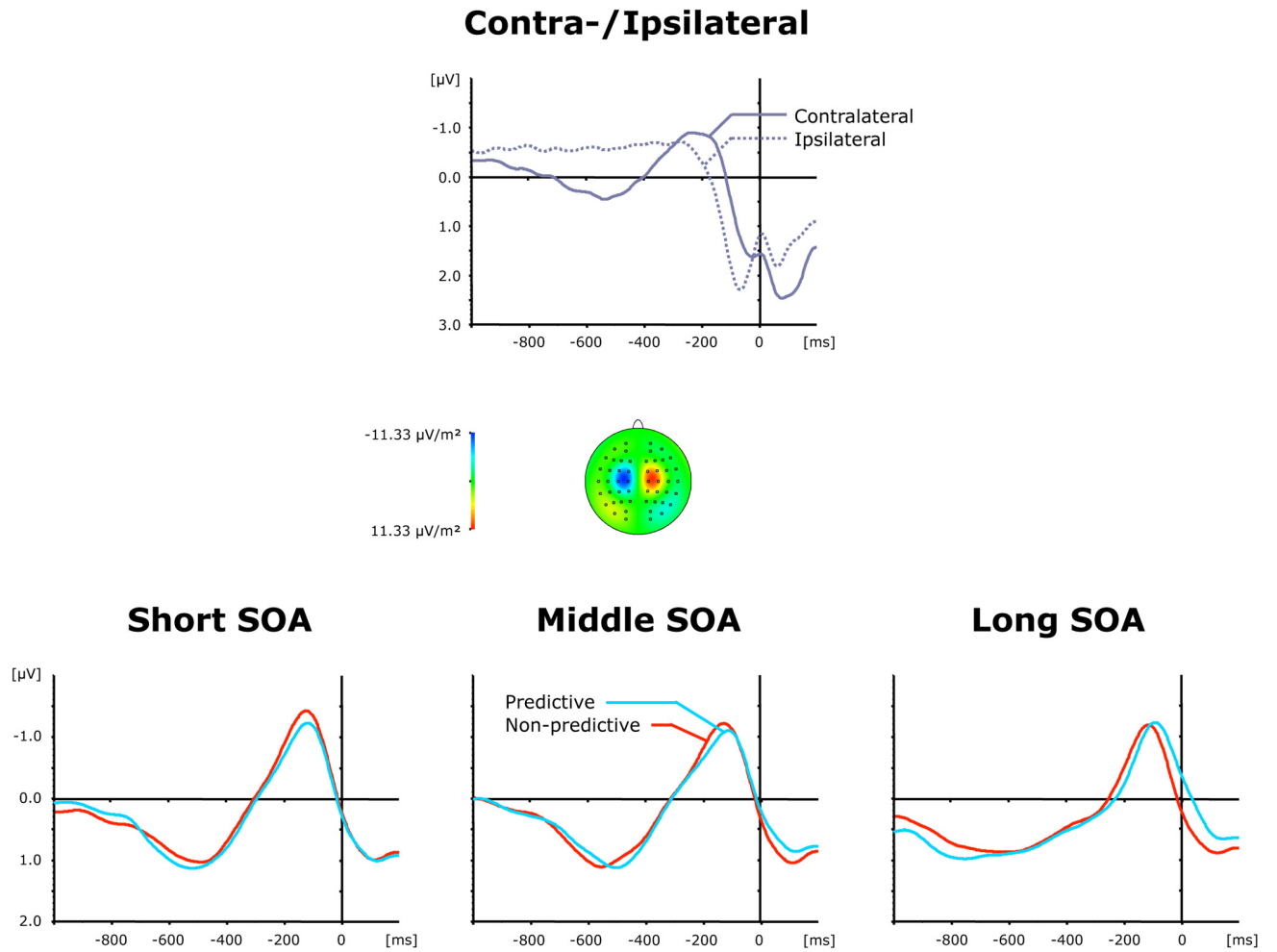


FIGURE 6 | Grand-averaged event-related brain responses time-locked to the motor response of the second task at electrodes C3/C4. Top panel: Waveforms contra—and ipsilateral to the response side. **Central panel**: Topographical map of the LRP scalp distribution at the point in time when the difference between contra—and ipsilateral waveforms reached its maximum. These maps were computed by mirroring the contra-minus-ipsilateral difference waves to obtain symmetrical values for both hemispheres (using spherical spline interpolation). **Bottom panel**: LRP difference waves obtained by subtracting ipsilateral from contralateral activity for each of the six TOP × SOA conditions.

inter-task SOAs when the order of the two upcoming (dual) tasks could not be predicted. At variance with classical PRP models—such as the passive bottleneck model (e.g., Pashler, 1994)—which do not envisage bottlenecks to arise at early sensory processing levels, this set of findings provides support for a recent proposal of Szameitat and colleagues (2006; see also De Jong, 1993), namely, that processes of task-order scheduling may not operate exclusively at a central level of tasks or task sets. Instead, task-order scheduling may already affect perceptual processes via *priming* the modality of the task expected to be presented first. In particular, Szameitat et al. (2006) studied a PRP paradigm in which the order of the two upcoming (dual) tasks changed randomly on a trial-by-trial basis. They found RT (to both the first and the second task) to be speeded markedly when participants had to perform the identical, relative to a different, task order as on the previous trial. As suggested by Szameitat et al., this RT benefit may result, at least in part, from pre-activating the respective sensory modality when participants could properly anticipate the correct task order. However, since their study used functional magnetic resonance imaging in relation to RT data, any suggestions regarding the temporal dynamics underlying this task-order repetition advantage had to remain speculative.

The notion of pre-activating sensory modalities as a function of the previous trial bears a close resemblance to the "Modality-Weighting" Account (MWA; Töllner et al., 2009), originally devised to explain the "modality-shift" effect in cross-modal attention studies (e.g., Spence et al., 2001). According to the MWA, the outcome of pre-attentive saliency computations guides the deployment of focal-attentional selection (analogous to the dimensional-weighting idea of Müller and colleagues for the visual modality; e.g., Gramann et al., 2007, 2010; Müller et al., 2010). In more detail, it is assumed that basic target (e.g., feature contrast) signals computed in separate

modality-specific processing modules (e.g., vision and audition) can be top-down weighted prior to being integrated by units of the attention-guiding overall-saliency map. Critically, however, the total amount of available weight is limited, so that an increase in the weight for one modality goes at the expense of the other modality. Following this logic, being able to predict the correct modality (i.e., with predictable task orders) might have led participants to top-down set themselves to the respective sensory modality of the first task (say: vision), leading to an enhanced coding of target signals—as indicated by stronger PCN activations—and, thus, faster selection of targets defined in this weighted (i.e., vision), relative to the non-weighted (in the example, audition), modality. At the same time, this would help participants shield processing for the first task against interference from (second target) signals defined in a sensory modality other than the first target (analogous to dimension-based shielding within a modality; e.g., Müller et al., 2009). This shielding effect would be most prominent under conditions in which the second target signal occurs in close succession to the first target, such as in the present short to intermediate SOA conditions. By contrast, when the task order is non-predictable, weighting of the correct modality is feasible only at chance level, as a result of which a temporally close signal defined in a modality other than the first signal would capture processing resources. However, when the temporal distance between the two targets becomes relatively large—as in the present long SOA condition—there is no longer any interference, so that processing of the (first) visual target can proceed as smoothly with non-predictable as with predictable task orders.

Regarding the *second* task, we again observed the timing (but not the magnitude) of the PCN to be interactively determined by both factors: for predictive task orders, the PCN was (on average ∼27 ms) delayed with short inter-task intervals, with a graded decrease in PCN latencies for visual targets presented following intermediate and long intervals after the first (auditory) target. By contrast, no SOA effect was evident for non-predictive task orders. This clearly demonstrates that already early, pre-attentive perceptual encoding processes contribute to the well-established SOA effect of increased behavioral RTs to the second task. Of theoretical significance, this finding indicates that there is a limit to the total amount of attentional resources not only within a given sensory modality, but also across modalities—the central assumption of the MWA (Töllner et al., 2009). Restated, perceptual processing resources are not confined to a single sensory modality, but can be shifted from one modality to another in order to optimize target processing. For the present study, this implies that participants needed more time to focally select the visual target item with short (relative to intermediate and long) inter-task SOAs because a significant amount of cross-modal attentional processing resources was already captured by, and still bound to, the first, auditory target stimulus. At intermediate and longer SOAs, however, these resources became available again, thus expediting focal target selection in the visual modality.

The absence of differences in PCN magnitude across different inter-task intervals appears to be at variance with Jolicœur and colleagues (e.g., Brisson and Jolicoeur, 2007a,b, see also Lien and Croswaite, 2011), who observed an SOA effect under predictable task order conditions. In particular, these authors reported attenuated PCN activations for shorter relative to longer SOAs, which they took to suggest that, at relatively short SOAs, participants could not deploy focal attention to the (second) visual target as efficiently when their *central* attention was still engaged on the (first) auditory target. One reason for the absence of such an SOA effect in the present study might lie in the particular SOA introduced, namely: 150, 300, and 600 ms. By contrast, Brisson and Jolicoeur (2007a) had also used conditions with much longer SOAs (300, 650, and 1000 ms), thus introducing greater differences between the SOAs. Accordingly, any differences among the SOAs used in the present study may have been too small to yield statistically significant effects. In fact, comparing the signal strength of the PCNs to the second task with those elicited by the first task discloses at least a numerical reduction (on average ∼0.4 μV) which—in line with Brisson and Jolicoeur (2007a,b)—might reflect the automatic engagement of attentional processing resources on the just previously presented auditory target stimulus.

ELECTROENCEPHALOGRAPHIC MEASUREMENT OF MOTOR CAPACITY LIMITATIONS IN DUAL-TASK

Evidence for capacity limitations existing at the motor stage of processing is revealed from the activation pattern of the LRP (e.g., Coles, 1989). Recall that, in the present study, all LRP activation values were derived time-locked to the onset of the respective motor response. As demonstrated by Miller (2007), amongst others, the onset latency of this lateralized brain response reflects the time required by the human motor system to activate and execute the motor response, whereas its activation strength indicates how forceful the response produced was. Accordingly, both LRP parameters mirror pure motor processes occurring after the completion of *central* stimulus-response translation (i.e., response selection) processes.

For the *first* task, we found response execution times—as indexed by faster LRP onset latencies—to be greatly expedited (on average ∼53 ms) when participants knew, rather than did not know, in advance the order of the two upcoming (dual) tasks. This finding can be explained by the operation of another weighting mechanism: one located at the stage of motor-response production. According to this notion of response weighting ("Response-Weighting" Account, RWA; e.g., Töllner et al., 2012b), participants might top-down set their response system to the respective motor effector (e.g., left index finger) used for the first task in advance when the task order can be correctly anticipated. This, in turn, would lead to the pre-activation of motor units that represent this response (relative to other response alternatives), so that, when the first task is presented, less additional motor evidence would be required to reach the threshold for response initiation and execution. In other words, the RWA assumes a limit to the total amount of processing resources that can be allocated to the various motor responses (i.e., effectors) on the motor production stage. Thus, weighting of one motor effector would lead to facilitated processing of this response, and goes at the expense of alternative motor responses. Note that this

notion is different from the MWA (see above), which assumes limited attentional resources that can be allocated to the various sensory modalities (e.g., vision, audition) at a pre-attentive level, affecting the integration of modality-specific feature contrast signals at the attention-guiding master map. As further evidenced by a recent compound-search study (Töllner et al., 2008; Wiegand et al., in press), weighting mechanisms within sensory modalities operate independently of response-based weighting dynamics.

With respect to the *second* task, we observed the timing of the LRP (relative to response onset) to be speeded (by on average ∼28 ms) for long, as compared to both intermediate and short, SOAs. This pattern of rLRP onset timing differences replicates the observations of Osman and Moore (1993) who had likewise observed such a numerical (though, in this study, statistically non-significant) pattern of speeded processing times for long relative to both short and intermediate inter-task intervals at the stage of response execution. This demonstrates that motor response dynamics can—together with perceptual and central processing dynamics— contribute to the classic, behavioral SOA effect on RTs to the second task, traditionally assumed to originate exclusively from central bottlenecks. Accordingly, the present findings indicate that the execution of simple (e.g., button press) responses can be affected by the motor system having had to execute another (prior) response at a short time beforehand (as in the present short and intermediate SOA conditions), even when this first response was initiated by the motor cortex of the contralateral hemisphere. By contrast, sufficient time in-between the two dual tasks (as in the present long SOA condition) substantially reduces the time demands required for executing the motor response of the second task. The present electrophysiological evidence of impaired response execution for the second task is also in line with a recent behavioral study by Ulrich et al. (2006), who systematically manipulated the temporal demands for executing the first response. Specifically, participants had to respond to the first task by performing a ballistic movement, namely, move a slider to one of two possible target locations indicated by the pitch of a tone (see Ulrich et al., 2006, for further methodological details). Crucially, Ulrich et al. found that RTs to the second task increased systematically with increasing response execution demands for the first task, which they took to suggest that response execution can be part of the processing bottleneck(s) in classical PRP paradigms.

CONCLUSION

In conclusion, the present study was designed to examine for dual-task interference arising at processing stages either prior or subsequent to central-stage processing. We found TOP to interact with inter-task SOA in determining the speed of (visual)

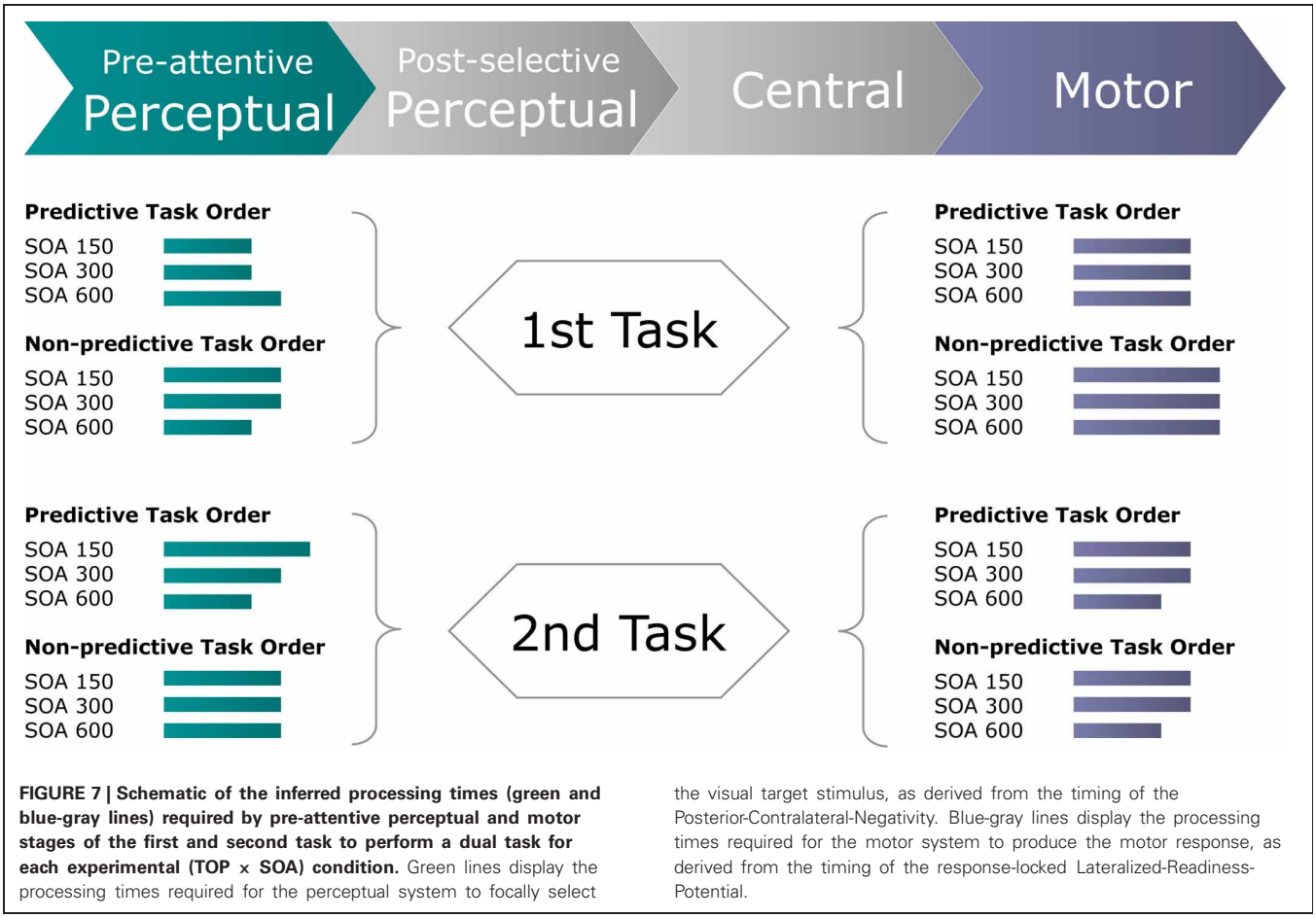


FIGURE 7 | Schematic of the inferred processing times (green and blue-gray lines) required by pre-attentive perceptual and motor stages of the first and second task to perform a dual task for each experimental (TOP × SOA) condition. Green lines display the processing times required for the perceptual system to focally select the visual target stimulus, as derived from the timing of the Posterior-Contralateral-Negativity. Blue-gray lines display the processing times required for the motor system to produce the motor response, as derived from the timing of the response-locked Lateralized-Readiness-Potential.

perceptual processes for both the first and the second task. By contrast, response execution times were influenced independently by TOP for the first, and by inter-task SOA for the second, task (see **Figure 7**). Together, this set of findings complements classical (e.g., Pashler, 1994) as well as advanced versions (e.g., Sigman and Dehaene, 2006) of the central bottleneck model by providing electrophysiological evidence for modulations of both perceptual and motor processing dynamics that, in summation with central capacity limitations, give rise to the well-known behavioral PRP outcome.

REFERENCES

American Electroencephalographic Society. (1994). American electroencephalographic society. Guideline thirteen: guideline for standard electrode position nomenclature. *J. Clin. Neurophysiol.* 11, 111–113.

Brisson, B., and Jolicoeur, P. (2007a). Electrophysiological evidence of central interference on the control of visual-spatial attention. *Psychon. Bull. Rev.* 14, 126–132.

Brisson, B., and Jolicoeur, P. (2007b). A psychological refractory period in access to visual short-term memory and the deployment of visual-spatial attention: multitasking processing deficits revealed by event-related potentials. *Psychophysiology* 44, 323–333.

Brisson, B., Robitaille, N., and Jolicoeur, P. (2007). Stimulus intensity affects the latency but not the amplitude of the N2pc. *Neuroreport* 18, 1627–1630.

Coles, M. G. H. (1989). Modern mind-brain reading: psychophysiology, physiology, and cognition. *Psychophysiology* 26, 251–269.

De Jong, R. (1993). Multiple bottlenecks in overlapping task performance. *J. Exp. Psychol. Hum. Percept. Perform.* 19, 965–980.

De Jong, R. (1995). The role of preparation in overlapping-task performance. *Q. J. Exp. Psychol.* 48, 2–25.

Eimer, M. (1996). The N2pc component as an indicator of attentional selectivity. *Electroencephalogr. Clin. Neurophysiol.* 99, 225–234.

Eimer, M. (1998). The lateralized readiness potential as an on-line measure of selective response activation. *Behav. Res. Methods Instrum. Comput.* 30, 146–156.

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.

Gramann, K., Töllner, T., Krummenacher, J., Eimer, M., and Müller, H. J. (2007). Brain electrical correlates of dimensional weighting: an ERP study. *Psychophysiology* 44, 277–292.

Gramann, K., Töllner, T., and Müller, H. J. (2010). Dimension-based attention modulates early visual processing. *Psychophysiology* 47, 968–978.

Jiang, Y., Saxe, R., and Kanwisher, N. (2004). Functional magnetic resonance imaging provides new constraints on theories of the psychological refractory period. *Psychol. Sci.* 15, 390–396.

Lien, M. C., and Croswaite, K. (2011). Controlling spatial attention without central attentional resources: evidence from event-related potentials. *Vis. Cogn.* 19, 37–78.

Lien, M. C., Schweickert, R., and Proctor, R. W. (2003). Task switching and response correspondence in the psychological refractory period paradigm. *J. Exp. Psychol. Hum. Percept. Perform.* 29, 692–712.

Liepelt, R., Strobach, T., Frensch, P. A., and Schubert, T. (2011). Improved inter-task coordination skills after extensive dual-task practice. *Q. J. Exp. Psychol.* 64, 1251–1272.

Luck, S. J. (1998). Sources of dual-task interference: evidence from human electrophysiology. *Psychol. Sci.* 9, 223–227.

Luck, S. J., and Hillyard, S. A. (1994). Electrophysiological correlates of feature analysis during visual search. *Psychophysiology* 31, 291–308.

Miller, J. (2007). Contralateral and ipsilateral motor activation in visual simple reaction time: a test of the hemispheric coactivation model. *Exp. Brain. Res.* 176, 539–558.

Müller, H. J., Krummenacher, K., Geyer, T., and Zehetleitner, M. (2009). Attentional capture. by salient color singleton distractors is modulated by top-down dimensional set. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 1–16.

Müller, H. J., Töllner, T., Zehetleitner, M., Geyer, T., Rangelov, D., and Krummenacher, J. (2010). Dimension-based attention modulates feed-forward visual processing. *Acta Psychol.* 135, 117–122.

Osman, A., and Moore, C. M. (1993). The locus of dual-task interference: psychological refractory effects on movement-related brain potentials. *J. Exp. Psychol. Hum. Percept. Perform.* 19, 1292–1312.

Pashler, H. (1984). Processing stages in overlapping tasks: evidence for a central bottleneck. *J. Exp. Psychol. Hum. Percept. Perform.* 10, 358–377.

Pashler, H. (1994). Dual-task interference in simple tasks: data and theory. *Psychol. Bull.* 116, 220–244.

Pashler, H., and Johnston, J. C. (1989). Chronometric evidence for central postponement in temporally overlapping tasks. *Q. J. Exp. Psychol.* 41, 19–45.

Schubert, T. (1996). Interferenzeffekte bei der gleichzeitigen Bearbeitung zweier Aufgaben. *Z. Exp. Psychol.* 4, 625–656.

Schubert, T. (1999). Processing differences between simple and choice reactions affect bottleneck localization in overlapping tasks. *J. Exp. Psychol. Hum. Percept. Perform.* 25, 1–18.

Schubert, T. (2008). The central attentional limitation and executive control. *Front. Biosci.* 13, 3569–3580.

Shedden, J. M., and Nordgaard, C. L. (2001). ERP time course of perceptual and post-perceptual mechanisms of spatial selection. *Cogn. Brain Res.* 111, 59–75.

Sigman, M., and Dehaene, S. (2006). Dynamics of the central bottleneck: dual-task and task uncertainty. *PLoS Biol.* 4:e220. doi: 10.1371/journal.pbio.0040220

Sommer, W., Leuthold, H., and Schubert, T. (2001). Multiple bottlenecks in information processing? An electrophysiological examination. *Psychon. Bull. Rev.* 8, 81–88.

Spence, C., Nicholls, M., and Driver, J. (2001). The cost of expecting events in the wrong sensory modality. *Percept. Psychophys.* 63, 330–336.

Strobach, T., Frensch, P. A., Müller, H. J., and Schubert, T. (2012). Age- and practice-related influences on dual-task costs and compensation mechanisms under optimal conditions of dual-task performance. *Aging Neuropsychol. Cogn.* 19, 222–247.

Strobach, T., Frensch, P. A., Soutschek, A., and Schubert, T. (in press). Investigation on the improvement and transfer of dual-task coordination skills. *Psychol. Res.* doi: 10.1007/s00426-011-0381-0. [Epub ahead of print].

Szameitat, A. J., Lepsien, J., von Cramon, D. Y., Sterr, A., and Schubert, T. (2006). Task-order coordination in dual-task performance and the lateral prefrontal cortex: an event-related fMRI study. *Psychol. Res.* 70, 541–552.

Szameitat, A. J., Schubert, T., Müller, K., and von Cramon, D. Y. (2002). Localization of executive functions in dual-task performance with fMRI. *J. Cogn. Neurosci.* 14, 1184–1199.

Töllner, T., Gramann, K., Müller, H. J., Kiss, M., and Eimer, M. (2008). Electrophysiological markers of visual dimension changes and response changes. *J. Exp. Psychol. Hum. Percept. Perform.* 34, 531–542.

Töllner, T., Gramann, M., Müller, H. J., and Eimer, M. (2009). The anterior N1 component as an index of modality shifting. *J. Cogn. Neurosci.* 21, 1653–1669.

Töllner, T., Müller, H. J., and Zehetleitner, M. (2012a). Top-down dimensional weight set determines the capture of visual attention: evidence from the PCN component. *Cereb. Cortex* 22, 1554–1563.

Töllner, T., Rangelov, D., and Müller, H. J. (2012b). How the speed of motor-response decisions, but not focal-attentional selection, differs as a function of task set and target prevalence. *Proc. Natl. Acad. Sci. U.S.A.* 109, E1990–E1999.

Töllner, T., Zehetleitner, M., Gramann, K., and Müller, H. J. (2010). Top-down weighting of visual dimensions: behavioural and electrophysiological evidence. *Vision Res.* 50, 1372–1381.

Töllner, T., Zehetleitner, M., Gramann, K., and Müller, H. J. (2011a). Stimulus saliency modulates pre-attentive processing speed in human visual cortex. *PLoS ONE* 6:e16276. doi: 10.1371/journal.pone.0016276

Töllner, T., Zehetleitner, M., Krummenacher, J., and Müller, H. J. (2011b). Perceptual basis of redundancy gains in visual pop-out search. *J. Cogn. Neurosci.* 23, 137–150.

Ulrich, R., Fernandez, S. R., Jentzsch, I., Rolke, B., Schröter, H., and Leuthold, H. (2006). Motor limitations in dual-task processing under ballistic movement conditions. *Psychol. Sci.* 17, 788–793.

Ulrich, R., and Miller, J. (2001). Using the jackknife-based scoring method for measuring LRP onset effects in factorial designs. *Psychophysiology* 38, 816–827.

Welford, A. T. (1952). The 'psychological refractory period' and the timing of high-speed performance - a review and a theory. *Br. J. Psychol.* 43, 2–19.

Wiegand, I., Finke, K., Müller, H. J., and Töllner, T. (in press). Event-related potentials dissociate perceptual from response-related age effects in visual search. *Neurobiol. Aging* doi: 10.1016/j.neurobiolaging.2012.08.002. [Epub ahead of print].

Woodman, G. F., and Luck, S. J. (1999). Electrophysiological measurement of rapid shifts of attention during visual search. *Nature* 400, 867–869.

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 May 2012; accepted: 27 August 2012; published online: 11 September 2012.

*Citation: Töllner T, Strobach T, Schubert T and Müller HJ (2012) The effect of task order predictability in audio-visual dual task performance: Just a central capacity limitation? Front. Integr. Neurosci. **6**:75. doi: 10.3389/fnint.2012.00075*

frontiers in
INTEGRATIVE NEUROSCIENCE

HYPOTHESIS AND THEORY ARTICLE
published: 29 August 2012
doi: 10.3389/fnint.2012.00062

Focused attention vs. crossmodal signals paradigm: deriving predictions from the time-window-of-integration model

***Hans Colonius*[1]* and *Adele Diederich*[2]**

[1] *Department of Psychology, Carl von Ossietzky Universitaet Oldenburg, Oldenburg, Germany*
[2] *School of Humanities and Social Sciences, Jacobs University Bremen, Bremen, Germany*

Edited by:
Zhuanghua Shi, Ludwig-Maximilians-Universität München, Germany

Reviewed by:
Benjamin A. Rowland, Wake Forest University, USA
Daniel Senkowski, University Medicine, Germany

***Correspondence:**
Hans Colonius, Department of Psychology, Carl von Ossietzky Universitaet Oldenburg, D-26129 Oldenburg, Germany.
e-mail: hans.colonius@uni-oldenburg.de

In the crossmodal signals paradigm (CSP) participants are instructed to respond to a set of stimuli from different modalities, presented more or less simultaneously, as soon as a stimulus from any modality has been detected. In the focused attention paradigm (FAP), on the other hand, responses should only be made to a stimulus from a pre-defined target modality and stimuli from non-target modalities should be ignored. Whichever paradigm is being applied, a typical result is that responses tend to be faster to crossmodal stimuli than to unimodal stimuli, a phenomenon often referred to as "crossmodal interaction." Here, we investigate predictions of the time-window-of-integration (TWIN) modeling framework previously proposed by the authors. It is shown that TWIN makes specific qualitative and quantitative predictions on how the two paradigms differ with respect to the probability of multisensory integration and the amount of response enhancement, including the effect of stimulus intensity ("inverse effectiveness"). Introducing a decision-theoretic framework for TWIN further allows comparing the two paradigms with respect to the predicted optimal time window size and its dependence on the prior probability that the crossmodal stimulus information refers to the same event. In order to test these predictions, experimental studies that systematically compare crossmodal effects under stimulus conditions that are identical except for the CSP-FAP instruction should be performed in the future.

Keywords: focused attention, cross-modal, time-window-of-integration, Bayesian decision theory, exponential distribution

1. INTRODUCTION

In the crossmodal signals paradigm[1] (CSP) participants are instructed to respond to a set of stimuli from different modalities, presented more or less simultaneously, as soon as a stimulus from any modality has been detected. In the focused attention paradigm (FAP), on the other hand, responses should only be made to a stimulus from a pre-defined target modality and stimuli from non-target modalities should be ignored. Thus in FAP, but not in CSP, participants are required to distinguish between target and non-target modality. Whichever paradigm is being applied, a typical result is that responses tend to be faster to crossmodal stimuli than to unimodal stimuli, a phenomenon often referred to as "intersensory (or crossmodal) interaction," already reported in Todd (1912). Many attempts have been made on both the behavioral and neurophysiological level to understand the dynamics of mechanisms that underlie these crossmodal effects (cf. Stein, 2012, for a recent overview). Up to now, however, reaction time models have predominantly been concerned with CSP. The purpose of this paper is to demonstrate how both types of paradigm can be accounted for within the time-window-of-integration (TWIN) modeling framework proposed by the authors (Colonius and Diederich, 2004; Diederich and Colonius, 2004). Moreover, we will extend the decision-making framework for TWIN to include both CSP and FAP. Under appropriate empirical restrictions, TWIN predicts crossmodal interaction effects in one of the paradigms (CSP, say) given crossmodal interaction effects observed in the other (FAP). While permitting a stringent test of this modeling framework by comparing the implementation of CSP and FAP in TWIN, we moreover strive to get a deeper understanding of the cognitive processes elicited by these two different crossmodal paradigms.

The classic explanation for a speed-up of responses to crossmodal stimuli in CSP has been that subjects start preparing a response as soon as the first stimulus has been detected (Raab, 1962). Taking detection times to be random variables and adding some technical assumptions, observed reaction time is represented as the minimum of the reaction times to, say, visual and auditory signals leading to a purely statistical facilitation effect (probability summation) in response speed. Numerous studies have shown that this *separate activation* or *race model* is not sufficient to explain the observed speedup in reaction time, however, (see Diederich and Colonius, 2004, for a review). Using the race model inequality (RMI) (Miller, 1982; Colonius and Diederich, 2006) as a benchmark test, responses to bimodal stimuli have

[1]The terms "redundant targets" or "redundant signals" paradigm are more common but do not explicitly refer to stimuli coming from different sensory modalities.

been found to be faster than predicted by statistical facilitation, in particular, when the stimuli were spatially aligned. Although the RMI test has sometimes been applied to data from both types of paradigm, its validity for FAP data seems problematic as long as no specific assumptions about the effect of a stimulus from the non-target modality winning the race are being made. Moreover, the race model gives no explanation for the decrease in facilitation observed with variations in many crossmodal stimulus properties, e.g., increasing spatial disparity between the stimuli.

An alternative model type *coactivation models* assumes that activation, raised in different sensory channels by presenting crossmodal stimuli, is combined to satisfy a single criterion for response initiation (Miller, 1982). Coactivation models predict faster average reaction time to multiple stimuli compared to single stimuli because the combined activation reaches that criterion faster. Mathematical instantiations of this model type include *superposition* or *counter models* (Schwarz, 1989; Diederich and Colonius, 1991; Diederich, 1995) and *diffusion models* (Schwarz, 1994; Diederich, 1995). Although these models have been quite successful in describing various empirical data sets for CSP, they have as yet no provision to deal with FAP. Note that neither coactivation nor race models can predict inhibition, i.e., sometimes responses to crossmodal stimuli are slower than to unimodal ones.

2. TIME WINDOW OF INTEGRATION MODELING FRAMEWORK: GENERAL DESCRIPTION

The *time-window hypothesis* holds that information from different sensory modalities must not be presented too far apart in time so that integration into a multisensory perceptual unit may occur. The concept, already mentioned over 20 years ago (Meredith et al., 1987; Stein and Meredith, 1993), recently enjoyed increasing popularity on both the neural and behavioral levels of observation (e.g., Lewald et al., 2001; Meredith, 2002; Lewald and Guski, 2003; Spence and Squire, 2003; Wallace et al., 2004; Bell et al., 2005, 2006; Navarra et al., 2005; Romei et al., 2007; Rowland and Stein, 2007; Rowland et al., 2007; Van Wassenhove et al., 2007; Musacchia and Schroeder, 2009; Powers et al., 2009; Royal et al., 2009). Although a "window of integration" has previously been defined for both spatial and temporal aspects of a crossmodal experiment (e.g., Wallace et al., 2004) and has even been suggested for higher-level aspects like semantic congruity (e.g., van Atteveldt et al., 2007), we will confine discussion to the temporal dimension within the reaction time context considered here. To the best of our knowledge, however, the TWIN model framework) is the only effort to develop an explicit quantitative rendering of a crossmodal time-window mechanism (Colonius and Diederich, 2004, 2012) and to introduce a decision-theoretic perspective on predicting an optimal time window (Colonius and Diederich, 2011).

Given that the basic concept of a "race" among neural activities elicited in separate peripheral sensory pathways, i.e., at a very early stage of processing, has considerable intuitive plausibility, the TWIN model retains this concept which is central to separate activation models. The first stage is complemented by a subsequent compound stage of converging processes which comprise neural integration of the input and preparation of a response. This second stage is defined by default: it includes all later, possibly temporally overlapping, processes that are not part of the peripheral processes in the first stage.

The central assumption of the model concerns the temporal configuration needed for multisensory integration to occur:

> [**TWIN assumption**] *Multisensory integration occurs only if all peripheral processes of the first stage terminate within a given temporal interval, the "time window of integration."*

Thus, the window acts as a "filter" determining whether or not afferent information delivered from different sensory organs is registered close enough in time to trigger multisensory integration. Passing the filter is necessary but not sufficient for crossmodal interaction to occur since the amount of interaction may also depend on several other aspects of the stimulus setting, like spatial configuration of the stimuli. The *amount* of crossmodal interaction manifests itself in an increase or decrease of second stage processing time but it is assumed not to depend on how far apart in time the stimuli have been presented (stimulus onset asynchrony, SOA).

For FAP, the TWIN assumption is further constrained in one important respect:

> [**FAP condition**] *Crossmodal interaction in FAP only occurs if (i) a non-target stimulus wins the race in the first stage opening the time window of integration, such that (ii) the termination of the target peripheral process falls into the window.*

One interpretation is that a winning non-target will keep the system in a state of heightened reactivity such that the upcoming target stimulus, if it falls into the time window, will trigger crossmodal interaction. For saccadic eye movements, for example, this may correspond to a gradual inhibition of fixation neurons (in *superior colliculus*) and/or *omnipause* neurons (in *midline pontine* brain stem). If a stimulus from the target modality is the winner of the race in the peripheral channels, second stage processing is initiated without any multisensory integration mechanism being involved.

Although these TWIN model assumptions clearly oversimplify matters, the framework generates several experimentally testable predictions, some of which have already found empirical support in recent studies (cf. Diederich and Colonius, 2007a,b, 2008a,b). Since physically identical stimuli can be presented in both FAP and CSP under the same spatiotemporal configuration, any systematic differences observed in the corresponding reaction times have to be due to the instructions being different. Thus, differences between the two paradigms may allow one to assess, and possibly separate from one another, the contribution of top-down processes and bottom-up processes in multisensory integration.

3. THE FORMAL PRESENTATION OF TWIN FOR FAP AND CSP

For the crossmodal condition, the race in the first stage is based on postulating statistically independent, non-negative continuous random variables representing the durations of the peripheral processes. With V and A denoting these visual and auditory

processing times [2], respectively, the central TWIN assumption introduced above translates into

$$|V - A| < \omega, \tag{1}$$

i.e., peripheral processes V and A terminate within an integration window of width ω. This inequality is the condition for the *event of integration* to occur in the case of CSP, denoted I_{CSP}, and it is obviously equivalent to the union of the events

$$\{V < A < V + \omega\} \cup \{A < V < A + \omega\} \equiv I_{\text{CSP}}.$$

For the FAP with, say, the visual as target modality, the condition for integration is, by translating the FAP condition stated above,

$$I_{\text{FAP}} = \{A < V < A + \omega\}.$$

Therefore, under identical stimulus conditions,

$$I_{\text{FAP}} \equiv I_{\text{CSP}} \cap \{A \text{ is the winner of the race}\}.$$

It follows that any realization of the peripheral processing times V and A that leads to an opening of the time window under the focused attention instruction also leads to that event under the crossmodal signals instruction, i.e., $I_{\text{FAP}} \subset I_{\text{CSP}}$. Thus, the probability of integration under crossmodal signals instruction can not be smaller than that under focused attention instruction: $\Pr(I_{\text{FAP}}) \leq \Pr(I_{\text{CSP}})$, given identical stimulus conditions hold.

3.1. EXPECTED CROSSMODAL REACTION TIME FOR FAP AND CSP

Although events I_{FAP} and I_{CSP} are not empirically observable, the numerical ordering of their associated probabilities leads to a corresponding prediction about mean crossmodal reaction times. Indeed, according to the two-stage assumption, total reaction time in the crossmodal condition can be written as a sum of two random variables:

$$\text{RT}_{VA} = W_1 + W_2, \tag{2}$$

where W_1 and W_2 refer to first and second stage processing times, respectively. With $\Pr(I)$ the probability that integration occurs in CSP or FAP, expected saccadic reaction time in the crossmodal condition ($\text{E}[RT_{VA}]$) then is:

$$\begin{aligned}\text{E}[RT_{VA}] &= \text{E}[W_1] + \text{E}[W_2] \\ &= \text{E}[W_1] + \Pr[I]\text{E}[W_2|I] + (1 - \Pr[I])\text{E}[W_2|\text{not-}I] \\ &= \text{E}[W_1] + \text{E}[W_2|\text{not-}I] \\ &\quad - \Pr[I](\text{E}[W_2|\text{not-}I] - \text{E}[W_2|I]),\end{aligned}$$

where $\text{E}[W_2|I]$ and $\text{E}[W_2|\text{not-}I]$ denote the expected second stage processing time conditioned on interaction occurring (I) or not occurring (not-I), respectively. Putting

$$\Delta \equiv \text{E}[W_2|\text{not-}I] - \text{E}[W_2|I],$$

this becomes

$$\text{E}[RT_{VA}] = \text{E}[W_1] + \text{E}[W_2|\text{not-}I] - \Pr[I] \cdot \Delta. \tag{3}$$

The term $\Pr[I] \cdot \Delta$ can be interpreted as a measure of the expected saccadic RT interaction effect in the second stage with positive Δ values corresponding to facilitation, negative ones to inhibition. The duration of the first stage, W_1, must be defined differently for CSP and FAP:

$$W_1 = \begin{cases} \min(V, A) & \text{for CSP,} \\ V & \text{for FAP,} \end{cases} \tag{4}$$

assuming the visual as target modality in FAP. Thus, for the expected overall reaction time in the crossmodal condition

$$\text{E}[RT_{VA}] = \begin{cases} \text{E}[\min(V, A)] + \mu - P(I_{\text{CSP}}) \cdot \Delta, & \text{for CSP,} \\ \text{E}[V] + \mu - P(I_{\text{FAP}}) \cdot \Delta, & \text{for FAP,} \end{cases} \tag{5}$$

with $\mu \equiv \text{E}[W_2|\text{not-}I]$.

The last equation allows to predict how (observable) mean reaction times for FAP and CSP may differ. In fact, under identical stimulus conditions and assuming facilitation occurs (i.e., $\Delta > 0$), expected crossmodal reaction time can never be longer in CSP than in FAP because both $\text{E}[\min(V, A)] \leq \text{E}[V]$ and $\Pr(I_{\text{FAP}}) \leq \Pr(I_{\text{CSP}})$. Thus,

$$\text{E}[RT_{VA}|\,\text{CSP}] \leq \text{E}[RT_{VA}|\,\text{FAP}].$$

Some empirical support for this prediction was found in an unpublished experiment from our lab, but further empirical testing is required.

3.2. CROSSMODAL RESPONSE ENHANCEMENT FOR FAP AND CSP

In the unimodal condition, no interaction is possible. Thus,

$$\text{E}[RT_{\text{unimodal}}] = \text{E}[W_1] + \text{E}[W_2|\text{not-}I]. \tag{6}$$

Note that in order to relate processing durations in the unimodal conditions to those occurring in the crossmodal conditions, one has to introduce a basic assumption, known as "context independence" or "context invariance" (cf. Ashby and Townsend, 1986; Luce, 1986; Colonius, 1990; Townsend and Eidels, 2011). Informally, it amounts to assuming that the (marginal) distributions of random variables (like V and A) occurring in the crossmodal conditions are identical to the distributions of the corresponding random variables occurring in the unimodal conditions. Although not empirically testable, context invariance has been widely accepted as a plausible modeling constraint and will be used here as well.

[2] For simplicity, we are using V and A for the crossmodal condition in the remainder of this paper, although this could be replaced by any other pair of modalities. Moreover, without losing much generality—since non-zero SOA values can be subsumed as additive constants under V or A—we suppress any reference to values of SOA different from zero.

In analogy to measuring multisensory enhancement in neural responses (cf. Meredith and Stein, 1986; Anastasio et al., 2000), the amount of crossmodal reaction time interaction is measured by relating mean RT in the crossmodal condition to that in the unimodal conditions. The following definition quantifies the percent RT enhancement (Diederich and Colonius, 2004). For visual, auditory, and visual-auditory stimuli with expected reaction times $\mathrm{E}[\mathrm{RT}_V]$, $\mathrm{E}[\mathrm{RT}_A]$, and $\mathrm{E}[\mathrm{RT}_{VA}]$, respectively, *crossmodal response enhancement* (CRE) is defined as

$$\mathrm{CRE} = \begin{cases} \dfrac{\min(\mathrm{E}[RT_V], \mathrm{E}[RT_A]) - \mathrm{E}[RT_{VA}]}{\min(\mathrm{E}[RT_V], \mathrm{E}[RT_A])} \cdot 100, & \text{for RTP,} \\ & \text{and} \\ \dfrac{\mathrm{E}[RT_V] - \mathrm{E}[RT_{VA}]}{\mathrm{E}[RT_V]} \cdot 100, & \text{for FAP,} \end{cases} \tag{7}$$

where the visual is again taken as target modality in the FAP case. Replacing the means by the corresponding expressions from the TWIN model Equation (5) results in

$$\mathrm{CRE} = \begin{cases} \dfrac{\min(\mathrm{E}[V], \mathrm{E}[A]) \\ -\mathrm{E}[\min(V, A)] + P(I_{\mathrm{CSP}}) \cdot \Delta}{\min(\mathrm{E}[V], \mathrm{E}[A]) + \mu} \cdot 100, & \text{for CSP,} \\ & \text{and} \\ \dfrac{P(I_{\mathrm{FAP}}) \cdot \Delta}{\mathrm{E}[V] + \mu} \cdot 100, & \text{for FAP.} \end{cases} \tag{8}$$

Assuming further that visual and auditory intensity are matched, such that $\mathrm{E}[A] = \mathrm{E}[V]$, yields identical denominators in the above ratios. Comparing the corresponding numerators then reveals that response enhancement for CSP is at least as large as that for FAP because (1) $P(I_{\mathrm{FAP}}) \leq P(I_{RTP})$ and (2) the term $\min(\mathrm{E}[V], \mathrm{E}[A]) - \mathrm{E}[\min(V, A)]$, the amount of statistical facilitation, is always non-negative. Therefore, we have

$$\mathrm{CRE(CSP)} \geq \mathrm{CRE(FAP)}. \tag{9}$$

This result holds if $\Delta > 0$, in analogy to the result derived above for crossmodal expected reaction time. Note that it is possible to have an observed CRE(CSP) of zero even if Δ is different from zero: it may have a negative amount just outweighing the statistical facilitation effect.

3.3. THE EFFECT OF INTENSITY VARIATION ON CROSSMODAL RESPONSE ENHANCEMENT

According to the TWIN model assumptions, a direct effect of stimulus intensity only occurs in the peripheral processing channels. In later processing stages, direction and amount of crossmodal interaction are assumed to be modulated by intensity only via the outcome of first-stage processing, i.e., whether or not integration takes place. Obviously, any intensity variation that increases the likelihood that the peripheral processes terminate within a time window will lead to an increase in the crossmodal effect. This prediction has found ample empirical support. For example, in CSP the largest RT facilitation is typically found when stimulus intensities for both modalities are matched ("physiological synchronicity"; e.g., Corneil et al., 2002). In FAP, intensity effects become a bit more complex: first, increasing the intensity of a relatively weak visual target stimulus will speed up visual peripheral processing up to some minimum level, thereby increasing the chance for the visual target to win the race. Thus, the probability that the window of integration opens decreases, predicting less crossmodal interaction. Increasing the intensity of a non-target auditory stimulus, on the other hand, leads to the opposite prediction: the auditory stimulus will have a better chance to win the race and to open the window of integration, hence predicting more crossmodal interaction, on average. If SOA is varied as well, further distinctions can be made that will not be considered here.

3.4. THE EMERGENCE OF INVERSE EFFECTIVENESS

In order to further examine the effect of intensity variation on CRE in the TWIN model, we introduce some distributional assumptions for the first stage processing times. These peripheral processing times, V for the visual and A for the visual stimulus, are assumed to have exponential probability distributions with positive-valued parameters λ_V and λ_A, respectively. That is,

$$f_V(t) = \lambda_V \exp[-\lambda_V t],$$
$$f_A(t) = \lambda_A \exp[-\lambda_A t]$$

for $t \geq 0$, and $f_V(t) = f_A(t) \equiv 0$ for $t < 0$. The exponential assumption is primarily motivated by its mathematical simplicity. Together with a Gaussian distribution assumption for second stage processing time[3] the resulting distribution is a mixture of ex-Gaussian distributions for total reaction time, which has been demonstrated to be a reasonably adequate description for many empirically observed reaction time data (cf. Van Zandt, 2002).

For the probability of integration in FAP, we get

$$\begin{aligned} \Pr(I_{\mathrm{FAP}}) &= \Pr(A < V < A + \omega) \\ &= \int_0^\infty f_A(t)[F_V(t + \omega) - F_V(t)]\, dt \\ &= \int_0^\infty \lambda_A \exp[-\lambda_A t]\{\exp[-\lambda_V t] \\ &\quad - \exp[-\lambda_V (t + \omega)]\, dt\} \\ &= \frac{\lambda_A}{\lambda_A + \lambda_V}\{1 - \exp[-\lambda_V \omega]\}. \end{aligned}$$

Similarly, for the probability of integration in CSP, we get

$$\begin{aligned} \Pr(I_{\mathrm{CSP}}) &= \Pr(A < V < A + \omega) + \Pr(V < A < V + \omega) \\ &= \frac{\lambda_A}{\lambda_A + \lambda_V}\{1 - \exp[-\lambda_V \omega] \\ &\quad + \frac{\lambda_V}{\lambda_A + \lambda_A}\{1 - \exp[-\lambda_A \omega] \end{aligned}$$

[3]That is, a convolution of an exponential and a Gaussian distribution.; for an alternative, replacing the Gaussian by the Wald distribution, see Schwarz, 2001.

Assuming matching intensity levels again (that is, $\lambda_V = \lambda_A \equiv \lambda$) this simplifies to

$$\Pr(I_{\mathrm{CSP}}) = 1 - \exp[-\lambda\omega] \equiv 2\,\Pr(I_{\mathrm{FAP}}). \tag{10}$$

It is now straightforward to compute the crossmodal response enhancement expressions,

$$\mathrm{CRE} = \begin{cases} \dfrac{(2\lambda)^{-1} + (1 - e^{-\lambda\omega}) \cdot \Delta}{\lambda^{-1} + \mu} \cdot 100, & \text{for CSP,} \\ & \text{and} \\ \dfrac{(1 - e^{-\lambda\omega}) \cdot \Delta}{2(\lambda^{-1} + \mu)} \cdot 100, & \text{for FAP.} \end{cases} \tag{11}$$

Inspection of these expressions reveals that crossmodal response enhancement, for both CSP and FAP, increases as a function of the facilitation parameter ($\Delta > 0$) and the window width (ω), but decreases as a function of second stage processing time without crossmodal interaction (μ), as one would expect.

Intriguingly, the effect of increasing intensity parameter λ is different for the two paradigms: For FAP, CRE *increases* with λ (for $\Delta > 0$) no matter the values of the remaining parameters. Note that this is no contradiction to the observations in the previous section since here we are assuming identical λ parameters for target and non-target.

For CSP, however, CRE *decreases* with λ for many plausible values of the other parameters. Thus, TWIN's prediction here concurs with the "principle of inverse effectiveness" according to which crossmodal facilitation is strongest when stimulus strengths are weak or close to threshold level (Meredith and Stein, 1986). **Figure 1** illustrates this finding for specific parameters and shows that it holds across all values of window width. Note that the difference between FAP and CSP with respect to "inverse effectiveness" is mainly due to an additional term in the numerator of the CRE equation (Equation 11) for CSP. This term, $\frac{1}{2\lambda}$, is the amount of statistical facilitation, $\min(\mathrm{E}[V], \mathrm{E}[A]) - \mathrm{E}[\min(V, A)]$. Thus, here the "principle of inverse effectiveness" is based on the fact that statistical facilitation becomes the smaller the higher the intensity levels of the stimuli are. This observation suggests that, at least in the domain of reaction time measurement, "inverse effectiveness" may in part be a purely statistical effect. Because this result has been derived under the auxiliary assumption of exponentially distributed peripheral processing durations and is limited to certain, though plausible, parameter combinations, its remains to be shown whether it can be generalized to a larger class of distributions.

4. OPTIMAL TIME WINDOWS FOR FAP AND CSP

The effect of adding information from another modality should be particularly strong in an adverse environment, i.e., with a low signal-to-noise ratio (SNR). The *prima facie* plausibility of the inverse effectiveness principle is actually based on this idea. Within the TWIN framework, this would correspond to adjusting the size of the time window with respect to SNR, i.e., widening it for lower SNR values. Note that this differs from the above discussion of the effect of stimulus intensity where time window size was assumed to be constant across trials. The perspective taken now is that the adjustment of the time window is a *top-down process* occurring only if there are long-term changes in the environment as measured by SNR or, possibly, as a consequence of changes in the cost/benefit of integration. This raises the question of how an appropriate window size should be determined.

Clearly, an infinitely large time window would lead to mandatory integration, and one could argue that this is what, e.g., a sufficiently low SNR would require. A more elaborate response, however, is based on the hypothesis that integrating crossmodal information always involves a possibly implicit decision about

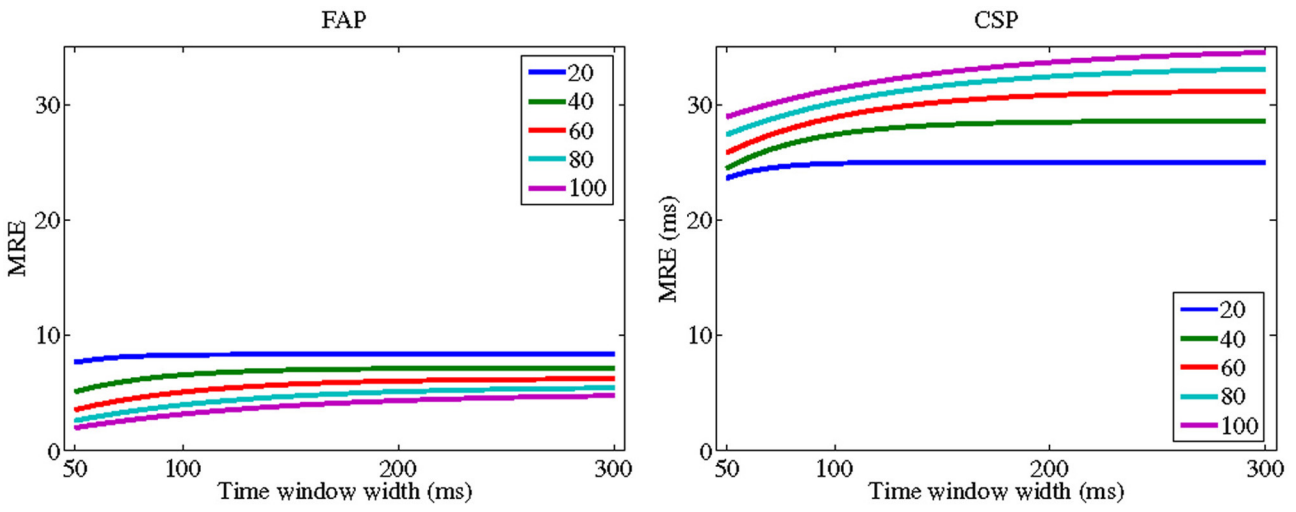


FIGURE 1 | TWIN predictions for crossmodal response enhancement (CRE) for focused attention paradigm (FAP) (left panel) and crossmodal signals paradigm (CSP) (right panel) as a function of time window width (ω). Each curve corresponds to a specific intensity parameter of the stimuli demonstrating a "inverse effectiveness" for CSP. The peripheral processing times for the auditory and visual stimuli are $1/\lambda_A = 1/\lambda_V$ equal to 20 (blue line); 40 (green); 60 (red); 80 (cyan); and 100 (magenta). Mean second stage processing time is $\mu = 100$. Interaction parameter is $\Delta = 20$.[all values in ms].

whether or not two (or more) sensory cues originate from the same event, i.e., have a common cause and that integration should only occur in that case (e.g., Stein and Meredith, 1993; Koerding et al., 2007). For example, in a predator-prey situation it may be vital for the potential prey to recognize whether a sudden movement in the dark is caused by a predator or a harmless wind gust. If visual information is accompanied by some vocalization from a similar direction, it may be adequate to respond to the potential threat by assuming that the visual and auditory information are caused by the same source, i.e., to perform multisensory integration leading to a speeded escape reaction. On the other hand, in a rich dynamic environment it may also be disadvantageous, e.g., leading to a depletion of resources, or even hazardous, to routinely combine information associated with sensory events which—in reality—may be entirely independent and unrelated.

Colonius and Diederich (2010) introduced a decision-theoretic approach for finding an optimal time window that is in line with this setup. Subsequently, we have derived an explicit expression for the optimal time window for the FAP case (Colonius and Diederich, 2011). Here, we present an optimal time window for CSP as well and discuss how predictions for MRE under optimal performance differ between the two paradigms. To keep this paper self-sustained, the next two sections summarize our previously obtained results.

4.1. BASIC DECISION SITUATION AND OPTIMAL DECISION RULE

The basic decision situation is presented in a schematic manner by the following *payoff matrix* (**Table 1**). It defines the gain (blue) or cost (red) function U associated with the *states of nature* (C) and the *action* (I) of audiovisual integration: Variable C indicates whether visual and auditory stimulus information are generated by a common source ($C = 1$), i.e., an *audiovisual event*, or by two separate sources ($C = 2$), i.e., auditory and visual stimuli are unrelated to each other. Variable I indicates whether or not integration occurs ($I = 1$ or $I = 0$, respectively). The values U_{11} and U_{20} correspond to correct decisions and will in general be assumed to be positive numbers, while U_{21} and U_{10}, corresponding to incorrect decisions, will be negative. The organism's task is to balance these costs and benefits of multisensory integration by an appropriate optimizing strategy.

In order to derive an optimal decision rule, we assume that *a-priori* probabilities for the events $\{C = i\}_{i=1,2}$ exist, with $\Pr(C = 1) = 1 - \Pr(C = 2)$. In general, an optimal strategy may involve many different aspects of the empirical situation, like spatial and temporal contiguity. As a simplifying starting point, the temporal disparity between the "arrival times" of the unimodal signals is assumed to be the *only* perceptual evidence utilized by the organism. Thus, computation of an optimal time window will be based on the prior probability of a common cause and the likelihood of temporal disparities between the unimodal signals; that is, we define the *likelihood function* $f(t|\,C)$, where f denotes the probability mass function or, if it exists, the density function of T given C takes on a value. Using Bayes' rule, we immediately have the *posterior* probability of a common cause given the occurrence of an arrival time difference t,

$$\Pr(C = 1|\,t) = \frac{f(t|\,C = 1)\Pr(C = 1)}{f(t|\,C = 1)\Pr(C = 1) + f(t|\,C = 2)\Pr(C = 2)}.$$

On each trial, in order to maximize the expected value $\mathrm{E}[U]$ of function U in the payoff matrix (**Table 1**), the decision-maker is to choose that action alternative (i.e., to integrate or not) which contributes, on the average, more to $\mathrm{E}[U]$ than the other action alternative. Introducing the *likelihood ratio* function

$$\mathrm{L}(t) = f(t|\,C = 1)/f(t|\,C = 2),$$

results in the following decision rule (cf. Colonius and Diederich, 2010):

$$\text{"}\textit{If}\quad \mathrm{L}(t) > \frac{\Pr(C = 2)}{\Pr(C = 1)} \times \frac{U_{20} - U_{21}}{U_{11} - U_{10}},$$
$$\textit{integrate, otherwise do not integrate."} \qquad (12)$$

This decision rule implicitly defines a window that is optimal in the sense of maximizing $\mathrm{E}[U]$:

> *The* optimal time window *is the set of all values of absolute arrival time differences* $\{T = t\}$ *satisfying the inequality in the above decision rule (12).*

The effect of the prior probability for a common cause on the time window is immediately predictable from this decision rule: Keeping the U-values constant, the expression on the right of inequality (12) will decrease as $P(C = 1)$ increases, implying an extension of the time window.

4.2. COMPUTING AN OPTIMAL TIME WINDOW FOR FAP

In order to compute the optimal time window, we must specify the likelihood function. For two separate sources we assume a uniform law,

$$f(t|\,C = 2) = \begin{cases} 1/(t_1 - t_0) & \text{if } t_0 < t < t_1, \\ 0 & \text{otherwise,} \end{cases} \qquad (13)$$

Here, t_0, t_1 are real numbers defining the *observation interval*, that is, the interval of time limiting all possible ATDs due to the construction of the trial length by the experimenter. Thus, under two separate sources any arrival time difference is assumed to occur with the same likelihood within the observation interval (t_0, t_1).

For a single source, we postulate[4] that the likelihood function is induced by the distribution of the peripheral processing times V

Table 1 | Payoff matrix for the basic decision situation.

Gain/Cost	Integration ($I = 1$)	No integration ($I = 0$)
Common source ($C = 1$)	U_{11}	U_{10}
Separate sources ($C = 2$)	U_{21}	U_{20}

[4]It is important to keep in mind that this is an additional assumption not directly following from the decision framework. It seems plausible, however, given that in a typical environment visual and auditory information deriving from a common source should occur more or less at the same point in time (cf. Leone and McCourt, 2012).

and A. For the FAP, given the independent exponential distribution assumption for V and A in TWIN, the distribution of arrival time differences under a common source, $V - A$, can be shown to be an *asymmetric Laplace distribution* (Colonius and Diederich, 2011):

$$f(t|\,C=1) = \frac{\lambda_V \lambda_A}{\lambda_V + \lambda_A} \times \begin{cases} \exp(-\lambda_V t) & \text{if } t \geq 0, \\ \exp(\lambda_A t) & \text{if } t < 0. \end{cases} \tag{14}$$

Note that the asymmetry derives from the asymmetry of the role of the modalities in FA tasks (target vs. non-target). For $t_0 \leq t \leq t_1$, the likelihood ratio becomes[5]

$$L(t) = f(t|\,C=1)/f(t|\,C=2) = \frac{\lambda_V \lambda_A}{\lambda_V + \lambda_A}(t_1 - t_0) \times \begin{cases} \exp(-\lambda_V t) & \text{if } t \in (t_0, t_1) \cap [0, t_1), \\ \exp(\lambda_A t) & \text{if } t \in (t_0, t_1) \cap (t_0, 0] \end{cases} \tag{15}$$

To simplify the exposition, in the following the ratio of utility differences occurring in Equation 12 will be set equal to one. Thus, according to the optimal decision rule, audiovisual integration should be performed if and only if

$$L(t) > \frac{1-p}{p},$$

with $p \equiv \Pr(C=1)$. Assuming matching intensity levels ($\lambda \equiv \lambda_A = \lambda_V$), inserting the expression for $L(t)$ from Equation 15, and solving for t yields the following *optimal* time window for $t \in (t_0, t_1)$:

$$\left\{ t \;\middle|\; \frac{1}{\lambda} \log\left[\frac{2\,(1-p)}{\lambda\,(t_1 - t_0)\,p}\right] \leq t \leq \frac{1}{\lambda} \log\left[\frac{\lambda\,(t_1 - t_0)\,p}{2\,(1-p)}\right] \right\} \tag{16}$$

provided that

$$\frac{2\,(1-p)}{\lambda\,(t_1 - t_0)\,p} \leq 1. \tag{17}$$

This latter condition guarantees that the left side of the interval is non-positive and the right side is non-negative. For the width of the optimal time window, we get immediately

$$\omega_{\text{opt}} = \left(\frac{2}{\lambda}\right) \log\left[\frac{\lambda\,(t_1 - t_0)}{2} \frac{p}{1-p}\right] \tag{18}$$

This is obviously an increasing function of the prior odds $p/(1-p)$ and of the observation interval (t_0, t_1). Increasing $P(C=1)$ leads to a widening of the time window, in this case approaching infinity in a non-linear fashion. Moreover, the optimal time window disappears for values of the prior below a certain positive threshold value. Although the exact threshold value depends on the experimental context (i.e., $t_1 - t_0$ and λ) and may get close to zero, this prediction provides a potentially strong model test: for a small enough value of $P(C=1)$ there should be no multisensory integration effect at all.

[5]Note that for t outside of the observation interval the likelihood ratio remains undefined.

4.3. COMPUTING AN OPTIMAL TIME WINDOW FOR CSP

The derivation of an optimal time window for CSP is analogous to the FAP case, except that now the likelihood is defined using the *absolute* arrival time difference of the unimodal signals, $T = |V - A|$. Given the assumption of independent exponential distribution for V and A in TWIN, the distribution of T under a common source, then turns out to be a mixture of exponential distributions:

$$\begin{aligned} \Pr(|V - A| \leq t) &= \Pr(V < A < V + t) + \Pr(A < V < A + t) \\ F_T(t|\,C=1) &= \int_0^\infty f_V(v)[F_A(v+t) - F_A(v)]\,dv \\ &\quad + \int_0^\infty f_A(a)[F_V(a+t) - F_V(a)]\,da \\ &= \frac{\lambda_V}{\lambda_A + \lambda_V}\left\{1 - \exp[-\lambda_A t]\right\} \\ &\quad + \frac{\lambda_A}{\lambda_A + \lambda_V}\left\{1 - \exp[-\lambda_V t]\right\}. \end{aligned}$$

Differentiation then yields the density for $|V - A|$:

$$f_T(t|\,C=1) = \frac{\lambda_V \lambda_A}{\lambda_A + \lambda_V}\left\{\exp[-\lambda_A t] + \exp[-\lambda_V t]\right\}, \tag{19}$$

from which the likelihood ratio follows:

$$L(t) = (t_1 - t_0) f_T(t) \tag{20}$$

which is defined for $t \in (t_0, t_1)$. It is easy to see that $L(t)$ is monotonically decreasing in t; thus, larger arrival time differences, positive or negative, provide evidence in favor of two separate sources rather than a single source, as is to be expected.

Inserting the expression for $L(t)$ from Equation 20 and solving for t yields the following *optimal* time window with $\lambda \equiv \lambda_A = \lambda_V$:

$$\left\{ t \;\middle|\; 0 \leq t < \frac{1}{\lambda} \log\left[\frac{p}{1-p}\,\lambda\,(t_1 - t_0)\right] \right\} \tag{21}$$

for $t \in (t_0, t_1)$. In order to exclude negative values of the logarithm,

$$p \geq [\lambda\,(t_1 - t_0) + 1]^{-1}$$

must hold. The upper bound of the optimal time window is identical to its length. As in FAP, it is obviously an increasing function of the prior odds $p/(1-p)$ and of the observation interval (t_0, t_1). Increasing $p = P(C=1)$ leads to a widening of the time window, approaching infinity in a non-linear fashion. Moreover, as before, the optimal time window disappears for values of the prior below a certain positive threshold value, providing a potential model test since for a small enough value of $P(C=1)$ there should be no multisensory integration effect at all.

4.4. CSP VS. FAP: COMPARING OPTIMAL TIME WINDOW WIDTH AND CRE

We are now in a position to compare both paradigms with respect to their optimal time window width and the magnitude of their multisensory response enhancement under optimality. For the optimal time window size, ω_{opt},

$$\omega_{\text{opt}} = \begin{cases} \frac{2}{\lambda} \log\left[\frac{\lambda}{2}\frac{p}{1-p}\, s\right] & \text{for FAP;} \\ \frac{1}{\lambda} \log\left[\lambda \frac{p}{1-p}\, s\right] & \text{for CSP} \end{cases} \tag{22}$$

under the provision that the logarithmic term does not become negative. Note that the length of the observation interval ($s \equiv t_1 - t_0$), being determined by the experimental setup, can be considered an inessential scaling factor. Not surprisingly, as observed before, both optimal window widths increase with increasing prior p for a common cause, approaching infinity for $p \to 1$. **Figure 2** shows optimal time window width for both FAP and CSP as a function of the prior p. The width for FAP is larger than for CSP nearly everywhere, except for rather small values (depending on the scaling factor s) of the prior, where the opposite holds. This make sense intuitively: the probability of integration in FAP is only half the size of the probability of integration in CSP (cf. Equation 10). Thus, for a fixed and not too small prior, window size in FAP must increase in order to match the probabilities of integration in both paradigms[6]. Inspection of ω_{opt} (Equation 22) reveals that the effect of intensity parameter λ is more complex. For small values of p it is non-monotonic (increasing, then decreasing) and for larger p values ω_{opt} it decreases for both FAP and CSP. The latter observation may reflect a moderating effect of intensity on window size once the window already is rather large.

[6]All other conditions being equal there is no *a-priori* reason why the optimal *probability* of integration should differ between FAP and CSP.

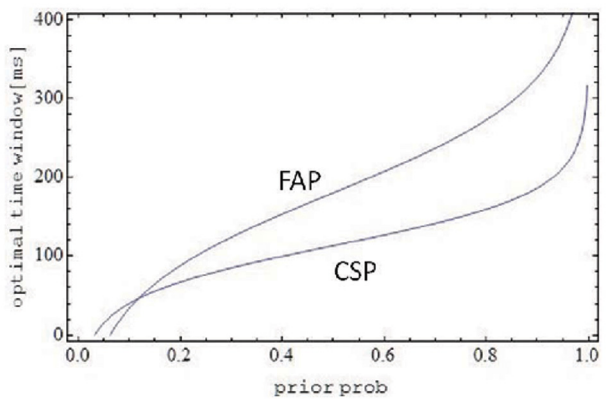


FIGURE 2 | Optimal time window as a function of prior probability *p* for a common source. Except for very small *p*, the optimal window size for FAP is larger than for CSP, compensating for the lower probability of integration in FAP compared to CSP. Parameters are $\lambda = 0.03$, $s = 1$ s.

Inserting ω_{opt} into the expressions for crossmodal response enhancement (CRE) yields

$$\text{CREopt} = \begin{cases} (1/\lambda + \mu)^{-1}\left[\frac{1}{2\lambda} + \Delta\left(1 - \frac{1-p}{\lambda p s}\right)\right] \times 100 \\ \qquad \text{for CSP,} \\ \qquad \text{and} \\ (1/\lambda + \mu)^{-1}\,\Delta/2\left[1 - \left(\frac{2(1-p)}{\lambda p s}\right)^2\right] \times 100, \\ \qquad \text{for FAP.} \end{cases} \tag{23}$$

We know from Ineq. 9 that CRE(FAP) cannot be larger than CRE(CSP) when the parameters ω, λ, Δ ($\Delta > 0$), and μ are all identical for the two paradigms. However, since now the optimal window widths are not identical for CSP and FAP, this ordering might no longer hold. Closer scrutiny of the above equations reveals, however, that crossmodal response enhancement in CSP still dominates the one in FAP when the other parameters are kept the same. Moreover, for λ increasing without bound, CRE(CSP) will become twice as large as CRE(FAP).

5. SUMMARY AND CONCLUSION

Assuming exponential arrival time distributions, the framework of the TWIN model has been specified here so that specific quantitative predictions could be made comparing the FAP and the CSP with respect to (1) the probability of multisensory integration and (2) expected crossmodal response enhancement (reaction time facilitation/inhibition). Moreover, introducing a decision-theoretic framework for TWIN, the investigation could be extended to comparing the CSP and FAP paradigms with respect to their predicted optimal time windows. Glossing over some of the required conditions concerning the specific parameter values, the main findings were:

- the probability of crossmodal integration for CSP is twice the probability of integration for FAP;
- crossmodal response enhancement (facilitation) for CSP is at least as large as for FAP;
- TWIN model is consistent with the occurrence of a "inverse effectiveness" under the CSP but not under FAP;
- within the decision-making framework for TWIN, explicit expressions for the computation of time windows of *optimal* width for both CSP and FAP have been derived;
- the optimal time window is larger for FAP than for CSP across (nearly) all values of the prior probability (of a common source for both modalities), thereby compensating for the smaller probability of integration in FAP (see first item on this list)
- optimal crossmodal response enhancement (facilitation) for CSP is larger than for FAP (or at least as large) even though their optimal window widths differ.

The obvious next step will be to test these predictions experimentally. Apart from a pilot study in our lab (cf. Colonius and Diederich, 2012), we are not aware of any systematic empirical studies comparing FAP and CSP under matching stimulus intensity levels. In particular, studies are needed varying the prior

probability of a common source in order to test the above predictions concerning optimality (for FAP, see Van Wanrooij et al., 2010). An unsolved issue, for example, is whether data that are not consistent with optimality indicate sub-optimal behavior or are simply due to participants' subjective priors deviating from the objective priors. Moreover, except for the first two items in the above list, the current predictions have been derived under the hypothesis of independent exponential arrival time distributions. It remains to probe by further analysis whether or not these predictions can be generalized to other plausible distributions, e.g., gamma distributions.

A fundamental difference between the tasks in FAP and CSP is that in the focused attention paradigm there must be a mechanism to distinguish a target- from a non-target-modality stimulus at a very early stage of processing, whereas in the CSP such a mechanism is not required. This difference between paradigms is in line with a recent suggestion in Kayser et al. (2010) of two different modes of multisensory integration, one occurring in a detection task where the response to weak stimuli is enhanced, and another occurring in discrimination and identification tasks where the precision and reliability of the responses are improved (see also the commentary by Ghazanfar and Lemus, 2010). This, in turn, suggests to probe whether, in focussed attention data, one effect of the non-target-modality stimulus is to diminish the variability of crossmodal reaction times, relative to the unimodal variability. In the TWIN model, no explicit mechanism to distinguish target- from non-target modalities has been implemented yet, but this may be called for if one attempts to investigate such hypotheses.

Given that the TWIN model predicts changes in the (optimal) time window as a function of the prior probability of a common source, the basic question about the malleability of the time window arises. There are a number of recent studies, using other experimental paradigms, that provide evidence for a dynamic adaptation of the time window to changes in context. For example, using a simultaneity judgment task, Powers and colleagues showed that significant and lasting changes of perceived simultaneity (40% narrowing in the width of the window) can be induced after a single day of training (Powers et al., 2009) and are accompanied by decreases in BOLD activity within a network of multisensory and unisensory areas (Powers et al., 2012)[7]. Nevertheless, direct evidence in the context of the reaction time paradigm will only be provided by the type of experimental tests suggested above.

[7]For a recent review of the general area of perception of synchrony, see Keetels and Vroomen (2012).

REFERENCES

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.

Ashby, F. G., and Townsend, J. T. (1986). Varieties of perceptual independence. *Psychol. Rev.* 93, 154–179.

Bell, A. H., Meredith, A., Van Opstal, A. J., and Munoz, D. P. (2005). Crossmodal integration in the primate superior colliculus underlying the preparation and initiation of saccadic eye movements. *J. Neurophysiol.* 93, 3659–3673.

Bell, A. H., Meredith, A., Van Opstal, A. J., and Munoz, D. P. (2006). Stimulus intensity modifies saccadic reaction time and visual response latency in the superior colliculus. *Exp. Brain Res.* 174, 53–59.

Colonius, H. (1990). Possibly dependent probability summation of reaction time. *J. Math. Psychol.* 34, 253–275.

Colonius, H., and Diederich, A. (2004). Multisensory interaction in saccadic reaction time: a time-window-of-integration model. *J. Cogn. Neurosci.* 16, 1000–1009.

Colonius, H., and Diederich, A. (2006). Race model inequality: interpreting a geometric measure of the amount of violation. *Psychol. Rev.* 113, 148–154.

Colonius, H., and Diederich, A. (2010). The optimal time window of visual-auditory integration: a reaction time analysis. *Front. Integr. Neurosci.* 4:11. doi: 10.338/fnint.2010.00011

Colonius, H., and Diederich, A. (2011). Computing an optimal time window of audiovisual integration in focused attention tasks: illustrated by studies on effect of age and prior knowledge. *Exp. Brain Res.* 212, 327–337.

Colonius, H., and Diederich, A. (2012). "Models of the time window of integration," in *The New Handbook of Multisensory Processes,* ed B. E. Stein (Cambridge, MA: MIT Press), 545–555.

Corneil, B. D., Van Wanrooij, M., Munoz, D. P., and Van Opstal, A. J. (2002). Auditory-visual interactions subserving goal-directed saccades in a complex scene. *J. Neurophysiol.* 88, 438–454.

Diederich, A. (1995). Intersensory facilitation of reaction time: evaluation of counter and diffusion coactivation models. *J. Math. Psychol.* 39, 197–215.

Diederich, A., and Colonius, H. (1991). A further test of the superposition model for the redundant-signals effect in bimodal detection. *Percept. Psychophys.* 50, 83–86.

Diederich, A., and Colonius, H. (2004). "Modeling the time course of multisensory interaction in manual and saccadic responses," in *Handbook of Multisensory Processes*, eds G. Calvert, C. Spence, and B. E. Stein (Cambridge, MA: MIT Press), 395–408.

Diederich, A., and Colonius, H. (2007a). Why two "distractors" are better then one: modeling the effect on non-target auditory and tactile stimuli on visual saccadic reaction time. *Exp. Brain Res.* 179, 43–54.

Diederich, A., and Colonius, H. (2007b). Modeling spatial effects in visual-tactile saccadic reaction time. *Percept. Psychophys.* 69, 56–67.

Diederich, A., and Colonius, H. (2008a). Crossmodal interaction in saccadic reaction time: separating multisensory from warning effects in the time window of integration model. *Exp. Brain Res.* 186, 1–22.

Diederich, A., and Colonius, H. (2008b). When a high-intensity "distractor" is better then a low-intensity one: modeling the effect of an auditory or tactile non-target stimulus on visual saccadic reaction time. *Brain Res.* 1242, 219–230.

Diederich, A., and Colonius, H. (2012). "Modeling multisensory processes in saccadic responses: time-window-of-integration model," in *The Neural Bases of Multisensory Processes,* eds M. T. Wallace and M. M. Murray (Boca Raton, FL: CRC Press), 253–276.

Ghazanfar, A. A., and Lemus, L. (2010). Multisensory integration: vision boosts information through suppression in auditory cortex. *Curr. Biol.* 20, R22–R23.

Kayser, C., Logothetis, N. K., and Panzeri, S. (2010). Visual enhancement of the information representation in auditory cortex. *Curr. Biol.* 20, 19–24.

Keetels, M., and Vroomen, J. (2012). "Perception of synchrony between the senses," in *The Neural Bases of Multisensory Processes*, eds M. T. Wallace and M. M. Murray (Boca Raton, FL: CRC Press), 147–177.

Koerding, 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

Leone, L., and McCourt, M. E. (2012). "The question of simultaneity in multisensory integration," in *Human Vision and Electronic Imaging XVII, Proceedings of SPIE-IS & T Electronic Imaging SPIE, Vol. 8291*, eds B. E. Rogowitz, T. N. Pappas, and H. de Ridder, 82910J-1–82910J-8.

Lewald, J., Ehrenstein, W. H., and Guski, R. (2001). Spatio-temporal constraints for auditory–visual integration. *Behav. Brain Res.* 121, 69–79.

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.

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

Meredith, M. A. (2002). On the neural basis for multisensory convergence: a brief overview. *Cogn. Brain Res.* 14, 31–40.

Meredith, M. A., Nemitz, J. W., and Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal facotrs. *J. Neurosci.* 10, 3215–3229.

Meredith, M. A., and Stein, B. E. (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *J. Neurophysiol.* 56, 640–662.

Miller, J. O. (1982). Divided attention: evidence for coactivation with redundant signals. *Cogn. Psychol.* 14, 247–279.

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.

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., and Spence, C. (2005). Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Cogn. Brain Res.* 25, 499–507.

Powers, A. R. 3rd, Hevey, M. A., and Wallace, M. T. (2012). Neural correlates of multisensory perceptual learning. *J. Neurosci.* 32, 6263–6274.

Powers, A. R. 3rd, Hillock, A. R., and Wallace, M. T. (2009). Perceptual training narrows the temporal window of multisensory binding. *J. Neurosci.* 29, 12265–12274.

Raab, D. H. (1962). Statistical facilitation of simple reaction times. *Trans. N.Y. Acad. Sci.* 24, 574–590.

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.* 7, 11465–11472.

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., Quessy, S., Stanford, T. R., and Stein, B. E. (2007). Multisensory integration shortens physiological response latencies. *J. Neurosci.* 27, 5879–5884.

Royal, D. W., Carriere, B. N., and Wallace, M. T. (2009). Spatiotemporal architecture of cortical receptive fields and its impact on multisensory interactions. *Exp. Brain Res.* 198, 127–136.

Schwarz, W. (1989). A new model to explain the redundant-signal effect. *Percept. Psychophys.* 46, 498–500.

Schwarz, W. (1994). Diffusion, superposition, and the redundant targets effect. *J. Math. Psychol.* 38, 504–520.

Schwarz, W. (2001). The ex-Wald distribution as a descriptive model of response times. *Behav. Res. Methods Instrum. Comput.* 33, 457–469.

Spence, C., and Squire, S. (2003). Multisensory integration: maintaining the perception of synchrony. *Curr. Biol.* 13, R519–R521.

Stein, B. E. (ed.). (2012). *The New Handbook of Multisensory Processing*. Cambridge, MA: The MIT Press.

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

Todd, J. W. (1912). "Reaction to multiple stimuli," in *Archives of Psychology, No. 25. Columbia Contributions to Philosophy and Psychology, Vol. XXI No. 8*, ed R. S. Woodworth (New York, NY: The Science Press), 1–65.

Townsend, J. T., and Eidels, A. (2011). Workload capacity spaces: a unified methodology for response time measures of efficiency as workload is varied. *Psychon. Bull. Rev.* 18, 659–681.

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.

Van Wanrooij, M. M., Bremen, P., and Van Opstal, A. J. (2010). Acquired prior knowledge modulates audiovisual integration. *Eur. J. Neurosci.* 31, 1763–1771.

Van Wassenhove, V., Grant, K. W., and Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia* 45, 598–607.

Van Zandt, T. (2002). "Analysis of response time distributions," in *Stevens' Handbook of Experimental Psychology*, 3rd Edn. Vol. 4, eds J. T. Wixted (Vol. Ed.) and H. Pashler (Series Ed.) (New York, NY: Wiley Press), 461–516.

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.

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: 08 June 2012; accepted: 05 August 2012; published online: 29 August 2012.

Citation: Colonius H and Diederich A (2012) Focused attention vs. crossmodal signals paradigm: deriving predictions from the time-window-of-integration model. Front. Integr. Neurosci. ***6****:62. doi: 10.3389/fnint.2012.00062*

frontiers in
INTEGRATIVE NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 05 November 2012
doi: 10.3389/fnint.2012.00100

Adaptation to visual or auditory time intervals modulates the perception of visual apparent motion

***Huihui Zhang*[1], *Lihan Chen*[1,2]* *and Xiaolin Zhou*[1,2]**

[1] *Department of Psychology, Center for Brain and Cognitive Sciences, Peking University, Beijing, China*
[2] *Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, China*

Edited by:
Zhuanghua Shi, Ludwig-Maximilians-Universität München, Germany

Reviewed by:
Matthew S. Matell, Villanova University, USA
Karin M. Bausenhart, University of Tuebingen, Germany

***Correspondence:**
Lihan Chen, Department of Psychology, Center for Brain and Cognitive Sciences, Peking University, Beijing 100871, China.
e-mail: clh20000@gmail.com

It is debated whether sub-second timing is subserved by a centralized mechanism or by the intrinsic properties of task-related neural activity in specific modalities (Ivry and Schlerf, 2008). By using a temporal adaptation task, we investigated whether adapting to different time intervals conveyed through stimuli in different modalities (i.e., frames of a visual Ternus display, visual blinking discs, or auditory beeps) would affect the subsequent implicit perception of visual timing, i.e., inter-stimulus interval (ISI) between two frames in a Ternus display. The Ternus display can induce two percepts of apparent motion (AM), depending on the ISI between the two frames: "element motion" for short ISIs, in which the endmost disc is seen as moving back and forth while the middle disc at the overlapping or central position remains stationary; "group motion" for longer ISIs, in which both discs appear to move in a manner of lateral displacement as a whole. In Experiment 1, participants adapted to either the typical "element motion" (ISI = 50 ms) or the typical "group motion" (ISI = 200 ms). In Experiments 2 and 3, participants adapted to a time interval of 50 or 200 ms through observing a series of two paired blinking discs at the center of the screen (Experiment 2) or hearing a sequence of two paired beeps (with pitch 1000 Hz). In Experiment 4, participants adapted to sequences of paired beeps with either low pitches (500 Hz) or high pitches (5000 Hz). After adaptation in each trial, participants were presented with a Ternus probe in which the ISI between the two frames was equal to the transitional threshold of the two types of motions, as determined by a pretest. Results showed that adapting to the short time interval in all the situations led to more reports of "group motion" in the subsequent Ternus probes; adapting to the long time interval, however, caused no aftereffect for visual adaptation but significantly more reports of group motion for auditory adaptation. These findings, suggesting amodal representation for sub-second timing across modalities, are interpreted in the framework of temporal pacemaker model.

Keywords: interval timing, adaptation, visual apparent motion, cross-modal interaction, Ternus display

INTRODUCTION

Timing is fundamental for the brain to process dynamically changing stimuli and interact with the environment. The neural system processes temporal information across a wide range of scales, from microseconds to circadian rhythms, with each scale corresponding to a specific underlying processing mechanism (Fraisse, 1963; Pöppel, 1988; Czeisler et al., 1999; Grothe, 2003). It has been revealed that sub-second timing is closely related to perceptual processing (Rammsayer, 1999; Wearden et al., 2007) and free of cognitive processing, although there is evidence showing that emotional arousal states, triggered by emotional stimuli such as emotional pictures, affect sub-second time perception in another modality (Shi et al., 2012). However, temporal processing above one second may involve more sophisticated cognitive processes (Rammsayer, 1999; Lewis and Miall, 2003; Mauk and Buonomano, 2004; Buhusi and Meck, 2005). The question remains as to whether sub-second interval timing in different modalities is subserved by a centralized mechanism ("central timer" or "central clock"; Grondin and Rousseau, 1991; Penton-Voak et al., 1996) or by the intrinsic properties of task-related neural activity in a particular modality (Ivry and Schlerf, 2008).

The traditional view toward sub-second temporal processing assumes that it is achieved by a centralized mechanism, independent of the specific sensory modality that conveys the temporal information. An implement of this idea is the "temporal pacemaker" model (Treisman, 1963; Treisman et al., 1990, 1994; Ivry et al., 2002), which consists of two major components. The first is a temporal oscillator that emits regular pulses at some fundamental frequency. These pulses are gated to a second component, a calibration or "gain control" or switch unit that can increase or decrease the frequency. The modulated pulses are counted and stored in working memory. In addition, temporal frequency of the repetitive, rhythmic stimuli could modulate the speed of pacemaker. Repetitive stimuli (clicks or flashes) of high temporal frequency may increase the speed of pacemaker, such that more pulses are accumulated in a given time; repetitive stimuli of low temporal frequency may decrease the speed of pacemaker,

with less accumulated pulses for a given time (Ono and Kitazawa, 2011).

This model is supported by an increasingly large body of evidence. Firstly, psychophysics studies on visual and auditory sub-second time perception all showed that the ability to discriminate two time intervals is determined by the ratio of just-discriminable time difference to the base interval, suggesting that there might be a common temporal mechanism to compute the time information (i.e., the number of pulses; Creelman, 1962; Allan and Kristofferson, 1974; Divenyi and Danner, 1977; Killeen and Weiss, 1987; Keele and Ivry, 1991; Ivry, 1993). Secondly, tasks differed in sensorimotor processing (time perception vs. time reproduction) and in modality of stimuli used to define the intervals (visual vs. auditory) all showed a linear increase in performance variability as a function of the interval duration, and individuals' performances in tasks related to perception and reproduction of time intervals were highly correlated. These findings can be adduced to support the existence of a centralized internal clock which functions in all the tasks (Keele et al., 1985; Ivry and Hazeltine, 1995; Merchant et al., 2008). Thirdly, cross-modal adaptation experiments showed that adaptation to intervals defined by audiovisual events affect the perceived direction of visual apparent motion (AM) (Freeman and Driver, 2008); learning studies also demonstrated that training in a timing context can be generalized to other timing behaviors. For instance, learning to discriminate time intervals in the tactile domain can affect the performance in a similar task in the auditory domain (Nagarajan et al., 1998) and vice versa (Meegan et al., 2000). Such crossmodal transfer in timing suggests that there might be amodal representation and centralized time mechanism across different sensory modalities.

However, recent studies challenged this view. Using a direct visual temporal interval discrimination task, Lapid and Ulrich (2009) found no transfer of the learned time interval from the auditory to the visual domain. Using an adaptation paradigm, Becker and Rasmussen (2007) found a robust auditory temporal rhythm aftereffect, but only when the adaptation and test stimuli came from the same modality. In this study, an auditory test rhythm (400 ms interval) was preceded by an either faster or slower auditory rhythm and participants were asked to replicate the rhythm by pressing a button. They found a significant negative aftereffect, i.e., after adaptation to a faster rhythm, the reproduced rhythm was slower than the test rhythm; after adaptation to a slower rhythm, the reproduced rhythm was faster than the test rhythm. This aftereffect vanished when the test rhythm was presented in the visual modality. The authors suggested distinct mechanisms for sub-second temporal processing in different modalities.

A problem with Becker and Rasmussen (2007) is that reproduction of auditory rhythms is generally more accurate than that of visual rhythms (Welch and Warren, 1980; Glenberg et al., 1989; Glenberg and Jona, 1991; Recanzone, 2003). It is possible that the null crossmodal adaptation aftereffect with the visual stimuli in Becker and Rasmussen (2007) was due to the inaccuracy in perceiving time intervals conveyed by visual flashes. Alternatively, reproduction of visual rhythm is less reliable due to this motor activity being tightly coupled with inaccurate visual temporal processing (Repp, 2003; Patel et al., 2005). Thus, it might be the unreliable perception of visual rhythm and/or inaccurate motor reproduction of the visual rhythm, rather than the lack of cross-modal adaptation, that caused the null effect in the reproduction task.

To avoid the potential pitfalls associated with the reproduction task which *explicitly* measures the time interval processing, here we used the visual Ternus display to *implicitly* measure the processing of time intervals in the sub-second range (**Figure 1**). A typical Ternus display is composed of two frames with a variable inter-stimulus interval (ISI) (Ternus, 1926; Petersik and Rice, 2006; Shi et al., 2010). The first frame of the display contains two discs, and the second frame contains the same two discs with the second disc of the first frame and the first disc of the second frame sharing the same location. Depending on the locations of the first and the second frames, the AM could be either rightward or leftward. The Ternus display is an ambiguous display of which two different kinds of AM can be perceived depending on the ISI. At a short ISI, observers see the "overlapping" disc of two frames remaining stationary (or just blink) and the outer disc moving back and forth; this is called "element motion." At a long ISI, observers see the discs of one frame moving as a whole; this is called "group motion." The classification of two percepts of Ternus AM is a function of the ISI between the two frames, and we can use the report of element vs. group motion to measure the change of the implicitly perceived time interval triggered by temporal adaptation.

Specifically, we carried out three experiments (plus a control experiment) in which the ISI of the probe Ternus display was set at a time interval (about 125 ms) in which the report of element

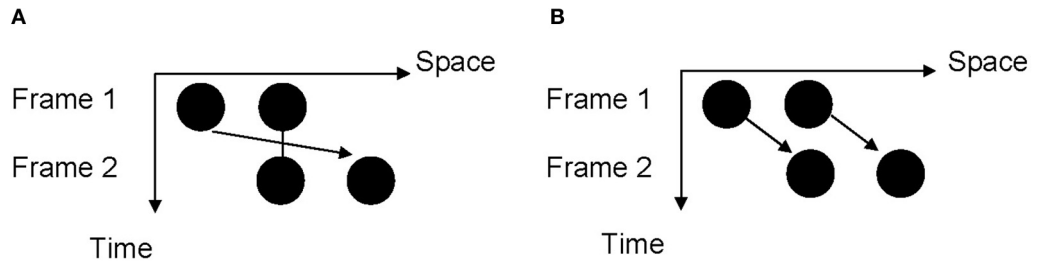


FIGURE 1 | Illustration of the Ternus display. (A) "Element motion" percept: the disc which occupies the same position in two frames is perceived to remain static or to blink at the same location while the "outer" discs are perceived to move from one location to the other. **(B)** "Group motion" percept: the two discs are perceived to move together in the manner of a coherent lateral displacement.

vs. group motion was equally probable (i.e., a bistable situation). The empirical question was whether (and how) this balance between the two types of percepts would be altered by the preceding adaptation procedures and whether different schemes of adaptation would differentially affect the perception of motion in the Ternus display. The adaptation scheme in Experiment 1 used visual Ternus displays in which the ISI between the two frames was set at either 50 ms (for element motion) or 200 ms (for group motion). If adaptation to the two time intervals is equally effective in affecting temporal processing, we would expect to observe more reports of group motion for the probe displays after adapting to the short ISI and more reports of element motion after adapting to the long ISI. Experiments 2 and 3 used, respectively, paired blinking discs and paired auditory beeps to demarcate the time intervals that the participants were supposed to adapt to. There were also two types of intervals for adaptation, 50 or 200 ms. Although the adaptation schemes in Experiments 1 and 2 were both presented in the visual modality, they differed in the extent to which the adaptation scheme was similar to the probe in perception and task structures. The adaptation schemes in Experiments 2 and 3 were similar in task structure, but differed in the presentation modality. If interval timing in the sub-second range relies on an amodal neurocognitive mechanism, we should observe an adaptation aftereffect not only for adaptation schemes sharing the modality with the probe (i.e., Experiments 1 and 2), but also for cross-modal adaptation schemes (i.e., Experiment 3); moreover, the pattern of the aftereffect should be similar across experiments, although the task structure may to a certain degree modulate this pattern. If, however, time intervals are encoded as intrinsic properties of stimulus processing, the earlier temporal processing at the adaptation stage should have different impacts upon (implicit) temporal processing at the probe stage, depending on the task structure and/or modality of adaptation schemes.

METHODS

PARTICIPANTS

Seventeen students (mean age 21.9 years old, 8 females), nineteen (mean age 20.7 years old, 12 females), and twenty-five (mean age 20.7 years old, 14 females) from Peking University participated in Experiments 1, 2, and 3, respectively. Twenty seven students (mean age 20.6 years old, 13 females) participated a control experiment (Experiment 4). They all had normal or corrected-to-normal vision and normal hearing and were naïve to the purpose of the research. We used different pools of participants for the four experiments because we were concerned with the possible contaminations across different tasks: for example, participants could adopt response strategies if each participant takes part in all the experiments (as we observed in pilot studies but not reported here). Informed consent was obtained from each participant as required by the Ethics Committee, Department of Psychology at Peking University.

STIMULI AND APPARATUS

Each probe Ternus display was composed of two frames, with each frame of two black discs (12.71 cd/ m^2 in luminance) presented horizontally on a gray background (16.11 cd/ m^2 in luminance). The two frames shared one disc location at the center of the screen and contained the other two discs on the horizontally opposite side of the center (**Figure 1**). The diameter of each black disc was 1.6° in visual angle, and the distance between the centers of the two adjacent discs was 3.1°. The duration of each frame was 30 ms. The ISI that yielded equally probable reports of element motion and group motion was determined individually for each participant in a pretest (see "Procedure").

For the three adaptation schemes, the Ternus display in Experiment 1 was structured in the same way as the probe display, except that the ISI between the two frames was set at either 50 or 200 ms. The time interval between the two paired blinking discs in Experiment 2 was also set at either 50 or 200 ms. The time interval between pairs of discs were 400 ms. Each disc had the same physical properties as the disc in the Ternus display. All the discs were presented consecutively at the center of the screen. The auditory beeps in Experiment 3 were presented binaurally, with the duration of each beep (65dB, 1000 Hz, sampled at 44.1 kHz for Experiment 3; 65 dB, 500 or 5000 Hz for control experiment) lasting 30 ms. Again, the time interval between the two paired beeps were either 50 or 200 ms and the interval between pairs of beeps was 400 ms.

The testing room was dimly lit with an average ambient luminance of 0.12 cd/m^2. Visual stimuli were presented on a 22-inch CRT monitor (1,024 × 768 pixels; 100 Hz) positioned at eye level. Viewing distance was set to 57 cm, maintained by using a chin-rest. A headset (Philips, SHM 1900) was used to emit sound stimuli. Stimulus presentation and data collection were implemented by computer programs which was developed with Matlab 7.1 (MathWorks Inc., Natick, MA) and Psychophysics Toolbox (Brainard, 1997; Pelli, 1997).

PROCEDURE

Prior to the formal experiment, participants underwent practice to be familiar with a Ternus display of either typical "element motion" (ISI = 50 ms) or typical "group motion" (ISI = 200 ms) percept. They were asked to discriminate the above two percepts by pressing the left and right mouse button to indicate responses for "element motion" and "group motion," respectively. The mapping between button and response type was counterbalanced across participants. When participants made an incorrect response, an immediate feedback appeared on the screen showing the percept (element motion or group motion) that should be reported. The practice session continued until the participant's accuracy of report was close to 100%. Almost all the participants could meet this standard within 120 trials. They then underwent the pretest which aimed to find out each participant's point of subjective equality (PSE), i.e., the time interval on which the probabilities of reporting "element motion" and "group motion" were equal (50% each).

Pretest

A typical visual Ternus display procedure was used. The ISI between the two visual frames of Ternus display was selected from one of the following six durations: 50, 80, 110, 140, 170, or 200 ms. Directions of AM (leftward or rightward) were balanced across trials. Each configuration (with ISI level and motion

direction) was presented 40 times. All the 240 trials (6 levels × 40 trials) were randomized in presentation order. These trials were divided into 4 blocks. Participants could take a short break between blocks.

A trial started with a fixation cross presented on the center of the screen for 300 ms. Next, a blank display (with a gray background) was shown for a random duration of 300–500 ms to reduce time-based expectations toward the next stimulus. Then the Ternus display with a variable ISI (50, 80, 110, 140, 170, or 200 ms) was presented. After a blank of 300 ms, participants were presented with a question mark until they made a two-alternative forced choice response indicating whether they had perceived "element motion" or "group motion." The inter-trial interval was 500 ms.

For each ISI condition, the percentage of "group motion" reports was collapsed over two motion directions. The six data points (one for each ISI) were fitted into the psychometric curve using a logistic function (Treutwein and Strasburger, 1999). The transitional ISI (PSE) at which the participant was equally likely to report the two percepts could be calculated by estimating the 50% of reporting "group motion" on the fitted curve. For each participant, we calculated his/her PSE immediately after the pretest session. The Ternus display with ISI equal to PSE would be used as a probe in the following adaptation session.

Comparisons were conducted for the PSEs derived for the three groups of participants. There were no significant differences between the "Visual-AM" (114.4 ± 13.8 ms), the "Visual-Blink" (117.1 ± 14.0 ms), and the "Beeps" (120.4 ± 13.0 ms) groups, $F_{(2,\ 58)} = 1.00$, $p > 0.1$. Comparisons were also made for the JNDs (just noticeable differences), which measured the task difficulty/participants' sensitivity of discriminating the two percepts in visual Ternus display. There were no differences between the three groups of participants (21.7 ± 5.2, 20.8 ± 6.0, and 22.2 ± 5.4 ms, respectively), $F_{(2,\ 58)} = 0.33$, $p > 0.1$. These results suggested that the three groups of randomly selected participants were well matched in their basic abilities in perceiving AM and in the implicit processing of time intervals between visual frames.

Adaptation

Each trial consisted of two phases: exposure and immediate probe test. In Experiment 1 ("Visual-AM"), the adaptation stimuli were Ternus displays of either typical "element motion" (short interval, ISI = 50 ms) or typical "group motion" (long interval, ISI = 200 ms). The probe test was a Ternus display with ISI equal to the PSE obtained in the pretest session, which rendered ambiguous percepts between "element motion" and "group motion." The trials for two types of adapting stimuli ("element motion" and "group motion") were arranged in blocks, the presentation order was pseudo-randomized. Each participant received 8 blocks (4 blocks for each adaptation type) with each block containing 20 target trials and 10 filler trials. We introduced filler trials with Ternus displays of typical "element motion" (ISI = 50 ms) or "group motion" (ISI = 200 ms) among probe trials to minimize potential response bias. The direction of Ternus AM (leftward or rightward) was same between exposure phase and probe test. For each trial, after a fixation of 300 ms, the exposure phase started. The exposure phase was composed of 7–9 repetitions of Ternus display. The time interval between consecutive presentations of the Ternus display was 400 ms, which was good enough to separate the adjacent adapting Ternus display clearly with a pilot test. After the presentation of adapting stimuli, followed by a 900 ms blank interval, the probe Ternus display was given. After a 1200 ms blank interval, a question mark appeared on the screen and remained until a two-alternative forced choice of either "element motion" or "group motion" was made. For each trial, participants were instructed to respond to the last presentation of Ternus display. The inter-trial interval was 500 ms. Participants could take a short break between blocks.

In Experiment 2 ("Visual-Blink"), the adapting time intervals (50 or 200 ms) were given by a sequence of two consecutively presented blink discs (the same central disc of Ternus display used in Experiment 1). Participants were asked to respond to the probe Ternus after viewing the blinking discs. The other arrangements of parameters and response method were the same as in Experiment 1.

In Experiment 3 ("Beeps"), the adapting time intervals (50 or 200 ms) were given by a sequence of paired beeps. During the exposure phase, participants were instructed to keep looking at the cross presented on the center of the screen while listening to the auditory beeps. This arrangement was used to make participants maintain their fixation on the location where the probe Ternus would be presented as in Experiment 1 and 2. Participants were asked to judge probe Ternus display after hearing the beeps. The other arrangements of temporal parameters and response method were the same as in Experiment 1.

In Experiment 4 ("Beeps"), the adapting time intervals (50 or 200 ms) were given by a sequence of paired beeps, different to Experiment 3 (auditory stimuli with fixed pitch: 1000 Hz), two pitches (one sequence of beeps were of the same pitches) were used: 500 or 5000 Hz. The procedures for adaptation and probe test were similar to those in Experiment 3, except that when the whole adaptation and the test probe trials were finished, participants took an additional subjective rating task, in which they were asked to rate on a 7-point Likert scale about the perceived degree of arousal for the following four types of auditory stimuli sequence: short interval-low pitch, short interval-high pitch, long interval-low pitch, long interval-high pitch. Each type was repeated three times and the presentation orders of the above types of sound sequences were randomized.

RESULTS

For filler trials in all the four experiments, the average proportion of reporting "group motion" was less than 10% of the filler trials with a ISI of 50 ms ("element motion" displays) and more than 90% of the filler trials with a ISI of 200 ms ("group motion" displays), indicating that participants generally had clear percepts of element motion and group motion. Each individual's performance accuracy on the filler trials were within the distribution of mean value for all the participants plus/minus three standard deviations.

We made statistical comparisons between Experiments 1–3 (**Figure 2**). For the critical trials, analysis of variance (ANOVA) was conducted, using the difference between the proportion of

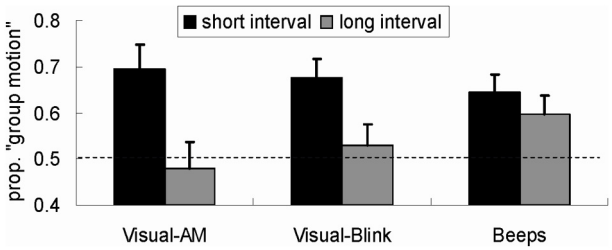


FIGURE 2 | Reports of group motion for the probe Ternus display (with ISI = the time corresponding to PSE) for the three experiments. The values in Y-axis represent the proportion of "group motion." The black bars represent "group motion" reports after short time intervals (50 ms) adaptation and the gray bars represent "group motion" reports after long time intervals (200 ms) adaptation. "Visual-AM," "Visual-Blink," and "Beeps" refer to, respectively, the three experiments in which the visual Ternus apparent motion, visual blinking discs and auditory beeps were used in different adaptation schemes. The error bar represents one standard error.

"group motion" report and the proportion (0.50) corresponding to the PSE as the dependent measure and with experiment as a between-participant factor. The main effect of time interval was significant, $F_{(1, 58)} = 29.99$, $p < 0.001$, with more reports of group motion after the adaptation to the short time interval (17.1%) than after adaptation to the long time interval (3.5%). Further tests showed that while overall the adaptation effect was significant for the short interval, $F_{(1, 58)} = 44.99$, $p < 0.001$, it was not for the long time interval, $F_{(1, 58)} = 1.63$, $p > 0.1$.

Importantly, the interaction between adaptation scheme time and experiment was significant, $F_{(2, 58)} = 4.074$, $p < 0.05$, indicating that the adaptation schemes had different impacts upon the report of group motion in different experiments. Further analysis was conducted to examine the interaction in detail. We first tested the adaptation aftereffect for the short or the long time interval, respectively, treating experiment as a between-subjects factor. This test found no significant differences between experiments for either the short interval adaptation, $F_{(2, 58)} = 0.35$, $p > 0.1$, or the long interval adaptation, $F_{(2, 58)} = 1.65$, $p > 0.1$. However, separate comparisons with 0.5 showed that all short interval adaptations in different experiments led to more reports of group motion, $ps < 0.01$; for the long interval adaptation, the aftereffect was not observed in Experiments 1 and 2 with visual modality, $ps > 0.1$, but in Experiment 3 with auditory adaptation, $p < 0.05$.

On the other hand, comparison between the adaptation effects after the short and long time interval adaptation in each experiment revealed a significant difference in Experiment 1, $F_{(1, 58)} = 21.62$, $p < 0.001$, in Experiment 2, $F_{(1, 58)} = 10.86$, $p < 0.01$, but not in Experiment 3, $F_{(1, 58)} = 1.44$, $p > 0.1$.

In Experiment 4, the averaged appraisal scores are: high pitch-short interval (5.8), high pitch-long interval (4.9), low pitch-short interval (5.2), low pitch-long interval (4.0). The main effect of pitch was significant, $F_{(1, 26)} = 12.676$, $p < 0.01$; The main effect of interval was also significant, $F_{(1, 26)} = 59.297$, $p < 0.001$. The auditory signals with higher pitch triggered more arousal (5.4) than the low pitch did (4.4). The sequences with shorter inter-intervals induced more arousal (5.4) than those with longer inter-intervals (4.4). However, the interaction between pitch and interval was not significant, $F_{(1, 26)} = 1.040$, $p > 0.1$. For reports of the proportion of "group motion," both the main effects of pitch [$F_{(1, 26)} = 0.076$, $p > 0.1$] and interval [$F_{(1, 26)} = 1.649$, $p > 0.1$] were not significant, and the interaction was also not significant, $F_{(1, 26)} = 1.342$, $p > 0.1$. *Post-hoc* T tests revealed that the percentages of "group motion" percentages, were significantly larger than 0.5 in all the four sub-conditions (short-high, $p < 0.01$; short-low, $p < 0.01$; long-high, $p < 0.01$; long-low, $p < 0.05$).

DISCUSSION

Using a temporal adaptation paradigm, we demonstrated that adaptation to the preceding short temporal interval (50 ms) induced significant negative aftereffects on perception of the subsequent visual Ternus AM, irrespective of whether the time interval was conveyed by events in the same modality (i.e., visual AM or blinking discs) or in a different modality (i.e., auditory beeps). This pattern of aftereffects suggests that there is a general "temporal pacemaker" mechanism (Treisman, 1963; Treisman et al., 1990, 1994) and amodal representation for sub-second interval time. Although adaptation to the preceding long temporal interval (200 ms) did not lead to unanimous significant aftereffects across the three tasks, the differences between experiments may reflect the differential impacts of temporal attending (see below) in the visual and auditory modalities, rather than distinct time interval representations in different modalities for the sub-second range.

The within-modality aftereffect for the short time interval adaptation replicated Becker and Rasmussen (2007); the significant between-modality adaptation aftereffect, however, contrasted sharply with the null effect in Becker and Rasmussen (2007), suggesting that the implicit task used here is more sensitive to the adaptation aftereffect than the explicit reproduction task. The existence of cross-modality adaptation effect is clearly inconsistent with the idea of distinct timers for different modalities (Keele et al., 1989; Ivry, 1996; Pashler, 2001), at least at the sub-second range. Instead, it suggests that there is amodal representation of internal clock and adapting to the repetitive stimuli in one modality can alter the speed of the internal clock, leading to a subjectively changed percept of the subsequent time interval in another modality. Specifically, according to the "temporal pacemaker" model, temporal frequency of preceding repetitive stimuli can influence the speed of internal clock and hence the perceived subsequent (target) time interval. A higher frequency can increase the speed of internal clock, rendering a given time interval being perceived longer; a lower frequency can decrease the speed of internal clock, making a given time interval being perceived shorter (Ono and Kitazawa, 2011).

On the other hand, the regular repetitive, rhythmic stimuli can trigger temporal attending—a shift of attentional focus to anticipate the onsets of subsequent events (Jones et al., 2002). In other words, the temporal attending mechanism, established after exposing to either visual or auditory sequences, guides the distribution of attentional resources around the time points the rhythmic stimuli are presented. The pattern of temporal distribution

of attentional resources over different time points can be applied to subsequent within-modal or cross-modal events, affecting the temporal processing of these events (Jones, 1976; Large and Jones, 1999; Jones et al., 2002). This effect of temporal attending is dependent on the reliability of perceiving the temporal regularity. Given that perception and reproduction of auditory rhythmic sequences are generally better than perception and reproduction of visual rhythmic sequences (Welch et al., 1986; Glenberg et al., 1989; Glenberg and Jona, 1991; Recanzone, 2003, 2009; Repp, 2003; Patel et al., 2005), it is possible that the effect of temporal attending is more potent in the auditory domain than in the visual domain.

We suggest that the change of speed of the internal clock by the repetitive adaptation stimuli with short or long time intervals and the efficiency of temporal attending in different modalities co-determined the patterns of adaptation aftereffects. For the short interval (50 ms) adaptation, the internal clock speed was accelerated by both visual and auditory adaptation stimulus sequences, potentially leading to more reports of group motion in the subsequent Ternus displays. However, the temporal attending, established after exposing to either visual or auditory sequences, affected the distribution of attentional resources around the time points that the two visual frames of the Ternus display were presented (Aydin et al., 2011). Specifically, although the first frame could be aligned with the first time point of the temporal attending, the second frame, located after the second time point of the temporal attending, could be "pulled" in time closer to the second time point (see Aydin et al., 2011; Keetels et al., 2007; Chen et al., 2010; Shi et al., 2010 for the effect of temporal attention on perceptual segregation), potentially leading to more reports of element motion. Thus, adaptation to rhythmic sequence of visual or auditory events had two potentially conflicting consequences; with the short interval (50 ms), the increase of clock speed could play a dominant role, leading to more reports of group motion overall.

For the long interval (200 ms) adaptation, the slowed-down clock would make the interval between the frames of the visual display being perceived shorter, and this should potentially lead to more reports of element motion. However, the temporal attending mechanism would "pull" the second frame, located before the second time point of temporal attending, toward the second point, potentially leading to more reports of group motion. These two conflicting effects could cancel each other for adaptation within the visual modality. For cross-modality adaptation, however, given that adapting to rhythmic auditory events could activate a stronger temporal attending mechanism than the adapting to rhythmic visual events (Welch et al., 1986; Jones et al., 2002; Recanzone, 2003, 2009), the overall effect was the stronger segregation of the two Ternus frames and more reports of group motion.

The finding of equivalent aftereffects for auditory beeps with short and long intervals in cross-modal adaptation is surprising. To replicate this effect and to rule out an alternative account which attributes the positive aftereffects to arousal evoked by auditory input, we conducted a further experiment similar to Experiment 3 except that the pitch of the auditory beeps was manipulated. Previous studies showed that the arousal state correlates with the fundamental frequency of sound and temporal rhythms (Banziger and Scherer, 2005; Bruck et al., 2008, 2009). We adopted a within-subject factorial design, with two levels of adaptation time intervals (50 or 200 ms) and two types of pitches (500 or 5000 Hz), plus an additional subjective rating task of perceived arousal (Edelman, 1970; Gudjonsson, 1981; Slomine et al., 1999; Cuthbert et al., 2000). The subjective ratings of the perceived arousal were differed among the four types of auditory sequences (short interval-low pitch, short interval-high pitch, long interval-low pitch, long interval-high pitch), both auditory sequences with higher pitch and short intervals were perceived as higher arousal. However, the perceived different arousal levels did not affect the pattern of adaptation aftereffects, both short interval and long interval adaptation lead to more reports of "group motion" and there was no statistical difference for the percentages of "group motion" in the two conditions (**Figure 3**). The less impact of arousal upon the temporal adaptation aftereffect might be due to three possible reasons: *First*, there is interplay between attention and arousal. The arousal effects are two sides of a coin. The (higher) arousal contributes to accumulate more pulses as implicated in the pacemaker model (Treisman, 1963), the perceived interval would be expanded and hence give rise to a dominant percept of "group motion"; on the other hand, the arousal auditory stimuli attract more attention, meanwhile less attention resources were allocated to time processing itself. The reduced attention on the pacemaker would lead to a loss of pulses and reduced interval, as indicated by a number of relevant studies (Fortin, 2003; Buhusi and Meck, 2006; Noulhiane et al., 2007; Buhusi and Meck, 2009). The reduced interval would give rise to dominant percept of "element motion". The above opposite influences interact and cancel out, imposing non-observable impact on the probe visual motion. *Second*, adaptation to temporal intervals with short temporal length (about average 4.5 seconds for an adaptation trial in our experiments) might trigger an "immediate" arousal (Coull, 1998; Del-Fava and Ribeiro-do-Valle, 2004), and the effect of the arousal perhaps dissipates after several hundred milliseconds (with 900 ms delay between the offset of the adaptation sequence and the probe in current study)

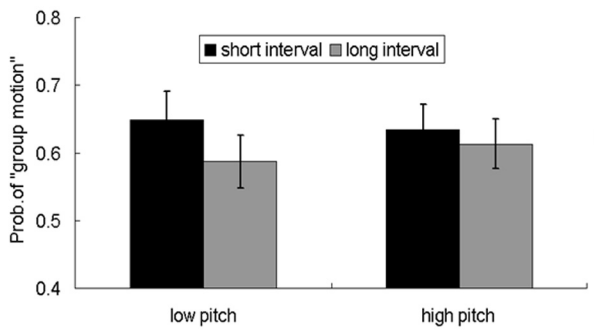


FIGURE 3 | Reports of group motion for the probe Ternus display after adapting to the four kinds of auditory sequences: low pitch-short interval, low pitch-long interval, high pitch-short interval, and high pitch-long interval. The error bar represents standard error.

(Ulrich and Mattes, 1996; Fernandez-Duque and Posner, 1997; Coull et al., 2000). *Third*, even the arousal effect still remains after a temporal delay, the arousal effect has been revealed to be less important and somehow inhibited by the entrained attention issued by the auditory sequence (Jones et al., 2002; Del-Fava and Ribeiro-do-Valle, 2004). Previous study using visual discrimination tasks, where auditory stimuli as (preceding) accessory stimuli, would speed up the response to a subsequent visual stimulus, however, for the accessory auditory stimuli, the expectancy (of temporal attention) is revealed to be more important and could inhibit the "immediate arousal" effect (Del-Fava and Ribeiro-do-Valle, 2004). This analogous mechanism might operate in the current investigation using short temporal range for adaptation.

To conclude, using an adaptation paradigm with implicit test of timing, the present study found that adapting to a short time interval conveyed by either visual or auditory stimuli leads to more report of group motion in the subsequent visual Ternus probes; adapting to a longer time interval, however, caused no aftereffect for visual adaptation but significantly more reports of group motion for auditory adaptation. These results suggest that there exists amodal representation for sub-second timing, but adaptation to repetitive, rhythmic sequence of stimuli in different modalities may elicit temporal attending of different strengths, affecting the manifestation of adaptation after effects.

ACKNOWLEDGMENTS

This study was supported by grants from the Natural Science Foundation of China (30770712, 30970889, 90920012) and the Ministry of Science and Technology of China (2010CB833904) to Xiaolin Zhou; from the China Postdoctoral Science Foundation (20100470, 201104032) to Lihan Chen.

REFERENCES

Allan, L. G., and Kristofferson, A. B. (1974). Psychophysical theories of duration discrimination. *Percept. Psychophys.* 16, 26–34.

Aydin, M., Herzog, M. H., and Öğmen, H. (2011). Attention modulates spatio-temporal grouping. *Vision Res.* 51, 435–446.

Banziger, T., and Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Commun.* 46, 252–267.

Becker, M. W., and Rasmussen, I. P. (2007). The rhythm aftereffect: support for time sensitive neurons with broad overlapping tuning curves. *Brain Cogn.* 64, 274–281.

Brainard, D. H. (1997). The psychophysics toolbox. *Spat. Vis.* 10, 433–436.

Bruck, D., Ball, M., and Thomas, I. (2008). Auditory arousal thresholds as a function of sounds of different pitches and pattern. *J. Sleep Res.* 17, 255–256.

Bruck, D., Ball, M., Thomas, I., and Rouillard, V. (2009). How does the pitch and pattern of a signal affect auditory arousal thresholds? *J. Sleep Res.* 18, 196–203.

Buhusi, C. V., and Meck, W. H. (2005). What makes us tick? Functional and neural mechanisms of interval timing. *Nat. Rev. Neurosci.* 6, 755–765.

Buhusi, C. V., and Meck, W. H. (2006). Interval timing with gaps and distracters: evaluation of the ambiguity, switch, and time-sharing hypotheses. *J. Exp. Psychol. Anim. Behav. Process.* 32, 329–338.

Buhusi, C. V., and Meck, W. H. (2009). Relative time sharing: new findings and an extension of the resource allocation model of temporal processing. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 1875–1885.

Chen, L., Shi, Z., and Müller, H. J. (2010). Influences of intra- and crossmodal grouping on visual and tactile Ternus apparent motion. *Brain Res.* 1354, 152–162.

Coull, J. T. (1998). Neural correlates of attention and arousal: insights from electrophysiology, functional neuroimaging and psychopharmacology. *Prog. Neurobiol.* 55, 343–361.

Coull, J. T., Frith, C. D., Büchel, C., and Nobre, A. C. (2000). Orienting attention in time: behavioural and neuroanatomical distinction between exogenous and endogenous shifts. *Neuropsychologia* 38, 808–819.

Creelman, C. D. (1962). Human discrimination of auditory duration. *J. Acoust. Soc. Am.* 34, 528–593.

Cuthbert, B. N., Schupp, H. T., Bradley, M. M., Birbaumer, N., and Lang, P. J. (2000). Brain potentials in affective picture processing: covariation with autonomic arousal and affective report. *Biol. Psychol.* 52, 95–111.

Czeisler, C. A., Duffy, J. F., Shanahan, T. L., Brown, E. N., Mitchell, J. F., Rimmer, D. W., et al. (1999). Stability, precision, and near-24-hour period of the human circadian pacemaker. *Science* 284, 2177–2181.

Del-Fava, F., and Ribeiro-do-Valle, L. E. (2004). Relative contribution of expectancy and immediate arousal to the facilitatory effect of an auditory accessory stimulus. *Braz. J. Med. Biol. Res.* 37, 1161–1174.

Divenyi, P. L., and Danner, W. F. (1977). Discrimination of time intervals marked by brief acoustic pulses of various intensities and spectra. *Percept. Psychophys.* 21, 125–142.

Edelman, R. I. (1970). Validity of verbal report as a prognosticator of physiological arousal to threat. *J. Abnorm. Psychol.* 76, 492.

Fernandez-Duque, D., and Posner, M. I. (1997). Relating the mechanisms of orienting and alerting. *Neuropsychologia* 35, 477–486.

Fortin, C. (2003). "Attentional time-sharing in interval timing," in *Functional and Neural Mechanisms of Interval Timing*, ed W. H. Meck (Boca Raton, FL: CRC Press), 235–259.

Fraisse, P. (1963). *The Psychology of Time.* New York, NY: Harper and Row.

Freeman, E., and Driver, J. (2008). Direction of visual apparent motion driven solely by timing of a static sound. *Curr. Biol.* 18, 262–1266.

Glenberg, A., and Jona, M. (1991). Temporal coding in rhythm tasks revealed by modality effects. *Mem. Cognit.* 19, 514–522.

Glenberg, A., Mann, S., Altman, L., Forman, T., and Procise, S. (1989). Modality effects in the coding reproduction of rhythms. *Mem. Cognit.* 17, 373–383.

Grondin, S., and Rousseau, R. (1991). Judging the relative duration of multimodal short empty time intervals. *Percept. Psychophys.* 49, 245–256.

Grothe, B. (2003). New roles for synaptic inhibition in sound localization. *Nat. Rev. Neurosci.* 4, 540–550.

Gudjonsson, G. H. (1981). Self-reported emotional disturbance and its relation to electrodermal reactivity, defensiveness and trait anxiety. *Pers. Indiv. Differ.* 2, 47–52.

Large, E. W., and Jones, M. R. (1999). The dynamics of attending: how people track time varying events. *Psychol. Rev.* 106, 119–159.

Ivry, R. B. (1993). "Cerebellar involvement in the explicit representation of temporal information," in *Temporal information Processing in the Nervous System*, eds P. Tallal, A. Gallaburda, R. R. Llinas, and C. von Euler (New York, NY: Annals of the New York Academy of Sciences), 214–230.

Ivry, R. B. (1996). The representation of temporal information in perception and motor control. *Curr. Opin. Neurobiol.* 6, 851–857.

Ivry, R. B., and Hazeltine, R. E. (1995). Perception and production of temporal intervals across a range of durations: evidence for a common timing mechanism. *J. Exp. Psychol. Hum. Percept. Perform.* 21, 3–18.

Ivry, R. B., and Schlerf, J. E. (2008). Dedicated and intrinsic models of time perception. *Trends Cogn. Sci.* 12, 273–280.

Ivry, R. B., Spencer, R. M., Zelaznik, H. N., and Diedrichsen, J. (2002). The cerebellum and event timing. *Ann. N.Y. Acad. Sci.* 978, 302–317.

Jones, M. R. (1976). Time, our lost dimension: toward a new theory of perception, attention, and memory. *Psychol. Rev.* 83, 323–355.

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.

Keele, S. W., and Ivry, R. B. (1991). "Does the cerebellum provide a common computation for diverse tasks?" in *The Development and Neural Bases of Higher Cognitive Functions*, ed A. Diamond (New York, NY: New York Academy of Sciences), 179–211.

Keele, S. W., Nicoletti, R., Ivry, R. I., and Pokorny, R. A. (1989). Mechanisms of perceptual timing: beat-based or interval-based judgements? *Psychol. Res.* 50, 251–256.

Keele, S. W., Pokorny, R. A., Corcos, D. M., and Ivry, R. (1985). Do perception and motor production share common timing mechanisms: a correlational analysis. *Acta Psychol.* 60, 173–191.

Keetels, M., Stekelenburg, J., and Vroomen, J. (2007). Auditory grouping occurs prior to intersensory pairing: evidence from temporal ventriloquism. *Exp. Brain Res.* 180, 449–456.

Killeen, P. R., and Weiss, N. A. (1987). Optimal timing and the Weber function. *Psychol. Rev.* 94, 455–468.

Lapid, E., and Ulrich, R. (2009). Perceptual learning in auditory temporal discrimination: no evidence for a cross-modal transfer to the visual modality. *Psychon. Bull. Rev.* 16, 382–389.

Lewis, P. A., and Miall, R. C. (2003). Distinct systems for automatic and cognitively controlled time measurement: evidence from neuroimaging. *Curr. Opin. Neurobiol.* 13, 250–255.

Mauk, M. D., and Buonomano, D. V. (2004). The neural basis of temporal processing. *Annu. Rev. Neurosci.* 27, 307–340.

Meegan, D. V., Aslin, R. N., and Jacobs, R. A. (2000). Motor timing learned without motor training. *Nat. Neurosci.* 3, 860–862.

Merchant, H., Zarco, W., and Prado, L. (2008). Do we have a common mechanism for measuring time in the hundreds of millisecond range? Evidence from multiple-interval timing tasks. *J. Neurophysiol.* 99, 939–949.

Nagarajan, S. S., Blake, D. T., Wright, B. A., Byl, N., and Merzenich, M. M. (1998). Practice-related improvements in somatosensory interval discrimination are temporally specific but generalize across skin location, hemisphere, and modality. *J. Neurosci.* 18, 1559–1570.

Noulhiane, M., Mella, N., Samson, S. S., Ragot, R. R., and Pouthas, V. V. (2007). How emotional auditory stimuli modulate time perception. *Emotion* 7, 697–704.

Ono, M., and Kitazawa, S. (2011). Shortening of subjective visual intervals followed by repetitive stimulation. *PLoS ONE* 6:e28722. doi: 10.1371/journal.pone.0028722

Pashler, H. (2001). Perception and production of brief durations: beat based versus interval-based timing. *J. Exp. Psychol. Hum. Percept. Perform.* 27, 485–493.

Patel, A. D., Iversen, J. R., Chen, Y., and Repp, B. H. (2005). The influence of metricality and modality on synchronization with a beat. *Exp. Brain Res.* 163, 226–238.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.* 10, 437–442.

Penton-Voak, I. S., Edwards, H., Percival, A., and Wearden, J. H. (1996). Speeding up an internal clock in humans? Effects of click trains on subjective duration. *J. Exp. Psychol. Anim. Behav. Process.* 22, 307–320.

Petersik, J. T., and Rice, C. M. (2006). The evolution of explanations of a perceptual phenomenon: a case history using the Ternus effect. *Perception* 35, 807–821.

Pöppel, E. (1988). *Mindworks: Time and Conscious Experience*. Boston, MA: Harcourt Brace Jovanovich.

Rammsayer, T. H. (1999). Neuropharmacological evidence for different timing mechanisms in humans. *Q. J. Exp. Psychol. B* 52, 273–286.

Recanzone, G. H. (2003). Auditory influences on visual temporal rate perception. *J. Neurophysiol.* 89, 1078–1093.

Recanzone, G. H. (2009). Interactions of auditory and visual stimuli in space and time. *Hear. Res.* 258, 89–99.

Repp, B. H. (2003). Rate limits in sensorimotor synchronization with auditory and visual sequences: the synchronization threshold and the benefits and costs of interval subdivision. *J. Mot. Behav.* 35, 355–370.

Shi, Z., Chen, L., and Müller, H. J. (2010). Auditory temporal modulation of the visual Ternus effect: the influence of time interval. *Exp. Brain Res.* 203, 723–735.

Shi, Z., Jia, L., and Müller, H. J. (2012). Modulation of tactile duration judgments by emotional pictures. *Front. Integr. Neurosci.* 6:24. doi: 10.3389/fnint.2012.00024

Slomine, B., Bowers, D., and Heilman, K. M. (1999). Dissociation between autonomic responding and verbal report in right and left hemisphere brain damage during anticipatory anxiety. *Neuropsychiatry Neuropsychol. Behav. Neurol.* 12, 143–148.

Ternus, J. (1926). Experimentelle Untersuchungen über phänomenale Identität. *Psychol. Res.* 7, 81–136.

Treisman, M. (1963). Temporal discrimination and the indifference interval: implications for a model of the "internal clock." *Psychol. Monogr.* 77, 1–31.

Treisman, M., Cook, N., Naish, P. L. N., and MacCrone, J. K. (1994). The internal clock: electroencephalographic evidence for oscillatory processes underlying time perception. *Q. J. Exp. Psychol.* A 47, 241–289.

Treisman, M., Faulkner, A., Naish, P. L. N., and Brogan, D. (1990). The internal clock: evidence for a temporal oscillator underlying time perception with some estimates of its characteristic frequency. *Perception.* 19, 705–743.

Treutwein, B., and Strasburger, H. (1999). Fitting the psychometric function. *Percept. Psychophys.* 61, 87–106.

Ulrich, R., and Mattes, S. (1996). Does immediate arousal enhance response force in simple reaction time? *Q. J. Exp. Psychol.* A 49, 972–990.

Wearden, J. H., Norton, R., Martin, S., and Montford-Bebb, O. (2007). Internal clock processes and the filled-duration illusion. *J. Exp. Psychol. Hum. Percept. Perform.* 33, 716–729.

Welch, R. B., Duttonhurt, L. D., and Warren, D. H. (1986). Contributions of audition and vision to temporal rate perception. *Percept. Psychophys.* 39, 294–300.

Welch, R. B., and Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychol. Bull.* 88, 638–667.

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: 30 May 2012; accepted: 16 October 2012; published online: 05 November 2012.

Citation: Zhang H, Chen L and Zhou X (2012) Adaptation to visual or auditory time intervals modulates the perception of visual apparent motion. Front. Integr. Neurosci. ***6****:100. doi: 10.3389/fnint.2012.00100*

frontiers in
INTEGRATIVE NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 01 October 2012
doi: 10.3389/fnint.2012.00071

Assessing the effect of physical differences in the articulation of consonants and vowels on audiovisual temporal perception

Argiro Vatakis[1]*, Petros Maragos[2], Isidoros Rodomagoulakis[2] and Charles Spence[3]

[1] *Cognitive Systems Research Institute, Athens, Greece*
[2] *Computer Vision, Speech Communication and Signal Processing Group, National Technical University of Athens, Athens, Greece*
[3] *Crossmodal Research Laboratory, Department of Experimental Psychology, University of Oxford, UK*

Edited by:
Zhuanghua Shi, Ludwig-Maximilians-Universität München, Germany

Reviewed by:
Virginie Van Wassenhove, Cognitive Neuroimaging Unit, France
Massimiliano Di Luca, University of Birmingham, UK

***Correspondence:**
Argiro Vatakis, Cognitive Systems Research Institute, 7 Makedonomaxou Prantouna, 11525 Athens, Greece.
e-mail: argiro.vatakis@gmail.com

We investigated how the physical differences associated with the articulation of speech affect the temporal aspects of audiovisual speech perception. Video clips of consonants and vowels uttered by three different speakers were presented. The video clips were analyzed using an auditory-visual signal saliency model in order to compare signal saliency and behavioral data. Participants made temporal order judgments (TOJs) regarding which speech-stream (auditory or visual) had been presented first. The sensitivity of participants' TOJs and the point of subjective simultaneity (PSS) were analyzed as a function of the place, manner of articulation, and voicing for consonants, and the height/backness of the tongue and lip-roundedness for vowels. We expected that in the case of the place of articulation and roundedness, where the visual-speech signal is more salient, temporal perception of speech would be modulated by the visual-speech signal. No such effect was expected for the manner of articulation or height. The results demonstrate that for place and manner of articulation, participants' temporal percept was affected (although not always significantly) by highly-salient speech-signals with the visual-signals requiring smaller visual-leads at the PSS. This was not the case when height was evaluated. These findings suggest that in the case of audiovisual speech perception, a highly salient visual-speech signal may lead to higher probabilities regarding the identity of the auditory-signal that modulate the temporal window of multisensory integration of the speech-stimulus.

Keywords: temporal perception, TOJs, articulatory features, speech, audiovisual, signal saliency, attentional modeling

INTRODUCTION

The optimal perception (i.e., the successful perception) of speech signals requires the contribution of both visual (i.e., articulatory gestures) and auditory inputs, with the visual signal often providing information that is complementary to that provided by the auditory signal (e.g., Sumby and Pollack, 1954; Erber, 1975; McGrath and Summerfield, 1985; Summerfield, 1987; Reisberg et al., 1987; Arnold and Hill, 2001; Davis and Kim, 2004; Ross et al., 2007; Arnal et al., 2009). Speech intelligibility has been shown to be fairly robust under conditions where a time discrepancy and/or a spatial displacement has been introduced between the auditory and/or visual stream of a given speech signal (e.g., Munhall et al., 1996; Jones and Jarick, 2006). The present study focuses on the former case, where a signal delay (either auditory or visual) is present in a congruent audiovisual speech stream. Such delays occur frequently in everyday life as the by-product of poor transmission rates often found in broadcasting or sensory processing delays (e.g., Spence and Squire, 2003; Vatakis and Spence, 2006a).

In order to understand how audiovisual speech perception is affected by the introduction of temporal asynchronies, researchers have evaluated the limits of the temporal window of audiovisual integration (i.e., the interval in which no temporal discrepancy between the signals is perceived; outside of this window, audiovisual stimuli are perceived as being desynchronized) and the specific factors that modulate the width of this temporal window (e.g., Vatakis and Spence, 2007, 2010). One of the first studies to investigate the temporal perception of speech stimuli was reported by Dixon and Spitz (1980). Participants in their study had to monitor a video of a man reading prose that started in synchrony and was gradually desynchronized at a rate of 51 ms/s (up to a maximum asynchrony of 500 ms) with either the auditory or visual stream leading. The participants had to respond as soon as they detected the asynchrony in the video. Dixon and Spitz reported that the auditory stream had to lag the visual stream by an average of 258 ms or lead by 131 ms before the asynchrony in the speech signal became noticeable (see also Conrey and Pisoni, 2003, 2006, for similar results using a simultaneity judgment, SJ, task; i.e., participants had to report whether the stimuli were synchronous or asynchronous). More recently, Grant et al. (2004), using a two-interval forced choice adaptive procedure, reported that participants in

their study only noticed the asynchrony in audiovisual sentences when the auditory-speech led the visual-speech signal by at least 50 ms or else lagged by 220 ms or more (see also Grant and Seitz, 1998; Miner and Caudell, 1998; Grant and Greenberg, 2001). Meanwhile, McGrath and Summerfield (1985) reported a study in which the intelligibility of audiovisual sentences presented in white noise deteriorated at much lower visual leads (160 ms; see also Pandey et al., 1986; Steinmetz, 1996) than those observed in the studies of Dixon and Spitz, Conrey and Pisoni, and Grant and colleagues. Based on these results, it would appear as though the perception of a continuous audiovisual speech signal remains intelligible across a wide range of signal delays (auditory or visual). It is not clear, however, what the exact interval range is since a high level of variability has been observed between the various studies that have been conducted to date (see **Figure 1A**).

In addition to the studies that have used continuous speech stimuli (i.e., passages, sentences), audiovisual temporal perception has also been evaluated for brief speech tokens using the McGurk effect (i.e., the visual influence on the perception of audiovisual speech; McGurk and MacDonald, 1976). For instance, Massaro et al. (1996) evaluated the temporal perception of consonant-vowel (CV) syllables under a wide range of different asynchronies using the fuzzy logic model of perception (FLMP). They found that audiovisual integration (as assessed by participants' reports of what was heard; i.e., speech identification task) was unaffected for auditory leads and lags of up to 250 ms (see also Massaro and Cohen, 1993). However, Munhall et al. (1996) reported results that were quite different. They presented vowel-consonant-vowel (VCV) stimuli and their results demonstrated that participants experienced the McGurk effect for auditory leads of 60 ms and lags of 240 ms. These values are similar to those that have been reported by Van Wassenhove et al. (2003, 2007) for CV stimuli (auditory leads from about 30 ms and lags of up to 200 ms; see **Figure 1B**, for a summary of these and other findings).

On the whole, the results of previous studies concerning the temporal perception of audiovisual speech stimuli have demonstrated that the audiovisual intelligibility of the speech signal remains high over a wide range of audiovisual temporal asynchronies. That said, this time-range (i.e., window) exhibits great variability across different studies (see **Figure 1**). This marked variation led us to investigate the possible factors that may affect the temporal perception of audiovisual speech (see Vatakis and Spence, 2006a–c, 2007, 2008, 2010). One factor that may help to explain the presence of variability in the temporal windows of multisensory integration previously observed for audiovisual speech stimuli relates to the particular speech stimuli utilized in the various studies. Specifically, the temporal window of integration for audiovisual speech has, in recent years, been shown

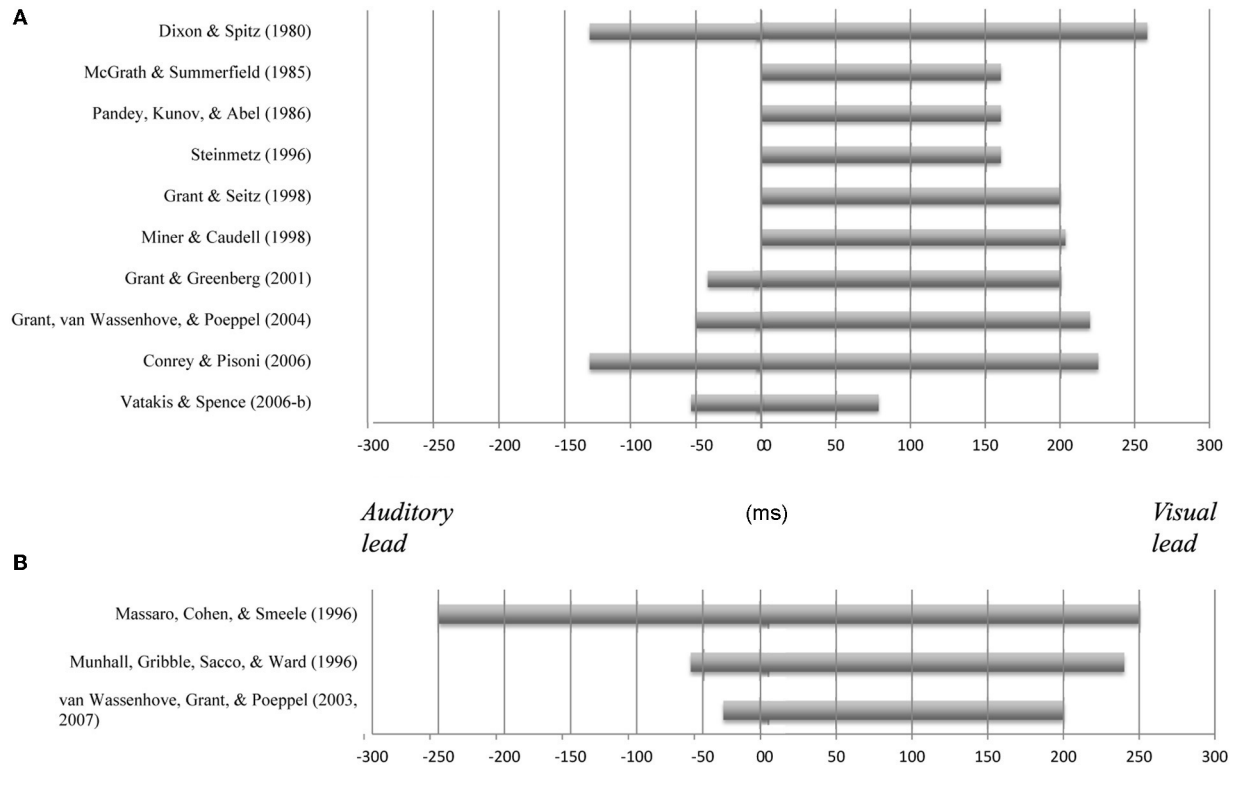


FIGURE 1 | Variability in the temporal window of multisensory integration observed in previous studies using continuous audiovisual speech stimuli and a variety of different response measures (including identification tasks, simultaneity judgment task, etc.; (A) and brief speech tokens in McGurk-type presentations (B).

to vary as a function of the physical parameters of the visual stimulus (e.g., inversion of the visual-speech stimulus promotes a wider window of integration; e.g., Vatakis and Spence, 2008) and the type of speech stimulus used (e.g., Vatakis and Spence, 2006b, 2007). Additionally, the temporal window of audiovisual integration appears to be wider for more complex stimuli (or more highly temporally correlated stimuli; e.g., syllables vs. words or sentences) than for simpler stimuli (such as light flashes and sound bursts; e.g., Navarra et al., 2005).

The present study focused on another possible factor that may affect the temporal perception of audiovisual speech, which is the effect that the physical changes due to the articulation of consonants (mainly characterized by the articulatory features of the place and manner of articulation and voicing; see Kent, 1997) and vowels (mainly characterized by the articulatory features of the height/backness of the tongue and roundedness of the lips; see Kent, 1997) may have on the parameters defining the temporal window for audiovisual integration. The optimal perception of speech stimuli requires the synergistic integration of auditory and visual inputs. However, according to the "information reliability hypothesis" in multisensory perception (whereby, the perception of a feature is dominated by the modality that provides the most reliable information), one could argue that the perception of a given speech token may, in certain cases, be dominated by the auditory-speech or the visual lip-movement that is more informative (e.g., Schwartz et al., 1998; Wada et al., 2003; Andersen et al., 2004; Traunmüller and Öhrström, 2007). Specifically, previous research has shown that auditory inputs are closely associated with the accurate detection of the manner of articulation and voicing of consonants, and the height/backness of vowels. Visual input, by contrast, provides essential cues regarding the accurate detection of the place of articulation of consonants and the roundedness of vowels (e.g., Miller and Nicely, 1955; Massaro and Cohen, 1993; Robert-Ribes et al., 1998; Girin et al., 2001; Mattys et al., 2002; Traunmüller and Öhrström, 2007). For example, Binnie et al. (1974) examine people's ability to identify speech by modulating the unimodal and bimodal contribution of vision and audition to speech using 16 CV syllables presented under noisy listening conditions. Their results indicated a large visual contribution to audiovisual speech perception (e.g., 41.4% visual dominance at −18 dB S/N), with the visual contribution being highly associated with the place of articulation of the CV syllables used. However, masking the auditory input has been shown to lead to a loss of information about the place of articulation, whereas information about the manner of articulation appears to be resistant to such masking (i.e., McGurk and MacDonald, 1976; Mills and Thiem, 1980; Summerfield and McGrath, 1984; Massaro and Cohen, 1993; Robert-Ribes et al., 1998; see Dodd, 1977, for a related study using CVCs; and Summerfield, 1983, for a study using vowels instead).

Previous studies of the effects of audiovisual asynchrony on speech perception have only been tested using a small number of syllables (e.g., Van Wassenhove et al., 2003; Vatakis and Spence, 2006b). It has therefore not been possible, on the basis of the results of such studies, to draw any detailed conclusions regarding the possible interactions of physical differences in speech articulation with audiovisual temporal perception. Additionally, given new findings indicating that high signal reliability leads to smaller temporal order thresholds (i.e., smaller thresholds imply high auditory- and visual-signal reliability; see Ley et al., 2009), further study of the temporal window of integration in audiovisual speech is necessary in order to possibly resolve the differences noted in previous studies. In the present study, we utilized a variety of different consonants (Experiments 1 and 2) and vowels (Experiment 3) in order to examine the possible effects that physical differences in articulation may have on the temporal perception of audiovisual speech stimuli. The stimuli used were selected according to the categorization of articulatory features established by the International Phonetic Alphabet (IPA) and they were sampled in such a way as to allow for comparison within and across different categories of articulation. We conducted a series of three experiments that focused on different articulatory features, and thus on the differential contribution of the visual- and auditory-speech signal. Specifically, in Experiment 1 (A–C), we focused on the place of articulation (i.e., the location in the vocal tract where the obstruction takes place; e.g., /p/ vs. /k/) and voicing features (i.e., the manner of vibration of the vocal folds; e.g., /p/ vs. /b/); in Experiment 2 (A–C), we looked at the manner of articulation (i.e., how the obstruction is made and the sound produced; e.g., /s/ vs. /t/); and, in Experiment 3, we explored the temporal perception of audiovisual speech as a function of the height/backness of the tongue and roundedness of the lips.

The temporal perception of the speech stimuli utilized in the present study was assessed using an audiovisual temporal order judgment (TOJ) task with a range of stimulus onset asynchronies (SOAs) using the method of constant stimuli (e.g., Spence et al., 2001). The TOJ task required the participants to decide on each trial whether the auditory-speech or the visual-speech stream had been presented first. Using the TOJ task permitted us to obtain two indices: the Just Noticeable Difference (JND) and the Point of Subjective Simultaneity (PSS). The JND provides a measure of the sensitivity with which participants could judge the temporal order of the auditory- and visual-speech streams. The PSS provides an estimate of the time interval by which the speech event in one sensory modality had to lead the speech event in the other modality in order for synchrony to be perceived (or rather, for the "visual-speech first" and "auditory-speech first" responses to be chosen equally often).

Overall, we expected that for the speech stimuli tested here (see **Table 1**) visual leads would be required for the synchrony of the auditory- and visual-signals to be perceived (i.e., PSS; except for the case of vowels, where auditory leads have been observed previously; Vatakis and Spence, 2006a). That is, during speech perception, people have access to visual information concerning the place of articulation before they have the relevant auditory information (e.g., Munhall et al., 1996). In part, this is due to the significant visual motion (e.g., the movement of facial muscles) that occurs prior to the auditory onset of a given syllable. In addition, according to the "information reliability hypothesis" (e.g., Schwartz et al., 1998; Traunmüller and Öhrström, 2007), we would expect that participants' TOJ responses would

Table 1 | The main articulatory features used to categorize the consonant and vowel stimuli used in Experiments 1–3, as a function of: (A) the place of articulation, (B) the manner of articulation, and (C) the height and backness of the tongue and roundedness of the lips in Experiments 1–3.

Place of articulation	Manner of articulation		
	Experiment 1A	Experiment 1B	Experiment 1C
	Stop	Fricative	Nasal
(A) CONSONANTS			
Bilabial	/b, p/	–	/m/
Labiodental	–	/v, f/	–
Dental	–	/ð, θ/	–
Alveolar	/d, t/	/z, s/	/n/
Velar	/g, k/	–	/ŋ/
Manner of articulation	**Place of articulation**		
	Experiment 2A	Experiment 2B	Experiment 2C
	Bilabial	Alveolar	Postalveolar
(B) CONSONANTS			
Stop	/b/	/d/	–
Fricative	–	/z/	/ʒ/
Nasal	/m/	/n/	–
Affricative	–	–	/ʤ/
Lateral approximant	–	/l/	/r/
Approximant	/w/	–	–
Height of tongue	**Backness/roundedness**		
	Front/unrounded	Back/rounded	
(C) VOWELS			
High	/i/	/u/	
Mid	/ɛ/	/ɔ/	
Low	/æ/	/ɒ/	

be differentially affected as a function of the "weight" placed on the auditory-speech or the visual lip-movement for the accurate detection of a particular speech token. That is, in the cases where the visual-speech signal is more salient (such as, for determining the place of articulation of consonants and the roundedness of vowels; such as, stimuli that involve high visible contrast with highly visible lip-movements; e.g., bilabial stimuli or rounded vowels; Experiments 1 and 3), we would expect participants' to be more sensitive to the presence of asynchrony (i.e., they should exhibit lower JNDs) as compared to less salient stimuli (such as, those involving tongue movement, as tongue movements are not always visible; e.g., as in the case of velar stimuli and unrounded vowels). No such effects would be expected for those cases where the auditory-speech input is more salient, such as, in the cases where the manner of articulation and voicing of consonants and the height/backness of vowels are evaluated (see Experiments 2 and 3). One must note, however, that in case auditory and visual signals are equally reliable, this should lead to smaller temporal order thresholds (i.e., JNDs; see Ley et al., 2009).

EXPERIMENTS 1–3

MATERIALS AND METHODS

Participants

All of the participants were naïve as to the purpose of the study and all reported having normal or corrected-to-normal hearing and visual acuity. The experiments were performed in accordance with the ethical standards laid down in the 1990 Declaration of Helsinki, as well as the ethical guidelines laid down by the Department of Experimental Psychology, University of Oxford. Each experiment took approximately 50 min to complete.

Apparatus and materials

The experiment was conducted in a completely dark sound-attenuated booth. During the experiment, the participants were seated facing straight-ahead. The visual stream was presented on a 17-inch (43.18 cm) TFT color LCD monitor (SXGA 1240 × 1024 pixel resolution; 60-Hz refresh rate), placed at eye level, approximately 68 cm in front of the participants. The auditory stream was presented by means of two Packard Bell Flat Panel 050 PC loudspeakers, one placed 25.4 cm to either side of the center of the monitor (i.e., the auditory- and visual-speech stimuli were presented from the same spatial location). The audiovisual stimuli consisted of black-and-white video clips presented on a black background, using Presentation (Version 10.0; Neurobehavioral Systems, Inc., CA). The video clips (300 × 280-pixel, Cinepak Codec video compression, 16-bit Audio Sample Size, Average pitch and amplitude (in Hz): 160 and 43, for consonants; 125 and 44, for vowels, respectively; 24-bit Video Sample Size, 30 frames/s) were processed using Adobe Premiere 6.0. The video clips consisted of the close-up views of the faces of a British male and two British females (visible from the chin to the top of the head), looking directly at the camera, and uttering a series of speech tokens (see **Table 1**). The open vowel /a/ was used for all of the articulated consonants in order to provide high levels of visible contrast relative to the closed mouth in the rest position. All of the audiovisual clips were 1400 and 2500 ms in duration (measured from the still frame before visual articulation of the speech token began to the last frame after articulation of the token had occurred) for consonants and vowels, respectively. All of the speech stimuli were recorded under the same conditions with the mouth starting and ending in a closed position. The articulation of all of the speech tokens was salient enough without our having to make the stimuli unnatural (i.e., by having the speakers exaggerate). In order to achieve accurate synchronization of the dubbed video clips, each original clip was re-encoded using XviD codec (single pass, quality mode of 100%).

At the beginning and end of each video clip, a still image and background acoustic noise was presented for a variable duration. The duration of the image and noise was unequal, with the difference in their duration being equivalent to the particular SOA tested (values reported below) in each condition. This aspect of the design ensured that the auditory and visual streams always started at the same time, thus ensuring that the participants were not cued as to the nature of the audiovisual

delay with which they were being presented. In order to achieve a smooth transition at the start and end of each video clip, a 33.33 ms cross-fade was added between the still image and the video clip (Note here that a newer methodology by Maier et al., 2011, allows for better control and, thus, more accurate measurement of the synchrony of the audiovisual stimulus presentation). The participants responded using a standard computer mouse, which they held with both hands, using their right thumb for "visual-speech first" responses and their left thumb for "speech-sound first" responses (or vice versa, the response buttons were counterbalanced across participants).

Design

Nine SOAs between the auditory and visual streams were used: ±300, ±200, ±133, ±66, and 0 ms (the negative sign indicates that the auditory stream was presented first, whereas the positive sign indicates that the visual stream was presented first). This particular range of SOAs was selected on the basis of previous research showing that people can typically discriminate the temporal order of briefly-presented audiovisual speech stimuli at 75% correct at SOAs of approximately 80 ms (e.g., McGrath and Summerfield, 1985; Vatakis and Spence, 2006a; see also Munhall and Vatikiotis-Bateson, 2004). The participants completed one block of practice trials before the main experimental session in order to familiarize themselves with the task and the video clips. The practice trials were followed by five blocks of experimental trials. Each block consisted of two presentations of each of the stimuli used at each of the nine SOAs (presented in a random order using the method of constant stimuli; see Spence et al., 2001).

Procedure

At the start of the experiment, the participants were informed that they would have to decide on each trial whether the auditory-speech or visual-speech stream appeared to have been presented first. They were informed that they would sometimes find this discrimination difficult, in which case they should make an informed guess as to the order of stimulus presentation. The participants were also informed that the task was self-paced, and that they should only respond when confident of their response. The participants were informed that they did not have to wait until the video clip had finished before making their response, but that a response had to be made before the experiment would advance on to the next trial. The participants were instructed prior to the experiment not to move their heads and to maintain their fixation on the center of the monitor throughout each block of trials.

ANALYSIS

The proportions of "visual-speech first" responses at each SOA were converted to their equivalent z-scores under the assumption of a cumulative normal distribution (Finney, 1964). The data of each participant and condition from the seven intermediate SOAs (±200, ±133, ±66, and 0 ms) were cumulated and converted in z-scores to be fitted with a straight line (values were limited between 0.1 and 0.9; 0 and 1 were weighted using $((n-(n-1))/n)$*100 and $((n-1)/n)$*100), respectively, where n is the number of trials). Slope values were used to calculate the JND (JND = 0.675/slope; since ± 0.675 represents the 75% and 25% point on the cumulative normal distribution) and intercepts were used to obtain PSSs (PSS = −intercept/slope; see Coren et al., 2004, for further details). The ±300 ms points were excluded from this computation due to the fact that most participants performed near-perfectly at this interval and therefore these data points did not provide significant information regarding our experimental manipulations (cf. Spence et al., 2001, for a similar approach). For all of the analyses reported here, repeated measures analysis of variance (ANOVA) and Bonferroni-corrected t-tests (where $p < 0.05$ prior to correction) were used.

Preliminary analysis of the JND and PSS data using a repeated measures ANOVA revealed no effects[1] attributable to the different speakers used to create the audiovisual stimuli, thus we combined the data from the three different speakers in order to simplify the statistical analysis (see Conrey and Gold, 2000, for a discussion of the effects of speaker differences on performance). The goodness of the data fits was significant for all conditions in all experiments conducted and the normality tests were also significant for all factors tested.

AUDIOVISUAL PHYSICAL SIGNAL SALIENCY ANALYSIS

Bottom-up attention or saliency is based on the sensory cues of a stimulus captured by its signal-level properties, such as spatial, temporal, and spectral contrast, complexity, scale, etc. Similar to competitive selection, saliency can be attributed on the feature level, the stream level or the modality level. Based on perceptual and computational attention modeling studies, efficient bottom-up models and signal analysis algorithms have been developed by Evangelopoulos et al. (2008) in order to measure the saliencies of both the auditory and visual streams in audiovisual videos of complex stimuli such as movie video clips. These saliencies can be integrated into a multimodal attention curve, in which the presence of salient events is signified by geometrical features such as local extrema and sharp transition points. By using level sets of this fused audiovisual attentional curve, a movie summarization algorithm was proposed and evaluated.

In the present study, we used the algorithms developed by Evangelopoulos et al. (2008) to separately compute two temporal curves indicating the saliencies of the auditory and visual streams for the stimuli presented (see **Figure 2** for an example). Auditory saliency was captured by bandpass filtering the acoustic signal into multiple frequency bands, modeling each bandpass component as a modulated sinusoid, and extracting features such as its instantaneous amplitude and frequency. These features were motivated by biological observations and psychophysical evidence that, modulated carriers seem more salient perceptually to human observers compared to stationary signals (e.g.,

[1]Experiments 1A, B—no significant interaction between Place, Voicing, and Speaker in either the JND or PSS data [$F_{(4,\ 52)} < 1$, *n.s.*], for both; Experiment 1C—no significant interaction of Place and Speaker for the JND and PSS data; $F_{(4,\ 48)} = 1.35$, $p = 0.27$; $F_{(4,\ 48)} = 2.44$, $p = 0.11$, respectively; Experiments 2A, C—no significant interaction of Manner of articulation and Speaker for the JND and PSS data; $F_{(4,\ 40)} < 1$, *n.s.*, for both; Experiment 2B—no significant interaction of Manner of articulation and Speaker for the JND and PSS data; $F_{(6,\ 60)} = 1.15$, $p = 0.20$; $F_{(6,\ 60)} = 1.27$, $p = 0.28$, respectively.

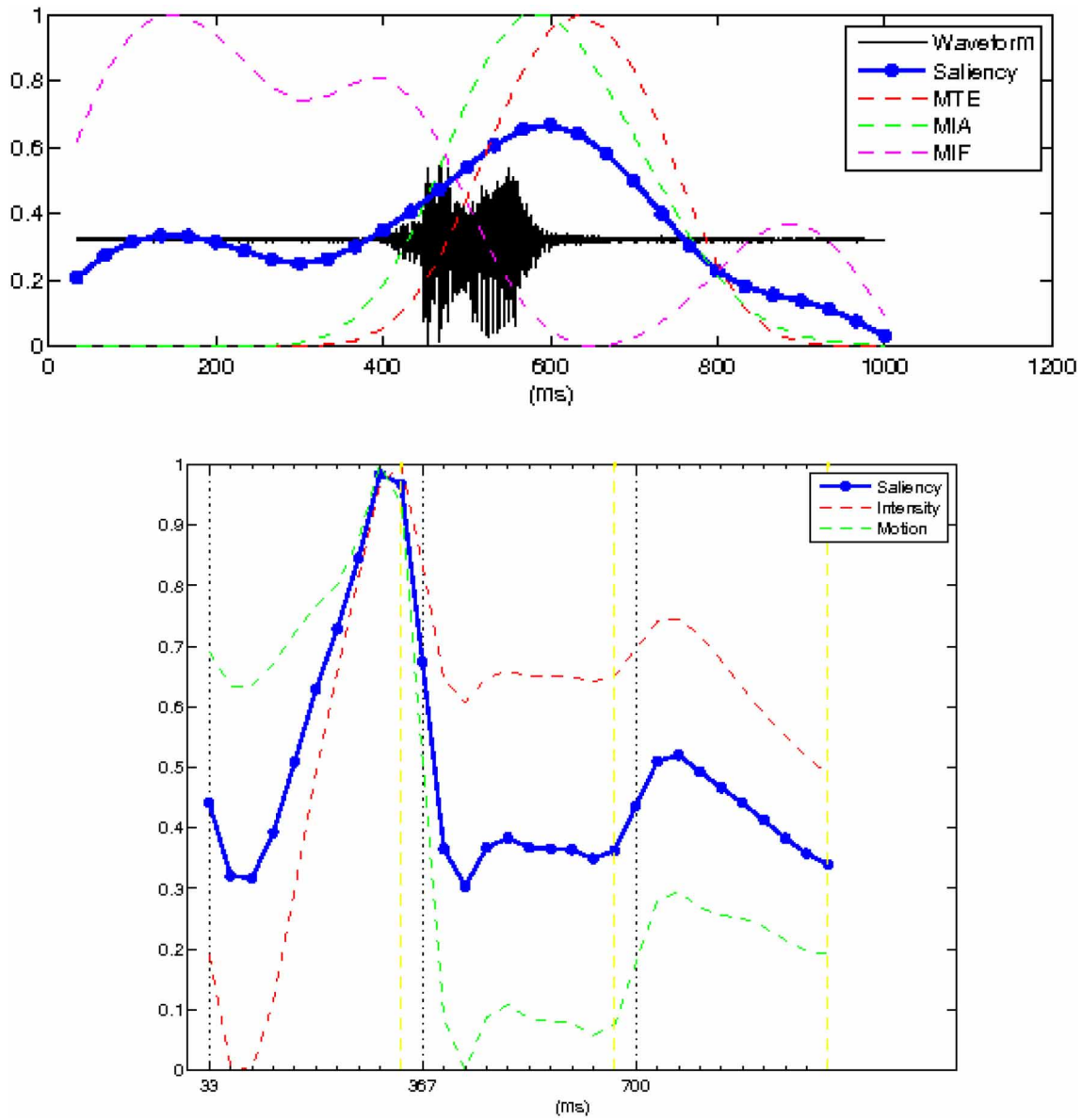


FIGURE 2 | Top panel shows the acoustic waveform (solid black line) of the speech utterance with the auditory salience superimposed (thick solid line). The superimposed dashed lines show the temporal evolution of the three auditory cues (mean instantaneous energy, MTE, amplitude, MIA, and the frequency of the dominant frequency channel, MIF) whose linear combination gives the saliency. **Bottom panel** shows the visual saliency curve (in thick solid line). The superimposed dash lines shows the two visual cues that contributed to the computation of the visual saliency.

Tsingos et al., 2004; Kayser et al., 2005). In our experiments, the audio signal is sampled at 16 kHz and the audio analysis frames usually vary between 10 and 25 ms. The auditory filterbank consists of symmetric zero-phase Gabor filters, which do not introduce any delays. In the frequency domain, the filters are linearly arranged in frequency steps of 200–400 Hz, yielding a tessellation of 20–40 filters (details of the auditory feature extraction process can be found in Evangelopoulos and Maragos, 2006, and Evangelopoulos et al., 2008). The final auditory saliency temporal curve was computed as a weighted linear combination of three acoustic features: the mean instantaneous energy of the most active filter and the mean instantaneous amplitude, and frequency of the output from this dominant filter.

The visual saliency computation module is based on the notion of a centralized saliency map (Koch and Ullman, 1985; Itti et al., 1998) computed through a feature competition scheme, which is motivated by the experimental evidence of a biological counterpart in the human visual system (interaction/competition among the different visual pathways related to motion/depth and gestalt/depth/color, respectively; Kandel et al., 2000). Thus, visual saliency was measured by means of this spatiotemporal attentional model, driven by three feature cues: intensity, color (this

feature was not used in our experiments given that the videos were presented in black and white), and motion. The spatiotemporal video volume (with time being the third dimension) was decomposed into a set of feature volumes, at multiple spatiotemporal scales (details on the visual feature extraction process can be found in Rapantzikos et al., 2009). By averaging over spatiotemporal neighborhoods, the feature (intensity and motion) volumes yielded a visual attentional curve whose value at each time instant represents the overall visual saliency of the corresponding video frame. The visual feature extraction process was synchronized with the respective auditory task on a frame-by-frame basis.

RESULTS AND DISCUSSION

PLACE OF ARTICULATION AND VOICING FOR STOP CONSONANTS (EXPERIMENT 1A)

In Experiment 1A, we evaluated whether (and how) the place of articulation and voicing of stop consonants (the manner of articulation was constant) influenced audiovisual TOJs. We categorized the data according to the factors of Place of articulation (three levels: bilabial, /b, p/; alveolar, /d, t/; velar, /g, k/) and Voicing (two levels: voiced, /b, d, g/; unvoiced, /p, t, k/; see **Table 1A**).

Fourteen participants (12 female; native English speakers) aged between 18 and 30 years (mean age of 24 years) took part in this experiment. A repeated measures ANOVA on the JND data revealed no significant main effect of Place of articulation [$F_{(2,\ 26)} = 2.10$, $p = 0.15$]. Although the participants were, numerically-speaking, more sensitive to the temporal order of the auditory- and visual-speech streams for bilabial stimuli (M = 55 ms) than for either alveolar (M = 67 ms) or velar (M = 68 ms) stimuli (see **Figure 3A**), this difference failed to reach statistical significance. There was also no significant main effect of Voicing [$F_{(1,\ 13)} < 1$, *n.s.*] (voiced, M = 64 ms; unvoiced, M = 63 ms; see **Figure 6C**), and the Place of Articulation by Voicing interaction was not significant either [$F_{(2,\ 26)} = 1.27$, $p = 0.30$].

The analysis of the PSS data revealed a significant main effect of Place of articulation [$F_{(2,\ 26)} = 6.72$, $p < 0.01$]. Large visual leads were required for the alveolar (M = 31 ms) and velar stimuli (M = 35 ms) as compared to the small auditory lead required for the bilabial (M = 3 ms) stimuli in order for the PSS to be reached ($p < 0.05$, for both comparisons; see **Figure 3B**). These results suggest that auditory leads were required when a visible place contrast was present for bilabial speech stimuli as compared to the large visual leads required for the invisible place contrast present in alveolar and velar stimuli (e.g., Girin et al., 2001). We also obtained a significant main effect of Voicing [$F_{(1,\ 13)} = 12.65$, $p < 0.01$], with voiced stimuli (M = 32 ms) requiring larger visual leads than unvoiced stimuli (M = 10 ms; see **Figure 6D**). There was no interaction between Place of articulation and Voicing [$F_{(2,\ 26)} < 1$, *n.s.*].

PLACE OF ARTICULATION AND VOICING FOR FRICATIVE CONSONANTS (EXPERIMENT 1B)

We further investigated the influence of the place of articulation and voicing on audiovisual temporal perception by testing fricative consonants. The data were categorized by the factors of Place of articulation (three levels: labiodental, /v, f/; dental, /ð, θ/; alveolar, /z, s/) and Voicing (two levels: voiced, /v, ð, z/; unvoiced, /f, θ, s/).

Fourteen new participants (10 female; native English speakers) aged between 18 and 34 years (mean age of 24 years) took part in this experiment. Analysis of the JND data revealed no significant main effect of Place of articulation [$F_{(2,\ 26)} = 1.40$, $p = 0.26$] or Voicing [$F_{(1,\ 13)} = 3.74$, $p = 0.10$], nor any interaction between these two factors [$F_{(2,\ 26)} < 1$, *n.s.*]. Participants' performance was very similar across the speech groups compared as a function of the Place of articulation and across the Voicing groups tested (i.e., Place of articulation: labiodental, M = 56 ms; dental, M = 58 ms; alveolar, M = 63 ms; Voicing: voiced, M = 57 ms; unvoiced, M = 61 ms; see **Figures 3A**,**6C**). Labiodental and dental stimuli are considered to be higher in visibility than alveolar stimuli (e.g., Binnie et al., 1974; Dodd, 1977; Cosi and Caldognetto, 1996). The JND values showed a trend toward higher visibility stimuli resulting in numerically smaller JNDs, however, this effect was not significant.

Analysis of the PSS data, however, revealed a significant main effect of Place of articulation [$F_{(2,\ 26)} = 8.51$, $p < 0.01$], with larger visual leads being required for the alveolar stimuli (M = 42 ms) than for the labiodental (M = 11 ms) or dental (M = 6 ms) stimuli ($p < 0.01$, for both comparisons; see **Figure 3B**). Given that labiodental and dental stimuli are considered to be higher in visibility than alveolar stimuli, the larger visual leads required for the alveolar stimuli provide similar results to those observed for the stop consonants tested earlier (Experiment 1A). There was no significant main effect for Voicing [$F_{(1,\ 13)} < 1$, *n.s.*] (see **Figure 6D**), nor was there any interaction between Place of articulation and Voicing [$F_{(2,\ 26)} = 2.84$, $p = 0.10$].

PLACE OF ARTICULATION FOR NASALS (EXPERIMENT 1C)

Finally, we evaluated the influence of the Place of articulation on audiovisual TOJs by testing nasal consonants (the voicing factor was not evaluated since nasals are voiced-only). The data were evaluated according to Place of articulation (three levels: bilabial, /m/; alveolar, /n/; velar, /ŋ/).

Thirteen new participants (nine female; native English speakers) aged between 19 and 34 years (mean age of 24 years) took part in the experiment. The analysis of the JND data resulted in a significant main effect of Place of articulation [$F_{(2,\ 24)} = 4.45$, $p < 0.05$], indicating that the participants were significantly more sensitive to the temporal order of the auditory- and visual-speech streams when evaluating bilabial stimuli (M = 51 ms) than when judging either alveolar (M = 60 ms) or velar (M = 64 ms) stimuli ($p < 0.05$ for both comparisons; see **Figure 3A**). These results are similar to the trend observed in Experiments 1A and B, with participants being more sensitive to the temporal order of the highly-visible speech tokens (e.g., Binnie et al., 1974; Sams et al., 1991; Robert-Ribes et al., 1998; Girin et al., 2001; see Massaro and Cohen, 1993, for evidence that people are better at identifying the syllable /ba/ as compared to the syllable /da/).

Analysis of the PSS data revealed a significant main effect of Place of articulation [$F_{(2,\ 24)} = 2.62$, $p < 0.05$], with the visual stream having to lead by a larger interval for the alveolar (M = 39 ms) and velar stimuli (M = 25 ms) than for the

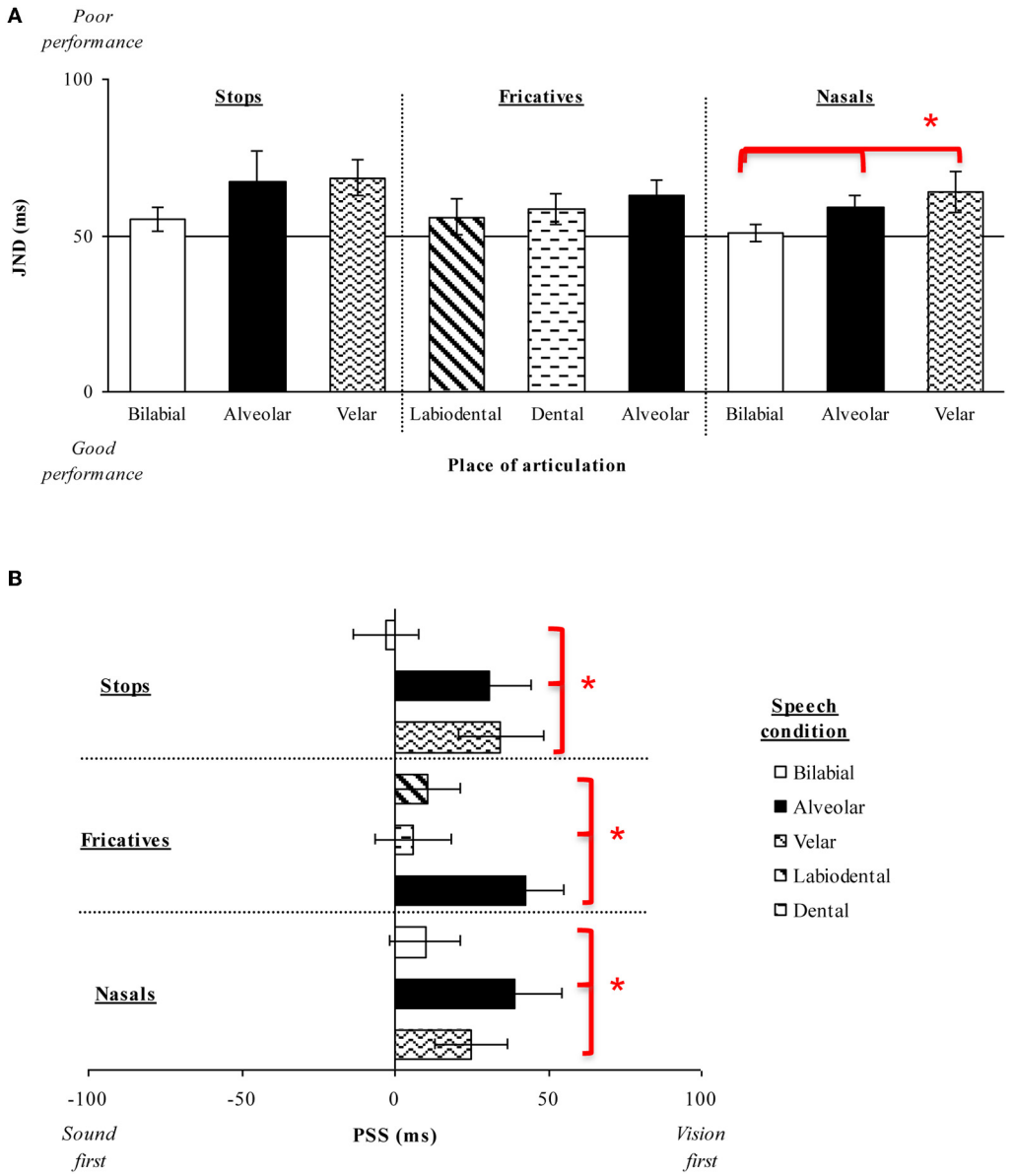


FIGURE 3 | (A) Average JNDs and **(B)** PSSs for the place of articulation of the consonant stimuli presented in Experiment 1. The error bars represent the standard errors of the mean. Asterisks indicate significant differences between the various stimuli presented.

bilabial (M = 10 ms) stimuli in order for the PSS to be reached ($p < 0.05$, for both comparisons; see **Figure 3B**). Once again, these results are similar to those obtained previously for the stop consonants (Experiment 1A), where alveolar and velar stimuli were shown to require greater visual leads as compared to bilabial stimuli.

Overall, therefore, the results of Experiments 1A–C demonstrate that the visual signal had to lead the auditory signal in order for the PSS to be reached for the speech stimuli tested here (see **Figure 3B**). The sole exception was the bilabial stimuli in Experiment 1A, where an auditory lead of 3 ms was required (although, note that this value was not significantly different from 0 ms; $[t_{(13)} < 1, n.s.]$). These findings are supported by prior research showing that one of the major features of audiovisual speech stimuli is that the temporal onset of the visual-speech often occurs prior to the onset of the associated auditory-speech (i.e., Munhall et al., 1996; Lebib et al., 2003; Van Wassenhove et al., 2003, 2005). More importantly for present purposes, the results of Experiments 1A–C also revealed that the amount of time by which the visual-speech stream had to lead the auditory-speech stream in order for the PSS to be reached was smaller in the presence of a highly-visible speech stimulus (e.g., bilabials) than when the speech stimulus was less visible (e.g., as in the case of alveolars; see **Figure 4A**). This finding is also compatible with the cluster responses that are often reported in studies of speech intelligibility that have utilized McGurk syllables. For

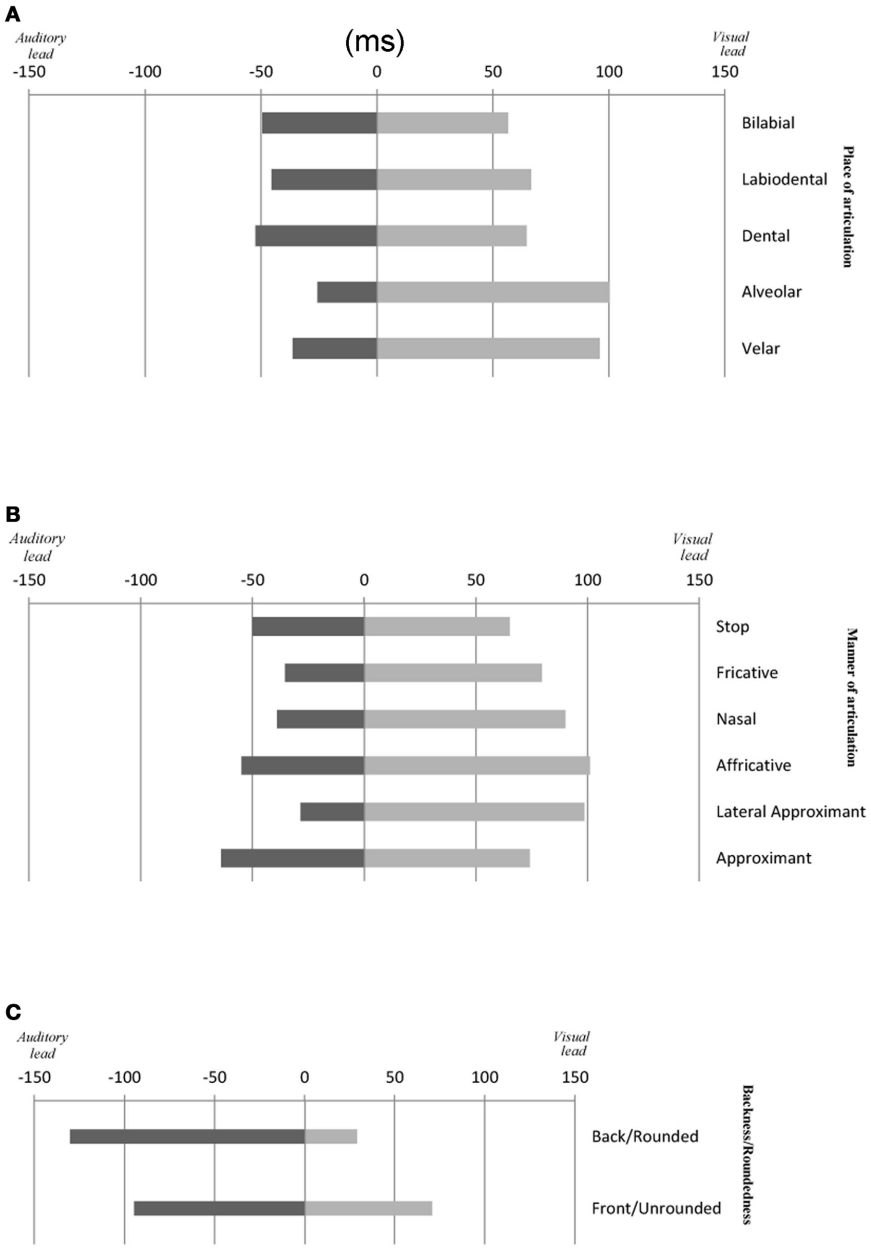


FIGURE 4 | Average temporal window of integration (PSS ± JND) for audiovisual speech as a function of: (A) the place of articulation and (B) the manner of articulation of consonant and (C) backness/roundedness of vowel stimuli used in this study.

example, the presentation of a visual /ba/ together with an auditory /da/ often produces the response /bda/. This is not, however, the case for the presentation of a visual /da/ and an auditory /ba/ (i.e., where no /dba/ cluster is observed). This result can partially be accounted for by the faster processing of the visual /ba/ as compared to the visual /da/ (e.g., Massaro and Cohen, 1993). It should also be noted that the sensitivity of our participants' audiovisual TOJ responses was only found to differ as a function of changes in the place of articulation (a visually-dominant feature) in Experiment 1C but not in Experiments 1A–B. Additionally, no differences were obtained in participants' sensitivity as a function of voicing, which is an auditorily-dominant feature (e.g., Massaro and Cohen, 1993; Girin et al., 2001).

In order to examine the relationship between the perceptual findings described above and the physical properties of the audiovisual stimuli utilized in Experiments 1A–C, we conducted an auditory- and visual-saliency analysis of the synchronous audiovisual stimuli by using the computational algorithms developed by Evangelopoulos et al. (2008) to compute audio-visual saliencies in multimodal video summarization. The saliency analysis allowed calculation of the saliency rise (i.e., beginning of the

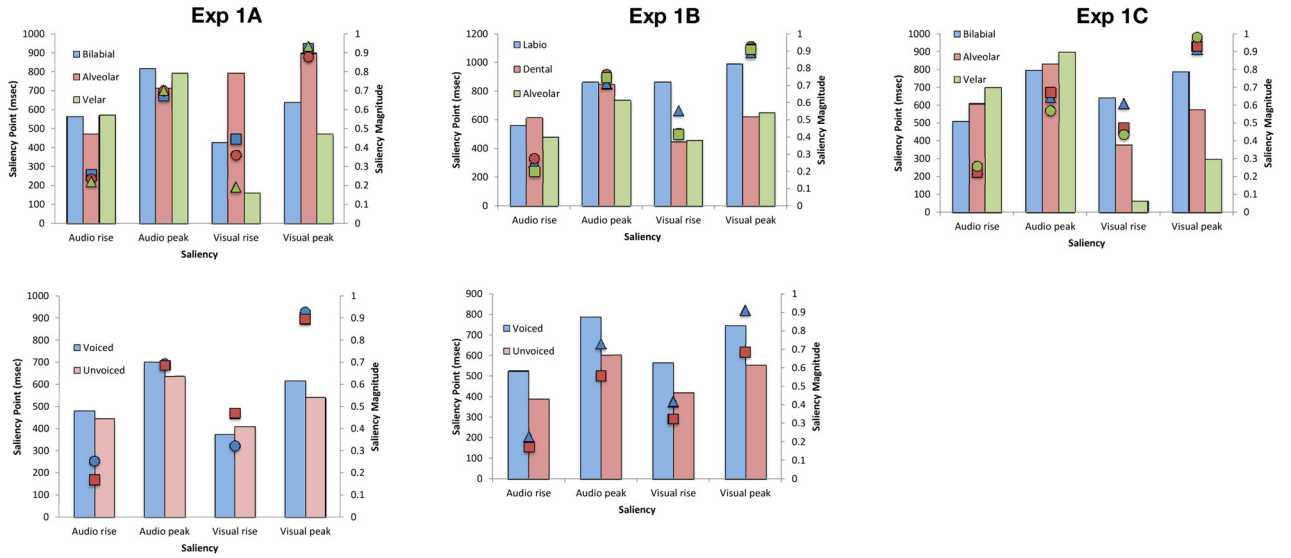


FIGURE 5 | Average saliency rise and peak (in ms) and saliency magnitude for each point for the audiovisual speech stimuli used in Experiments 1A–C as a function of the place of articulation and voicing.

saliency increase) and peak of each modality stream (in ms) and the magnitude of each saliency point (see **Figure 5**). In terms of the place of articulation, the saliency rise and peak occurred earlier for the visual stream as compared to the auditory stream for all stimuli except for the alveolar (Experiment 1A), labiodental (Experiment 1B), and bilabial (Experiment 1C) stimuli, where the reverse pattern was noted. The magnitude for each saliency rise and peak point, highlighted a clear trend for all stimuli with the magnitude being approximately the same for all points except for that of visual rise. Specifically, the highest saliency magnitude of visual rise was found for bilabials (Experiments 1A, C) and labiodentals (Experiment 1B).

Comparison of the physical and perceptual data revealed a trend whereby better TOJ performance coincided with visual rises that were larger in magnitude, thus, suggesting that higher in saliency stimuli lead to better detection of temporal order. In terms of PSS, the physical and perceptual data also exhibited a trend in terms of magnitude with larger visual leads being required for stimuli of lower magnitude (except for the case of dental stimuli in Experiment 1B), implying that lower magnitude stimulation is less salient, in terms of signal saliency, as compared to high in magnitude saliency points.

The saliency analysis for voicing did not reveal a consistent pattern, which might be due to the fact that voicing constitutes an auditorily-dominant feature. Specifically, the PSS-saliency magnitude pattern observed earlier was also present in Experiment 1A but not in 1B, where voiced stimuli were higher in magnitude in all saliency points.

The results of Experiments 1A–C therefore demonstrate that higher in saliency visual-speech stimuli lead to higher temporal discrimination sensitivity and smaller visual-stream leads for the speech signal. Previous studies support the view that visual-speech may act as a cue for the detection of speech sounds when the temporal onset of the speech signal is uncertain (e.g., Barker et al., 1998; Grant and Seitz, 1998, 2000; Arnold and Hill, 2001, though, see also Bernstein et al., 2004). Therefore, in the present study, it may be that the less visually salient speech stimuli required a greater visual lead in order to provide complementary information for the appropriate speech sound. We conducted a second set of experiments in order to explore how the manner of articulation of consonants affects audiovisual temporal perception. As mentioned already, the manner of articulation is an auditorily-dominant feature, thus we would not expect the visual-speech signal to modulate the temporal perception of consonants in the same manner as that observed in Experiment 1. The apparatus, stimuli, design, and procedure were exactly the same as in Experiment 1 with the sole exception that different groups of audiovisual speech stimuli were tested that now focused solely on the articulatory feature of the manner of articulation of consonants. All the stimuli tested in Experiments 2A–C were composed of voiced consonants with a constant place of articulation (see **Table 1B**).

MANNER OF ARTICULATION FOR BILABIALS (EXPERIMENT 2A)

We were interested in the influence that the manner of articulation of voiced bilabials has on the temporal aspects of audiovisual speech perception. We categorized the data according to the factor of Manner of articulation (three levels: stop, /b/; nasal, /m/; and approximant, /w/).

Eleven new participants (six female; native English speakers) aged between 18 and 30 years (mean age of 24 years) took part in the experiment. The participants were numerically somewhat more sensitive to the temporal order of the stop (Mean JND = 63 ms) and approximant (M = 69 ms) stimuli than for the nasal stimuli (M = 72 ms), although the main effect of the Manner of articulation was not significant [($F_{(2,\ 20)} < 1$, *n.s.*); see

Figure 6A]. The analysis of the PSS data, however, revealed a significant main effect of the Manner of articulation ($F_{(2, 20)} = 5.92$, $p < 0.05$), with significantly larger visual leads being required for the nasal stimuli (M = 27 ms) in order for the PSS to be reached as compared to the much smaller visual leads required for the stop (M = 3 ms) and approximant (M = 5 ms) stimuli ($p < 0.05$, for both comparisons; see **Figure 6B**). The results obtained here are similar to those reported in Experiments 1A–C in terms of the PSS data, where significantly smaller visual leads were required for the highly-visible stop and approximant stimuli as compared to the less visible nasal stimuli.

MANNER OF ARTICULATION FOR ALVEOLARS (EXPERIMENT 2B)

We were also interested in what role, if any, the manner of articulation of voiced alveolars would play in the temporal aspects of audiovisual speech perception. We evaluated the data based on the factor of Manner of articulation (four levels: stop, /d/; fricative, /z/; nasal, /n/; and lateral approximant, /l/).

Eleven new participants (six female; native English speakers) aged between 19 and 30 years (mean age of 24 years) took part in the experiment. The participants were slightly more sensitive to the temporal order of stop (M = 52 ms) and lateral approximant (M = 53 ms) stimuli than to the temporal order of the fricative (M = 57 ms) and nasal (M = 57 ms) stimuli (see **Figure 6A**). However, the analysis of the JND data revealed no significant main effect of the Manner of articulation [$F_{(3, 30)} = 1.23$, $p = 0.32$]. The analysis of the PSS data highlighted a significant main effect of the Manner of articulation [$F_{(3, 30)} = 9.13$, $p < 0.01$], with significantly larger visual leads being required for the fricative stimuli (M = 47 ms) as compared to the visual leads required for the stop stimuli (M = 12 ms), and the auditory leads required for the lateral approximant (M = 3 ms) stimuli ($p < 0.05$; $p < 0.01$, respectively).

MANNER OF ARTICULATION FOR POSTALVEOLARS (EXPERIMENT 2C)

Finally, we evaluated how the manner of articulation of voiced postalveolars influences the temporal aspects of audiovisual speech perception by varying the stimuli used as a function of the Manner of articulation (three levels: fricative, /ʒ/; affricative, /ʤ/; and lateral approximant, /r/).

Eleven new participants (five female; native English speakers) aged between 18 and 34 years (mean age of 24 years) took part in this experiment. The analysis of the JND data revealed a significant main effect of the Manner of articulation [$F_{(2, 20)} = 4.60$, $p < 0.05$], with the participants being significantly more sensitive to the temporal order of fricative stimuli (M = 58 ms) than of affricative (M = 78 ms) or lateral approximant stimuli (M = 74 ms; $p < 0.05$, for all comparisons; see **Figure 6A**). A similar analysis of the PSS data also revealed a significant main effect of the Manner of articulation [$F_{(2, 20)} = 12.24$, $p < 0.01$]. Fricative stimuli (M = 3 ms) required auditory leads for the PSS to be reached as compared to the visual leads required for the affricative (M = 23 ms) and lateral approximant (M = 73 ms) stimuli ($p < 0.05$, for all comparisons; see **Figure 6B**). The results obtained with the postalveolar stimuli tested here agree with the general findings of Experiment 1, whereby stimuli with a lower JND value (i.e., stimuli where participants are more sensitive to the temporal order of the presentation of the auditory and visual stimuli) also required smaller visual leads (i.e., fricatives). However, lateral approximant stimuli are generally considered to be more visible than fricative stimuli, therefore the higher sensitivity (in terms of the lower JNDs) observed here for fricative stimuli does not agree with the idea that highly-visible stimuli result in improved sensitivity to temporal order (i.e., lower JNDs).

The saliency analysis of the auditory and visual signals for the stimuli presented in Experiments 2A–C (see **Figure 7**) once again revealed saliency changes of greater magnitude for the points of visual rise, while the visual rise was not reached earlier as consistently as found in Experiment 1. Specifically, visual rise was earlier for stops and approximants in Experiment 2A, stop and lateral approximant in Experiment 2B, and fricative and lateral approximant in Experiment 2C. This earlier visual rise also coincides with the previously-noted trend toward better sensitivity to temporal order for these stimuli (which, however, only reached significance in the behavioral data for fricatives in Experiment 2C). In terms of saliency magnitude, no specific trend was observed (as with voicing in Experiment 1). This null result might be driven by the fact that the manner of articulation is an auditorily-driven feature. Specifically, in Experiments 2A and 2C, the participants required larger visual leads for nasals and affricatives, respectively, while physically those stimuli were higher in saliency magnitude for visual rise but saliency was reached earlier for auditory rise. Fricatives and lateral approximants in Experiments 2B and 2C, respectively, required perceptually visual leads for synchrony to be perceived, while the saliency magnitude was high and the saliency rise was reached earlier for the visual stream.

The results of Experiments 2A–C demonstrate similar results to those observed in Experiments 1A–C in terms of the PSS data. That is, the amount of time by which the visual-speech stream had to lead the auditory-speech stream in order for the PSS to be reached was smaller in the presence of highly-visible speech stimuli as compared to less-visible speech stimuli (see **Figure 4B**). There was no consistent pattern of the behavioral and the physical data, however, this result may be accounted for by the fact that the manner of articulation is a feature (just like voicing) that is highly associated with the auditory input (Massaro and Cohen, 1993; Girin et al., 2001). The results of Experiments 2A–C also revealed (just as had been highlighted in Experiments 1A–C) that the visual signal had to precede the auditory signal in order for the PSS to be achieved (except in the case of fricative and lateral approximant stimuli where a small auditory lead was observed; Experiments 2C and 2B, respectively; However, once again, this value was not significantly different from 0 ms; [$t_{(10)} = 1.10$, $p = 0.32$]; [$t_{(10)} < 1$, *n.s.*], respectively).

By themselves, the results of Experiments 2A–C suggest that visual-speech saliency influences the temporal perception of audiovisual speech signals mainly in terms of the PSS data. The perceptual and physical data do not, however, exhibit a consistent pattern. This may reflect the fact that the manner of articulation represents a feature that is largely dependent on the auditory signal for successful extraction of the speech signal, thus making the visible identification of all voiced consonants particularly difficult (due to the fact that neither the movements of the velum nor those of the vocal folds are visible; see Cosi and Caldognetto,

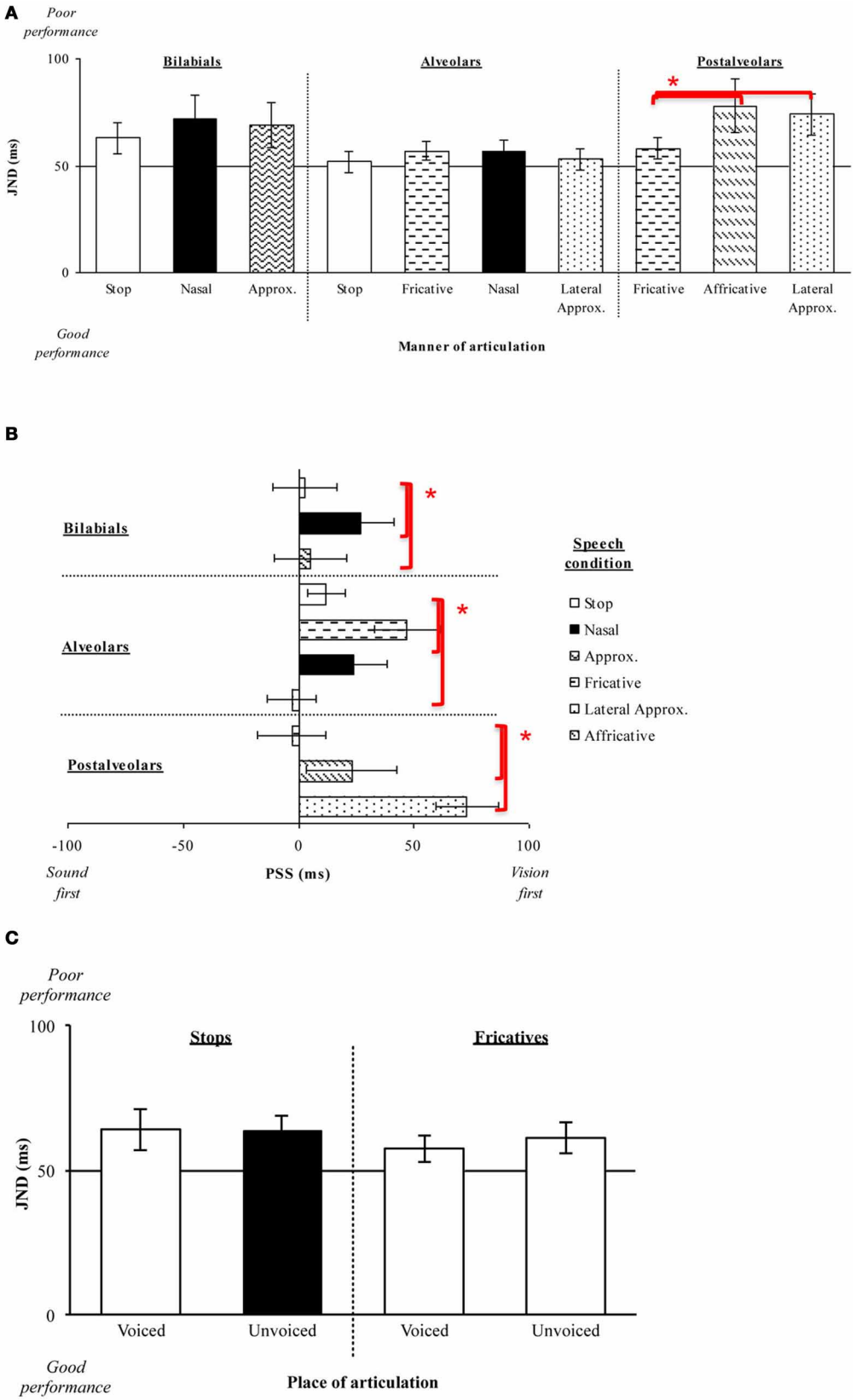


FIGURE 6 | Continued

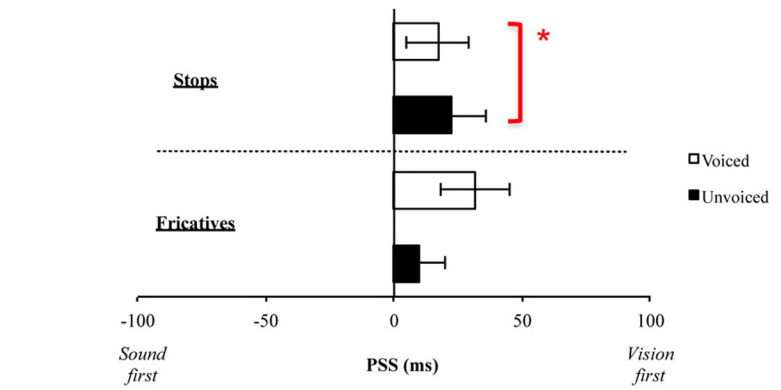


FIGURE 6 | (A) Average JNDs and **(B)** PSSs for the manner of articulation of the consonant stimuli presented in Experiment 2. The error bars represent the standard errors of the mean. Asterisks indicate significant differences between the various stimuli presented. **(C)** Average JNDs and **(D)** PSSs for the voicing of the stimuli presented in Experiment 1.

1996), thus supporting the "information reliability hypothesis" (e.g., Schwartz et al., 1998; Wada et al., 2003; Andersen et al., 2004; Traunmüller and Öhrström, 2007). The majority of previous research on speech perception has focused on the use of CV combinations as their main stimuli. In our third experiment, therefore, we further explored how physical differences in the articulation of vowels (in a non-consonant context) affect the temporal aspects of audiovisual speech perception. Here, we would expect that visual-speech should influence the JND data (in a similar way as that observed in Experiment 1) as a function of the roundedness of vowels, since this is the visually-dominant feature for vowels.

BACKNESS/ROUNDEDNESS AND HEIGHT FOR VOWELS (EXPERIMENT 3)

In our third and final experiment, we were interested in what role, if any, the backness/roundedness and height of articulation of vowels would play in the temporal aspects of audiovisual speech perception. The data were categorized according to the factors of Height (three levels: High, /i, u/; Mid, /ɛ, ɔ/; and Low, /æ, ɒ/) and Backness/Roundedness of articulation (two levels: front/unrounded, /i, ɛ, æ/ and back/rounded, /u, ɔ, ɒ/; see **Table 1C**).

Eleven new participants (eight female; native English speakers) aged 19–30 years (mean age of 23 years) took part in this experiment. Analysis of the JND data revealed in a significant main effect of Backness/ Roundedness [$F_{(1,\ 10)} = 4.75$, $p = 0.05$], with participants' being significantly more sensitive to the temporal order of the audiovisual speech stimuli when judging back/rounded stimuli (M = 73 ms) as compared to front/unrounded stimuli (M = 89 ms; see **Figure 8A**). No significant main effect of Height was obtained [$F_{(2,\ 20)} < 1$, *n.s.*], nor any interaction between Height and Backness/Roundedness in vowels [$F_{(2,\ 20)} < 1$, *n.s.*]. A similar analysis of the PSS data revealed a significant main effect of Backness/Roundedness [$F_{(1,\ 10)} = 18.60$, $p < 0.01$], with larger auditory leads being required for rounded vowels articulated at the back of the tongue (M = 51 ms) than for unrounded vowels articulated at the front (M = 12 ms; see **Figure 8B**). The large auditory leads observed for Roundedness agrees with research showing that the recognition of rounded stimuli is difficult for both automatic speech recognition systems, with the systems being blind to roundedness, and humans who recruit more subtle physical cues and possibly more complex operations along the auditory pathway in perceiving rounded vowels (e.g., Eulitz and Obleser, 2007).

The saliency analysis of the stimuli used in Experiment 3 (see **Figure 9**) showed a similar trend to that observed in Experiment 1. Specifically, the analysis revealed that, for back/rounded vowels, the saliency for both rise and peak was reached earlier for the visual stream and participants were better in their TOJ performance, the reverse pattern was observed for front/unrounded vowels. In terms of PSS, front/unrounded vowels were found to require large auditory leads with the saliency being noted earlier for the auditory stream (i.e., earlier auditory rise and peak) but was of lower magnitude (i.e., the highest magnitude was noted for the visual rise and peak). No specific trend was observed for height, a highly auditory feature (similar to Experiment 2).

Overall, the results of Experiment 3 replicate the patterns of JND and PSS results obtained in Experiments 1A–C and the PSS findings obtained in Experiments 2A–C. Specifically, larger auditory leads were observed for the highly salient rounded vowels as compared to the lower in saliency unrounded vowels (e.g., see Massaro and Cohen, 1993, for a comparison of /i/ and /u/ vowels and the /ui/ cluster; Traunmüller and Öhrström, 2007). Additionally, the participants were also more sensitive to the temporal order of the rounded vowels as compared to the unrounded vowels. It should, however, be noted that differences in the sensitivity to temporal order were only found as a function of roundedness/backness, while no such differences were observed as a function of the height of the tongue positions, which happens to be a highly auditory-dominant feature. The fact that auditory leads were required for all of the vowels tested here is consistent with similar findings reported previously by Vatakis and Spence (2006a).

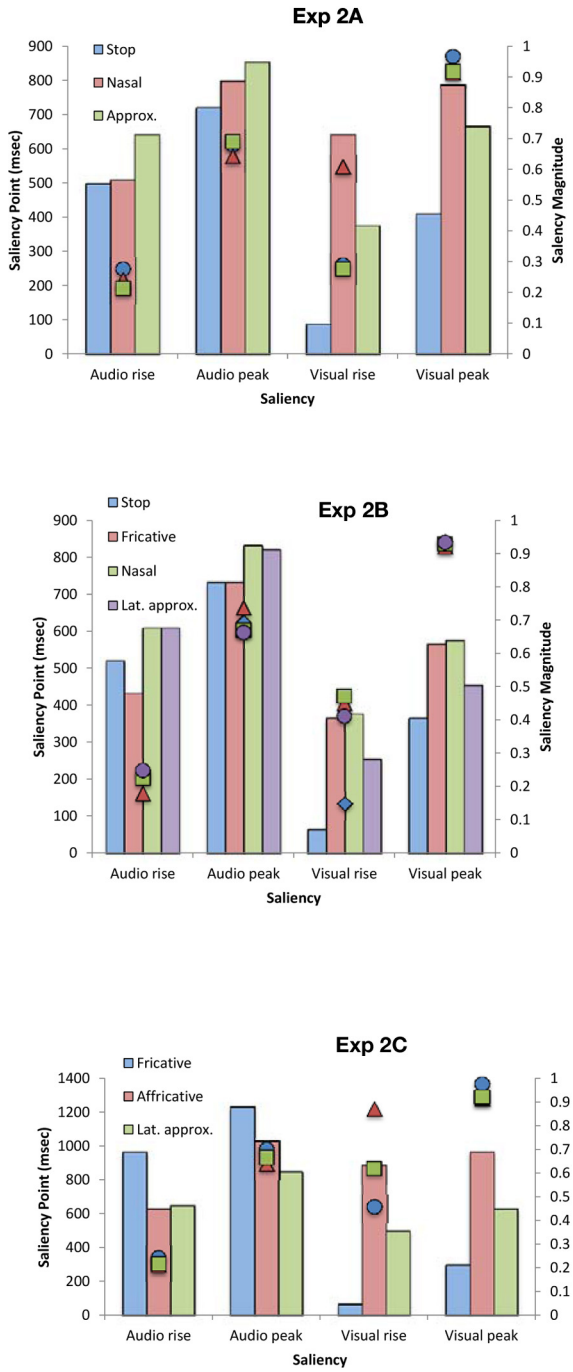


FIGURE 7 | Average saliency rise and peak (in ms) and saliency magnitude for each point for the audiovisual speech stimuli used in Experiments 2A–C as a function of the manner of articulation.

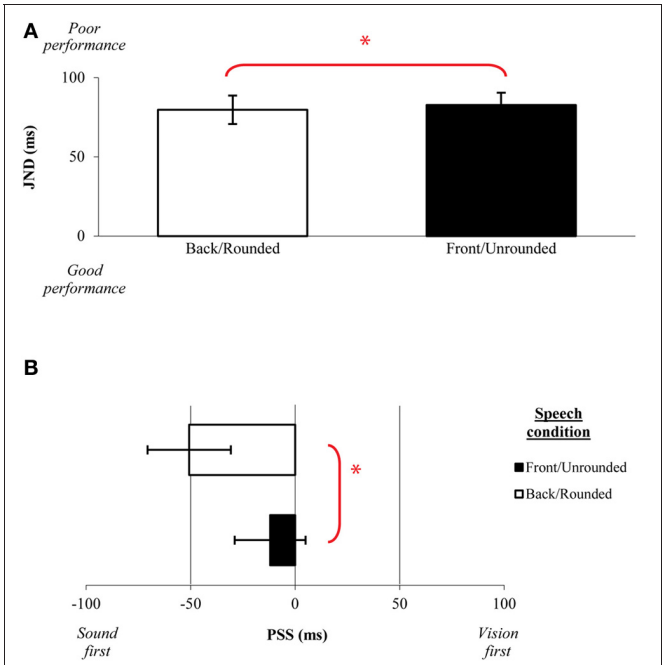


FIGURE 8 | Average (A) JNDs and (B) PSSs for the backness/roundedness of the vowel stimuli presented in Experiment 3. The error bars represent the standard errors of the mean. Asterisks indicate significant differences between the various stimuli presented.

GENERAL DISCUSSION

The three set of experiments reported in the present study provide empirical evidence regarding how physical differences in the articulation of different speech stimuli can affect audiovisual temporal perception utilizing a range of different consonant and vowel stimuli. The speech stimuli used here were compatible (i.e., both the visual-speech and auditory-speech referred to the same speech event). This contrasts with the large number of previous studies of the relative contribution of audition and vision to speech perception that have utilized incompatible speech signals (as in the McGurk effect; McGurk and MacDonald, 1976). The stimuli were also presented in the absence of any acoustic noise. This was done in order to explore how participants weight differently the auditory and visual information in speech given that surely the system weights the reliability of the modality information even under quiet settings (e.g., Andersen et al., 2004). Additionally, we utilized speech stimuli from three different speakers, while the majority of previous studies have used different tokens uttered by the same speaker (e.g., see Conrey and Gold, 2000, for a discussion of this point). The use of different speakers strengthens the present study since it takes account of the possible variability that may be present during the articulation of speech tokens by different individuals. Additionally, an audiovisual saliency analysis of the stimuli was conducted in order to make comparisons between the physical signal data and the behavioral data collected. Taken together, the results of the experiments reported here demonstrate (but see Maier et al., 2011 for different control of stimulus synchronous presentation) that the onset of the visual-speech signal had to precede that of the onset of the auditory-speech for the PSS to be reached for all the consonant stimuli tested (see Lebib et al., 2003; Van Wassenhove et al., 2003, 2005).

We hypothesize that the results of the present study show evidence that integration is being dominated by the modality stream

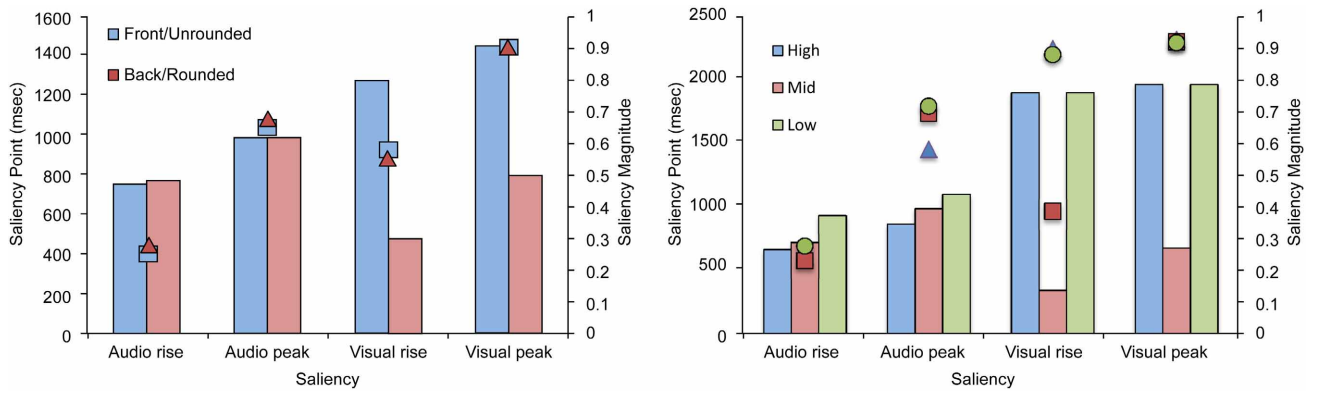


FIGURE 9 | Average saliency rise and peak (in ms) and saliency magnitude for each point for the audiovisual speech stimuli used in Experiment 3 as a function of the roundedness and height.

that provides the more salient information (e.g., place vs. manner of articulation of consonants; Schwartz et al., 1998; Wada et al., 2003). Our results also support the idea that the degree of saliency of the visual-speech signal can modulate the visual lead required for the two stimuli to be perceived as simultaneous. That is, the more visible (i.e., greater in saliency magnitude) the visual signal, the smaller the visual lead that is required for the PSS to be reached. These findings accord well with (Van Wassenhove et al.'s, 2005, p. 1183) statement that "… the more salient and predictable the visual input, the more the auditory processing is facilitated (or, the more visual and auditory information are redundant, the more facilitated auditory processing."

Visual speech signals represent a valuable source of input for audiovisual speech perception (i.e., McGrath and Summerfield, 1985; Dodd and Campbell, 1987) that can influence the acoustic perception of speech in both noisy and quiet conditions (e.g., Dodd, 1977; Calvert et al., 1997; Barker et al., 1998; Arnold and Hill, 2001; Girin et al., 2001; Möttönen et al., 2002). The visual input can also reduce the temporal and spectral uncertainty of the speech signal by directing auditory attention to the speech signal (Grant and Seitz, 2000), and can, in certain cases, serve as a cue that facilitates the listener's ability to make predictions about the upcoming speech sound and assist in the successful extraction of the relevant auditory signal (see Barker et al., 1998; Van Wassenhove et al., 2003, 2005). The idea that the visual signal serves as a cue that may help to identify the auditory signal is supported by the results of Experiments 1 and 3, where the visual signal had to lead the auditory signal (even for the cases of manner of articulation and voicing where the auditory input has a dominance over visual input; Massaro and Cohen, 1993; Girin et al., 2001; Van Wassenhove et al., 2005) for synchrony to be perceived depending on the degree of saliency of the speech stimulus presented.

The complementarity of vision and audition in the case of speech perception is more evident in those cases where the phonetic elements that are less robust in the auditory domain (in the presence of auditory noise) are the ones that are the most salient in the visual domain (i.e., Binnie et al., 1974; Summerfield, 1979, 1983, 1987; Grant et al., 1985; Robert-Ribes et al., 1998; De Gelder and Bertelson, 2003). It appears that those speech features that are hardest to discern on the basis of their auditory input benefit most from the addition of the visual inputs and vice versa. According to our results, highly salient speech contrasts (such as bilabial stimuli) lead to relatively shorter processing latencies for the speech signal, while lower in saliency (i.e., less visible) visual inputs lead to longer processing latencies. These findings are supported by the results of imaging studies reported by Van Wassenhove et al. (2003, 2005). There it was argued that salient visual inputs (as in /pa/) affect auditory speech processing (at very early stages of processing: i.e., within 50–100 ms of stimulus onset) by enabling observers to make a prediction concerning the about-to-be-presented auditory input. Additional support for this conclusion comes from the results of a study by Grant and Greenberg (2001) in which the introduction of even small auditory leads (of as little as 40 ms) in the audiovisual speech signal resulted in a significant decline in speech intelligibility while intelligibility remained high when the visual signal led by as much as 200 ms.

Previous research on the topic of audiovisual synchrony perception has demonstrated that the human perceptual system can recalibrate to the temporal discrepancies introduced between auditory and visual signals and that this recalibration appears to vary as a function of the type of stimuli being presented (i.e., Navarra et al., 2005; Vatakis and Spence, 2007). It has been shown that when people are presented with simple transitory stimuli (such as, light flashes and sound bursts) smaller discrepancies between the temporal order of the two signals can be perceived (e.g., Hirsh and Sherrick, 1961; Zampini et al., 2003), as compared to more complex events (such as speech, object actions, or musical stimuli) where audiovisual asynchrony appears to be harder to detect (e.g., Dixon and Spitz, 1980; Grant et al., 2004; Navarra et al., 2005; Vatakis and Spence, 2006a,b). For instance, studies using simple audiovisual stimuli (such as, sound bursts and light flashes) have typically shown that auditory and visual signals need to be separated by approximately 60–70 ms in order for participants to be able to accurately judge which sensory modality was presented first (e.g., Zampini et al., 2003). While studies using more complex

stimuli, such as audiovisual speech, have shown that the asynchrony of the audiovisual signals (i.e., visual- and auditory-speech) that can be tolerated can reach auditory leads of 100 ms or more, or auditory lags of at least 200 ms (e.g., Dixon and Spitz, 1980; Grant and Greenberg, 2001; Grant et al., 2004; Vatakis and Spence, 2006a,b, 2007). As discussed in the Introduction, the purported size of the temporal window of integration for audiovisual speech (a stimulus that is highly complex) exhibits great variability between published studies. The present findings highlight one important factor underlining this variability, which relates to the physical differences that are naturally present in the articulation of different consonants and vowels. The results of this study show that visual-speech has to lead auditory-speech in order for the two to be judged as synchronous, and the fact that larger visual lead times were required for lower saliency visual-speech signals, could provide one possible account for the human perceptual system's higher tolerance to asynchrony for the case of speech as compared to for simpler stimuli.

Overall, therefore, the results of the three sets of experiments reported here replicate previous findings that visual speech signals typically precede the onset of the speech sound signal in audiovisual speech perception (e.g., Munhall et al., 1996). In addition, our findings also extend previous research by showing that this precedence of the visual signal changes as a function of the physical characteristics in the articulation of the visual signal. That is, highly-salient visual-speech signals require less of a lead over auditory signals than visual-speech signals that are lower in saliency. Finally, our results support the analysis-by-synthesis model, whereby the precedence of the visual signal leads the speech-processing system to form a prediction regarding the auditory signal. This prediction is directly dependent on the saliency of the visual signal, with higher saliency signals resulting in a better prediction of the auditory signal (e.g., Van Wassenhove et al., 2005). It would be interesting in future research to explore how coarticulation cues affect the temporal relationship between auditory- and visual-speech signals observed in this study, since the oral and extra-ocular movements of a particular speech token are known to change depending on the context in which they are uttered (e.g., from syllable to word; Abry et al., 1994). In closing, future studies should further explore the relationship between the physical characteristics of the audiovisual speech signal (as explored by Chandrasekaran et al. (2009), for labial speech stimuli and in this manuscript in terms of saliency) and the behavioral data obtained in terms of temporal synchrony.

ACKNOWLEDGMENTS

We would like to thank Polly Dalton and two students from the University of Oxford for their willingness to participate in the recordings of our video clips and their patience throughout this process. Argiro Vatakis was supported by a Newton Abraham Studentship from the Medical Sciences Division, University of Oxford. We would also like to thank G. Evangelopoulos and K. Rapantzikos for providing the audio- and visual-saliency computation software developed by Evangelopoulos et al. (2008).

REFERENCES

Abry, C., Cathiard, M.-A., Robert-Ribès, J., and Schwartz, J.-L. (1994). The coherence of speech in audio-visual integration. *Curr. Psychol. Cogn.* 13, 52–59.

Andersen, T. S., Tiippana, K., and Sams, M. (2004). Factors influencing audiovisual fission and fusion illusions. *Cogn. Brain Res.* 21, 301–308.

Arnal, L. H., Morillon, B., Kell, C., A., and Giraud, A.-L. (2009). Dual neural routing of visual facilitation in speech processing. *J. Neurosci.* 29, 13445–13453.

Arnold, P., and Hill, F. (2001). Bisensory augmentation: a speechreading advantage when speech is clearly audible and intact. *Br. J. Psychol.* 92, 339–355.

Barker, J. P., Berthommier, F., and Schwartz, J. L. (1998). "Is primitive AV coherence an aid to segment the scene?" in *Proceedings of the Workshop on Audio Visual Speech Processing,* (Sydney, Australia: Terrigal), December 4–6, 103–108.

Bernstein, L. E., Auer, E. T., and Moore, J. K. (2004). "Audiovisual speech binding: convergence or association?" in *The Handbook of Multisensory Processing*, eds G. A. Calvert, C. Spence, and B. E. Stein (Cambridge, MA: MIT Press), 203–223.

Binnie, C. A., Montgomery, A. A., and Jackson, P. L. (1974). Auditory and visual contributions to the perception of consonants. *J. Speech Hear. Sci.* 17, 619–630.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., Woodruff, P. W., Iversen, S. D., and David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science* 276, 593–596.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., and Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Comput. Biol.* 5:e1000436. doi: 10.1371/journal.pcbi.1000436

Conrey, B., and Gold, J. M. (2000). An ideal observer analysis of variability in visual-only speech. *Vision Res.* 46, 3243–3258.

Conrey, B., and Pisoni, D. B. (2003). "Detection of auditory-visual asynchrony in speech and nonspeech signals," in *Research on Spoken Language Processing Progress Report No. 26,* (Bloomington, IN: Speech Research Laboratory, Indiana University), 71–94.

Conrey, B., and Pisoni, D. B. (2006). Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. *J. Acoust. Soc. Am.* 119, 4065–4073.

Coren, S., Ward, L. M., and Enns, J. T. (2004). *Sensation and Perception, 6th Edn.* Fort Worth, TX: Harcourt Brace.

Cosi, P., and Caldognetto, M. (1996). "Lip and jaw movements for vowels and consonants: spatio-temporal characteristics and bimodal recognition applications," in *Speechreading by Humans and Machine: Models, Systems and Applications, NATO ASI Series, Series F: Computer and Systems Sciences*, Vol. 150, eds D. G. Storke and M. E. Henneke (Berlin: Springer-Verlag), 291–313.

Davis, C., and Kim, J. (2004). Audio-visual interactions with intact clearly audible speech. *Q. J. Exp. Psychol. A* 57, 1103–1121.

De Gelder, B., and Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends Cogn. Sci.* 7, 460–467.

Dixon, N. F., and Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception* 9, 719–721.

Dodd, B. (1977). The role of vision in the perception of speech. *Perception* 6, 31–40.

Dodd, B., and Campbell, R. (eds.). (1987). *Hearing by Eye: The Psychology of Lip-Reading*. Hillsdale, NJ: LEA.

Erber, N. P. (1975). Auditory-visual perception of speech. *J. Speech Hear. Sci.* 40, 481–492.

Eulitz, C., and Obleser, J. (2007). Perception of acoustically complex phonological features in vowels is reflected in the induced brain-magnetic activity. *Behav. Brain Funct.* 3, 26–35.

Evangelopoulos, G., and Maragos, P. (2006). Multiband modulation energy tracking for noisy speech detection. *IEEE Trans. Audio Speech Lang. Process.* 14, 2024–2038.

Evangelopoulos, G., Rapantzikos, K., Maragos, P., Avrithis, Y., and Potamianos, A. (2008). "Audiovisual attention modeling and salient event detection," in *Multimodal Processing and Interaction: Audio, Video, Text,* eds P. Maragos, A. Potamianos, and P. Gros, (Berlin, Heidelberg: Springer-Verlag), 179–199.

Finney, D. J. (1964). *Probit Analysis: Statistical Treatment of the Sigmoid Response Curve*. London,

UK: Cambridge University Press.

Girin, L., Schwartz, J. L., and Feng, G. (2001). Audiovisual enhancement of speech in noise. *J. Acoust. Soc. Am.* 109, 3007–3020.

Grant, K. W., Ardell, L. H., Kuhl, P. K., and Sparks, D. W. (1985). The contribution of fundamental frequency, amplitude envelope, and voicing duration cues to speechreading in normal-hearing subjects. *J. Acoust. Soc. Am.* 77, 671–677.

Grant, K. W., and Greenberg, S. (2001). "Speech intelligibility derived from asynchronous processing of auditory-visual speech information," in *Proceedings of the Workshop on Audio Visual Speech Processing,* (Denmark: Scheelsminde), September 7–9, 132–137.

Grant, K. W., and Seitz, P. F. (1998). "The use of visible speech cues (speechreading) for directing auditory attention: reducing temporal, and spectral uncertainty in auditory detection of spoken sentences," in *Proceedings of the 16th International Congress on Acoustics and the 135th Meeting of the Acoustical Society of America*, Vol. 3, eds P. K. Kuhl and L. A. Crum, (New York, NY: ASA), 2335–2336.

Grant, K. W., and Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *J. Acoust. Soc. Am.* 108, 1197–1208.

Grant, K. W., van Wassenhove, V., and Poeppel, D. (2004). Detection of auditory (cross-spectral) and auditory-visual (cross-modal) synchrony. *Speech Commun.* 44, 43–53.

Hirsh, I. J., and Sherrick, C. E. Jr. (1961). Perceived order in different sense modalities. *J. Exp. Psychol.* 62, 423–432.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 1254–1259.

Jones, J. A., and Jarick, M. (2006). Multisensory integration of speech signals: the relationship between space and time. *Exp. Brain Res.* 174, 588–594.

Kandel, E., Schwartz, J., and Jessell, T. (2000). *Principles of Neural Science*, *4th Edn*. New York, NY: McGraw-Hill.

Kayser, C., Petkov, C., Lippert, M., and Logothetis, N. (2005). Mechanisms for allocating auditory attention: an auditory saliency map. *Curr. Biol.* 15, 1943–1947.

Kent, R. D. (1997). *The Speech Sciences*. San Diego, CA: Singular.

Koch, C., and Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.* 4, 219–227.

Lebib, R., Papo, D., de Bode, S., and Baudonniere, P. M. (2003). Evidence of a visual-to-auditory cross-modal sensory gating phenomenon as reflected by the human P50 event-related brain potential modulation. *Neurosci. Lett.* 341, 185–188.

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.

Maier, J., X., Di Luca, M., and Noppeney, U. (2011). Audiovisual asynchrony detection in human speech. *J. Exp. Psychol. Hum. Percept. Perform.* 37, 245–256.

Massaro, D. W., and Cohen, M. M. (1993). Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. *Speech Commun.* 13, 127–134.

Massaro, D. W., Cohen, M. M., and Smeele, P. M. T. (1996). Perception of asynchronous and conflicting visual and auditory speech. *J. Acoust. Soc. Am.* 100, 1777–1786.

Mattys, S. L., Bernstein, L. E., and Auer, E. T. Jr. (2002). Stimulus-based lexical distinctiveness as a general word-recognition mechanism. *Percept. Psychophys.* 64, 667–679.

McGrath, M., and Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *J. Acoust. Soc. Am.* 77, 678–685.

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748.

Miller, G. A., and Nicely, N. (1955). An analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am.* 27, 338–352.

Mills, A. E., and Thiem, R. (1980). Auditory-visual fusions and illusions in speech perception. *Linguistische Berichte* 68, 85–108.

Miner, N., and Caudell, T. (1998). Computational requirements and synchronization issues of virtual acoustic displays. *Presence Teleop. Virt. Environ.* 7, 396–409.

Möttönen, R., Krause, C. M., Tiippana, K., and Sams, M. (2002). Processing of changes in visual speech in the human auditory cortex. *Cogn. Brain Res.* 13, 417–425.

Munhall, K. G., Gribble, P., Sacco, L., and Ward, M. (1996). Temporal constraints on the McGurk effect. *Percept. Psychophys.* 58, 351–362.

Munhall, K., and Vatikiotis-Bateson, E. (2004). "Specialized spatiotemporal integration constraints of speech," in *The Handbook of Multisensory Processing,* eds G. Calvert, C. Spence, and B. E. Stein (Cambridge, MA: MIT Press), 177–188.

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., and Spence, C. (2005). Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Cogn. Brain Res.* 25, 499–507.

Pandey, C. P., Kunov, H., and Abel, M. S. (1986). Disruptive effects of auditory signal delay on speech perception with lip-reading. *J. Aud. Res.* 26, 27–41.

Rapantzikos, K., Tsapatsoulis, N., Avrithis, Y., and Kollias, S. (2009). Spatiotemporal saliency for video classification. *Sig. Process. Image Commun.* 24, 557–571.

Reisberg, D., McLean, J., and Goldfield, A. (1987). "Easy to hear but hard to understand: a lip-reading advantage with intact auditory stimuli," in *Hearing by Eye: The Psychology of Lipreading,* eds B. Dodd and R. Campbell (London, UK: Erlbaum Associates), 97–114.

Robert-Ribes, J., Schwartz, J. L., Lallouache, T., and Escudier, E. (1998). Complementarity and synergy in bimodal speech: auditory, visual, and audio-visual identification of French oral vowels in noise. *J. Acoust. Soc. Am.* 103, 3677–3688.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., and Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cereb. Cortex* 17, 1147–1153.

Sams, M., Aulanko, R., Hamalainen, M., Hari, R., Lounasmaa, O. V., Lu, S. T., and Simola, J. (1991). Seeing speech: visual information from lip movements modifies the activity in the human auditory cortex. *Neurosci. Lett.* 127, 141–145.

Schwartz, J.-L., Robert-Ribes, J., and Escudier, P. (1998). "Ten years after Summerfield: a taxonomy of models for audio-visual fusion in speech perception," in *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech,* ed D. Burnham (Hove, UK: Psychology Press), 85–108.

Spence, C., Shore, D. I., and Klein, R. M. (2001). Multisensory prior entry. *J. Exp. Psychol. Gen.* 130, 799–832.

Spence, C., and Squire, S. B. (2003). Multisensory integration: maintaining the perception of synchrony. *Curr. Biol.* 13, R519–R521.

Steinmetz, R. (1996). Human perception of jitter and media synchronization. *IEEE J. Sel. Areas Commun.* 14, 61–72.

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215.

Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetica* 36, 314–331.

Summerfield, Q. (1983). "Audio-visual speech perception, lipreading and artificial stimulation," in *Hearing Science and Hearing Disorders,* eds M. E. Lutman and M. P. Haggard (London, UK: Academic), 131–182.

Summerfield, Q. (1987). "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by Eye: The Psychology of Lip-Reading,* eds B. Dodd and R. Campbell (Hillsdale, NJ: Lawrence Erlbaum Associates), 3–51.

Summerfield, Q., and McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *Q. J. Exp. Psychol.* 36, 51–74.

Traunmüller, H., and Öhrström, N. (2007). Audiovisual perception of openness and lip rounding in front vowels. *J. Phon.* 35, 244–258.

Tsingos, N., Gallo, E., and Drettakis, G. (2004). "Perceptual audio rendering of complex virtual environments," in *Proceedings of the SIGGRAPH2004.* (Los Angeles, CA), August 8–12.

Van Wassenhove, V., Grant, K. W., and Poeppel, D. (2003). "Electrophysiology of auditory-visual speech integration," in *Proceedings of the Workshop on Audio Visual Speech Processing.* (St. Jorioz, France), September 31–35, 37–42.

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.

Van Wassenhove, V., Grant, K. W., and Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia* 45, 598–607.

Vatakis, A., and Spence, C. (2006a). Audiovisual synchrony perception for speech and music using a temporal order judgment task. *Neurosci. Lett.* 393, 40–44.

Vatakis, A., and Spence, C. (2006b). Audiovisual synchrony perception for music, speech, and object actions. *Brain Res.* 1111, 134–142.

Vatakis, A., and Spence, C. (2006c). Evaluating the influence of frame rate on the temporal aspects of audiovisual speech perception. *Neurosci. Lett.* 405, 132–136.

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

Vatakis, A., and Spence, C. (2008). Investigating the effects of inversion on configural processing using an audiovisual temporal order judgment task. *Perception* 37, 143–160.

Vatakis, A., and Spence, C. (2010). "Audiovisual temporal integration for complex speech, object-action, animal call, and musical stimuli," in *Multisensory Object Perception in the Primate Brain,* eds M. J. Naumer, and J. Kaiser (New York, NY: Springer), 95–121.

Wada, Y., Kitagawa, N., and Noguchi, K. (2003). Audio-visual integration in temporal perception. *Int. J. Psychophys.* 50, 117–124.

Zampini, M., Shore, D. I., and Spence, C. (2003). Audiovisual temporal order judgments. *Exp. Brain Res.* 152, 198–210.

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 March 2012; accepted: 22 August 2012; published online: 01 October 2012.

Citation: Vatakis A, Maragos P, Rodomagoulakis I and Spence C (2012) Assessing the effect of physical differences in the articulation of consonants and vowels on audiovisual temporal perception. Front. Integr. Neurosci. ***6****:71. doi: 10.3389/fnint.2012.00071*

frontiers in
INTEGRATIVE NEUROSCIENCE

METHODS ARTICLE
published: 20 January 2012
doi: 10.3389/fnint.2012.00001

A novel, variable angle guide grid for neuronal activity studies

Thomas Talbot[1]*, David Ide[2], Ning Liu[3] and Janita Turchi[4]

[1] Laboratory of Cellular and Synaptic Neurophysiology, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD, USA
[2] Section on Instrumentation, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA
[3] Section on Neurocircuitry, Laboratory of Brain and Cognition, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA
[4] Section on Cognitive Neuroscience, Laboratory of Neuropsychology, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA

Edited by:
Hermann J. Mueller, University of Munich, Germany

Reviewed by:
Toemme Noesselt, Otto-von-Guericke-University, Germany
Patrizia Fattori, University of Bologna, Italy

***Correspondence:**
Thomas Talbot, Laboratory of Cellular and Synaptic Neurophysiology, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, 13 South Drive Room G360, Bethesda, MD 20892-5712, USA. e-mail: talbott@mail.nih.gov

Introduction: Surgically implanted chambers with removable grids are routinely used for studying patterns of neuronal activity in primate brains; however, accessing target tissues is significantly constrained by standard grid designs. Typically, grids are configured with a series of guide holes drilled vertically, parallel to the walls of the chamber, thus targeted sites are limited to those in line vertically with one of the guide holes. **Methods:** By using the three-dimensional modeling software, a novel grid was designed to reach the targeted sites far beyond the standard reach of the chamber. The grid was fabricated using conventional machining techniques and three-dimensional printing. **Results:** A pilot study involving microinjection of the magnetic resonance (MR) contrast agent gadolinium into the discrete regions of interest (ROIs) in the temporal cortex of an awake, behaving monkey demonstrated the effectiveness of this new design of the guide grid. Using multiple different angles of approach, we were readily able to access 10 injection sites, which were up to 5 mm outside the traditional, orthogonal reach of the chamber.

Keywords: neuronal activity, recording chamber, guide grid, solid modeling, three-dimensional printing

INTRODUCTION

In order to examine a wide range of neuronal functions *in vivo*, several techniques have emerged which allow for controlled manipulations, and measurements, within circumscribed regions of interest (ROIs), including (1) local delivery of pharmacological agents, and (2) mapping the electrophysiological responses of neurons to behaviorally relevant stimuli or neurochemical modulation, and (3) electrical stimulation of discrete populations of neurons (e.g., Dias and Segraves, 1999; Nichols and Newsome, 2002; Pickens et al., 2009; Eifuku et al., 2010; Watanabe and Munoz, 2010). Additionally, recent studies have used local injections of magnetic resonance (MR)-visible tracers (e.g., Mn^{2+}) to determine anatomical connectivity *in vivo* (e.g., Saleem et al., 2002; Simmons et al., 2008). All of these research approaches, when utilized in non-human primates, may require the implantation of a recording chamber that can contain a grid to guide the recording/microstimulation electrodes, infusion cannulae, injection needles, etc., into the brain ROI (e.g., Evarts, 1968).

Traditionally, a chamber is chronically affixed to the skull above the ROIs via surgical screws and acrylic dental cement. During the subsequent experimental sessions (e.g., electrophysiological recording/microstimulation/focal drug delivery), a removable grid, with holes parallel to the walls of the recording chamber, is used to guide one or multiple electrodes/cannulae through the craniotomy and the dura mater into the targeted tissue (see also Crist et al., 1988).

There are, however, several disadvantages to this traditional approach. First, on the skull, the space for placing the chamber is limited. This limited space may be further reduced by other associated mechanical attachments, such as the headpost. As a result, it may not be possible to place the chamber at the most advantageous position, nor to have a single chamber be as large as the study might dictate (e.g., covering both hemispheres and both medial and lateral target areas). Second, in studies that require reaching the outermost lateral regions of the brain, the chamber must be implanted in a vicinal region. This necessitates larger muscle retractions in order to place the chamber, which increases the risk of collateral damage to the animal (e.g., damaging the temporalis muscle). Third, as in the case where a vessel passes through the top of the ROI, if using a straight grid, the investigators must either take the risk of hitting the vessel and damaging the brain tissue, or relocate the cannula/electrode a sufficient number of grid holes away to avoid the vessel, but possibly missing the critical ROI; clearly both of these solutions are less than ideal.

Although there are some angled guide grids commercially available, most of them have only one specified angle, which may allow the investigator to reach a single ROI but at the same time possibly preclude reaching other ROIs (e.g., when reaching the different ROIs requires different angles, as might occur with bilateral ROIs). For such cases, the investigator would have to do the experiment in serial fashion, changing to differently angled grids to access each ROI, prohibitively protracting the overall experimental time, and rendering simultaneous study of multi-ROI neuronal activities impossible. Furthermore, a single angle grid actually only shifts the reach of the chamber but cannot increase it. Actually, it

will even decrease it in some cases because the wall of the chamber may block some grid holes from reaching the target.

In order to solve the problems described above, we pursued development of a novel angled guide grid system. This new guide grid system would permit each point of interest to be reached at the same time, irrespective of the angle(s) required to target that point. Additionally, it would permit a larger target area to be reached by the same size of chamber.

METHODS AND RESULTS

SUBJECTS AND GENERAL PROCEDURES

One adult male macaque monkey (*Macaca mulatta*, 9 years old, and 7 kg) was used. All experimental and surgical procedures followed the Institute of Laboratory Animal Research (part of the National Research Council of the National Academy of Sciences) guidelines and were approved by the National Institute of Mental Health (NIMH) Animal Care and Use Committee (ACUC). The monkey was anesthetized and surgically implanted, using aseptic techniques, with a plastic headpost and custom rectangular chamber (58×32 mm, made of Ultem®; see **Figure 1**).

After a two-week recovery period, we inserted the traditional straight guide grid ($52 \times 25 \times 10$ mm, made of Ultem®) into the chamber and filled the chamber with gadolinium (Magnevist, Berlex Pharmaceuticals; 1:1200 dilution in sterile saline, pH 7.0–7.5) to illuminate the grids holes in the MR images (**Figure 2**). Then, a high-resolution T1-weighted whole-brain anatomical scans (voxel size: $0.5 \times 0.5 \times 0.5\ mm^3$) was acquired to map the grid and the brain structure by using a 4.7T Bruker scanner with a Modified Driven Equilibrium Fourier Transform (MDEFT) sequence.

TARGETING THE ROIs

Five ROIs were selected from each hemisphere. Only the right side is shown for clarity (see **Figure 2A**). Using the AFNI (Analysis of Functional NeuroImages) software (Cox, 1996), we determined the corresponding grid hole that we should use if we wanted to use the straight grid (**Figure 2B**). The grid holes in red are those of the straight grid. As shown in this figure, the straight grid could reach only three targets, (yellow, orange, and red), inside of the bounding box that represented the perimeter of the straight grid.

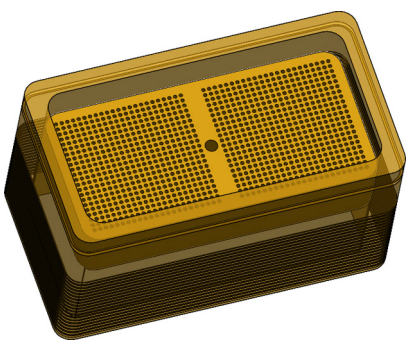


FIGURE 1 | Chamber with the removable straight grid insert. The base of the chamber will be shaped to the contours of the monkey's skull and cemented in place.

Obviously, a single angled grid could not reach all of the bilateral targets. Moreover, even a grid with two separate angles (one for each hemisphere) was not adequate to solve our problems because the angle in anterior-posterior direction would reduce the area we could reach due to intersection with the walls of the recording chamber such that there were several targets which remained unreachable (**Figure 3**). To overcome these limitations, a new type of grid was designed as follows.

SOLID MODELING

For the example presented here, 10 points were determined by depth, distance right to left, and distance anterior to posterior referenced to the top center of the guide grid. Each point was assigned a color to aid in identification. A three-dimensional model of the guide grid, recording chamber, and targets of interest was created using the software SolidWorks® (Dassault Systemes SolidWorks, Concord: MA). Projection axes were created between each of the 10 target points and corresponding points on the

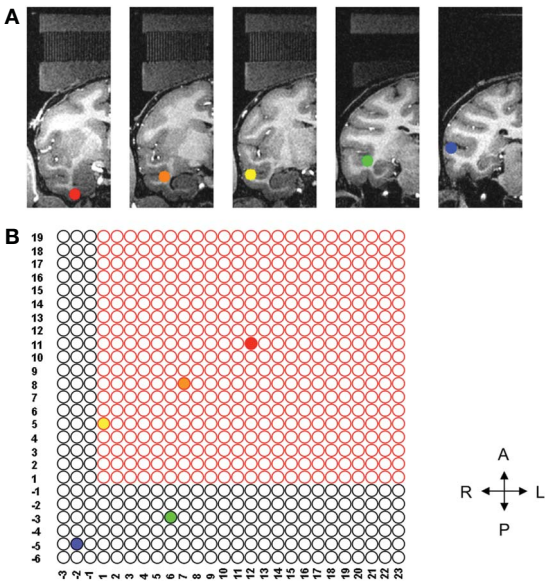


FIGURE 2 | Location of each of five ROIs and corresponding grid holes. **(A)** Five ROIs in the coronal plane, each of them marked by different color. **(B)** Illustrates the five sites needed for the study on the grid. Only the three sites inside the bounded box can be reached using the conventional straight grid; the remaining two sites are beyond the reach of the straight grid. The left hemisphere (not shown) also had three sites that could be reached with the straight grid and two sites out of reach of the straight grid.

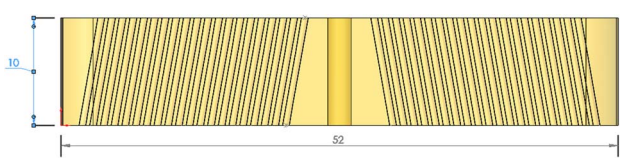


FIGURE 3 | Grid insert with two distinct angles. This scheme leaves a large trapezoidal section in the middle of the brain ROI unreachable. Dimensions in millimeter.

surface of the grid. Sketch planes were created perpendicular to each axis at the depth of the target points (**Figure 4**). Each mapped point served as the center of a 5 mm by 5 mm square sketch that was used to create an extruded cut up through the surface of the guide grid (**Figure 5**). By moving the positions of the corresponding points on the surface of the grid, the final position of each of the extruded cuts could be adjusted to ensure that the 10 different cuts did not interfere with each other or intersect the recording chamber. The guide grid was now comprised of a 52 mm by 25 mm by 10 mm block with ten 5 mm by 5 mm extruded cuts, and each of these extruded cuts was angled exactly toward the point of interest from which it was derived.

MASTER GRID FABRICATION

This varied angle guide grid was saved in the stereolithography ".STL" file format and printed using a Dimension Elite 3D printer (Stratasys, Eden Prairie: MN). This is a significant salient feature that rendered this method feasible as conventional machining of such a grid insert would have been prohibitively expensive. Three-dimensional printing is readily available and affordable as an outside service for those institutions that do not have their own printer.

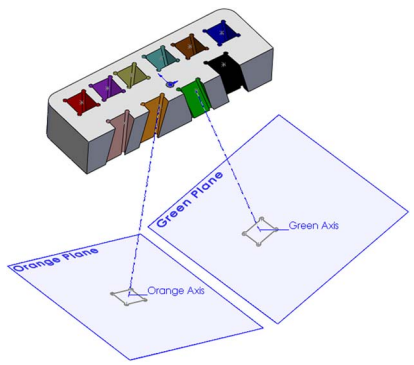


FIGURE 4 | Sectional view of the grid insert. A plane at each injection site was created perpendicular to an axis that connected the target site and a point on the surface of the grid. Each target site was assigned a color to distinguish them one from another. For clarity, only two are shown here.

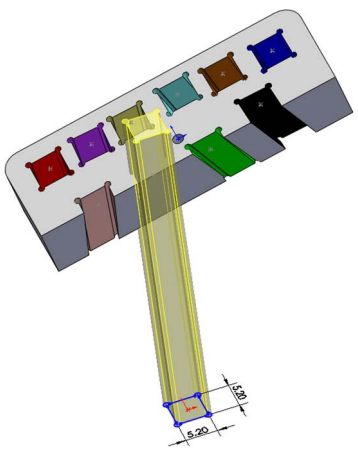


FIGURE 5 | The 5 × 5 mm sketch used to project an extruded cut up through the grid insert. The sketch was slightly oversized (5.2 × 5.2 mm), and circular grooves at the corners were added to facilitate introduction of adhesive.

Small guide grids, 5 mm by 5 mm by 10 mm were fabricated from Ultem® (**Figure 6**). Each small guide grid had twenty-five 635 micron thru holes drilled on 1 mm centers. This size was decided upon to give the investigator some leeway to avoid anatomical obstructions, such as blood vessels, while still permitting a reasonable number of points to be reached by each varied angle guide grid. These small guide grids were cemented in flush with the bottom surface of the varied angle guide grid (**Figure 6**). This new grid system provided a $5 \times 5\ mm^2$ area for each cannula or electrode insertion site that guaranteed all 10 targets could be reached with one master grid placed in a single chamber.

PILOT STUDY RESULTS

The goal of our initial study was to assess targeting accuracy through this new type of grid. As described in Sections "Solid Modeling and Master Grid Fabrication" above, a three-dimensional model was designed, based on AFNI mapping of target regions, for the master varied angled grid with ten individual square cutouts to house the individual small Ultem® guide grids. Once this master grid was fabricated, and the small guide grids had been inserted, we placed the grid system into the chamber and filled the chamber with gadolinium (1:1200 dilution in sterile saline) to illuminate the grids holes in the MR images (**Figure 7A**). We again acquired high-resolution T1-weighted whole-brain anatomical scans (voxel size: $0.5 \times 0.5 \times 0.5\ mm^3$). Using AFNI software, the ideal grid holes, as well as the lengths of the guide and infusion cannulae, required to reach the target sites, were established. Based on the calculation, all the 10 ROIs can be reached by using this new grid. These theoretical calculations were verified by the following experiment. Following aseptic protocol, a sterile field was prepared and using sterile gastight Hamilton syringes, mounted on the microliter rack of an infusion pump (Harvard Apparatus) and connected to the infusion cannula (30 gauge stainless steel). After inserting guide tubes (24 gauge stainless steel) into the intended grid holes, the infusion cannulae were lowered into place and we made microinfusions (2.7–3.75 μl/site, at rates ranging from 0.18 to 0.25 μl/min) of the MR contrast agent gadolinium (1:75 dilution in sterile saline, pH 7.0–7.5) into the circumscribed ROIs in the temporal cortex of the awake monkey prior to MR scan sessions. Based on the T1-weighted scans collected following these injection sessions, we were successfully able to inject the gadolinium solution into each site in an efficacious manner, demonstrating the viability of this new targeting technique (see **Figure 7B**).

DISCUSSION

In the present study, a new type of 3D printed grid insert system, which is capable of accommodating multiple angles simultaneously, was designed and proven to be a productive scheme for expanding the reach of the electrodes and/or injection cannulae aimed at sites deep in the brain. This type of grid insert system would have applications in a variety of experiments (see target areas in Hernández et al., 2010; Maior et al., 2010;

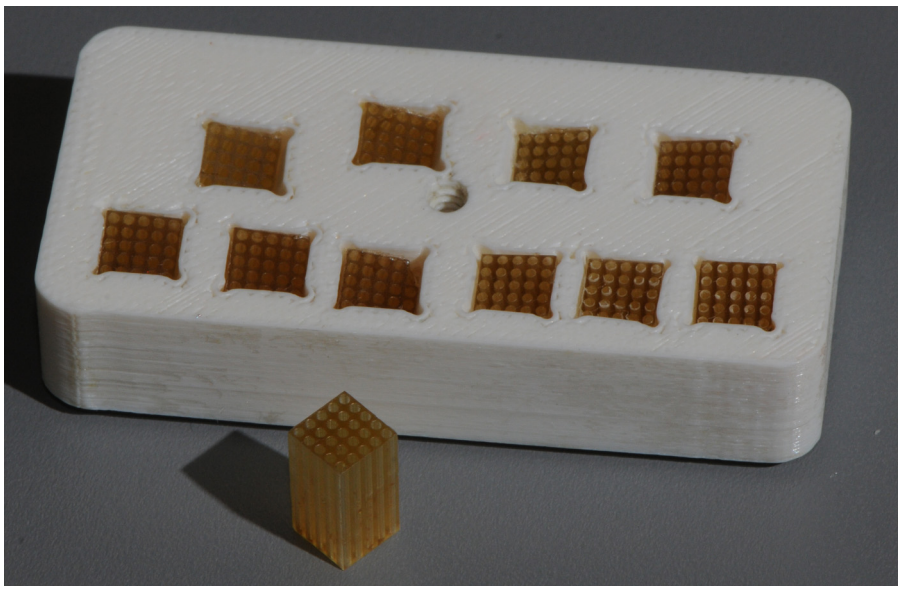


FIGURE 6 | The 5 × 5 × 10 mm insert. Each insert contained 25 holes (diameter: 635 micron; 1 mm on center). Final master grid with 10 inserts each aimed at a specific ROI.

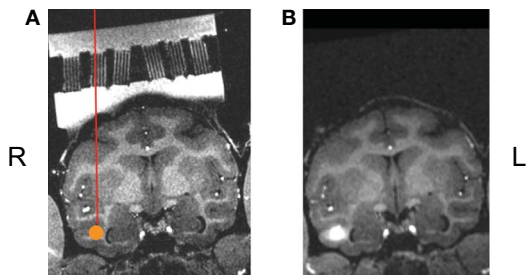


FIGURE 7 | (A) One ROI, shown here in the coronal plane, rotated into the coordinate system of the guide grid, the ROI marked as an orange disk. **(B)** The actual injection results made though this angled grid.

Vallentin and Nieder, 2010), allowing concurrent access to several ROIs, including bilateral targeting, from a single-stage surgical preparation.

ALTERNATIVES AND CONSIDERATIONS

An equally valid but different approach to the problem of simultaneous multi-angle targeting of multiple brain regions would be the use of an array of permanent indwelling individual guide cannulae (as available through Plastics One, Roanoke: VA, for instance). Such cannulae can be obtained in MR compatible materials (fused silica), allowing for anatomical scanning to validate target acquisition, as well as functional imaging experiments. One caveat, however, is that this method is far less flexible than a grid based system, as modification of the target acquisition may require additional intervals for placement surgeries. Both methods could be combined effectively of course, for targeting of tissues by trajectories originating from more caudal or lateral points. For example, such combination would retain the capacity for flexible, simultaneous bilateral targeting of ROIs in frontal as well as temporal cortices via this new multi-angle grid system, while hippocampal tissue could be approached longitudinally (see Hampton et al., 2004), via chronic indwelling guide cannulae.

FUTURE DEVELOPMENTS

Future work will test the feasibility of adapting this technique for use with recording electrodes. This will involve modification of a microdrive (Nichols et al., 1998) to control the placement of the electrode to conform to the appropriate angle. The device and techniques described in the current study provide a useful method to reach injection sites well beyond of the vertical reach of a single recording chamber. This multi-angle guide grid system takes advantage of the capabilities offered by the Dimension Elite 3D printer (Stratasys, Eden Prairie: MN) and is an affordable alternative to conventional machining options. These factors combine to maximize flexibility during the course of an experiment, allowing issues that might arise, such as vascular obstacles, or gliosis, to be circumvented by changing an angle of approach and printing a new guide grid to achieve placements in the target ROI. Such flexibility would be beneficial for target trajectories passing near ventricles (see Bouret and Richmond, 2009; Maior et al., 2010), bilateral targeting (see Dunn and Colby, 2010), and regions with rich vascular challenges (see Singer and Sheinberg, 2010; Hayden et al., 2011).

ACKNOWLEDGMENTS

We thank the NIF facility including Dr. Frank Ye and Charles Zhu for assistance with scanning. This project was fully funded by the Intramural Research Program of NIMH, NICHD, and NINDS/NIH/DHHS.

REFERENCES

Bouret, S., and Richmond, B. J. (2009). Relation of locus coeruleus neurons in monkeys to Pavlovian and operant behaviors. *J. Neurophysiol.* 101, 898–911.

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* 29, 162–173.

Crist, C. F., Yamasaki, D. S. G., Komatsu, H., and Wurtz, R. H. (1988). A grid system and a microsyringe for single cell recording. *J. Neurosci. Methods* 26, 117–122.

Dias, E. C., and Segraves, M. A. (1999). Muscimol-induced inactivation of monkey frontal eye field: effects on visually and memory-guided saccades. *J. Neurophysiol.* 81, 2191–2214.

Dunn, C. A., and Colby, C. L. (2010). Representation of the ipsilateral visual field by neurons in the macaque lateral intraparietal cortex depends on the forebrain commissures. *J. Neurophysiol.* 104, 2624–2633.

Eifuku, S., Nakata, R., Sugimori, M., Ono, T., and Tamura, R. (2010). Neural correlates of associative face memory in the anterior inferior temporal cortex of monkeys. *J. Neurosci.* 30, 15085–15096.

Evarts, E. V. (1968). A technique for recording activity of subcortical neurons in moving animals. *Electroencephalogr. Clin. Neurophysiol.* 24, 83–86.

Hampton, R. R., Buckmaster, C. A., Anuszkiewicz-Lundgren, D., and Murray, E. A. (2004). Method for making selective lesions of the hippocampus in macaque monkeys using NMDA and a longitudinal surgical approach. *Hippocampus* 14, 9–18.

Hayden, B. Y., Heilbronner, S. R., Pearson, J. M., and Platt, M. L. (2011). Surprise signals in anterior cingulated cortex: neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *J. Neurosci.* 31, 4178–4187.

Hernández, A., Nácher, V., Luna, R., Zainos, A., Lemus, L., Alvarez, M., Vázquez, Y., Camarillo, L., and Romo, R. (2010). Decoding a perceptual decision process across cortex. *Neuron* 66, 300–314.

Maior, R. S., Hori, E., Tomaz, C., Ono, T., and Nishijo, H. (2010). The monkey pulvinar neurons differentially respond to emotional expressions of human faces. *Behav. Brain Res.* 215, 129–135.

Nichols, M. J., and Newsome, W. T. (2002). Middle temporal visual area microstimulation influences veridical judgments of motion direction. *J. Neurosci.* 22, 9530–9540.

Nichols, A. H., Ruffner, T. W., Sommer, M. A., and Wurtz, R. H. (1998). A screw microdrive for adjustable chronic unit recording in monkeys. *J. Neurosci. Methods* 81, 185–188.

Pickens, C. L., Adams-Deutsch, T., Nair, S. G., Navarre, B. M., Heilig, M., and Shaham, Y. (2009). Effect of pharmacological manipulations of neuropeptide Y and corticotropin-releasing factor neurotransmission on incubation of conditioned fear. *Neuroscience* 164, 1398–1406.

Saleem, K. S., Pauls, J. M., Augath, M., Trinath, T., Prause, B. A., Hashikawa, T., and Logothetis, N. K. (2002). Magnetic resonance imaging of neuronal connections in the macaque monkey. *Neuron* 34, 685–700.

Simmons, J. M., Saad, Z. S., Lizak, M. J., Ortiz, M., Koretsky, A. P., and Richmond, B. J. (2008). Mapping prefrontal circuits *in vivo* with manganese-enhanced magnetic resonance imaging in monkeys. *J. Neurosci.* 28, 7637–7647.

Singer, J. M., and Sheinberg, D. L. (2010). Temporal cortex neurons encode articulated actions as slow sequences of integrated poses. *J. Neurosci.* 30, 3133–3145.

Vallentin, D., and Nieder, A. (2010). Representations of visual proportions in the primate posterior parietal and prefrontal cortices. *Eur. J. Neurosci.* 32, 1380–1387.

Watanabe, M., and Munoz, D. P. (2010). Saccade suppression by electrical microstimulation in monkey caudate nucleus. *J. Neurosci.* 30, 2700–2709.

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: 31 October 2011; paper pending published: 04 December 2011; accepted: 05 January 2012; published online: 20 January 2012.

Citation: Talbot T, Ide D, Liu N and Turchi J (2012) A novel, variable angle guide grid for neuronal activity studies. Front. Integr. Neurosci. ***6****:1. doi: 10.3389/fnint.2012.00001*