Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Behavioral Sciences

Multisensory binding: is the contribution of synchrony and semantic congruency obligatory?

Efthymios Tsilionis[1,2] and Argiro Vatakis[1,3]

We perceive the world as a unified whole with multisensory events being 'aligned' in every possible sense. This 'aligned' sense is a complex orchestration of multiple factors and underlying mechanisms, here we focus on two: synchrony and semantic (or informational) congruency. These factors, the former structural and the latter cognitive, appear to favor the binding of multisensory stimuli, leading in a coherent unified percept. Furthermore, the strong binding of the senses affects our perception of synchrony by making us tolerant to large temporal discrepancies between the input sensory streams. A longstanding debate in the field concerns the contribution of low- and high-level factors in the merging operation (i.e., unity assumption). Recent neuroimaging studies propose the existence of a brain network responsible for multisensory integration, consisting of frontal, temporal, and primary sensory areas, each responding to different stimulus properties. Converging evidence suggests the dissociation of integration and synchrony perception, which is consistent with the view that these processes entail distinct mechanisms, both anatomically and functionally.

Addresses
[1] Postgraduate Program Basic and Applied Cognitive Science, Department of Philosophy and History of Science, University of Athens, Athens, Greece
[2] National Technical University of Athens, Athens, Greece
[3] Cognitive Systems Research Institute, Athens, Greece

Corresponding author: Vatakis, Argiro (argiro.vatakis@gmail.com)

Current Opinion in Behavioral Sciences 2016, **8**:7–13

This review comes from a themed issue on **Time in perception and action**

Edited by **Warren H Meck** and **Richard B Ivry**

For a complete overview see the Issue and the Editorial

Available online 21st January 2016

http://dx.doi.org/10.1016/j.cobeha.2016.01.002

Introduction

In our everyday interaction, we experience a multitude of events, the majority of which are multisensory in nature. Our sensory channels receive modality-specific information from the environment and the brain must process these inputs in order to form a coherent percept. The merging of these sensory inputs, known as multisensory integration, has been repeatedly shown to provide significant advantages (e.g., faster and more accurate detection) over unisensory processing (e.g., [1–3]) and has been the subject of vigorous behavioral and neuroscientific research for over 60 years now.

One longstanding question in multisensory perception relates to how the sensory systems cooperate and achieve this unified representation of the world. At the neural level, physiological studies involving a number of species and brain regions have illustrated three important principles that govern the function of multisensory neurons (e.g., [4,5]). The spatial principle, where two input modalities produce an enhanced neuronal activation given that their spatial point-of-origin is enclosed by overlapping receptive fields. The temporal principle, where an increase in neuronal activity can be achieved when two stimuli occur in close temporal proximity, even if this inter-stimulus temporal distance is several hundred of milliseconds apart. The third and final principle is that of inverse effectiveness, where low saliency stimuli together cause a superadditive neuronal response as compared to the individual responses. These principles, although they deepen our understanding of neuronal processing, cannot explain why our perceptual system 'chooses' to bind particular unimodal events when presented close in time and/or space.

The aforementioned principles and the whole notion of integration have also been the subject of many behavioral investigations. A series of studies have shown that integration depends on numerous low- and high-level factors [6–12]. Low-level structural factors refer to temporal synchrony and spatial location, as well as any temporal correlation between the signal modalities (see [13] for a review), while high-level cognitive factors refer to prior knowledge or top-down control (i.e., both in terms of mechanisms influencing perceptual processes and constraints imposing a unified structure) and include semantic congruency (see [14] for a review), perceptual grouping, and phenomenal causality ([15]; see [7] for a review). A theory combining these factors is the so called 'unity assumption' [8–12,16], which proposes that when two stimuli share many common amodal properties, an observer is more likely to perceive them as referring to the same multisensory event rather than multiple unisensory events. The dissociation between structural and cognitive factors, however, is sometimes difficult given that many factors that promote a bottom-up multisensory integration (e.g., spatial-temporal coincidence) are also likely to result in a top-down assumption of unity [7]. Probably the most important structural property for multisensory binding is temporal (physical) coincidence and,

more generally, the perception of synchrony [16,17]. This review will focus on the latter in association with semantic (or informational) congruency.

Multisensory binding: a behavioral view

Semantic congruency in multisensory biding has been explored through the cross-modal illusion of spatial ventriloquism during which the perceived location of a sound is mislocalized toward the visual stimulation, provided that they are presented close in time. Jackson [2] showed that the illusion was evident over larger spatial disparities for realistic (i.e., viewing a steaming kettle and hearing the steam whistle of the kettle) than for artificial (i.e., viewing a LED flashing and hearing the sound of a bell) stimuli. However, Jackson's study along with other sound localization studies (e.g., [18]) have been potentially confounded by response biases [8,9]. That is, the participants may have assumed that if they heard a steam whistle at the same time as when they saw a steaming kettle then these two events 'ought' to go together. A later study by Warren *et al.* [18] showed larger spatial ventriloquist effects for face-speech pairings as compared to abstract visual-speech pairings and argued that the higher degree of 'compellingness' that characterized the speaker's face and matching voice led to an enhanced integration. It could, however, also be the result of the high temporal coherence of audiovisual speech rather than the rich informational content of the pairings [6].

Recently, a number of studies on the 'unity assumption' have managed to eliminate the above-mentioned confounds. Specifically, these studies focused on the prediction that binding for semantically congruent stimuli is stronger than for incongruent pairings. Thus, temporal order discrimination is expected to be more difficult for the former as compared to the latter pairing [8–10,19]. For instance, van Wassenhove and colleagues [20] presented congruent audiovisual syllables or incongruent McGurk syllables [21] under different stimulus onset asynchronies (SOAs) in a simultaneity judgment (SJ) task. The estimation of the width of the temporal window of integration revealed higher asynchrony tolerance for congruent as compared to incongruent pairs (203 vs. 159 ms). Additionally, Vatakis and Spence [8] found that it was easier to judge the temporal order of mismatched (in gender) as opposed to matched audiovisual speech (see also [22]). These studies provide the first unequivocal demonstration of cognitive factors modulating spatiotemporal integration (but see [9,19]).

Facilitation of cross-modal binding due to semantic congruency, with no time modulation, was demonstrated by Laurienti and colleagues [3]. Specifically, they measured unimodal (a blue/red circle or the auditory word for blue/red) and bimodal speeded color discrimination (detect auditory, visual or audiovisual blue or red color) in the presence of irrelevant stimulation (i.e., green). The results showed faster and more accurate discrimination for congruent-bimodal than incongruent and distractor trials. This study, however, utilized identical response keys for congruent auditory-visual stimuli, thus confounding faster responding with perceptual and decisional effects (i.e., response redundancy). Chen and Spence [23], in a follow-up study, utilized an unspeeded identification task for a series of briefly presented masked pictures. The pictures were sometimes accompanied with a synchronous- or lagging-sound that was semantically congruent or incongruent or else neutral. Results showed improved picture identification in the presence of congruent over incongruent sounds in relation to the control condition. This effect was maintained even for auditory lags of around 300 ms, while it disappeared for longer asynchronies.

A critique on these studies, however, is that comparisons were made across stimulus classes that were not equalized across the visual-acoustic dimensions (but see [8,24]). Vroomen and Stekelenburg [25] eliminated this potential structural confound by using matched sine-wave speech (SWS) replicas of pseudowords with lip-read information and participant groups instructed to perceive SWS as speech or non-speech. The two groups were equally sensitive in their order judgments, thus suggesting the dissociation of multisensory integration from cognitive factors. Additional evidence against the assumption of unity comes from studies exploring dominance of vision and touch in size estimation [26,27]. In these studies, participants were either presented with visual feedback of their haptic exploration or visual feedback was withheld [27]. For the latter conditions, knowledge of source stimulation was manipulated (i.e., common or different signal origin) [26]. Size estimation was found to be more dependent on haptic information in the absence of feedback *regardless of source origin* and on both modalities when feedback was available.

A resolution of the above-mentioned contradictions may be a recently suggested dissociation of timing and semantic congruency in integration. Specifically, Thomas and Shiffrar [28], using point-light displays, demonstrated that footstep sounds enhanced walker identification regardless of presentation time. A finding further strengthened by the fact that random footstep sounds enhanced sensitivity relative to simultaneously presented simple tones. Thus, it could be said, at least under some conditions, that meaningful associations may drive multisensory integration across wide temporal windows. This hypothesis is further supported by two recent clinical studies [29••,30•]. Specifically, Freeman and colleagues [29••] investigated whether synchrony perception and integration depended on distinct rather than common synchronization mechanisms. They tested neurologically healthy participants and patient PH who had lesions in the pons and basal ganglia, and who, when listening to