Cambridge University Press
978-1-108-46822-0 — Multisensory Interactions in the Real World
Salvador Soto-Faraco , Daria Kvasova , Emmanuel Biau , Nara Ikumi , Manuela Ruzzoli , Luis Morís-Fernández , Mireia Torralba
Excerpt
More Information

1 Introduction and Scope

The rich diversity of our sensory systems allows us to sample the environment from a variety of independent channels. Although the number of human senses is still under debate, it is undoubtedly far greater than the traditional five identified by Aristotle (Durie, 2005). The different senses, or sensory modalities, allow us to process complementary attributes of an object, extract correlations between its various features, and segregate some particular sensory properties from others. The impact of this multisensory capacity on cognition and brain functions has been extensively acknowledged in cognitive neuroscience literature. As multisensory processes have become a popular subject of scientific enquiry, they have also raised interest in other domains such as design, consumer sciences, philosophy, marketing, and even gastronomy. For example, at Hospital Sant Joan de Déu in Barcelona, researchers use multisensory principles to create more appetising menus for children undergoing chemotherapy, who often suffer taste alterations that lead to nutritional problems (Puigcerver et al., 2018).

One of the reasons for this surge of interest in multisensory processing is its relevance in understanding, explaining, and modulating human perception amidst the sensory complexity of real-world environments. Achieving this understanding would be difficult by studying each sense one at a time (e.g., Churchland, Ramachandran, & Sejnowski, 2005; De Gelder & Bertelson, 2003; Driver & Noesselt, 2008; Driver & Spence, 1998, 2000; Ernst & Banks, 2002; Ghazanfar & Schroeder, 2006; Macaluso & Driver, 2005; Shams & Seitz, 2008; Stein & Meredith, 1993). In contrast with this initial motivation, however, most of the research within the multisensory literature has used idealised, simplified laboratory tasks. The advantage of laboratory tasks is that they permit tight experimental control; that is to say, the phenomena under study can be carefully manipulated in the absence of extraneous variables that would complicate interpretation. The disadvantage, however, is that the deceptively simple models used in the laboratory are much unlike real-world environments. Because research in real-world-relevant conditions has been sparse, most basic laboratory findings in multisensory research, whilst valuable, have not been confirmed in ecologically valid conditions.

1.1 Why Is Real-World Cognitive Neuroscience Interesting?

There is ample agreement that generalising the findings discovered within controlled laboratory setups to real-life contexts faces substantial challenges. One of these challenges is the trade-off between rigorous experimental control and generalisation (ecological validity) (e.g., Blanton & Jaccard, 2006; Burgess et al., 2006; Kayser, Körding, & König, 2004; Kingstone et al., 2003; see also Neisser, 1976, 1982). In short, the higher the ecological validity of a study, the less the degree of experimental control that can be achieved over all relevant variables. So, if we are to trade experimental control for ecological validity, what is the gain of addressing real-world research?

One obvious motivation to address the generalisation of findings outside the laboratory is transference, viz. the application of basic principles discovered in experimental settings to practical problems. Transference has historically fuelled some multisensory studies to improve or assist human performance in real-life, multisensory tasks. For example, Sumby and Pollack's classic work, which demonstrated that watching a speaker's lips in a noisy environment enhances speech comprehension, was initially driven by the purely practical need of improving communication in industrial and military environments (Sumby & Pollack, 1954). The area of sensory substitution uses multisensory principles to induce cross-modal plasticity with the hope of improving perceptual capacities in sensory-deprived people (e.g., Bach-y-Rita & W. Kercel, 2003; Martolini et al., 2018; Vercillo, Tonelli, Goodale, & Gori, 2017). One example of this approach is a device called *BrainPort*, developed by Paul Bach-y-Rita. This device can transform visual shapes picked up by a video camera into tactile patterns presented to the tongue of blind persons, allowing them to perform some basic discrimination tasks (Bach-y-Rita, Danilov, Tyler, & Grimm, 2005; Danilov & Tyler, 2005). Other examples of multisensory approaches to practical problems are packaging design (Spence, 2016) and road safety (Ho, Reed, & Spence, 2007; Spence & Ho, 2008). In all these cases, the generalisation of basic laboratory findings is fundamental for the process of transference. Basic research must not only characterise multisensory phenomena by carefully controlled laboratory work and theoretical analysis, but it must also provide a good understanding of how these multisensory phenomena play out in complex, realistic environments (e.g., Maidenbaum & Abboud, 2014).

Besides practical motivations, there are also important theoretical and empirical lessons to be learned from extrapolating laboratory findings to real-world conditions. For one, as pointed out by E. A. Maguire, it would appear highly relevant to study the brain in the real-world conditions under which it has evolved to function (Maguire, 2012). These real-world conditions might disclose unforeseen gaps in our current knowledge, bring to the fore complex behaviours which might only occur (and therefore be studied) in these naturalistic settings, or even alter the outcomes of known laboratory-based results in significant ways (Hasson, Malach, & Heeger, 2010; Smilek, Birmingham, Cameron, Bischof, & Kingstone, 2006; Smilek, Eastwood, Reynolds, & Kingstone, 2007). One example of how laboratory findings might change when tested under real-life conditions comes from visual attention research. Wolfe et al. (Wolfe, Horowitz, & Kenner, 2005) found that when people searched for easy-to-see but very infrequent targets (<1 per cent prevalence) under naturalistic conditions, such as guns in airport baggage screening, miss rates were unexpectedly high. Visual search is one of the most widely used psychophysical tasks in experimental psychology, but for practical reasons, the typical protocol, until Wolfe et al.'s study, had used high target prevalence (i.e., a large proportion of trials bear the target the observer is instructed to search for). In this case, even a modest step towards realistic scenarios, such as varying target prevalence, resulted in substantial changes in the outcome.

A further incentive to pursue real-world generalisation is to avoid the risk of atomisation; that is, laboratory approaches tend to isolate the object of study from all other possible confounding variables, sometimes neglecting the fact that studying isolated components of a complex system can obscure its emergent properties (Kitano, 2002; Ward, 2002). Brain mechanisms might be grossly mischaracterised when they are singled out in controlled laboratory experiments, compared to when they operate intertwined with each other in real-world conditions. For example, in the field of visual attention, the dissociation between endogenous and reflexive orienting mechanisms has been perpetuated by the mainstream use of idealised spatial cueing protocols. Endogenous orienting is often probed with central symbolic cues such as arrows, whereas exogenous orienting is usually triggered with eccentric salient cues such as lateralised flashes. Some studies, however, have highlighted that this strict dissociation between reflexive and endogenous attention protocols could prevent us from understanding attention orienting during everyday social interactions (Birmingham & Kingstone, 2009; Kingstone et al., 2003; Risko, Laidlaw, Freeth, Foulsham, & Kingstone, 2012).

The ability of human observers to find people amongst the complexity of real-world scenes provides another example of the potential risk of atomisation. Traditionally, visual attention research has emphasised a division between parallel feature-based search mechanisms (efficient) and serial conjunction search (inefficient). According to this classical framework, finding complex objects defined by the conjunction of various simple attributes requires an effortful, serial process. New findings emerging from search tasks in naturalistic scenes, however, show that finding people (fairly complex visual objects) in photos of cluttered and heterogeneous everyday life scenes is unexpectedly fast and efficient (Peelen & Kastner, 2014; Papeo, Groupil & Soto-Faraco, 2019).

1.2 Multisensory Processing in a Complex World

The 1990s saw increased motivation to study the interaction between the senses instead of each sense in isolation. This was in part because of a concern regarding ecological validity. It is our view, however, that this research did not go far enough in solving the concern. For the most part, efforts were directed towards understanding multisensory processes themselves. They were less concerned with understanding how these multisensory processes play out in real-world contexts or how they mesh with other brain processes under complex task demands. In real world conditions, multisensory interactions happen amidst, or even as part of, other processes such as attention, expectation, meaning integration, executive control, and sensorimotor integration. What transformations of multisensory phenomena discovered in simple environments will arise when they are brought to these more complex, close-to-real-life situations? Furthermore, which new questions might emerge from studying these scenarios?

The interest in approaching real-world conditions in multisensory research is now taking off (e.g., Mastroberardino, Santangelo, & Macaluso, 2015; Matusz, Dikker, Huth, & Perrodin, 2018; Nardo, Santangelo, & Macaluso, 2014). This new interest is not unique to multisensory research. It sparked investigations in other fields such as spatial navigation (Spiers & Maguire, 2006), episodic memory (e.g., Santangelo, Di Francesco, Mastroberardino, & Macaluso, 2015), event perception (e.g., Hasson, Nir, Levy, Fuhrmann, & Malach, 2004), and sensorimotor decision making (e.g., Gallivan, Chapman, Wolpert, & Flanagan, 2018), as well as in the visuospatial attention domain as we mentioned earlier (Kingstone et al., 2003; Nardo, Santangelo, & Macaluso, 2011; Peelen & Kastner, 2014). In the case of multisensory perception, the knowledge gap between laboratory and real life is still significant. For example, although there is a sizable body of evidence for cross-modal enhancement (i.e., perception in one sensory modality is more accurate and faster when complementary information is available in another modality), the impact of this multisensory benefit in real-world conditions is still largely unknown. Based on the examples from visual research discussed above, one can infer that understanding the interplay between multisensory processes and other mechanisms under complex conditions would appear to be important.

Unfortunately, multisensory effects discovered in the laboratory are rarely put to the test under real-life conditions. These tests are not only difficult to realise, and unfeasible in some cases, but very often their data is hard to interpret. As a compromise solution, some studies addressing intermediate steps between laboratory and real-life conditions have begun to emerge. This is referred to as 'naturalistic laboratory research' (by Matusz et al., 2018). This Element brings up some examples from the multisensory literature in which known laboratory findings have been put to the test in complex situations, i.e., when multisensory mechanisms are studied in interaction with other cognitive processes. The focus will be placed on the interplay between multisensory processes and attention, prediction, temporal organisation, and conflict processing mechanisms. These illustrative examples have a bias towards the authors' own research interests and are mostly confined to human studies with healthy, young adults. Other ramifications of multisensory research, which might also be relevant to real-world generalisation, are left out of this review. (More extensive reviews about multisensory processes can be found, for example, in Calvert, Spence, & Stein, 2004; Spence, 2018; and Stein, 2012.)

We will first consider the issue of the limited capacity of human information-processing machinery. Sensory-rich, complex scenarios, typical of real-life environments, pose a serious problem of selection to the limited-capacity cognitive system (e.g., Desimone & Duncan, 1995; Huang, Treisman, & Pashler, 2007). The term *selective attention* generally refers to a variety of processes and mechanisms that help select, parse, and organise information in order to allocate resources efficiently. The interplay between these attention mechanisms and multisensory processes lies at the core of any attempt to address perception in real-world environments. Do multisensory interactions break down under such high selection pressure? This issue has been the object of intense debate in the last few years (De Meo, Murray, Clarke, & Matusz, 2015; Hartcher-O'Brien, Soto-Faraco, & Adam, 2017; Hartcher-O'Brien et al., 2016; Koelewijn, Bronkhorst, & Theeuwes, 2010; Navarra, Alsius, Soto-Faraco, & Spence, 2010; Talsma, Senkowski, Soto-Faraco, & Woldorff, 2010; ten Oever et al., 2016). Section 2 presents a brief up-to-date review of this debate with a focus on complex, real-world scenarios.

Next, we will discuss the impact of prediction and temporal organisation in parsing complex, multisensory environments. Real-world environments are often structured across various scales, from the smallest spatial and temporal patterns to an intricate web of semantic relationships. Multisensory perception exploits this structure via several mechanisms that help anticipate and organise sensory inputs. These mechanisms can involve phase reset, entrainment of neural activity to rhythmic sensory input (Lakatos, Chen, O'Connell, Mills, & Schroeder, 2007; Schroeder & Lakatos, 2009), grouping events in time (Ikumi & Soto-Faraco, 2014; Lewald & Guski, 2003; Vatakis & Spence, 2006), anticipating properties from one modality to another via semantic association (Chen & Spence, 2010, 2011; Iordanescu, Guzman-Martinez, Grabowecky, & Suzuki, 2008; Parise & Spence, 2009), and inferring causal structure between sensory inputs via priors built from experience (Gau & Noppeney, 2016; Kayser & Shams, 2015; Noppeney & Lee, 2018). Section 3 discusses some examples of prediction and temporal organisation in multisensory processing and their potential consequences for real-world perception.

Third and lastly, we will consider the interplay between multisensory interactions and conflict processing mechanisms. Brain mechanisms of conflict processing are set in motion when incompatible mental representations are activated and compete (Botvinick, Braver, Barch, Carter, & Cohen, 2001). We will discuss two examples of this interplay. The first example regards sensorimotor conflict; that is, when alternative courses of action in response to a stimulus compete to drive behaviour. This kind of conflict can arise when events in different sensory modalities trigger alternative spatial representations for action. The second example of the interplay between multisensory and conflict processes relates to perception of inter-sensory conflict. Multisensory research has frequently resorted to inter-sensory conflict as a model to study general principles of multisensory integration (Bertelson, 1998; De Gelder & Bertelson, 2003) even though most real-world objects provide congruent (i.e., correlated) information across the senses. Therefore, it is surprising that the relationship between multisensory and conflict processes has rarely been addressed explicitly. Section 4 presents some recent findings that bring to the fore the interplay between conflict mechanisms and multisensory interactions.

2 Multisensory Processing and Attention: Real-World Environments Require a Multifaceted Interplay

Imagine a stranger coming to talk to you at a party at the precise moment when the buzz is at its loudest. The hearing is tough, so to figure out what the stranger is trying to say you must listen carefully while watching his lip movements. In this case, the effort of integrating visual and auditory information demands our fully devoted attention. In other cases, however, multisensory interactions appear to happen in an effortless, unavoidable fashion. An instance of this is when the olfactory and gustatory properties of foods create the unified experience of taste. When do we need to focus attention to capitalise on the benefits of multisensory interactions, and when do these benefits arise effortlessly? The relationship between attention and multisensory interactions has been in the limelight for nearly two decades (Alsius, Navarra, Campbell, & Soto-Faraco, 2005; Busse, Roberts, Crist, Weissman, & Woldorff, 2005; De Gelder & Bertelson, 2003; Driver, 1996; McDonald, Teder-Sälejärvi, & Ward, 2001; Spence & Driver, 2004; Talsma et al., 2010).

Early views regarding the relationship between multisensory processes and attention were positioned along two opposing extremes. At one end of the spectrum were researchers who regarded multisensory processes as automatic and pre-attentive. At the other end of the spectrum were researchers who emphasised the need for selective attention as a prerequisite for multisensory integration. Under the former view, multisensory integration does not only happen irrespective of whether the involved sensory inputs are attended or not, but its outcome can summon attention itself (Bertelson, Vroomen, De Gelder, & Driver, 2000; Driver, 1996; Van Der Burg, Olivers, Bronkhorst, & Theeuwes, 2008; Vroomen, Bertelson, & de Gelder, 2001). Under the latter view, however, attentional selection must be deployed prior to (and often is a condition for) multisensory integration (Alsius, Navarra, Campbell, & Soto-Faraco, 2005; Duncan, Martens, & Ward, 1997; Talsma & Woldorff, 2005). These opposing views often put the emphasis on different conceptions about the functional architecture of multisensory interactions in the brain. Feedforward architectures allow bottom-up convergence of sensory information, which is associated with fast and automatic interactions. In contrast, feedback (or recurrent) architectures are amenable to attention mediation via top-down processes (Driver & Noesselt, 2008; Driver & Spence, 2000; Foxe & Schroeder, 2005).

2.1 The Real-World Relevance of the Interplay between Attention and Multisensory Processes

The either–or debate about the role of attention in multisensory processes sketched above has important implications for real-world perception (Hartcher-O'Brien et al., 2017, 2016; Koelewijn et al., 2010; Matusz et al., 2018; Talsma, 2015; Talsma et al., 2010; ten Oever et al., 2016). If we accept the hypothesis that multisensory interactions occur automatically, in a purely bottom-up fashion, then multisensory phenomena discovered under simplified laboratory conditions should still work well in more complex real-world scenarios. In this case, multisensory interactions can bring about very relevant benefits[1] without any cognitive cost. If, on the other hand, we reject the

---

[1] For example, automatic integration mechanisms can furnish multisensory events with increased salience (Noesselt et al., 2010; Van Der Burg et al., 2008), leading to faster and more precise saccadic reactions to imperative events in the environment (Colonius & Arndt, 2001; Corneil, Van Wanrooij, Munoz, & Van Opstal, 2002; Diederich & Colonius, 2004); improve sensitivity to stimuli that are hard to notice (Caclin et al., 2011; Frassinetti et al., 2002; Gleiss & Kayser, 2013, 2014; Jaekl & Harris, 2009; Jaekl & Soto-Faraco, 2010; Noesselt et al., 2010); increase precision when estimating the properties of objects (Ernst & Banks, 2002; Fetsch, Pouget, DeAngelis, & Angelaki, 2012; Yau, Olenczak, Dammann, & Bensmaia, 2009); and make perception in noisy environments more accurate (Grant & Seitz, 2000; Jaekl et al., 2015; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2006; Sumby & Pollack, 1954).

automaticity hypothesis, the complexity of real-world conditions might impose strong limitations on, if not compromise altogether, the multisensory phenomena discovered under simpler laboratory conditions. In this case, attentional selection can become a bottleneck to achieving the perceptual benefits of multisensory interactions.

Research findings offering evidence for automatic multisensory interactions have been matched by equally compelling results supporting attentional mediation. According to recent reviews on the topic, both types of mechanism do play a role, and the debate has boiled down to a more nuanced question: what is the balance of power between bottom-up automatic processes and top-down mediation (De Meo et al., 2015; Hartcher-O'Brien et al., 2016; Talsma, 2015; ten Oever et al., 2016)? The answer to this question seems to depend on a variety of factors, including the level(s) of processing involved in the representation of a multisensory event (from low-level spatio-temporal attributes to higher-level semantic properties), the physical salience and task-relevance of the stimuli, and the perceptual load[2] of the scenario, amongst others. Similar to what was proposed in the biased competition framework of attention a couple of decades ago (Desimone & Duncan, 1995), the outcome of multisensory interactions might emerge from a competitive process between bottom-up evidence and top-down, endogenous biases.

One study by Fujisaki and colleagues (Fujisaki, Koene, Arnold, Johnston, & Nishida, 2006) provides a demonstration of this competition between bottom-up and top-down processes during multisensory perception. Participants saw a display populated by blobs flashing randomly plus a sound varying in amplitude. They had to find the only blob whose flashing was synchronised with the amplitude changes of the sound. Hence, subjects were to individuate the one audiovisually congruent object amongst other visual-only distractors. Fujisaki et al. discovered that search efficiency for audiovisual synchrony was determined both by exogenous factors, triggering bottom-up interactions, as well as by endogenous processes relying on top-down mediation. Audiovisual targets whose single-modality components were salient (by making them spatially and temporally unique) could be detected automatically. When the salience of single-modality components was reduced because the inputs were embedded in more cluttered displays, endogenous attention became necessary for the detection of audiovisual synchrony. A similar outcome was obtained with speaking faces and speech sounds in subsequent experiments, which will be discussed later in the context of language (Alsius & Soto-Faraco, 2011).

---

[2] Perceptual load relates to the number of different items in a display that need to be perceived and/or the amount of resources required for the perceptual identification of each item.

The competition involving stimulus-driven bottom-up processes and endogenously driven top-down biases can play out throughout the different levels of representation that characterise multisensory objects (e.g., spatial location, temporal correspondence, semantics, action plans). This framework of competition at multiple stages, called ‘multifaceted interplay’ (Talsma et al., 2010), can be especially suitable to understand multisensory interactions in complex scenarios; that is, scenarios where goal-directed behaviours likely involve parsing information at various levels of representation in a high-load, but structured environment. The multifaceted interplay hypothesis can account for the fact that sometimes multisensory processes occur in a bottom-up, automatic manner, even influencing attention, whereas at other times multisensory processes are mediated by attention. That is, attention and multisensory interactions can mutually influence each other or at times even be indistinguishable from each other. Below we discuss the evidence for either type of influence, with some examples relevant to real-world multisensory perception.

2.2 What, and How Much, Can We Expect from Bottom-Up Multisensory Interactions?

One recurrent finding supporting an automatic view of multisensory interactions is that a multisensory singleton embedded amongst unisensory events stands out and can therefore capture attention. The straightforward interpretation of this type of finding is that correlated sensory inputs are automatically bound into a multisensory representation, increasing the salience of that object in the scene by making it unique. For example, it has been claimed that a sound synchronised with a visual event can make it seem brighter[3] (Stein, London, Wilkinson, & Price, 1996), last longer (Vroomen & Gelder, 2000), and make it easier to detect (Andersen & Mamassian, 2008; Bolognini, Frassinetti, Serino, & Làdavas, 2005; Frassinetti et al., 2002; Jaekl & Soto-Faraco, 2010) and faster to respond to (e.g., Murray et al., 2005; Pérez-Bellido, Soto-Faraco, & López-Moliner, 2013). A popular study by Van der Burg et al. (Van Der Burg et al., 2008) showed that in crowded dynamic visual search displays, a spatially uninformative sound synchronised with an irrelevant colour change led to pop-out in an otherwise difficult serial search task. This phenomenon has been dubbed the ‘pip and pop’. Similarly, Maddox et al. (Maddox, Atilgan, Bizley, & Lee, 2015) showed that irrelevant visual events could aid auditory selective attention. Some of these phenomena have been linked to physiological interactions in subcortical or primary sensory areas – defined as ‘early’, sensory-based interaction (Driver & Noesselt, 2008; Shams & Kim, 2010; Stein & Stanford, 2008).

---

[3] This interpretation, however, has been disputed (Odgaard, Arieh, & Marks, 2003).

Complementing the results discussed above, Santangelo and Spence (2007) reported that multisensory events are less prone to be neglected in high-perceptual-load conditions than are unisensory events. This result could be explained by an automaticity account whereby multisensory interactions occur via bottom-up mechanisms based on a feedforward architecture. Some neuroimaging findings support this view because they reveal that the brain correlates of multisensory interactions can be expressed at early stages of sensory processing, in terms of both latency and functional anatomy (Foxe et al., 2000; Matusz & Eimer, 2011; Molholm et al., 2002; Murray et al., 2005; Van der Burg, Talsma, Olivers, Hickey, & Theeuwes, 2011). Some authors have related these fast, automatic multisensory interactions to the discovery of direct (i.e., monosynaptic) cortico-cortical connections between sensory areas of different modalities (Falchier, Clavagnier, Barone, & Kennedy, 2002; Rockland & Ojima, 2003).

2.2.1 Multisensory Warning and Interference in Real-World Environments

The putative automaticity of multisensory enhancement effects, such as the ones described above, conveys a ‘privileged’ attentional status to multisensory stimuli. This has potential real-life implications for the design of warning signals in demanding environments. For example, multisensory events appear to be effective at summoning the drivers’ attention to road hazards, eliciting fast braking responses (Ho et al., 2007; Spence & Santangelo, 2009). Remarkably, some findings suggest that the capacity of multisensory events for attracting attention is not limited solely to spatio-temporal congruence between abrupt stimulus onsets (typical flash-beep stimuli). They indicate, as well, that this multisensory benefit extends to cross-modal congruence between higher-level attributes such as semantics (Iordanescu, Grabowecky, Franconeri, Theeuwes, & Suzuki, 2010; Iordanescu, Grabowecky, & Suzuki, 2011; Iordanescu et al., 2008; Pesquita, Brennan, Enns, & Soto-Faraco, 2013). This research shows that finding one predefined visual target object (e.g., cell-phone) amongst other ordinary everyday life objects is faster if the person hears its characteristic sound (e.g., ringtone). This could have important implications for real-life contexts, where we are usually surrounded by familiar objects interconnected by a rich web of semantic associations. This case will be addressed further in Section 3.3. In other cases, applied research concerning real-world scenarios (e.g., in the context of driving) has concentrated on the potential consequences of synergy or interference when a variety of spatial cues are delivered to different sensory modalities (Ho & Spence, 2014; Spence & Ho, 2015b, 2015a, Spence & Soto-Faraco, in press).