Synthese (2021) 198:2101–2127
https://doi.org/10.1007/s11229-019-02195-9

STRUCTURE OF PERCEPTUAL OBJECTS

The structure of audio–visual consciousness

Błażej Skrzypulec[1]

Received: 28 July 2018 / Accepted: 27 March 2019 / Published online: 6 April 2019

Abstract

It is commonly believed that human perceptual experiences can be, and usually are, multimodal. What is more, a stronger thesis is often proposed that some perceptual multimodal characters cannot be described simply as a conjunction of unimodal phenomenal elements. If it is the case, then a question arises: what is the additional mode of combination that is required to adequately describe the phenomenal structure of multimodal experiences? The paper investigates what types of audio–visual experiences have phenomenal character that cannot be analysed as a mere conjunction of visual and auditory elements; and how can we properly characterise the required, additional mode of perceptual combination. Three main modes of combination are considered: (a) instantiation, (b) parthood, and (c) grouping. It is argued that some phenomena involving intermodal relations, like spatial and temporal ventriloquism, can be analysed in terms of audio–visual, perceptual grouping. On the other hand, cases of intermodal binding need a different treatment. Experiences involving audio–visual binding should be analysed as experiences presenting objects or events which instantiate, or which have a proper part instantiating, both visually and auditorily determined properties.

Keywords Perception · Multimodal · Vision · Audition · Binding · Intermodal · Perceptual objects · Instantiation · Parthood · Grouping

In contemporary philosophy of perception, it is commonly believed that human perceptual experiences can be, and usually are, multimodal (see Briscoe 2016;

✉ Błażej Skrzypulec
blazej.skrzypulec@gmail.com

[1] Institute of Philosophy and Sociology, Polish Academy of Sciences, ul. Nowy Świat 72, 00-330 Warsaw, Poland

Macpherson 2011; O'Callaghan 2012). It seems plausible that within the same conscious episode we may see something, for example a tree in a park, hear something, for instance passing cars, and smell something, like bread from a nearby bakery.[1] In more technical terms, it can be stated that the phenomenal character of perceptual experiences is often a combination of various unimodal phenomenal elements.

What is more, a stronger thesis is often proposed that perceptual phenomenal character not only is commonly multimodal, but at least some perceptual multimodal characters cannot be described simply as a conjunction of unimodal phenomenal elements (e.g., Kubovy and Schutz 2010; O'Callaghan 2015a; Richardson 2014; Stevenson 2000). Common examples are experiences involving "intermodal binding".[2] Such experiences present something as associated with phenomenal elements related to different perceptual modalities. For instance, one can have an experience as of a dog that looks a certain way and also barks. An initial idea may be to analyse such an experience as a mere conjunction of visual and auditory elements by claiming that (a) it visually presents that something is a dog *and* (b) it auditorily presents that something barks. However, such an analysis does not allow us to differentiate the phenomenal character of this experience from the phenomenal character of an experience which presents a distinct combination of the same elements (i.e., a visually presented dog and an auditorily presented barking). In particular, the same analysis in terms of a conjunction (that something looks like a dog *and* that something barks) can also be applied to an experience as of two dogs: one visible and silent and a second barking dog. This experience seems to be phenomenally different from an experience presenting a single barking dog despite the fact that the involved auditory and visual elements may be the same.

If it is the case that not every multimodal phenomenal character can be analysed as a conjunction of unimodal elements, such as in the above audio–visual example, then we should investigate what mode of combination is proper in such cases. Let's again consider the case of an experience presenting a barking dog. For instance, one may propose that what is presented in such an experience is a single object possessing, or technically speaking 'instantiating', visual and auditory properties (e.g., the 'being a dog' property and the 'producing a barking sound' property). However, this is not the only available option. Another idea may be to interpret such an experience not in terms of a subject and properties, but in mereological terms of parts and wholes. In this case, what is presented is a whole that has a visual part and an auditory part (e.g., a dog part and a barking part). Yet another approach would be to apply the notion of perceptual grouping. From this perspective, what the considered

[1] However, see Spence and Bayne (2014) for an alternative interpretation that such conscious episodes may be in fact a series of short-lived unimodal experiences.

[2] In the philosophical and empirical literature there is no convention governing the use of terms such as 'multimodal', 'intermodal', 'multisensory', or 'cross-modal'. Here, I use 'multimodal' to cover all experiences that involve elements related to more than one modality, and 'intermodal binding' and 'intermodal relations' to refer to two more specific phenomena that are present in some multimodal experiences (in this I follow O'Callaghan, whose works are frequently referred to in this paper, for instance see O'Callaghan 2016).

experience presents should be described as a group with visual and auditory elements (e.g., a dog and a sound).

Such various modes of perceptual combination differ in terms of the perceptual capacities required for applying them in organising the structure of an experience. For example, representing edges between spatially and/or temporally connected elements is crucial for part-perception, while to achieve perceptual grouping it is more important to recognise similarity relations between spatially or temporally disjoint items (see Elder and Goldberg 2002; Hoffman and Richards 1984; Palmer and Rock 1994; Xu and Singh 2002).

The goal of this paper is to investigate the following questions: (a) what types of audio–visual experiences have phenomenal character that cannot be analysed as a mere conjunction of visual and auditory elements; and (b) how can we properly characterise the required, additional mode of perceptual combination? I focus on audio–visual experiences, as there is a rich empirical literature which may help to determine the proper mode of combination and there are already interesting philosophical proposals regarding audio–visual consciousness that can serve as a basis for a discussion (e.g., O'Callaghan 2014b, 2015a, b; Macpherson 2011; Matthen 2010). Relying on obtained results, further investigation may reveal whether the same modes of combination are also applicable in cases of multimodal experiences combining elements related to different modalities.

The paper starts by introducing the notion of 'conjunctive multimodality' and discusses types of phenomena whose occurrence entail that some experiences are not conjunctively multimodal (Sect. 1). Subsequently (Sect. 2), it is argued that not all audio–visual experiences are conjunctively multimodal due to the presence of intermodal relations and intermodal binding. In Sect. 3, I present several perceptual modes of combination that may be useful in analysing the phenomenal character of audio–visual experiences that are not conjunctively multimodal. Finally, in Sects. 4, 5, and 6, I analyse, by referring to the described modes of combination, how the phenomenal character of audio–visual experiences involving intermodal relations and intermodal binding can be described using the notions of instantiation, parthood, and grouping.

1 Conjunctive multimodality and beyond

I assume that there exist human multimodal perceptual experiences, including audio–visual ones, whose phenomenal character can be described as a conjunction of unimodal phenomenal elements. Furthermore, I believe that such conjunctive multimodality is properly grasped by the notion of minimal multimodality presented by O'Callaghan (2015b: p. 560):

> (Minimal Multimodality) *The phenomenal character of each perceptual episode is exhausted by that which could be instantiated by a corresponding merely visual, merely auditory, merely tactual, merely gustatory, or merely olfactory experience, plus whatever accrues thanks to simple co-consciousness.*

Later, I refer to phenomenal character *that could be instantiated by a corresponding merely visual experience* using the term "visual elements of an experience", and analogously to phenomenal character *that could be instantiated by a corresponding merely auditory experience* with the term "auditory elements of an experience". While O'Callaghan's characterization of minimal multimodality concerns more modalities than just vision and audition, its scope can easily be restricted to fit the topic of the paper simply by removing elements that do not refer to the two considered senses (see the definition of *Conjunctive Multimodality* on p. 5).

According to the notion of minimal multimodality, the phenomenal character of a multimodal experience is nothing more than a co-occurrence of unimodal phenomenal elements related to all modalities contributing to this experience. This mode of combination can be represented as a logical conjunction of elements associated with various modalities. The only additional aspects are those that arise merely due to the co-occurrence of unimodal elements. For instance, when having a multimodal experience one may also have an additional feeling of unity arising from the fact that different unimodal elements occur together within a single conscious episode (e.g., Bayne 2008; Macpherson 2011; Tye 2003).

The notion of minimal multimodality is able to accommodate the existence of common sensibles, i.e. properties that can be experienced using more than one modality (Tye 2007). For example, one may have an experience in which a square shape is presented both by visual and tactile modality without additionally experiencing that both modalities present the shape of the same object. Such a case can be analysed in conjunctive terms because visual squareness and tactile squareness can be treated as phenomenally distinct, co-occurring elements of an experience.

By adopting the notion of minimal multimodality, I treat all experiences involving at least two phenomenal elements associated with different modalities as multimodal experiences. Some authors postulate a more restricted notion of multimodal experiences according to which genuine multimodality requires that the elements associated with various modalities are fused into a single, multimodal element (see Connolly 2014; Fulkerson 2014 for discussion of this topic). However, in this paper I use a broader notion presented in the definition of *Minimal Multimodality*.

By using the resources provided by the notion of minimal multimodality, we may define what it means for an audio–visual experience to be conjunctively multimodal:

> (Conjunctive Multimodality) *An audio-visual perceptual experience is conjunctively multimodal if and only if its phenomenal character is exhausted by that which could be instantiated by a corresponding merely visual and merely auditory experience, plus whatever accrues thanks to simple co-consciousness.*

In the philosophical literature, there are two main ideas concerning phenomena whose occurrence would entail that there are multimodal experiences that are not conjunctively multimodal.[3] The first such postulated type of phenomena is the

[3] There are also other ideas, for instance concerning intermodal completion (O'Callaghan 2015a, b; see also Spence and Bayne 2014), but it is less plausible that such phenomena actually occur in human perceptual experiences.

presence of spatial or temporal intermodal relations, understood as relations that connect elements associated with different modalities (Briscoe 2017; O'Callaghan 2014a, 2015a; Richardson 2014). Let's consider two multimodal experiences involving elements *A* and *B*, associated with different modalities. In the case of the first experience, *A* is presented as positioned on the left of *B* within an egocentric frame of reference, and in the second experience the situation is reversed, such that *A* is presented as located on the right of *B*. The difference between the phenomenal characters of such experiences cannot be expressed by referring solely to the conjunctive mode of combination. This is because in each case the conjunction will be the same: *A and B*. In consequence, if audio–visual experiences present such intermodal relations between visual and auditory elements, then not all audio–visual experiences are conjunctively multimodal and some additional mode of combination has to be introduced.

One may try to resist the above conclusion by stating that elements associated with different modalities cannot co-occur within an experience without standing in some spatial and temporal relation, and so the notion of conjunctive multimodality already accommodates the presence of intermodal relations. According to this view, intermodal relations constitute that which *accrues thanks to simple co-consciousness*, as stated in definitions of *Minimal Multimodality* and *Conjunctive Multimodality*. However, such a statement concerning the connection between co-occurrence and spatiotemporal intermodal relations is too strong. In particular, it seems possible to have a multimodal experience in which some elements associated with different modalities are presented as temporally co-occurring, while not being able to recognise the spatial relations between those elements. Even if such experiences are rare, their possibility is sufficient to demonstrate that multimodal co-occurrence does not entail the presence of spatial intermodal relations.

An analogous point can be made in the case of temporal intermodal relations if one accepts that a single perceptual experience does not have to be an experience as of an instantaneous moment without any temporal extension (see Arstila 2018; Benovsky 2013; Power 2012 for a discussion). In the case of experiences presenting only instantaneous moments, the co-occurrence of elements associated with different modalities indeed seems to entail that they stand in a temporal relation of intermodal synchrony. However, if a single experience can be an experience presenting some time-period, then it is possible to have an experience of multimodal co-occurrence without being able to recognise the temporal order of elements associated with different modalities. In fact, the lack of certainty concerning temporal order is a well-recognised phenomenon in the case of unimodal auditory and visual experiences (e.g., Block and Gruber 2013; Kanabuse et al. 2002; Warren and Obusek 1972) as well as in the case of multimodal, audio–visual experiences (Vatakis and Spence 2007). It has been observed that people have difficulty recognising which stimulus was presented earlier and which later if presentation time is short and the stimuli are complex. Furthermore, robust effects concerning uncertainty of temporal order have been obtained by using continuous multimodal stimuli, like a speech recording desynchronized with a video of a talking person. (Vatakis and Spence 2006).

The second type of phenomena whose occurrence is incompatible with the thesis that all multimodal experiences are conjunctively multimodal is intermodal,

or cross-modal, binding (Briscoe 2017; Clark 2001; Kubovy and Schutz 2010; Kubovy and Yu 2012; Macpherson 2011; O'Callaghan 2006, 2008, 2011, 2012, 2014a, 2015a, b, 2016, 2017). Multimodal experiences involving intermodal binding are those that present a single entity, usually an object or an event, as combined with elements associated with different modalities. The term 'binding' may suggest that the visual and auditory elements are properties instantiated by some subject. It also has a specific meaning in psychological and neuroscientific works, where it does not primarily concern the structure of experiences but rather the mechanisms responsible for combining data about separately processed features (e.g., Holcombe and Cavanagh 2001; Usher and Donnelly 1998). However, for the moment I use 'binding' in a neutral way, following a general philosophical intuition that it describes situations in which elements associated with different modalities are experienced as somehow combined within a single object or event. Later, in Sects. 5 and 6, I analyse how intermodal binding should be understood, for instance whether it should be characterised in terms of instantiation, parthood relation, or perceptual grouping.

For instance, it seems that one may have an experience of a ball hitting a surface involving visual elements (like a ball and a surface) and auditory elements (like a sound generated by the impact). However, it seems the same elements can also occur in a different experience that does not involve intermodal binding even if it may involve some intermodal relations. For example, one may visually experience a ball hitting a surface and auditorily experience an impact sound without experiencing them as belonging to the same event. If one believes that such differences between experiences are possible in the case of human perception, then the thesis that all multimodal experiences are conjunctively multimodal has to be rejected. As in the case of intermodal relations, the analysis made in conjunctive terms cannot express a difference between the structures of the above experiences because in each of them the conjunction will be the same: *visually experienced ball hitting a surface and auditorily experienced impact sound*. Again, the insufficiency of conjunctive analysis suggests that an additional mode of combination should be postulated.

As shown above, the thesis that not all multimodal experiences are conjunctively multimodal is entailed by the presence of intermodal binding and by the presence of intermodal relations. Despite this similarity, it should be noted that they are different phenomena, as not every experience presenting some intermodal relations is also an experience involving intermodal binding. This is because elements associated with different modalities may be experienced as spatiotemporally related without being experienced as bound with a single object or event. Referring to the above example, a ball hitting a surface and an impact sound may be experienced as standing in some spatiotemporal relation without also being experienced as related to the same event.

In the next section, I characterise more precisely phenomena that are plausible candidates for involving audio–visual intermodal relations and audio–visual intermodal binding. I argue that both intermodal relations and intermodal binding occur in the case of human audio–visual experiences and that not all such experiences are conjunctively multimodal.

2 Audio–visual binding and audio–visual relations

It seems very intuitive that visual experiences present objects as being bound with many properties simultaneously. For instance, we may have an experience presenting an object as being square and red. Furthermore, we may have visual experiences that present the same properties and differ only in how the presented elements are combined. For example, an experience as of a red square and a green triangle is clearly different from an experience as of a green square and a red triangle. This observation has motivated a belief that the phenomenal character of visual states cannot be analysed simply as a conjunction of elements because then the phenomenology of two above-mentioned experiences could not be distinguished (see Clark 2004). This is because no matter whether an experience presents (a) a red square and a green triangle or (a) a green square and a red triangle, the conjunctive description would be the same: *object*$_1$ *and object*$_2$ *and red and green and square and triangular*.

It is plausible that an analogous binding occurs in the case of audio–visual experiences, and so not all of them can be analysed conjunctively. In particular, it seems commonly the case that we experience a single object or event as having a certain look and simultaneously as making a sound. In psychological works such experiences are characterised as involving a *phenomenal fusion* of elements associated with distinct modalities or as giving the impression of a *common origin* (see Bertelson 1999; Radeau and Bertelson 1977).[4] Sometimes cases of such intermodal binding are illusory. For instance (O'Callaghan 2012, 2015a), when participating in a ventriloquist performance we may have an experience as of a speaking puppet, and when watching a movie it seems to us that people on a screen are speaking even if in fact sounds are coming from nearby speakers (see Chen and Spence 2017; Deroy et al. 2016; Jordan et al. 2010; Kubovy and Schutz 2010; Kubovy and Yu 2012; O'Callaghan 2017; Vatakis and Spence 2007 for discussion concerning spatiotemporal and other possible factors that lead to such illusory cases of binding). It seems that such examples of audio–visual intermodal binding demonstrate that not all audio–visual experiences are conjunctively multimodal. This is because they present not only that some visual and auditory elements co-occur but additionally that they are related to the same object or event. For instance, if while watching a movie a proper synchronization between visual and auditory stimuli is not achieved, then while the same visual and auditory elements can be experienced, they are no longer experienced as associated with the same speaking-event.

Furthermore, it has been shown that similar experimental results are obtained both in cases of audio–visual binding and in unimodal visual binding (see O'Callaghan 2017). In particular, some psychological works support the idea of an object-specific preview benefit, i.e. a more efficient recognition of features when they reappear as properties of the same object, which occurs in the case of audio–visual binding

[4] It should be noted that *perceptual fusion* is a different effect to audio–visual cross-modal bias, i.e. a phenomenon where the location of an auditory stimulus is biased towards the location of synchronous visual stimulus (see Alais and Burr 2004; Bertelson and Aschersleben 1998 for description of results and methodological considerations about the genuine perceptual character of this effect).

(Jordan et al. 2010). Such results suggest that bound visual and auditory elements are represented as composing a single item, probably by virtue of maintaining a multimodal object-file storing both visual and auditory data (Zmigrod et al. 2009). Nevertheless, despite these similarities there are also studies that demonstrate differences between cases in which binding occurs in unimodal and multimodal contexts. For example, there are data suggesting that multimodal cues have a greater ability to attract spatial attention than unimodal ones (Santangelo et al. 2008). Furthermore, stimulus (e.g., a blue figure) is recognised faster when multiple cues are presented through multiple modalities (e.g., blue hue and a spoken word 'blue') than when several cues are presented through the same modality (e.g., blue hue and a written word 'blue', Laurienti et al. 2004).

In fact, some authors remain sceptical regarding whether experiences in which one seems to perceive visual and auditory elements as characterising a single object or event are genuine examples of audio–visual, intermodal binding. In particular, the relation between the behavioural results regarding intermodal processing of stimuli, like those concerning object-files and object-specific preview benefit, and the structure of phenomenal character of audio–visual states, is not obvious (Briscoe 2017; Deroy et al. 2014; Spence and Bayne 2014). For instance, some experiments show that phenomenal character may not match the information stored in object-files (e.g., Mitroff et al. 2005). If behavioural results are not conclusive in considerations regarding phenomenal character, then one may postulate that we should rather rely on analysing phenomenal contrast cases concerning audio–visual experiences (O'Callaghan 2014a, b). However, it seems that by comparing the phenomenology of states that involve binding (e.g., an experience presenting that a person in a movie is talking) and in which binding is not achieved (e.g., an experience in which the sound is not well-synchronized with what happens on the screen), we cannot refute the hypothesis, given the important role of spatiotemporal factors in experiencing binding, that an experience of an audio–visual binding is in fact nothing more than an experience presenting a spatiotemporal co-location, or at least spatiotemporal coordination, of auditory and visual elements (Briscoe 2017).

Nevertheless, even if the above arguments are accurate, and experiences involving audio–visual binding, i.e. which seem to present that visual and auditory elements belong to the same object or event, can be adequately characterised in terms of a spatiotemporal co-localisation or coordination, the occurrence of such co-localisation is still sufficient to refute the hypothesis that all audio–visual experiences are conjunctively multimodal.[5] This has been shown by Austen Clark in the case of visual experiences, but the same reasoning is also applicable to audio–visual, multimodal ones (see Clark 2000, 2004). The crucial observation is that experiencing elements as co-located involves experiencing them as sharing the same location, i.e. the location of the first element has to be identical to the location of the second one (in the case of coordination these locations should be at least proximal or overlapping). However, an analysis in terms of a conjunctive co-occurrence cannot express

[5] In section five, relying on ideas developed by Nudds (2009, 2010, 2014), I argue that in fact the analysis in terms of co-location is not the most plausible one.

the required relation between locations. Conjunctive analysis tells us that there are some visual elements, some auditory elements, and some spatiotemporal locations, but does not provide any information about which locations are related to which elements and thus which elements have identical or proximal locations. Hence, even if audio–visual binding is mere spatiotemporal co-location or coordination, some additional mode of combination has to be provided in addition to the simple conjunction.

While some experiences caused by real-life ventriloquist performances may be good intuitive examples of audio–visual binding, in experimental psychological works the term 'ventriloquist effect' is mainly used to refer to a different phenomenon that does not usually involve experiencing visual and auditory elements as belonging to a single entity. What such studies show is not an experience of common origin of visual and auditory elements, but rather a cross-modal bias concerning spatial or temporal localization of visual or auditory stimuli (see Bertelson 1999 for a review). More specifically, two forms of ventriloquism, spatial and temporal, are distinguished (Bertelson and de Gelder 2004; Colin et al. 2001; de Gelder and Bertelson 2003). When an auditory element is presented in synchrony with a visual element, there is a tendency to experience the auditory element as positioned closer to the visual element than in cases in which these elements are not presented at the same time. This effect is called 'spatial ventriloquism', as it concerns the modification of an auditory element's spatial location (e.g., Bruns and Getzmann 2008). Analogously, temporal ventriloquism is a phenomenon in which the experienced temporal distance is modified (e.g., Bertelson and Aschersleben 2003). When successive visual and auditory elements are presented in spatial proximity, there is a tendency to experience the visual element as temporally closer to the auditory one, in comparison with cases where the spatial distance between successive elements is larger. In fact, the strongest effect is obtained when visual and auditory stimuli are presented at the same position what often leads to an impression of their simultaneity (see Zampini et al. 2005).

The existence of such phenomena suggests that not only are there audio–visual experiences involving intermodal binding, there is also a broad class of audio–visual experiences presenting intermodal relations. This is because phenomena of spatial and temporal ventriloquism involve experiencing visual and auditory elements as being positioned at some spatial and/or temporal distance. In particular, in some studies participants are explicitly asked to form a judgement relying on perceived spatiotemporal relations between visual and auditory stimuli. For instance, participants are asked whether visual or auditory stimuli were presented first (Bertelson and Aschersleben 2003) or for the spatial position of a sound in relation to a visible element (Bertelson and Aschersleben 1998). In addition, the concerned phenomena demonstrate that experienced spatial relations may modify experienced temporal relations and vice versa. The presence of such intermodal relations provides another reason for believing that not all audio–visual experiences are conjunctively multimodal. As shown in the first section, the complexity introduced by intermodal relations cannot be grasped by characterising the structure of experiences simply as a co-occurrence of auditory and visual elements.

Beyond spatial and temporal ventriloquism, there are also other audio–visual phenomena that suggest the presence of intermodal relations. For instance, it has

been observed that the direction of apparent visual motion is able to influence the direction of the apparent auditory motion (an effect known as 'cross-modal dynamic capture', Sanabria et al. 2005). When visual apparent motion occurs in temporal synchrony with auditory apparent motion (in fact, the best effects are achieved if visual stimuli start to be present slightly earlier than auditory stimuli, Sanabria et al. 2004), it is likely that the direction of auditory motion will be experienced as the same as the direction of the visual motion even if the actual direction of the auditory motion is different. It seems that in such cases one experiences visual and auditory elements as standing in spatiotemporal relations determining that these elements move in the same direction.[6] In addition, there is evidence for the presence of audio–visual priming and synchronous facilitation (see Clark 2011; Vroomen and de Gelder 2000). Audio–visual priming occurs when an element from one modality is processed more efficiently if earlier an element from the second modality was presented in a similar location. Synchronous facilitation concern cases when processing of element from one modality is enhanced by synchronous presentation of an element from the other modality. Such results suggest that human perception is able to represent subsequent auditory and visual elements as positioned in spatial and temporal proximity.[7] Furthermore, it is well-established that the experienced number of visual elements may be influenced by the experienced number of auditory elements. In particular, in the so-called 'sound-induced flash illusion', the presentation of a single visual flash together with two sounds often leads to an experience as of two flashes corresponding to each of the sounds (e.g., Andersen et al. 2004). It seems that in case of such an illusion, experiencing the correspondence between the visual and auditory elements involves experiencing relations of temporal proximity.[8]

One may oppose by proposing that the phenomenal character associated with the above experiences can be analysed without referring to intermodal relations, as it is enough to specify the locations and spatiotemporal properties of objects and events that are visually or auditorily experienced (see O'Callaghan 2015a for a discussion). For instance, it may be claimed that an experience related to spatial ventriloquism in which auditory and visual elements are presented as being close to each other can be properly analysed in terms of the auditorily experienced location L_A, in which the auditory element is positioned, and a distinct, visually experienced location L_V, which is filled in by the visual element. However, such an analysis lacks crucial information regarding the precise spatial relationship between locations L_A and L_V. From the fact that an auditory element occupies a different location to a visual element one can only infer that these elements are not co-located. Nevertheless, the distinctiveness of location is not sufficient to determine whether these locations are

[6] It should be noted that the claim that visual and auditory elements are experienced as moving in the same direction does not entail a more controversial claim that in such a case one experiences an intermodal, audio–visual motion that cannot be reduced to the motions of the involved visual and auditory elements (see Spence 2015; Spence and Bayne 2014).

[7] See Spence (2013) for a review concerning whether various experimental designs provide results suggesting an important influence of spatial relations on multimodal processing.

[8] For further examples concerning audio–visual rhythm perception and intermodal experiences of causality see O'Callaghan (2015a).

close or far from each other. In consequence, distinguishing between phenomenally different experiences presenting auditory and visual elements as being spatially close or far requires referring to intermodal relations in analysing their phenomenal character. Alternatively, one may postulate not intermodal relations but intermodal relational properties. However, the content of such properties already specifies that elements not only jointly co-occur but co-occur in a specific arrangement.

The above considerations show that among audio–visual experiences there are some that present intermodal binding or intermodal relations. The occurrence of these phenomena entails that not all audio–visual experiences can be analysed by referring solely to the conjunctive mode of combination. Relying on this result, we can now consider the additional modes of combination that have to be postulated in order to properly describe the structure of audio–visual phenomenology.

3 Perceptual modes of combination

In this section, I describe four perceptual modes of combination that may be used in characterising the structure of audio–visual experiences involving intermodal binding and intermodal relations. The first is instantiation occurring between individuals and properties. The second is a parthood relation combining elements into mereological wholes. Third is the grouping relation that unifies experienced elements into perceptual groups. Finally, the fourth mode is the relation of perceptual infusion used by O'Dea (2008) to characterise symmetric relations between experienced properties.

Instantiation is a mode of combination by virtue of which individuals, in particular objects and places, posses properties (see Orilla and Swoyer 2016). This mode of combination is common for both unimodal visual and auditory experiences, as both these modalities usually present the environment as containing property-bearing individuals. For instance, visual experiences present figures as having a certain colour and shape, while auditory experiences present sounds as having a certain loudness and pitch (Cohen 2010; Matthen 2004; O'Callaghan 2008). Instantiation is an asymmetric relation: it is a sound that is experienced as possessing a pitch and not the other way around. Furthermore, instantiation involves a mutual dependence,[9] as neither vision nor audition presents uninstantiated, free-floating properties (like colours that do not characterise any objects or places) or propertyless individuals (like sounds without any pitch or loudness).[10] In addition, it is usually the case that many properties of a single individual are co-located at the same place and time. For

[9] More precisely, it is general, and not specific, dependence. This is because properties have to be instantiated by something, but the same property, like colour or shape, can be instantiated by more than one individual.

[10] There are psychological theories suggesting that at early stages of perceptual processing features are represented as unrelated to objects (see Treisman and Gelade 1980: p. 98), or that objects are represented as featureless (see Pylyshyn 2007: p. 52). However, this happens at a subpersonal level and is not reflected in the structure of conscious experiences.

instance, the colour, shape, and size of a red square figure seem to be simultaneously positioned at exactly same location.[11]

An individual, such as object or an event, together with its properties, constitutes a perceptual unit on which various perceptual processes may operate. In particular, an individual together with instantiated properties can be easily chosen by attentional mechanisms and attention tends to automatically spread to fill in the borders of such a perceptual unit (Richard et al. 2008; Scholl 2001). Furthermore, individuals instantiating features can be tracked and re-identified despite movement and qualitative changes (von Marle and Scholl 2003; Pylyshyn 2007; Scholl 2007).

However, human perceptual experiences present not only individuals instantiating properties but also mereological wholes that are constituted by simpler perceptual units (e.g., Hoffman and Richards 1984; Palmer and Rock 1994; Xu and Singh 2002). Such experiences involve a second perceptual mode of combination that organises elements using parthood relations. In the case of visual experiences, parthood mainly organises spatial mereological wholes. For instance, we may experience a rectangular figure as being composed of two square-parts connected by edges. Furthermore, philosophers of perception often claim that in auditory experiences we are presented with sounds that have temporal and not spatial parts (Matthen 2010; O'Callaghan 2008). For example, a complex sound such as a melody is experienced as composed of temporal parts that are simpler sounds differing in features such as pitch or timbre. Elements combined by parthood relations also, as in instantiation, create perceptual units on which attentional processes may operate. However, such operations are likely to be more demanding, as wholes united by parthood have significant structural complexity (e.g., Balaban and Luria 2016; Xu 2006).

Similarly to instantiation, parthood is also an asymmetric relation. When one visually experiences a square as a part of a rectangle, it is not the case that the rectangle is also experienced as a part of a given square. However, parthood does not involve mutual dependence. One cannot visually experience a rectangle as composed of squares without experiencing the presence of these squares, but the squares can be presented in an experience even if they are not experienced as constituting a rectangle (for instance, they may be spatially disjoint).

As stated earlier, properties instantiated by an individual are often co-located by occupying the same place at the same time. This is not typically the case with parts of perceptual mereological wholes that occupy different, but usually proximal, spatial or temporal regions and are separated from neighbouring parts by qualitative edges (Palmer and Rock 1994; Singh and Hoffman 2001). Within visual experiences parts are typically distinguished by relying on discontinuities in surface features like colour or by recognizing points of convexity created by the spatial layout of edges (like in the case of an hour-glass figure, Hoffman and Richards 1984; Tse 1999). In the case of the mereology of auditory experiences, the most important source

[11] The considerations about dependency of instantiation entail only that sounds are always experienced as having some properties. In consequence, they are consistent with results suggesting that sounds may be perceived as unlocalised (see Spence and Driver 2000).

of qualitative discontinuities are differences in pitch between subsequent temporal fragments of complex sounds (Bregman 1994; O'Callaghan 2008).

The third mode of perceptual combination is perceptual grouping. This is most salient in the case of visual experiences where several disjoint elements are likely to be perceived as a single group if they obey some Gestalt laws concerning proximity and similarity (Elder and Goldberg 2002; Kubovy et al. 1998).[12] As in the case of parthood and instantiation, the relation of being a member of a group is asymmetric: perceptual groups are not experienced as being elements of their members. In addition, the relation of being an element of a group does not involve a mutual dependence. This is because the same elements may be experienced both as grouped and as ungrouped depending, for instance, on their spatial layout (Kubovy and Wagemans 1995). However, in contrast to the case of mereological wholes, perceptual groups are not usually composed of spatially or temporally connected elements distinguished by qualitative edges, but by disjoint elements that stand in relations of spatiotemporal proximity. Furthermore, parts composing a mereological whole do not need to obey any of the usual Gestalt principles. For instance, an object can be visually experienced as having parts that are not similar to each other and do not create a symmetric shape. The situation is different in the case of perceptual grouping. To be perceptually grouped, elements usually have to be represented as similar in virtue of sharing some properties, as having a symmetrical spatial layout, or as moving in the same direction with the same velocity (e.g., Ben-Av and Sagi 1995; Hon et al. 1999; Treisman 1982).

Finally, the fourth mode of perceptual combination has been proposed by O'Dea (2008). O'Dea observes that when in unimodal experiences two properties are instantiated by the same individual, they are often also related to each other. For instance, if a figure is square and red, then these properties mutually characterise each other such that squareness is red and redness is square-shaped. This symmetric relation between co-instantiated properties has been called "infusion". In addition to symmetricity, it seems to involve mutual dependence, as properties cannot be experienced without being characterised by certain other properties. For example, one cannot have a visual experience as of a colourful object without experiencing that its colour fills in some shape.

To sum up, it seems that there are at least four perceptual modes of combination that commonly organise unimodal experiences and are different from mere conjunctive co-occurrence. The first is asymmetric instantiation, which creates perceptual units from individuals and properties. It involves a mutual dependence and allows for several properties to be spatiotemporally co-located. The second mode of combination consists in creating mereological wholes through a parthood relation. Similarly to instantiation, it is asymmetric but does not involve mutual dependence. In

[12] It is more difficult to distinguish grouping from parthood in the case of unimodal auditory experiences. Usual psychological examples of auditory grouping concern cases in which successive, simpler sounds are combined into a single, more complex sound (e.g., Kubovy and Schutz 2010). However, such simpler sounds composing a more complex one are also interpreted as temporal parts of a sound (Matthen 2010; O'Callaghan 2008).

addition, parts of perceptual wholes rarely, if ever, occupy exactly the same spatiotemporal location. Usually they are positioned in proximal locations separated by qualitative edges. Perceptual grouping, the third mode of combination, is also asymmetric and like parthood does not involve mutual dependence. However, perceptual groups are usually composed not from connected, but from disjoint elements that obey some Gestalt principles. Finally, infusion is a symmetric relation involving mutual dependence, by virtue of which co-instantiated properties characterise each other.

It should be noted that the perceptual modes of combination allow that elements may be combined in a stronger or a weaker way. For instance, the strength of perceptual grouping depends on the similarity of the experienced elements, and the part-status of an object's fragment can be more or less salient in virtue of a specific arrangement of perceived edges. Even in cases of instantiation, one can experience an object as having a property in a weaker or stronger fashion. For example, phenomena of visual modal and amodal completion provide plausible examples of instantiation in a weaker form. This ability to accommodate combinations of different strengths is important when considering audio–visual experiences as it is likely that intermodal binding can be experienced as being stronger or weaker depending on the specific arrangement of visual and auditory elements.

In the following sections, I consider whether the above modes of perceptual combination can be used in analysing the structure of audio–visual experiences involving intermodal relations and intermodal binding.

4 Audio–visual relations and perceptual grouping

Let's start by considering the structure of audio–visual experiences involving intermodal relations. As argued in Sect. 2, these are experiences involving such phenomena as spatial and temporal ventriloquism, the sound-induced flash illusion, and interactions between auditory and visual apparent motions known as cross-modal dynamic capture. Experiences presenting intermodal relations are not conjunctively multimodal, and so we may ask what additional mode of combination has to be postulated. Of course, it may be claimed that being connected by a spatiotemporal, intermodal relation is in itself a mode of combination different from conjunctive co-occurrence. While this is true, below I argue for the stronger thesis that at least some of the considered phenomena involving intermodal relations are also examples of audio–visual, perceptual grouping. I do not claim that this is necessarily the case about all audio–visual experiences involving intermodal relations, but it is likely to be true about phenomena such as spatial and temporal ventriloquism or cross-modal dynamic capture. It should be noted that the term "grouping" (or "pairing") has been already applied in psychological works to describe these phenomena (e.g., Bertelson 1999; Radeau and Bertelson 1977; Sanabria et al. 2004, see also Spence et al. 2007 for a chronological review of various experimental results interpreted in terms of intermodal grouping). However, it has not been explicitly discussed whether an interpretation in terms of grouping has a stronger justification than an interpretation referring to some other perceptual mode of combination. Below, relying on the

intuitions presented in such psychological works, I present arguments for the thesis that analysis in terms of grouping is more plausible than the alternatives.

The main reason to believe that the considered phenomena are examples of audio–visual grouping is that the involved elements are related in a way that satisfies the general characteristics associated with perceptual grouping. As stated in the previous section, perceptual grouping has several characteristics that jointly distinguish it from other perceptual modes of combinations: (a) grouping usually combines disjoint but proximal elements; (b) it does not involve mutual dependence; (c) the relation between a group and its elements is asymmetric; and (d) strong grouping often demands the presence of Gestalt-like relations between grouped elements.

The first characteristic, concerning the proximity of disjoint elements, is satisfied by the phenomena of ventriloquism and audio–visual cross-modal dynamic capture. In spatial ventriloquism, we experience visual and auditory elements that are spatially disjoint but positioned not far away from each other. Similarly, in temporal ventriloquism, elements are typically not experienced as synchronous but as appearing in succession. Similarly, when visual and auditory apparent motions interact, what is experienced is a series of successively occurring auditory and visual elements, such that subsequent elements are experienced as positioned in proximal, but distinct, locations than the earlier ones.

Furthermore, the mode of combination that occurs in the considered audio–visual phenomena does not involve mutual dependence. For example, the same auditory and visual elements which, when presented in temporal synchrony give rise to spatial ventriloquism, can also figure in an experience separately, without constituting the considered phenomenon. Analogously, visual and auditory elements involved in cross-modal dynamic capture can be experienced independently of each other, in unimodal experiences. It is also very plausible that the relation between the considered audio–visual phenomena and their visual and auditory elements is asymmetric. Quite obviously, if a visual element *A* and an auditory element *B* are elements of a complex audio–visual phenomenon *C*, it is not the case that *C* itself is also an element of *A* or *B*.

Finally, to obtain a strong perceptual grouping, elements should not only be spatiotemporally proximal but should also obey some Gestalt-like laws regarding similarity, or regularities like symmetry or common motion. There is evidence that this condition is also satisfied in the case of some phenomena involving audio–visual intermodal relations. For instance, the strength of influences between auditorily and visually apparent motions depends on whether the number of auditory elements equals the number of visual elements, which constitutes a form of symmetry between two element types (e.g., Sanabria et al. 2004; Spence and Chen 2012). In addition, the congruency between visual and auditory elements also influences audio–visual phenomena involving intermodal relations. In particular, the temporal ventriloquist effect is stronger when the concerned elements are congruent, like a female voice paired with an image of a speaking female face, such that participants are often uncertain as to whether the visual stimuli precedes the auditory one or vice versa (see Vatakis and Spence 2007).

In addition to the fact that the considered audio–visual phenomena satisfy the general characteristics of perceptual grouping, there are two more specific similarities

between them and unimodal cases of grouping. First, the spatial ventriloquist effect observed when auditory and visual elements are experienced in synchrony is not specific to audio–visual phenomena but constitutes a more general feature of perceptual grouping. In particular, it has been shown that as a result of visual perceptual grouping the involved elements are experienced as being spatially closer (Coren and Girgus 1980). Second, alternative unimodal perceptual groupings can compete with each other (e.g., Ben-Av and Sagi 1995). For instance, let's consider a situation in which three types of elements are visually experienced: red circles, red squares, and black circles. In this case, at least two types of grouping can occur: by red colour and by circular shape. However, they are likely to compete with each other such that only one of them can be experienced at a given time. A similar form of competition is present in the case of the considered audio–visual phenomena. For instance, it has been observed that the temporal ventriloquist effect is weaker if the auditory element is grouped with some additional sounds (Bruns and Getzmann 2008).

The above considerations show that phenomena such as spatial/temporal ventriloquism and cross-modal dynamic capture can be plausibly interpreted as cases of intermodal, perceptual grouping. In consequence, not only are there audio–visual experiences that are not conjunctively multimodal, the structure of some of such experiences is actually organised according to grouping principles. In the following sections, I extend this picture by analysing the case of intermodal binding.

5 Instantiation and audio–visual binding

Audio–visual intermodal binding occurs when one has an experience as of visual and auditory elements in some sense belonging to the same entity. For instance, when one experiences a ball hitting the ground, such an event is perceived as involving both auditory and visual elements. While it is very plausible that binding involves some intermodal relations, audio–visual binding, in contrast to the phenomena analysed in the previous section, cannot be easily interpreted as organised according to principles of perceptual grouping. First, a characteristic feature of perceptual grouping is that is occurs between disjoint elements. However, in the case of an intermodal binding, visual and auditory elements are typically experienced as spatiotemporally overlapping (Briscoe 2017; O'Callaghan 2015b; Spence and Bayne 2014). Second, phenomenal reports obtained during psychological experiments differentiate intermodal binding from audio–visual grouping phenomena. In cases of binding, we do not experience the presence of some related but separate auditory and visual elements, but their "audio–visual fusion" or a "common origin" (e.g., Bertelson 1999; Radeau and Bertelson 1977). These reports are supported by neuroscientific investigations showing that the neural correlates of perceptual fusion are significantly different from those responsible for the assessment of spatiotemporal relations crucial for grouping phenomena (Miller and D'Esposito 2005). Third, the empirical data showing the presence of object-specific preview benefits in the case of intermodal binding suggest that binding phenomena involve a greater level of perceptual unity than is associated with perceptual grouping (Jordan et al. 2010; Zmigrod et al. 2009).

Another idea is to characterise audio–visual binding in terms of instantiation (see O'Callaghan 2014b, 2015a, b; Macpherson 2011). A straightforward specification of this idea would be to claim that in the case of intermodal binding, auditory elements such as sounds or visual elements such as objects are properties instantiated by some subject-element. For instance, it has been proposed that in an audio–visual experience a high-pitched note can be experienced as instantiated by a silver cylinder (O'Callaghan 2008). In this case a sound is characterised as the property of a cylinder. In a different example, visually experienced moving lips and a speech sound are characterised as instantiated by a common speech event (O'Callaghan 2015a). In the structure of such an experience both the visually presented object and the auditorily presented sound are experienced as co-instantiated properties of the same event.

Nevertheless, the above proposition has a significant disadvantage. The crucial characteristic of instantiation is that it involves mutual dependence. Properties cannot be experienced as uninstantiated and individuals cannot be experienced as propertyless. However, such dependence is not present in cases of audio–visual binding because the visual and auditory elements may be present without being involved in a binding. For instance, while one can have an audio–visual experience of a ball hitting a surface and making a noise, the two involved elements can also be experienced separately. There is a possible experience as of a hitting noise accompanied by a static visual scene as well as an experience as of a ball hitting a surface without any sound corresponding to the impact.

Similarly, the lack of dependence makes it less plausible that audio–visual binding can be interpreted in terms of perceptual infusion.[13] As stated above, a ball hitting a surface can be visually experienced without an accompanying sound, and vice versa. Furthermore, as suggested by O'Dea (2008), it is less plausible that bound multimodal elements determine each other in a way that is characteristic for unimodal features. For instance, when a colour stands in a relation of perceptual infusion to different shapes, each of these shapes determines the spatial properties of a colour. However, it is less obvious whether the properties of an impact sound significantly modify how a ball hitting a surface is visually experienced.

O'Callaghan recognises that it is problematic to treat auditory elements involved in intermodal binding, like sounds, as properties possessed by some individuals. Quite oppositely, auditory elements themselves seem to be subjects which possess properties (like pitch or timbre). Thus, in some works he characterises audio–visual binding not in terms of subjects instantiating properties, but in terms of parts and wholes (see O'Callaghan (2014a, 2016) for such mereological interpretation). According to this approach, what is experienced in case of binding is an event with a visual proper part and an auditory proper part. For example, when an experience presents a ball hitting a surface, the visually experienced impact is a visual proper part of a hitting event and the produced sound is an auditory proper part of the same event. Adopting this position allows for avoiding the problem connected with mutual dependence that makes the analysis in terms of instantiation implausible.

[13] "Perceptual infusion" is a technical term introduced by O'Dea to name a symmetric relation between co-instantiated properties, see Sect. 3.

It is a characteristic feature of both visual and auditory parts that they can also be experienced as separate entities that do not constitute a larger mereological whole.

However, treating bound visual and auditory elements as proper parts of a complex event has another disadvantage. Let's reconsider an experience in which a ball hits a surface and makes a noise. The visually experienced impact and auditorily experienced noise seem to be spatiotemporally co-located. They are experienced as happening at the same time and the noise seems to be spatially positioned at the place of the impact. However, this is inconsistent with the way in which perceptual parthood typically creates wholes from simpler elements. Proper parts of perceptually experienced wholes are not co-located but are separated by qualitative borders. In the case of visual modality, one experiences spatial parts that are separated by edges designating points of convexity (like in the case of an hour-glass shape) or changes in qualities concerning, *inter alia*, colour or texture (Hoffman and Richards 1984; Paler and Rock 1994; Xu and Singh 2002). Similarly, complex sounds are experienced as having temporal parts that are positioned at successive moments and separated by differences in pitch and other auditory features (Matthen 2010; O'Callaghan 2008).

Nevertheless, I believe that the structure of audio–visual binding can be properly analysed by using the notions of instantiation and parthood in a more nuanced way. In developing my solution, I rely on an intuition expressed in works on auditory perception, namely that audition presents not only sounds and their properties but also properties of sound-producing events and objects (in particular, see Nudds 2009, 2010, 2014). For instance, when an auditory experience presents a rolling sound, then by virtue of experiencing properties of this sound we also experience some properties of a rolling object, like those concerning its velocity and size (Nudds 2014). Let's once again consider our example with an auditory element, namely the impact sound made by a ball hitting a surface. When having an experience involving such a sound, we experience the sound as having some properties, for instance a certain loudness. However, relying on sound's properties, like loudness and spatial features, we also experience the impact-event as having certain properties, for example we experience it as happening with certain force and involving entities with certain sizes. Such relationships between experiencing sounds and experiencing properties of events and objects are common for environmental sounds, and a lack of them is postulated as a characteristic that distinguishes musical pieces as a special category of sounds that can be experienced without experiencing the properties of sound-producing objects and events (e.g., Scruton 2009).

According to the above perspective, it is not the sound itself that is instantiated by an object or an event. Furthermore, the properties instantiated by an object or an event do not have to be the same as the usual properties of sounds such as pitch or loudness. For instance, an auditorily experienced property of a rolling object may be its size or rotation frequency, but not pitch or the rolling sound itself. Nevertheless, the auditorily experienced properties of an object or an event are not independent of the properties of an experienced sound. Rather, they are determined by relying on the auditorily presented properties of a sound. For example, the frequency and periodicity experienced as properties of a sound allow us to determine the speed of a rolling object (see Nudds 2014).

The above approach regarding sound-perception can be easily extended to cases of intermodal, audio–visual binding. When such binding occurs, like when a ball hitting a surface is both seen and heard, there are properties experienced as possessed by the hitting event which are determined relying on the properties of involved auditory and visual elements. The impact sound is experienced as having some properties such as pitch and loudness, and by virtue of them the hitting event is experienced as having certain properties, for instance those related to the force of the impact. Analogously, visually presented elements, such as a moving ball and a surface, are experienced as having certain properties such as size and velocity. Relying on these properties, the impact-event itself seems to be experienced as having properties characterising its spatiotemporal position or its force determined, *inter alia*, by the ball's velocity.

I believe that there are good reasons to postulate that in many cases the attribution of properties to a common event or object, relying on characteristics of visual and auditory elements, happens by virtue of perceptual mechanisms and not only through some higher-order reasoning about the connection between properties. In some cases a connection may indeed be merely conventional and rely on postperceptual reasoning, like in the case of a fire alarm sound that serves as a cue for forming the belief that there is a fire. However, in many situations, with a rolling object serving as a useful example (again, see Nudds 2014), the properties of auditory and visual elements are determined by the properties of a common event or object. For instance, the frequency at which an object rolls determines the temporal regularities in the experienced sound and in the visually perceived motion. Such systematic connections make it more likely that attribution of properties to a common event or object may happen in virtue of mechanisms that do not rely on higher-order reasoning. First, in some cases the perceptually represented properties of auditory and visual elements can be directly attributed to the underlying object or event. For instance, the periodicity of sound and the periodicity of visual motion may be treated as being the same as the periodicity of a movement involved within a common event. Second, there may be statistical correlations between the properties of auditory and visual elements and the properties of common objects or events. For example, in the literature concerning cross-modal correspondences, it is often postulated that the intensity of auditory and visual stimuli is associated with size, or that changes in pitch are associated with an upwards or downwards movement (see Spence 2011). While the mechanisms responsible for cross-modal correspondences are not completely clear, it is likely that there are perceptual mechanisms that can predict the occurrence of a property by relying on the detection of a different property given the data about statistical regularities (see Spence and Deroy 2013 for a review of possible pre-attentive and attentional mechanisms and Spence 2011 for overview of theories that explain such abilities in terms of Bayesian reasoning and early developmental neuroplasticity). Finally, the relevant attributions may happen in virtue of mechanisms of perceptual learning and categorization. Relying on frequently observed relations between visual and auditory properties, the perceptual system may develop categories of events that are likely to be a source of certain combinations of visual and auditory elements (see Lyons 2005; Skrzypulec 2018). While of course it is difficult to strictly delineate high-level perception

from postperceptual reasoning, may authors believe that the representation of some category-related properties lies within the realm of perception (e.g., Di Bona 2013; Siegel 2006).

According to the above view, audio–visual binding consists in experiencing a numerically same entity (event or an object) that instantiates some properties determined by relying on the visual and auditory elements of an experience. In other words, the structure of audio–visual binding is as follows: (a) there are auditory elements instantiating auditory properties; (b) there are visual elements instantiating visual properties; and (c) there is a common entity instantiating properties that are determined by properties of visual and auditory elements.

The fact that the qualitative character of auditory and visual elements involved in intermodal binding allows us to experience the properties of objects and events provides a reason to believe that audio–visual binding cannot be reduced to a spatiotemporal co-localization of visual and auditory elements. This is because in the case of intermodal binding, we do not merely experience auditory and visual elements as positioned in the same place, but also an additional entity with properties determined by relying on the qualitative characters of visual and auditory elements.

The above analysis has several advantages. First, it does not postulate that auditory and visual elements are experienced as properties instantiated by a common entity. In consequence, it preserves an intuition that sounds and objects are auditorily and visually presented as subjects of properties and not as properties of something else. According to the proposed view, what is experienced as instantiated by a common entity are properties determined by relying on qualitative characters of auditory and visual elements. Second, the proposed analysis allows that not all experienced properties of auditory and visual elements are also experienced as properties of a common entity. For instance, in case of a ball hitting a surface, the ball is a visual element with a certain colour, but this colour is not presented as a property of the impact event. Analogously, the impact sound is presented as having a certain pitch but the impact event is not. Third, while visual and auditory elements are not treated as instantiated properties, the analysis preserves the mutual dependence between a common entity and its properties. For example, an impact event cannot be experienced as not having any properties and event-related properties, like the force of an impact, are not experienced as uninstantiated by any event. Fourth, because auditory and visual elements involved in binding are not interpreted as proper parts of a common entity, the fact that they are often experienced as spatiotemporally co-located is not problematic.

Furthermore, the proposed approach allows us to express a difference between weaker and stronger cases of audio–visual binding. Let's consider two experiences: one in which a speech sound is produced by a person with appropriate lip movements, and a second where the same speech sound comes from a loudspeaker. It is likely that the experience of binding would be stronger in the first case, as the visual element is more congruent with the auditory one (see Chen and Spence 2017; Deroy et al. 2016; Jordan et al. 2010; Kubovy and Schutz 2010; Laurienti et al. 2004; O'Callaghan 2015a; Palmer and Ramsey 2012; Vatakis and Spence 2007). When such congruency is present, both visual elements like lip movements, and auditory elements like speech sounds, allow for determining various properties

of the experienced event of speech-production. For instance, both visual and auditory elements provide data concerning how parts of the event are temporally organised. However, when speech sounds come from a loudspeaker, then nearly all data concerning the speech-production event are determined by the auditory element, as the visual properties of a loudspeaker merely allow us to determine the event's spatial location. In consequence, the multimodal character of lips-speech binding is stronger than loudspeaker-speech binding, as only in the first case do both visual and auditory elements significantly contribute to determining the properties of the speech-production event.

6 Parthood and audio–visual binding

While the above analysis was conducted in terms of instantiation, there are reasons to believe that a full account of audio–visual binding also has to include a reference to parthood relations. However, similarly to the case of instantiation, a proper analysis should not simply postulate that the visual and auditory elements are experienced as proper parts of multimodal entity. Let's consider a fast moving object that produces a significant disturbance in the air as its moves and then hits a surface. In such a case one can have an experience as of a complex event involving something travelling with high velocity and producing a whizzing sound and then hitting a surface and producing an impact sound. According to philosophical investigations concerning auditory perception, auditory experiences present complex sounds as having temporal proper parts (e.g., Matthen 2010). In analogy to visually experienced spatial proper parts, which are fragments of objects distinguished by virtue of representing edges and qualitative discontinuities (Hoffman and Richards 1984; Palmer and Rock 1994; Xu and Singh 2002), such temporal proper parts are temporal fragments of sounds which are distinguished by virtue of represented discontinuities in auditory properties (in particular pitch, see O'Callaghan, 2008). It should be noted that according to psychological models of part perception, not every experienced spatial or temporal fragment of an entity can be distinguished as its proper part because distinguishing parts has to rely on some represented discontinuities. For instance, the white interior of a sheet of paper has many spatial fragments, like a circular fragment around its centre, which lack properties that would allow them to be visually discriminated as perceptual proper parts.

Given this account of perceptual parts, the experienced moving/hitting event can be analysed as composed of two temporal proper parts: first when the objects moves and second when it hits the surface. Similarly, as in the case of unimodal auditory and visual parts, these proper parts are fragments distinguished by relying on represented discontinuities in properties determined by the qualitative character of an auditory element, like changes in the sound's characteristics, and by the qualitative character of a visual element, like changes in the way an object moves. This is because each of these two parts is associated with an auditory element (a whizzing sound or an impact sound) and a visual element (a moving object or an object hitting a surface). As a result, each of the event's parts is experienced as possessing some

properties determined by the qualitative characters of respective visual and auditory elements.

The application of the mereological mode of combination is justified in the above case as the event's experiential structure satisfies the most important characteristics of perceptual parthood. First, both proper temporal parts are spatiotemporally proximal but not spatiotemporally co-located as they occur one after the other. Second, these parts are divided by a qualitative border marked by a significant change in properties, determined by the characters of some visual and auditory elements. When an object hits the surface, both the auditory and visual elements included within the experience change rapidly, and in consequence features instantiated by an event are immediately modified. Third, there is no dependence between the event's parts. Movement with a whizzing sound can be experienced without experiencing a subsequent impact and vice versa.

The event considered above is experienced as composed of at least two multimodal, audio–visual proper parts, as each of the parts have properties determined by both the visual and auditory elements of the experience. However, it is not the only mereological variant of audio–visually experienced events. Let's imagine a slightly different case in which an object is at first visually experienced as moving but does not make any sound, and later hits a surface with an impact sound. This event is also experienced as composed of two temporal proper parts, but only the second one is experienced as a multimodal, audio–visual part. The first part is purely visual, as none of its properties are determined by any associated auditory element. Furthermore, there seem to be experiences involving multimodal events with only unimodal parts. For instance, one may have an experience as of a moving object that does not produce any sound. However, just after the object leaves the visual field one hears an impact sound from the appropriate location. In such a case an event is experienced as having two unimodal temporal proper parts associated with distinct modalities: one purely visual part and one purely auditory. Due to this, in some sense it can also be characterised as a multimodal, audio–visual event while it is not experienced as having any multimodal, audio–visual part.

While the mereological analysis seems more natural in the case of audio–visually experienced events than audio–visually experienced objects, there is no a priori reason to refute the presence of objects experienced as having visual and auditory parts. For instance, there may be an object experienced as having two spatial parts such that each of them has properties associated with different visual and auditory elements (e.g., parts look different and simultaneously make distinct sounds). Another variant may be an object that is only partly visible, such that the visible part is not associated with any sound, but the second, non-visible part is experienced by virtue of a produced sound.

The above considerations suggest that the general feature of audio–visually experienced events and objects is that they are experienced as having a part instantiating properties determined by the qualitative character of a visual element and as having a part instantiating properties determined by the qualitative character of an auditory element. This condition can be satisfied by several types of structures. First, there may be events or objects that are experienced as having only one part, i.e. a part identical to the event itself, but this sole part

is multimodal by instantiating properties determined both by visual and auditory elements. For instance, one may have an experience as of a spinning-event involving a uniform, rotating sphere that also produces a constant noise. While such an experience involves an audio–visual binding, there seem to be no qualitative borders between the successive stages of an event, due to a lack of variation in the experienced visual and auditory properties, and so there is no foundation for perceptually dividing such an event into proper temporal parts. Second, there are experiences presenting entities as having proper parts such that at least some of these parts are multimodal. An example may be the event described earlier, in which an object is first visually experienced as moving and auditorily experienced as making a whizzing sound, and is then visually experienced as hitting the surface and auditorily experienced as making an impact sound. Finally, one may have an experience as of an event or object that has unimodal proper parts, but nevertheless these parts are associated with different modalities. An event in which a moving object is experienced in a purely visual fashion and then an impact sound is heard may serve as an example. This last category does not involve an audio–visual binding in the usual sense, as there are no proper parts with both visually and auditorily determined properties. However, such audio–visual experiences can thus still be regarded as multimodal as they present entities with some properties determined by relying on visual elements of the experience and some determined by relying on the auditory elements.

7 Conclusions

In investigating the structure of audio–visual experiences, I have argued that not all such experiences are conjunctively multimodal. This is because there are audio–visual phenomena, involving the presence of intermodal relations and intermodal binding, such that they cannot be properly analysed as merely co-occurrences of auditory and visual elements. Relying on this result, one might ask what the additional mode of combination is that has to be postulated to account for the structure of audio–visual experiences. I claim that some phenomena involving intermodal relations, like spatial and temporal ventriloquism, can be expressed in terms of audio–visual, perceptual grouping. However, cases of intermodal binding need a different treatment. Experiences involving audio–visual binding should be analysed as experiences presenting objects or events which instantiate, or which have a proper part instantiating, both visually and auditorily determined properties.

Acknowledgements The author would like to thank two anonymous reviewers for their comments concerning the earlier versions of the paper. The work was supported by the National Science Center (Poland) Grant 2016/20/S/HS1/00090.

Compliance with ethical standards

Conflict of interest The author declares that he has no conflict of interest.

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology, 14*, 257–262.

Andersen, T. S., Tiippana, K., & Sams, M. (2004). Factors influencing audiovisual fission and fusion illusions. *Cognitive Brain Research, 21,* 301–308.

Arstila, V. (2018). Temporal experiences without the specious present. *Australasian Journal of Philosophy, 96*(2), 287–302.

Balaban, H., & Luria, R. (2016). Integration of distinct objects in visual working memory depends on strong objecthood cues even for different-dimension conjunctions. *Cerebral Cortex, 26*(5), 2093–2104.

Bayne, T. (2008). The unity of consciousness and the split-brain syndrome. *Journal of Philosophy, 105*(6), 277–300.

Ben-Av, M. B., & Sagi, D. (1995). Perceptual grouping by similarity and proximity: experimental results can be predicted by intensity autocorrelations. *Vision Research, 35*(6), 853–866.

Benovsky, J. (2013). The present vs. the specious present. *Review of Philosophy and Psychology, 4*(2), 193–203.

Bertelson, P. (1999). Ventriloquism: A case of crossmodal perceptual grouping. In G. Aschersleben, T. Bachmann, & J. Müsseler (Eds.), *Cognitive contributions to the perception of spatial and temporal events* (pp. 347–362). Amsterdam: Elsevier Science B.V.

Bertelson, P., & Aschersleben, G. (1998). Automatic visual bias of perceived auditory location. *Psychonomic Bulletin and Review, 5*(3), 482–489.

Bertelson, P., & Aschersleben, G. (2003). Temporal ventriloquism: crossmodal interaction on the time dimension 1. Evidence from auditory–visual temporal order judgment. *International Journal of Psychophysiology, 50,* 147–155.

Bertelson, P., & de Gelder, B. (2004). The psychology of multimodal perception. In C. Spence & J. Driver (Eds.), *Crossmodal space and crossmodal attention* (pp. 151–177). Oxford: Oxford University Press.

Block, R. A., & Gruber, R. P. (2013). Time perception, attention, and memory: A selective review. *Acta Psychologica, 149,* 129–133.

Bregman, A. S. (1994). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: The MIT Press.

Briscoe, R. E. (2016). Multisensory processing and perceptual consciousness: Part I. *Philosophy Compass, 11*(2), 121–133.

Briscoe, R. E. (2017). Multisensory processing and perceptual consciousness: Part II. *Philosophy Compass, 12*(12), e12423.

Bruns, P., & Getzmann, S. (2008). Audiovisual influences on the perception of visual apparent motion: Exploring the effect of a single sound. *Acta Psychologica, 129,* 273–283.

Chen, Y.-C., & Spence, C. (2017). Assessing the role of the 'unity assumption' on multisensory integration: A review. *Frontiers in Psychology, 8,* 445. https://doi.org/10.3389/fpsyg.2017.00445.

Clark, A. (2000). *A theory of sentience*. Oxford: Oxford University Press.

Clark, A. (2001). Some logical features of feature integration. In W. Backhaus (Ed.), *Neuronal coding of perceptual systems* (Vol. 9, pp. 3–20)., Series on biophysics and biocybernetics New Jersey: World Scientific.

Clark, A. (2004). Feature-placing and proto-objects. *Philosophical Psychology, 17*(4), 443–469.

Clark, A. (2011). Cross-modal cuing and selective attention. In F. Macpherson (Ed.), *The senses: Classic and contemporary philosophical perspectives* (pp. 375–396). Oxford: Oxford University Press.

Cohen, J. (2010). Sounds and temporality. *Oxford Studies in Metaphysics, 5,* 303–320.

Colin, C., Radeau, M., Deltenre, P., & Morais, J. (2001). Rules of intersensory integration in spatial scene analysis and speechreading. *Psychologica Belgica, 41*(3), 131–144.

Connolly, K. (2014). Making sense of multiple senses. In R. Brown (Ed.), *Consciousness inside and out: Phenomenology, neuroscience, and the nature of experience* (pp. 351–364). Dordrecht: Springer.
Coren, S., & Girgus, J. S. (1980). Principles of perceptual organization and spatial distortion: The gestalt illusions. *Journal of Experimental Psychology: Human Perception and Performance, 6*(3), 404–412.
de Gelder, B., & Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends in Cognitive Sciences, 7*(10), 460–467.
Deroy, O., Chen, Y., & Spence, C. (2014). Multisensory constraints on awareness. *Philosophical Transactions of the Royal Society B, 369,* 20130207.
Deroy, O., Spence, C., & Noppeney, U. (2016). Causal metacognition: Monitoring uncertainty about the causal structure of the world. *Trends in Cognitive Sciences, 20,* 736–747.
Di Bona, E. (2013). Towards a rich view of auditory experience. *Philosophical Studies, 174*(11), 2629–2643.
Elder, J. H., & Goldberg, R. M. (2002). Ecological statistics of gestalt laws for the perceptual organization of contours. *Journal of Vision, 2*(4), 324–353.
Fulkerson, M. (2014). Explaining multisensory experience. In R. Brown (Ed.), *Consciousness inside and out: Phenomenology, neuroscience, and the nature of experience* (pp. 365–373). Dordrecht: Springer.
Hoffman, D. D., & Richards, W. A. (1984). Parts of recognition. *Cognition, 18*(1–3), 65–96.
Holcombe, A. O., & Cavanagh, (2001). Early binding of feature pairs for visual perception. *Nature Neuroscience, 4,* 127–128.
Hon, S., Humphreys, G. W., & Chen, L. (1999). Uniform connectedness and classical gestalt principles of perceptual grouping. *Perception and Psychophysics, 61*(4), 601–674.
Jordan, K., Clark, K., & Mitroff, S. (2010). See an object, hear an object file: Object correspondence transcends sensory modality. *Visual Cognition, 18,* 492–503.
Kanabus, M., Szelag, E., Rojek, E., & Pöppel, E. (2002). Temporal order judgement for auditory and visual stimuli. *Acta Neurobiologae Experimentalis, 62*(4), 263–270.
Kubovy, M., Holcombe, A. O., & Wagemans, J. (1998). On the lawfulness of grouping by proximity. *Cognitive Psychology, 35*(1), 71–98.
Kubovy, M., & Schutz, M. (2010). Audio–visual objects. *Review of Philosophy and Psychology, 1,* 41–61.
Kubovy, M., & Wagemans, J. (1995). Grouping by proximity and multistability in dot lattices: A quantitative gestalt theory. *Psychological Science, 6*(4), 225–234.
Kubovy, M., & Yu, M. (2012). Multistability, cross-modal binding and the additivity of conjoined grouping principles. *Philosophical Transactions of The Royal Society B Biological Sciences, 367,* 954–964.
Laurienti, P. J., Kraft, R. A., Maldjian, J. A., Burdette, J. H., & Wallace, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research, 158,* 405–414.
Lyons, J. (2005). Perceptual beliefs and nonexperiential looks. *Philosophical Perspectives, 19*, 237–256.
Macpherson, F. (2011). Cross-modal experiences. *Proceedings of the Aristotelian Society, 111,* 429–468.
Matthen, M. P. (2004). Features, places, and things: Reflections on Austen Clarks's theory of sentience. *Philosophical Psychology, 17*(4), 497–518.
Matthen, M. (2010). On the diversity of auditory objects. *Review of Philosophy and Psychology, 1,* 63–89.
Miller, L. M., & D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *The Journal of Neuroscience, 25*(25), 5884–5893.
Mitroff, S., Scholl, B., & Wynn, K. (2005). The relationship between object files and conscious perception. *Cognition, 96,* 67–92.
Nudds, M. (2009). Sounds and space. In M. Nudds & C. O'Callaghan (Eds.), *Sounds and perception. New philosophical essays* (pp. 69–96). Oxford: Oxford University Press.
Nudds, M. (2010). What are auditory objects? *Review of Philosophy and Psychology, 1*(1), 105–122.
Nudds, M. (2014). Auditory appearances. *Ratio, 27,* 462–482.
O'Callaghan, C. (2006). Shared content across perceptual modalities: Lessons from cross-modal illusions. *Electroneurobiología, 14*(2), 211–224.
O'Callaghan, C. (2008). Seeing what you hear: Cross-modal illusions and perception. *Philosophical Issues, 18,* 317–338.

O'Callaghan, C. (2011). Lessons from beyond vision (sounds and audition). *Philosophical Studies, 153,* 143–160.
O'Callaghan, C. (2012). Perception and multimodality. In R. Samuels & S. Stich (Eds.), *The Oxford handbook of philosophy of cognitive science* (pp. 92–117). Oxford: Oxford University Press.
O'Callaghan, C. (2014a). Audible independence and binding. In D. J. Bennett & C. S. Hill (Eds.), *Sensory integration and the unity of consciousness* (pp. 73–103). Cambridge: The MIT Press.
O'Callaghan, C. (2014b). Intermodal binding awareness. In D. J. Bennett & C. Hill (Eds.), *Sensory integration and the unity of consciousness* (pp. 73–103). Cambridge, MA: MIT Press.
O'Callaghan, C. (2015a). Not all perceptual experience is modality specific. In D. Stokes, M. Matthen, & S. Biggs (Eds.), *Perception and its modalities* (pp. 133–165). Oxford: Oxford University Press.
O'Callaghan, C. (2015b). The multisensory character of perception. *Journal of Philosophy, 112*(10), 551–569.
O'Callaghan, C. (2016). Objects for multisensory perception. *Philosophical Studies, 173,* 1269–1289.
O'Callaghan, C. (2017). Grades of multisensory awareness. *Mind and Language, 32*(2), 155–181.
O'Dea, J. (2008). Transparency and the unity of experience. In E. Wright (Ed.), *The case for qualia* (pp. 299–308). Cambridge, MA: The MIT Press.
Orilla, F. & Swoyer, C. (2016). Properties. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2016 Edition), https://plato.stanford.edu/archives/win2016/entries/properties/. Accessed 20 July 2018.
Palmer, T. D., & Ramsey, A. K. (2012). The function of consciousness in multisensory integration. *Cognition, 125,* 353–364.
Palmer, S., & Rock, I. (1994). Rethinking perceptual organization: The role of uniform connectedness. *Psychonomic Bulletin and Review, 1*(1), 29–55.
Power, S. E. (2012). The metaphysics of the 'specious' present. *Erkenntnis, 77*(1), 121–132.
Pylyshyn, Z. W. (2007). *Things and places. How the mind connects with the world?*. Cambridge: The MIT Press.
Radeau, M., & Bertelson, P. (1977). Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations. *Perception and Psychophysics, 22*(2), 137–146.
Richard, A. M., Lee, H., & Vecera, S. P. (2008). Attentional spreading in object-based attention. *Journal of Experimental Psychology: Human Perception and Performance, 34*(4), 842–853.
Richardson, L. (2014). Non sense-specific perception and the distinction between the senses. *Res Philosophica, 91*(2), 215–239.
Sanabria, D., Soto-Faraco, S., Chan, J. S., & Spence, C. (2004). When does visual perceptual grouping affect multisensory integration? *Cognitive, Affective, and Behavioral Neuroscience, 4*(2), 218–229.
Sanabria, D., Soto-Faraco, S., Chan, J. S., & Spence, C. (2005). Intramodal perceptual grouping modulates multisensory integration: evidence from the crossmodal dynamic capture task. *Neuroscience Letters, 377,* 59–64.
Santangelo, V., Ho, C., & Spence, C. (2008). Capturing spatial attention with multisensory cues. *Psychonomic Bulletin and Review, 15*(2), 398–403.
Scholl, B. J. (2001). Objects and attention: The state of art. *Cognition, 80*(1–2), 1–46.
Scholl, B. J. (2007). Object persistence in philosophy and psychology. *Mind and Language, 22*(5), 563–591.
Scruton, R. (2009). Sounds as secondary objects and pure events. In M. Nudds & C. O'Callaghan (Eds.), *Sounds and perception. New philosophical essays* (pp. 50–68). Oxford: Oxford University Press.
Siegel, S. (2006). Which properties are represented in perception. In T. S. Gendler & J. Hawthorne (Eds.), *Perceptual experience* (pp. 481–503). Oxford: Oxford University Press.
Singh, M., & Hoffman, D. D. (2001). Part-based representations of visual shape and implications for visual cognition. In P. Kellman & T. Shipley (Eds.), *From fragments to objects: segmentation and grouping in vision* (pp. 401–459). Amsterdam: North-Holland, Elsevier Science.
Skrzypulec, B. (2018). Perceptual kinds as supervening sortals. *Pacific Philosophical Quarterly*. https://doi.org/10.1111/papq.12253.
Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, and Psychophysics, 73*(4), 971–995.
Spence, C. (2013). Just how important is spatial coincidence to multisensory integration? Evaluating the spatial rule. *Annals of the New York Academy of Sciences, 1296,* 31–49.

Spence, C. (2015). Cross-modal perceptual organization. In J. Wagemans (Ed.), *The Oxford handbook of perceptual organization* (pp. 639–654). Oxford: Oxford University Press.
Spence, C., & Bayne, T. (2014). Is consciousnes multisensory? In D. Stokes, M. Matthen, & S. Biggs (Eds.), *Perception and its modalities* (pp. 95–132). New York: Oxford University Press.
Spence, C., & Chen, Y.-C. (2012). Intramodal and crossmodal perceptual grouping. In B. E. Stein (Ed.), *The new handbook of multisensory processing* (pp. 265–282). Cambridge, MA: MIT Press.
Spence, C., & Deroy, O. (2013). How automatic are crossmodal correspondences? *Consciousness and Cognition, 22,* 245–260.
Spence, C., & Driver, J. (2000). Attracting attention to the illusory location of a sound: reflexive cross-modal orienting and ventriloquism. *NeuroReport, 11,* 2057–2061.
Spence, C., Sanabria, D., & Soto-Faraco, S. (2007). Intersensory Gestalten and crossmodal scene perception. In K. Noguchi (Ed.), *Psychology of beauty and Kansei: New horizons of Gestalt perception* (pp. 519–579). Fuzanbo International: Tokyo.
Stevenson, L. (2000). Synthetic unities of experience. *Philosophy and Phenomenological Research, 60*(2), 281–305.
Treisman, A. (1982). Perceptual grouping and attention in visual search for features and for objects. *Journal of Experimental Psychology: Human Perception and Performance, 8*(2), 194–214.
Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136.
Tse, P. U. (1999). Complete mergeability and amodal completion. *Acta Psychologica, 102*(2–3), 165–201.
Tye, M. (2003). *Consciousness and persons*. Cambridge, MA: MIT Press.
Tye, M. (2007). The problem of common sensibles. *Erkenntnis, 66*(1–2), 287–303.
Usher, M., & Donnelly, N. (1998). Visual synchrony affects binding and segmentation in perception. *Nature, 394,* 179–182.
Vatakis, A., & Spence, C. (2006). Audiovisual synchrony perception for music, speech, and object actions. *Brain Research, 1111,* 134–142.
Vatakis, A., & Spence, C. (2007). Crossmodal binding: Evaluating the 'unity assumption' using audio-visual speech stimuli. *Perception and Psychophysics, 69*(5), 744–756.
von Marle, K., & Scholl, B. J. (2003). Attentive tracking of objects versus substances. *Psychological Science, 14*(5), 498–504.
Vroomen, J., & de Gelder, B. (2000). Sound enhances visual perception: Cross-modal effects of auditory organization on vision. *Journal of Experimental Psychology: Human Perception and Performance, 26*(5), 1583–1590.
Warren, R. M., & Obusek, C. J. (1972). Identification of temporal order within auditory sequences. *Perception and Psychophysics, 12,* 86–90.
Xu, Y. (2006). Understanding the object benefit in visual short-term memory: The roles of feature proximity and connectedness. *Perception and Psychophysics, 68*(5), 815–828.
Xu, Y., & Singh, M. (2002). Early computation of part structure: evidence from visual search. *Perception and Psychophysics, 64*(7), 1039–1054.
Zampini, M., Guest, S., Shore, D. I., & Spence, C. (2005). Audio–visual simultaneity judgments. *Perception and Psychophysics, 67*(3), 531–544.
Zmigrod, S., Spapé, M., & Hommel, B. (2009). Intermodal event files: integrating features across vision, audition, taction, and action. *Psychological Research, 73,* 674–684.