An Object-based Interpretation of Audiovisual Processing

Adrian KC Lee[1], Ross K Maddox[2], and Jennifer K Bizley[3]

[1]Department of Speech and Hearing Sciences, Institute for Learning & Brain Sciences (I-LABS), 1715 Columbia Road NE, Portage Bay Bldg. Room 206, University of Washington, Box 357988, Seattle, WA 98195-7988, USA; akclee@uw.edu; +1 206-616-0102

[2] Departments of Neuroscience and Biomedical Engineering, and Del Monte Institute for Neuroscience, University of Rochester, 201 Robert B. Goergen Hall, Box 270168, Rochester, NY, 14627, USA; ross.maddox@rochester.edu; +1 585-275-1835

[3] Ear Institute, University College London, 332 Gray's Inn Road, London, WC1X 8EE, UK; j.bizley@ucl.ac.uk; +44 77 7914 8804

Keywords: multisensory, crossmodal, sensory integration, binding, object-based attention, scene analysis, ventriloquism, sound-induced flash illusion, McGurk illusion

Corresponding author: Adrian KC Lee

Running title: Audiovisual objects

Abstract

Visual cues help listeners to follow conversation in a complex acoustic environment. Many audiovisual research studies focus on how sensory cues are combined to optimize perception, either in terms of minimizing the uncertainty in the sensory estimate or maximizing intelligibility, particularly in speech understanding. From an auditory perception perspective, a fundamental question that has not been fully addressed is how visual information aids the ability to select and focus on one auditory object in the presence of competing sounds in a busy auditory scene. In this chapter, audiovisual integration is presented from an object-based attention viewpoint. In particular, it is argued that a stricter delineation of the concepts of multisensory integration versus binding would facilitate a deeper understanding of the nature of how information is combined across senses. Furthermore, using an object-based theoretical framework to distinguish binding as a distinct form of multisensory integration generates testable hypotheses with behavioral predictions that can account for different aspects of multisensory interactions. In this chapter, classic multisensory illusion paradigms are revisited and discussed in the context of multisensory binding. The chapter also describes multisensory experiments that focus on addressing how visual stimuli help listeners parse complex auditory scenes. Finally, it concludes with a discussion of the potential mechanisms by which audiovisual processing might resolve competition between concurrent sounds in order to solve the cocktail party problem.

1 Introduction

There are many different perspectives on how auditory and visual information can be combined to influence our perception of the world. Chapter 2 by Alais and Burr focuses on how the cues in each of these sensory modalities could be optimally combined in order to maximize perceptual precision through the lens of Bayesian modeling. From a communication perspective, Grant and Bernstein in Chap. 3 describe how speech intelligibility could be altered by the presence of visual information, especially in situations where the auditory signal is embedded in masking noise. The focus of this chapter takes on yet another perspective: does visual information help to segregate sounds in a mixture, and if so: how?

1.1 Multisensory Cocktail Party: Disambiguating Sound Mixtures using Visual Cues

Most normal-hearing listeners can recognize what one person is saying when others are speaking at the same time. This is the classic “cocktail party problem” as defined by Cherry more than six decades ago (Cherry 1953; Middlebrooks et al. 2017). Current state-of-the-art machine learning algorithms still struggle with this auditory scene analysis problem, yet the brain exploits the statistics of natural sound to accomplish this task comparatively easily. Psychoacoustic studies in the past decades have shown that sound elements are likely to be grouped together to form an auditory stream when they are harmonically related to each other (Culling and Stone 2017), temporally coherent with one another (Shamma et al. 2011), or share common spatial cues across time (Maddox and Shinn-Cunningham 2012). All of these past studies examined how the brain solves the cocktail party problem using auditory cues alone (and see Lee 2017 for a summary of the human neuroimaging efforts to understand the listening brain). It is noteworthy to point out that Cherry in his original paper highlighted “lip-reading” as a potential component of the cocktail party problem’s solution

(Cherry 1953). Even though visual cues are usually present and can potentially be used to separate sound mixtures, this process is considerably less explored in behavioral listening experiments. Since visual cues in conversation are often intimately linked to speech reading—using a talker's lip and articulator movements and other facial expressions to better understand conversation (see Grant and Bernstein, Chap. 3)—many studies have focused on characterizing the increase in speech intelligibility when a listener can see a talker's face. This chapter, instead, focuses on how visual cues help a listener to better segregate and select a sound of interest from a mixture.

1.2 Object-based Attention

Auditory and visual information propagate in different physical forms, and reach the brain via coding in different sensory epithelia, but features across these sensory modalities are seamlessly bound to create a coherent percept. Binding stimulus features from a common source is not a unique problem across sensory modalities—within a modality, independently encoded perceptual features (e.g., color, shape and orientation in vision; pitch, timbre, and spatial cues in audition) must also be combined to form a single perceptual object. These perceptual objects are the "units" on which attention operates, both in vision (Desimone and Duncan 1995) and audition (Shinn-Cunningham et al. 2017). Given that there is a central limitation in the amount of information the brain can process, attention helps to determine what object(s) the brain analyzes in order to make sense of a complex scene.

As elaborated below, audiovisual binding can be viewed through the lens of object-based attention, extending theories that have been well developed in each sensory modality. For example, watching the bowing action of the first violin in an orchestra can help a listener pick out the string melody. Conversely, it is more difficult to follow the violin melody if one instead

watches the timpanist's mallets. This example illustrates two fundamental aspects of object-based attention, namely 1) attending to one feature of an object (visual motion of the bow) automatically enhances another feature of the same object (the auditory melody produced by the bow), and 2) there is a cost associated with dividing attention across objects (listening to the violin melody while watching the timpanist). An *a priori* problem the brain must solve is determining which sets of features belong to which object.

1.3 The Auditory Perspective

From an auditory perspective, one of the most important questions to address is how and to what extent visual information can help someone listen, especially in a crowded environment. In this chapter, an object-based perspective of multisensory processing will be presented to account for the visual benefits in auditory perception, especially when the auditory signal of interest is embedded in other competing sounds. In particular, multisensory integration will be defined as any integration of information across sensory domains, while the term multisensory binding will be reserved for the binding of auditory and visual information into a single multisensory object. We propose that such a multisensory object is more easily segregable from competing stimuli, which allows us to distinguish binding as a distinct form of multisensory integration. While resolving competition between sensory objects is a basic problem the brain solves readily in the natural environment, laboratory procedures that employ competing stimuli are rare, with most focusing on characterizing the interactions between two individual (or, rarely, streams of) auditory and visual stimuli.

The most common way to study multisensory interaction is with multisensory illusions: experimenters present multisensory stimuli with a conflict between the modalities, and measure the change in perception compare to presenting either of these stimuli in a single

modality. Reports of these illusory percepts have been often discussed as a demonstration of multisensory integration and multisensory binding, without consideration of their differences. Among the most widely used paradigms, the first is the ventriloquist illusion whereby the location of a sound is “captured” by the visual stimulus (Howard and Templeton 1966). The second is the sound-induced flash illusion (Shams et al. 2000) in which the number of visual flashes reported is influenced by the number of rapidly presented auditory stimuli. Finally, the McGurk illusion (McGurk and MacDonald 1976) in which visual mouth movements for a given syllable are paired with an incongruent auditory syllable, resulting in the percept of a third syllable (e.g., a visual /ga/ and an auditory /ba/ can result in a /da/ percept). The findings from experiments eliciting these illusions are discussed in the context of multisensory binding in Sect. 3.

Furthermore, in contrast to the illusion paradigms outlined above that generally employ brief auditory and visual stimuli, this chapter highlights the benefits of using longer, dynamic stimuli in multisensory experiments. It is hoped that such experiments will provide a new perspective on how visual information can help listeners solve the cocktail party problem.

2 Visual, Auditory and Audiovisual Objects

A red car moving to the left, a trumpet riff in a jazz band—these are examples of visual and auditory objects that are not difficult to conceptualize. Intuitively, a visual or auditory object is a perceptual entity that is subjectively perceived to originate from one physical source. Yet, a precise (and concise) definition of what makes a visual (Feldman 2003) or auditory (Bizley and Cohen 2013) object is difficult to pin down, perhaps owing to sensitivities to specific stimulus factors and context (as described in Sect. 2.1.2) that contribute to how

objects are formed. Nevertheless, it is generally accepted that visual and auditory attention operates on objects (Desimone and Duncan 1995; Shinn-Cunningham et al. 2017).

Two classic studies elegantly demonstrate the fundamental attributes of object-based attention introduced above. In a functional magnetic resonance imaging (fMRI) study, O'Craven and colleagues (1999) showed visual stimuli consisting of a face transparently superimposed on a house, with one moving and the other stationary and asked the subjects to attend to either the house, the face or the motion. They showed that attending to one feature of an object (e.g., the motion of a moving house) enhanced not only the neural representation of that feature (i.e., motion) but also of the other feature of the same object (i.e., the house), compared with features of the other object (i.e., the face).

In a psychophysical study, Blaser and colleagues (2000) asked subjects to track and make judgments about a Gabor patch—a staple stimulus in vision studies consisting of a sinusoidal alternation in space between high and low luminance, also known as gratings, smoothed by a 2-D Gaussian window—that dynamically changed its features (viz., orientation, spatial frequency, and color saturation) in the presence of another competing Gabor patch at the same location, but with its features changed according to different temporal trajectories. Not only could observers track one Gabor patch in the presence of the other, they also reported that the target Gabor patch was more salient than the competing one, not dissimilar to figure-ground segmentation. Furthermore, when observers were asked to make two judgments on feature perturbations introduced to specific features of these stimuli, they performed worse when these perturbations were divided *across* the two Gabor patches (e.g., reporting a color perturbation in one Gabor patch and the orientation perturbation of the other) compared to when they were *within* the same Gabor patch (e.g., reporting the color and orientation perturbation of the same Gabor patch).

Figure 1A provides a sketched illustration of the benefits bestowed on reporting features from one object compared to across two objects, modified from the original study by Behrmann and colleagues (1998). The displays are made up of two overlapping rectangles, with a set of “cuttings” (or features) that looks as if one or two triangles have been cut away from one of the four possible edges of the X figure. These cuttings appear either at the ends within the same rectangle (Fig. 1A top row) or across two rectangles (Fig. 1A bottom row). The features had either the same (Fig. 1A left column) or different (Fig. 1A right column) number of cuttings. Consistent with object-based attention, subjects could perform this same/different psychophysics task faster and without loss of accuracy when the “cuttings” appeared on the same object (Fig. 1A top row) compared with features spread across two objects (bottom row).

These concepts translate intuitively to the auditory domain. For example, selectively listening to a female talker in the presence of a male talker will enhance all the features of the female’s voice (e.g., prosody and phonation). The ability to judge temporal relationships across two sounds is impaired when those sound elements belong to two separate streams (Fig 1.B, bottom row) instead of a single one (Cusack and Roberts 2000). Indeed, the ability to identify deviations from an isochronous (i.e., equally paced) rhythm within the same auditory stream but not across two separate streams is often leveraged in psychoacoustics paradigms to objectively measure stream segregation (Micheyl and Oxenham 2010).

=== Figure 1 near here ===

The notion that attending to a feature belonging to one object will enhance that object’s other features leads to a difficult question: how are these features bound to form an object in the first place? This feature-binding problem still vexes both psychophysicists and

neurophysiologists. One of many hypotheses is the temporal coherence theory of object formation (Shamma et al. 2011). This theory is based on the observation that the sound features (e.g., pitch, timbre, loudness, spatial location) associated with one source would be present whenever the source is active and absent when it is silent—features within the same source will be modulated coherently through time. Furthermore, different sound sources (and their associated features) will fluctuate according to their own time courses, which are independent of those of other sources. However, it is still unclear whether and how the coherence between neural populations is computed on a neural level. Nevertheless, temporal coherence across neural populations as a way to solve the binding problem can also be extended to other sensory domains where there is a natural temporal fluctuation of objects in the scene, but a caveat must be applied for the case of static visual scenes (which an observer is well able to segment into its constituent objects; Treisman 1998). The temporal coherence model could account for how observers process dynamic visual scenes (Alais et al. 1998; Blake and Lee 2005), or even multisensory stimuli.

An audiovisual object can functionally be defined as “a perceptual construct which occurs when a constellation of features are bound within the brain” (Bizley et al. 2016a). Most naturally occurring audiovisual objects have auditory and visual features that evolve coherently in time, dictated by the physical nature of the source. For example, mouth shape must covary with the dynamic acoustics of speech because it is the physical configuration of the speech articulators that determines the sounds being produced (Chandrasekaran et al. 2009). If one is watching the trumpet player generating the riff in the jazz band it is likely that one will see the player move the instrument and body in rhythm with the playing, providing temporal coherence between the visual and auditory scenes.

2.1 Integration Versus Binding

While many multisensory investigators use the terms “integration” and “binding” interchangeably (Stein and Stanford 2008), defining and distinguishing these two terms can provide clarity when attempting to distinguish changes in sensory representation from other interactions, such as combining independent sensory information at the decision-making stage—such clarity will be important in behavioral as well as neurophysiological studies (Bizley et al. 2016a). Specifically, multisensory integration can be defined as any process in which information across sensory modalities is combined to make a perceptual judgment, whereas multisensory binding should be reserved to describe a specific form of integration in which perceptual features are grouped into a unified multisensory object. In other words, binding is a form of integration; however, integrating information at the decision stage, for example, is not a form of binding.

Here is an example to illustrate these two concepts: at a 100-meter dash, the starting pistol is an important audiovisual stimulus that marks the beginning of the race. The runners in the race are concerned with *when* the gun was fired—they are thus likely to jump off the starting block when they hear the sound of the gun, rather than see the flash, because auditory stimuli generally provide more precise temporal information.[1] However, someone in the audience who couldn’t see *where* the pistol was before it was fired would be cued to its exact location by the visual flash, because visual stimuli provide much better spatial information. As discussed by Alais and Burr in Chap. 2, weighing evidence from each sensory system by their reliability—specifically temporal in audition, spatial in vision—to

[1] This example is best understood if acoustic propagation delay is ignored—modern track and field competitions use a loudspeaker mounted on each starting block, making that a practical reality.

reach a decision is an example of how multisensory integration is shaped by current behavioral demands.

From the above example, it is unlikely that an observer would ever perceive the sound of the gunshot and the motion of the athlete as features of a unified, multisensory object. However, an observer would likely associate the gun's sound and flash as sensory events that "go together"—after all, these two pieces of sensory information originate from the same location at the same time. What factors influence an observer to report that different sensory events "go together" and how can experimenters test whether perceptual features across sensory modalities truly are bound together into a multisensory object?

2.1.1 Unity Assumption

In the multisensory literature, a hypothesis known as the "unity assumption" posits a process in which an observer considers whether various unisensory stimuli originate from the same object or event (Welch and Warren 1980; Chen and Spence 2017). The degree to which observers infer these unisensory inputs as belonging together can be influenced by stimulus statistics, such as spatial and temporal coincidence, and other top-down influences, such as prior knowledge, context and expectations. Conceptually, the "unity assumption" provides an intuitive way to probe multisensory binding—based on one's belief, is there evidence that different sensory information should be grouped together to form a cohesive object? However, empirical evidence to support the unity effect is contentious, owing in part to a confluence of factors listed above. Furthermore, it remains unclear whether the unity assumption requires conscious belief of the observer or just an implicit assessment that the multisensory inputs belong together. Instead, many studies in the past few decades have focused on the individual factors that influence this unity assumption.

2.1.2 Stimulus Factors Guiding Multisensory Integration

Based on the findings of electrophysiological studies at the neuronal level in the deep layers of the superior colliculus (SC)—a convergence zone of multisensory inputs—three stimulus factors are thought to influence multisensory integration. The first two factors are concerned with whether sensory inputs are spatially and temporally proximal. Typically, multisensory stimuli that are close in space and time would lead to the largest enhancement in neuronal response (Stein and Stanford 2008) and these guiding principles are often referred to as the spatial and temporal rule, respectively. The third factor—inverse effectiveness—postulates that the crossmodal effect is maximal when at least one of the unisensory inputs is only weakly informative when presented on its own.

On the surface, behavioral studies seem to demonstrate that these neural observations extend well to the perceptual domain. For example, in agreement with the inverse effectiveness principle, visual cues are most useful when auditory targets are embedded in environments at low signal-to-noise ratios. There are also many studies that show behavioral facilitations when stimuli are presented close together in time and space (see Wallace and Stevenson 2014 for a review). However, upon closer inspection, the effects of spatial colocation and temporal coincidence across modalities can be both subtle and highly task-dependent at the psychophysical level.

2.1.2.1 Spatial Colocation

According to the spatial rule, multisensory integration is maximal when stimuli from different sensory modalities are presented in the same spatial location, i.e., spatial coincidence facilitates multisensory integration. From a neuronal perspective, particularly in relation to the orienting role of the SC with respect to multisensory integration, this spatial rule

makes intuitive sense—each multisensory neuron has multiple excitatory receptive fields and maximal neuronal gain would occur when these receptive fields align spatially (but see Willett, Groh, and Maddox, Chap. 5 about the issue of reference frame alignment). However, evidence from behavioral studies suggests that spatial colocation has more of a consistent effect on tasks involving spatial attention or tasks in which space is somehow relevant to the participant's task compared to other non-spatial tasks (Spence 2013). For example, Harrington and Peck (1998) found that human saccadic reaction time was faster when bimodal auditory and visual stimuli were presented together compared to when they were presented alone in each modality suggesting that there was an enhancement in multisensory integration. Furthermore, they found that saccadic latency increased as spatial distance between the auditory and visual targets increased, supporting the idea that behavioral enhancement is maximum when there is spatial correspondence across sensory modalities. This behavioral benefit extends to spatial cueing studies in which subjects are cued to covertly attend (i.e., without saccading to the target location and hold fixation). In general, subjects respond more rapidly and more accurately when the cue and target are presented from the same rather than opposite sides of fixation (Spence and McDonald 2004) and these results can be interpreted either in terms of the spatial rule, or that there is a robust link in crossmodal spatial attention (Spence and Driver 2004).

However, when subjects perform a non-spatial task in which they had to either identify the target stimuli and / or report the temporal content, spatial colocation seems to become unimportant for multisensory integration. For example, in a visual shape discrimination task, performance of the subjects improved when a sound was presented simultaneously along with the visual stimulus, but this improvement was present regardless of whether the location of the sound matched that of the visual stimulus (Doyle and Snowden 2001). Spatial

colocation also generally seems not to play a significant role in modulating multisensory integration in many of the classic audiovisual illusion paradigms such as the McGurk effect (e.g., Colin et al. 2001), and flash-beep illusion (e.g., Innes-Brown and Crewther 2009; Kumpik et al. 2014). While there are examples of non-spatial tasks where integration is modulated by spatial co-location (e.g., Bizley et al. 2012), many of these exceptions required that subjects deploy spatial attention to resolve stimulus competition.

2.1.2.2 Temporal Coincidence

Similar to the spatial rule, the temporal rule was derived from the temporal tuning functions of individual SC neurons (Meredith et al. 1987)—the gain of a multisensory unit is maximal when the stimulus onset asynchrony is minimal (i.e., temporally coincident). Behaviorally, multisensory enhancement has been shown by an increase of stimulus sensitivity when accompanied by a task-irrelevant, but synchronous, stimulus presented in another sensory modality. In one study, auditory signals were better detected when accompanied by a synchronous, though task-irrelevant, light flash (Lovelace et al. 2003). Analogously, in another study, visual sensitivity was only enhanced when the accompanied sound was presented simultaneously and not when the acoustic stimulus was presented 500 ms preceding the visual stimulus. Synchrony between a non-spatialized tone ‘pip’ can also make a visual target ‘pop’ out in cluttered displays. In the ‘pip and pop’ paradigm, subjects are tasked to search for a visual target (e.g., defined by an orientation) when an array of visual elements is flickered repeatedly and asynchronously with respect to one another. This is often a difficult visual search task because of the clutter of surrounding elements. However, when a tone pip is presented in synchrony with an abrupt temporal change of the visual target, subjectively this makes the target pop out and the visual search becomes quick

and efficient (Van der Burg et al. 2008), even if the auditory signal is not spatialized and provides no information about where to search in the visual scene.

Is perfect temporal alignment a prerequisite for these multisensory enhancements? So long as crossmodal stimulus pairs are presented in close temporal proximity, audiovisual integration can accelerate reaction time (Colonius and Diederich 2004) as well as improve speech perception (see Grant and Bernstein, Chap. 3). Furthermore, when subjects are asked to make subjective simultaneity judgments of an auditory-visual stimulus pair that is presented with various stimulus onset asynchronies, they are likely to report that these stimuli are simultaneous even with delays of a hundred milliseconds or more (Wallace and Stevenson 2014). This leads to the concept of temporal window of integration (also known as temporal binding window) and see Baum and Wallace, Chap. 12 using this construct to probe multisensory dysregulation in different developmental disorders. On a descriptive level, this time window describes probabilistically whether information from different sensory modalities will be integrated (Colonius and Diederich 2010). This temporal window differs in width depending on the stimuli, with it being narrowest for simple flash-beep stimuli and widest for complex multisensory speech stimuli (Wallace and Stevenson 2014). Estimation of the width of these temporal windows also varies markedly across subjects and its variability can be linked to individuals' susceptibility to audiovisual illusions (Stevenson et al. 2012).

2.1.2.3 Context Influencing Multisensory Integration

When multisensory stimuli have congruent low-level cues, as when each sensory stimulus is either spatially or temporally proximal, many studies have observed behavioral benefits, mostly attributed to the process of multisensory integration. But is there a crossmodal benefit when multisensory stimuli are congruent at a higher-level of cognitive

representations; for example, does showing a picture of a dog influence one's perception of a barking sound compared to showing a picture of a guitar? Many behavioral and electrophysiological studies have shown some form of behavioral enhancement or modulated brain activity by this type of semantic congruency (Laurienti et al. 2004; Thelen et al. 2015). These studies generally postulate that semantic congruency can lead to a bound multisensory object due to the learned associations between the individual sensory elements of a single event based on the unity assumption argument. However, even proponents of the unity assumption argument point out that most studies of semantic congruency's effect on multisensory integration have used unrealistic stimuli lacking ecological validity. Furthermore, the semantic congruency effect seems to be maximal when the auditory stimulus precedes the visual by a few hundred milliseconds as opposed to when they are presented simultaneously (Chen and Spence 2013). Thus, rather than attributing the congruency effect to binding, a more parsimonious explanation is simply semantic priming of one stimulus by the other, for example, hearing a barking sound primes one to react to a picture of a dog (Chen and Spence 2017).

Context can also rapidly modulate multisensory integration on a trial-by-trial basis. For example, the McGurk illusion is reduced in subjects who were first exposed to repeated presentations of incongruent visual lip movement and speech sounds (Nahorna et al. 2015).

2.2 Strong Test of Multisensory Binding and Multisensory Objecthood

The evidence presented above illustrates the difficulty in teasing apart the way in which information from multiple senses interacts. Distinguishing binding from integration experimentally is non-trivial. Returning to the example of the 100-meter dash, if one is listening for the judge's gunshot in a noisy stadium, it may be easier to achieve with eyes

open (i.e., with crossmodal input) than closed (i.e., without crossmodal input). Is this caused by the visual and auditory events being perceived as a unified object? It may be, but it is also possible that visual input simply biases the observer towards reporting hearing the shot. Experimenters often use signal detection theory to decouple the change in detection sensitivity (that comes from a perceptual change) from a shift in decision bias. However, if the observer uses one criterion for the gunshot with her eyes opened and another criterion with her eyes closed, a signal detection analysis may incorrectly conclude zero bias (because the bias shifts in equal amounts in opposite directions for the two conditions) and an increase in sensitivity and thus an underlying change in the sensory representation. This error occurs because the decision model used in standard applications of signal detection theory assumes a fixed, unchanging criterion across conditions (Green and Swets 1966; Durlach and Braida 1969).

To circumvent the experimental confound of bias versus enhancement through binding, Bizley and colleagues (2016a) suggested that binding can be identified behaviorally by observing crossmodal effects on a stimulus feature that is *orthogonal* to the features that create the binding. In other words, if a subject is presented with an audiovisual stimulus with temporally coherent changes in sound and light intensity we might expect that these two stimuli would be bound. To demonstrate this, subjects should perform a perceptual judgment on some other feature such as pitch or saturation that changes independently of the intensity (see Fig. 2). If the multisensory binding features are task irrelevant (i.e., they provide no information that could aid in a decision about the task-relevant feature), they cannot (meaningfully) influence the decision criterion, and any measured changes in behavior can be assumed to result not from a simple criterion shift, but from changes in sensory representation.

=== Figure 2 near here ===

2.3 Models of Audiovisual Integration and the Role of Attention

Behaviorally, multisensory integration of auditory and visual stimuli clearly makes an impact on decision making, but two questions remain to be answered regarding the neural basis of such multisensory processing: (i) where is multisensory information integrated, and (ii) does attention play a role in multisensory processing? Theoretically, there are two models that represent the extremes of a spectrum. In one extreme, the late integration model postulates that sensory information is processed separately (e.g., in the sensory cortices) and those unisensory sources of evidence are integrated at a later stage (higher-order cortical areas). In this framework, auditory and visual information can be weighted through unimodal attention at the integration stage. Alternatively, the early integration model postulates that multisensory integration begins early at the unisensory cortices (or before, in subcortical areas) with crossmodal inputs modifying the representations of incoming stimuli. In this early integration framework, integrated sensory information across modalities contributes to the accumulation of sensory evidence, and decision-making in higher-order cortical areas is thus based on an already multisensory representation (Bizley et al. 2016b). This bottom-up framework suggests that early multisensory binding can occur independently of attention, even though selective attention can act to further shape and define this representation.

Whether multisensory information is integrated early in the sensory cortices and / or at later stages by combining independent unisensory information may depend on the specific task. However, the early integration model provides the necessary neural substrate for multisensory binding and the formation of a multisensory object. Hence, the early integration model provides a theoretical conceptualization of how multisensory binding should be

realized. Since attention operates at the level of objects, were attention to be applied to this multisensory representation, this would imply that all crossmodal features associated with the multisensory object would also be enhanced.

While there is substantial physiological (Bizley et al. 2007; Lakatos et al. 2007) and anatomical (Bizley et al. 2007; Falchier et al. 2010) evidence to demonstrate that multisensory processing occurs in primary and non-primary sensory cortices, behavioral paradigms (Raposo et al. 2012) and neuroimaging (Rohe and Noppeney 2015) have provided evidence in favor of integration occurring in higher brain areas. Generally speaking, the neural basis for different kinds of multisensory integration remains underexplored. Therefore, when performing behavioral experiments, it is important to conceptually separate multisensory binding from general multisensory integration so that the findings can better inform neurophysiologists on discovering the different neural architectures that support multisensory processing. Even though this distinction is not often made (with some exceptions, e.g., Odgaard et al. 2004), previous work can be reinterpreted in this framework: Section 3 does just that.

3 Reinterpreting Classic Audiovisual Illusions: Binding or Multisensory Integration?

Many multisensory behavioral studies focus on illusions that place cross-sensory information in conflict in order to understand how the brain normally integrates sensory information. Often the effects are so perceptually salient that researchers assume not only that information has been integrated across the sensory modalities concerned but that it has also been bound to form a strong cohesive multisensory object. In other cases, authors have used the terms integration and binding interchangeably. In this section, three well-known

multisensory paradigms are examined to see whether there is evidence that these illusions pass the strong test of multisensory binding as previously outlined in Sect. 2.2.

3.1 Ventriloquism

In the ventriloquism illusion, the observer's perception of a sound source's location is 'captured' by a visual stimulus. However, does this illusion demonstrate binding of the auditory and visual signals into a multisensory object as often stated? It has been demonstrated that observers combine the visual and auditory location cues in an optimum Bayesian manner. In fact, the ventriloquism effect can be reversed when the visual stimuli used are so blurred that their spatial estimate is less reliable than that of the auditory cues. If observers are asked to provide a location estimate to both the auditory and visual sources, the location of the sound is less biased than if only one location was asked from the subject (see Alais and Burr, Chap. 2). These findings support the late processing model, suggesting that independent estimates are made for each modality and a task-modulated decision-making stage integrates and weighs evidence across sensory modalities. This contrasts with the scenario where the auditory and visual sources are bound in early processing, resulting in a single location associated with the unified multisensory object, independent of whether the observers are providing an auditory or a visual location estimate. Furthermore, behavioral modeling using casual inference suggests that these sensory estimates are maintained separately (Körding et al. 2007). Finally, reward expectation (Bruns et al. 2014) and emotional valence (Maiworm et al. 2012) can also modulate the ventriloquist effect suggesting that, at least in part, top-down factors could modulate decision-making, consistent with the late integration model.

Evidence from a recent functional magnetic resonance imaging (fMRI) study shows that in primary visual cortices, spatial disparity in a ventriloquist paradigm controlled the influence of auditory signals on the formation of spatial estimates (Rohe and Noppeney 2015). Only in higher parietal cortices were auditory and visual signals integrated and weighted by their bottom-up sensory reliabilities and top-down task relevance. This suggests that multisensory interactions are pervasive but governed by different computational principles across the cortical hierarchy. Future studies should further separate out whether the sensory cortex modulation in the ventriloquist illusion is primarily due to amodal attention modulation from the higher cortical areas or specific multisensory binding effects that can exist independent of attention.

3.2 Sound-Induced Flash Illusion

In the sound-induced flash illusion, brief auditory and visual stimuli are presented rapidly in succession and the number of auditory stimuli can influence the reported number of visual stimuli. The nature of these illusory flashes is not totally clear—subjectively, observers often report the illusionary flashes are different from the real flashes. Indeed, if the experimenter offers a third ‘not-one, not-two’ option, many subjects choose that instead (van Erp et al. 2013). Nonetheless, using signal detection theory, it has been demonstrated that the illusory flashes affect sensitivity (and not only bias) to the number of flashes perceived, suggesting that the illusion is due in part to a change in the multisensory sensory representation (McCormick and Mamassian 2008; Kumpik et al. 2014). However, the caveat discussed above must be applied here—if the number of sounds systematically shifts the decision criteria towards the number of perceived sounds, what appears to be a sensitivity change could, in fact, be a systematic switching of the decision criteria. In contrast to the majority of sound-induced flash illusion experiments that do not fulfill the aforementioned strong test of

multisensory binding, a few studies do test perception of another stimulus dimension in the context of the illusion. Mishra and colleagues (2013) asked observers to report not only the number of flashes but also their color, which is an orthogonal feature dimension. Another study tested contrast perception—again an orthogonal dimension—in addition to the number of events and found that the illusion is likely explained by both an early perceptual change as well as a late criterion shift in the decision-making process (McCormick and Mamassian 2008).

Human neurophysiological studies provide further support for the early integration model playing a key role in the sound-induced flash illusion (Mishra et al. 2007). Specifically, a difference in event-related potentials from electroencephalographical recording derived to isolate neural activity associated with the illusory flash revealed an early modulation of activities in the visual cortex after the second sound. Furthermore, the amplitude of this difference waveform is larger in the groups of subjects who saw the illusory flash more frequently, pointing to consistent individual differences that underlie this multisensory integration. Similar to the behavioral observation, the overall pattern of cortical activity associated with the induced illusory flash differed markedly from the pattern evoked by a real second flash. There is evidence that this illusion is generated by a complex interaction between the primary sensory cortices and the multimodal superior temporal areas (see Beauchamp Chap. 8 for review). Perhaps future animal studies may shed more light on the neural basis of multisensory integration or binding underlying this illusion, although to do so would require that investigations be made in the context of a behavioral paradigm to ensure that there was a single-trial read out of whether the illusion was perceived on that trial.

3.3 McGurk Effect

The McGurk effect is often striking—watching a video of a mouth movement that does not match the auditory syllable presented can lead to a percept that is neither of the veridical unisensory percepts, but is instead a third one. In its original report (McGurk and MacDonald 1976), the investigators reported a “fused” percept arising out of a pair of incongruent auditory and visual speech syllables. This illusion has been widely used to understand different aspects of audiovisual processing, in part because the illusion can be measured by a few repetitions of a simple language stimulus. Not generally discussed, however, is the inhomogeneity of this effect across individuals, as well as the efficacy of the illusion across different stimuli (Magnotti and Beauchamp 2015). In fact, while 98% of adult subjects in the original study responded an intermediate /da/ percept when an auditory /ba/ and a visual /ga/ stimuli were presented, only 81% gave an intermediate /ta/ percept when the unvoiced counterparts were presented (i.e., an auditory /pa/ and a visual /ka/). Across individuals, some participants almost always perceive the McGurk effect, while others rarely do (Mallick et al. 2015). Are these discrepancies caused by differences in multisensory binding across individuals, or differences in how they integrate sensory information?

Studying the individual differences across subjects in susceptibility to the McGurk illusion can be more revealing about the nature of the underlying multiple processes than interpreting data at the group level. Meta-analyses across different studies using the same paradigm but with different stimulus parameters are equally important. One potential source of variability across experiments using the McGurk illusion is that the contribution of the unisensory components is not explicitly measured (Tiippana 2014). In fact, McGurk and MacDonold (1976) commented on their original paper that by their own observations, lip movements for /ga/ are frequently misread as /da/ in the absence of auditory input, although they did not

measure speech reading performance in that study. Similarly, the variations in the specific acoustic samples used in different experiments have also not been examined with respect to their phoneme categorization. Nevertheless, perceptually assessing the phonemic feature is still not orthogonal to the feature that links the auditory and visual stimuli.

Instead of asking subjects about the phonemic percept of congruent versus incongruent auditory and visual stimuli, one could ask the subjects to judge the pair's temporal synchrony. In this way, a temporal window of integration can be measured for both the congruent and incongruent multisensory stimuli (c.f., Sect. 2.1.2.2). Importantly, temporal synchrony is an orthogonal feature to the phonemic judgment that links the auditory and visual stimuli and would satisfy the strong test of multisensory binding as outlined in Sect. 2.2. Furthermore, it has been shown that the temporal window of integration correlates well with the amount of McGurk illusion perceived across subjects, as well as other illusion such as the sound-induced flash illusion as described in Sect. 3.2 (Stevenson et al. 2012). In one study, the temporal window was measured to be much narrower for incongruent pairs compared to congruent stimuli. In other words, subjects were more sensitive to asynchronies in incongruent audiovisual syllables than in congruent ones. This suggests that the McGurk-incongruent stimuli are not integrated as strongly as the congruent stimuli when the subjects were asked only to attend to the simultaneity of the stimuli and not the content of the utterances (van Wassenhove et al. 2007). This finding is also suggestive of binding at least of congruent McGurk stimuli, but leaves questions about binding of incongruent stimuli when illusions are or not perceived. Higher-level contextual effects (Nahorna et al. 2012, 2015) as well as visual attention (Tiippana et al. 2004) can also influence the strength of the McGurk effect, casting further doubt on binding as the sole explanation for this illusion.

Converging neurophysiological and computational modeling evidence suggests that audiovisual speech processing is best modeled as a two-stage process (Peelle and Sommers 2015; Magnotti and Beauchamp 2017). Visual input could alter the processing of auditory information through early integration instantiated by a crossmodal perturbation of low-frequency neural oscillations in auditory cortex (Mercier et al. 2015 and also see Keil and Senkowski, Chap. 10 for an in-depth discussion). However, the cues related to speech gestures are better modeled as a late integration process, with posterior superior temporal sulcus likely playing a role in weighting individual auditory and visual inputs (Nath and Beauchamp 2012 and also see Beauchamp, Chap. 8). In summary, an illusory trial whereby an intermediate percept was reported using McGurk stimuli does not necessarily show that the auditory and visual information were bound (even though this is often referred to as the "fused" percept, implicitly assuming binding). Rather, the report of the third percept is evidence that auditory and visual information have been integrated and influenced the decision-making in the syllable classification task. Paradoxically, the non-illusionary (or the non-fused) trials, especially if the auditory and visual stimuli were presented with relatively low asynchrony, could have elicited a bound percept even though the integrated audiovisual information did not result in a third syllable categorization, possibly due to the relative strength of the unisensory evidence. In other words, the presence of the McGurk effect is evidence for multisensory integration (but maybe not binding). Furthermore, the absence of the McGurk effect is not necessarily evidence for two distinct objects. Future studies should aim to explicitly separate how the early and late integration models could affect McGurk perception.

4 Competing Objects in the Audiovisual Scene

Even though most naturalistic environments comprise numerous auditory and visual objects vying for our limited attentional resources, stimulus competition is not often examined in the laboratory setting. One exception is the ‘pip and pop’ paradigm as discussed in Sect. 2.1.2.2. In that case, an auditory stimulus helps resolve visual competition. However, there are relatively fewer multisensory experiments that address how visual information resolves competition between auditory stimuli.

4.1 Prediction from Unisensory Object-Based Attention Theory

Introducing stimulus competition allows experimenters to draw upon the object-based attention literature. Doing so leads to testable hypotheses about the processing advantages offered by binding auditory and visual stimuli into a single audiovisual object. Stimulus competition also provides a more naturalistic and taxing environment making the perceptual benefit of multisensory binding easier to detect.

Based on the theory developed in object-based attention literature (c.f. Sect. 2.1), the expectation would be that if the auditory and visual stimuli were bound as a multisensory object, features in both sensory modalities belonging to the same object would be enhanced. Conversely, if the auditory and visual stimuli came from different objects, even though they can be integrated to make a judgment, there should be a measurable behavioral cost associated with dividing attention across two objects.

Chamber music provides a situation in which we can test how visual input shapes the ability of listeners to parse a complex auditory scene. A string quartet comprises four instruments—first violin, second violin, viola, and cello. The bowing action of each player yields some temporal information about each part of the music. If the observer were to look at

the movement of the first violinist while listening to the cello part, a behavioral cost would be expected due to attention spreading across two objects (here, players). Conversely, if the observer were to both watch the movements of and listen to the cellist, a behavioral benefit would be expected. The bow would provide information about the timing of the cellist's notes—therefore to test binding, an experimenter asks the observer to pick when the music is modulated to another key, something not discernible from the bowing alone. It is temporal coherence that binds the auditory and visual representations of the cellist, and judging temporal aspects of the music is susceptible to bias from the visual modality as vision directly informs the timing of the notes. However, listening for a specific change in key tests an orthogonal dimension to the temporal feature underlying binding, because the player's bowing motion provides no information about pitch. In this hypothetical experiment, if the observer were better at parsing the specific cello notes played when looking at the cellist, then it meets the strong test of binding and suggests that the observer was attending to a bound multisensory object.

4.2 Effect of Spatial Cues

In a previous sound-induced flash illusion experiment, it was concluded that the probability of an illusory percept was not influenced by the degree of spatial separation between the auditory and visual stimuli (Innes-Brown and Crewther 2009). A similar study suggested that the visual sensitivity, not the audiovisual spatial proximity, was the determining factor of the illusory percept (Kumpik et al. 2014). However, one recent study that used competing stimuli showed that the sound-induced flash illusion can be modulated by spatial congruence between the auditory and visual stimuli (Bizley et al. 2012). Subjects were asked to report number of flashes and beeps perceived—as opposed to an orthogonal feature like color. Nevertheless, the results show that stimulus competition can provide

different outcomes—in this case, the influence of spatial congruence for multisensory processing in a complex auditory scene.

4.3 Effect of Temporal Coherence

By using artificial stimuli with naturalistic dynamics, it is possible to test the potential benefits of temporal coherence between auditory and visual stimuli when performing a task based on an orthogonal feature. One such study used competing stimuli to examine whether a visual stimulus being temporally coherent with the target auditory stream would show better behavioral performance compared to when the visual stream was temporally coherent with the masker auditory stream (Maddox et al. 2015). Subjects were asked to report brief pitch or timbre deviants in one of two ongoing independently amplitude-modulated sound-streams; the timing of the brief pitch or timbre deviants was independent of the amplitude modulation imposed on each of the sound streams. They were also instructed to attend a radius-modulated disk that changed either coherently with the amplitude of the target stream or the masker stream, and were also asked to report occasional color changes of the visual disk. Performance was better when the visual disk was temporally coherent with the target auditory stream compared to when it was coherent with the masker stream. Importantly, since the modulations of the visual stimulus were orthogonal to the pitch or timbre deviants and offered no information to their timing (c.f., Fig. 2), the authors suggested that the behavioral benefit observed was through binding of the temporally coherent audiovisual streams forming an audiovisual object whose properties were subsequently enhanced (thus satisfying the “strong test” for binding). In other words, when the auditory target stream and the visual stream were bound into a single multisensory object, performance was improved since the observers no longer had to divide attention across two sensory objects (Fig. 3). While not tested, they also hypothesized that future experiments should be able to show an equivalent auditory

enhancement of visual perception, hinted at already by the “pip and pop” illusion as discussed in Sect. 2.1.2.2.

=== Figure 3 near here ===

5 Summary

The temporal evolution of the auditory and visual information available in real-life cocktail party environments is inevitably complex. Experimental paradigms used in laboratories should strive to expand from the canonical multisensory studies to address this complexity. As presented above, the collective findings from these well-studied illusions still leave much to be learned regarding the nature by which multisensory stimuli are grouped and processed and to what extent binding is a prerequisite for such phenomena.

One particular issue is the use of brief stimuli not representative of naturalistic signals. While this is sometimes seen as a necessary sacrifice for experimental control, this was not always the case. In one of the classic ventriloquism illusion experiments (Jackson 1953), it was shown that the bias in the perceived location of the sound of a steam whistle accompanied by the sight of steam was larger than the bias of a bell sound accompanied by a light pulse. This finding has been interpreted over the years to mean that arbitrary combinations of auditory and visual stimuli with no strong assumption of unity (see Sect. 2.1.1) lead to less effective multisensory integration. Alternatively, the temporally rich and naturally occurring combination of the sight and sound of a kettle may promote object formation due to their temporal coherence. The original experimental setup of realistic and temporally rich stimuli—steam coming out of a singing kettle—might be too elaborate to replicate in most psychophysics labs (Jackson 1953):

> “Three brass tubes each containing a 50-watt soldering iron element were inserted into the rear of each whistle. Water was led from a header tank through a battery of taps, controllable by the experimenter, to a fine jet which was pressed into a tufnol sleeve round the central of the three heater units in each whistle. Thus, when the heater units had been allowed to attain their working temperature, momentary release of one tap caused a visible cloud of steam to rise from the corresponding whistle.”

However, modern experimenters need not resort to such drastic measures. Digitally presenting coherent stimuli across time that can be parametrically manipulated is a relatively recent capability, especially if synchronicity of the auditory and visual stimuli were to be guaranteed. These technical challenges should no longer limit experimental possibilities, meaning researchers can now explore temporally rich stimuli, and move away from the canonical paradigms involving only brief stimuli.

Despite Cherry’s listing of visual cues as a potential solution to the cocktail party problem more than half a century ago (Cherry 1953), only recent audiovisual paradigms have started addressing how visual information can help listeners segregate sounds in complex auditory scenes. With new experimental paradigms and more specific frameworks for delineating audiovisual integration and binding, the field is poised to gain substantial insights into how humans communicate in noisy everyday environments.

Compliance with Ethics Requirements

Adrian K. C. Lee has no conflicts of interest.

Ross Maddox has no conflicts of interest.

Jennifer Bizley has no conflicts of interest.

Bibliography

Alais, D., Blake, R., & Lee, S. H. (1998). Visual features that vary together over time group together over space. *Nature Neuroscience, 1*(2), 160-164.

Behrmann, M., Zemel, R. S., & Mozer, M. C. (1998). Object-based attention and occlusion: evidence from normal participants and a computational model. *Journal of Experimental Psychology-Human Perception and Performance, 24*(4), 1011-1036.

Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature Reviews Neuroscience, 14*(10), 693-707.

Bizley, J. K., Shinn-Cunningham, B. G., & Lee, A. K. C. (2012). Nothing is irrelevant in a noisy world: Sensory illusions reveal obligatory within-and across-modality integration. *Journal of Neuroscience, 32*(39), 13402-13410.

Bizley, J. K., Maddox, R. K., & Lee, A. K. C. (2016a). Defining auditory-visual objects: Behavioral tests and physiological mechanisms. *Trends in Neuroscience, 39*(2), 74-85.

Bizley, J. K., Jones, G. P., & Town, S. M. (2016b). Where are multisensory signals combined for perceptual decision-making? *Current Opinion in Neurobiology, 40*, 31-37.

Bizley, J. K., Nodal, F. R., Bajo, V. M., Nelken, I., & King, A. J. (2007). Physiological and anatomical evidence for multisensory interactions in auditory cortex. *Cerebral Cortex, 17*(9), 2172-2189.

Blake, R., & Lee, S.-H. (2005). The role of temporal structure in human vision. *Behavioral and Cognitve Neuroscience Reviews, 4*(1), 21-42.

Blaser, E., Pylyshyn, Z. W., & Holcombe, A. O. (2000). Tracking an object through feature space. *Nature, 408*(6809), 196-199.

Bruns, P., Maiworm, M., & Röder, B. (2014). Reward expectation influences audiovisual spatial integration. *Attention Perception & Psychophysics, 76*(6), 1815-1827.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The Natural Statistics of Audiovisual Speech. *PLoS Computational Biololgy, 5*(7), e1000436.

Chen, Y.-C., & Spence, C. (2017). Assessing the role of the 'unity assumption' on multisensory integration: A review. *Frontiers in Psychology, 8*, Article 445.

Chen, Y. C., & Spence, C. (2013). The time-course of the cross-modal semantic modulation of visual picture processing by naturalistic sounds and spoken words. *Multisensory Research, 26*, 371-386.

Cherry, E. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America, 25*(5), 975-979.

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

Colonius, H., & Diederich, A. (2004). Multisensory interaction in saccadic reaction time: a time-window-of-integration model. *Journal of Cognitve Neuroscience, 16*(6), 1000-1009.

Colonius, H., & Diederich, A. (2010). The optimal time window of visual-auditory integration: a reaction time analysis. *Frontiers in Integrative Neuroscience, 4*, Article 11.

Culling, J. F., & Stone, M. A. (2017). Energetic masking and masking release. In J. C. Middlebrooks, J. Z. Simon, A. N. Popper & R. R. Fay (Eds.), *The Auditory System at the Cocktail Party* (Vol. 60, pp. 41-74). New York: Springer International Publishing.

Cusack, R., & Roberts, B. (2000). Effects of differences in timbre on sequential grouping. *Perception & Psychophysics, 62*(5), 1112-1120.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*(1), 193-222.

Doyle, M. C., & Snowden, R. J. (2001). Identification of visual stimuli is improved by accompanying auditory stimuli: The role of eye movements and sound location. *Perception, 30*(7), 795-810.

Durlach, N. I., & Braida, L. D. (1969). Intensity perception. I. Preliminary theory of intensity resolution. *The Journal of the Acoustical Society of America, 46*(2), 372-383.

Falchier, A., Schroeder, C. E., Hackett, T. A., Lakatos, P., Nascimento-Silva, S., Ulbert, I., Karmos, G., & Smiley, J. F. (2010). Projection from visual areas V2 and prostriata to caudal auditory cortex in the monkey. *Cerebral Cortex, 20*(7), 1529-1538.

Feldman, J. (2003). What is a visual object? *Trends in Cognitive Sciences, 7*(6), 252-256.

Green, D., & Swets, J. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.

Harrington, L. K., & Peck, C. K. (1998). Spatial disparity affects visual-auditory interactions in human sensorimotor processing. *Experimental Brain Research, 122*(2), 247-252.

Howard, I., & Templeton, W. (1966). *Human Spatial Orientation*. New York: Wiley.

Innes-Brown, H., & Crewther, D. (2009). The impact of spatial incongruence on an auditory-visual Illusion. *PLOS ONE, 4*(7), e6450.

Jackson, C. V. (1953). Visual factors in auditory localization. *Quarterly Journal of Experimental Psychology, 5*(2), 52-65.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLOS ONE, 2*(9), e943.

Kumpik, D. P., Roberts, H. E., King, A. J., & Bizley, J. K. (2014). Visual sensitivity is a stronger determinant of illusory processes than auditory cue parameters in the sound-induced flash illusion. *Journal of Vision, 14*(7), 12-12.

Lakatos, P., Chen, C.-M., O'Connell, M. N., Mills, A., & Schroeder, C. E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron, 53*(2), 279-292.

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 performancej. *Experimental Brain Research, 158*(4), 405-414.

Lee, A. K. C. (2017). Imaging the listening brain. *Acoustics Today, 13*(3), 35-42.

Lovelace, C. T., Stein, B. E., & Wallace, M. T. (2003). An irrelevant light enhances auditory detection in humans: a psychophysical analysis of multisensory integration in stimulus detection. *Cognitive Brain Research, 17*(2), 447-453.

Maddox, R. K., & Shinn-Cunningham, B. G. (2012). Influence of Task-Relevant and Task-Irrelevant Feature Continuity on Selective Auditory Attention. *Journal of the Association for Research in Otolaryngology, 13*(1), 119-129.

Maddox, R. K., Atilgan, H., Bizley, J. K., & Lee, A. K. C. (2015). Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. *eLife, 4*, e04995.

Magnotti, J. F., & Beauchamp, M. S. (2015). The noisy encoding of disparity model of the McGurk effect. *Psychonomic Bulletin & Review, 22*, 701-709.

Magnotti, J. F., & Beauchamp, M. S. (2017). A causal inference model explains perception of the McGurk effect and other incongruent audiovisual speech. *PLoS Computional Biology, 13*(2), e1005229.

Maiworm, M., Bellantoni, M., Spence, C., & Röder, B. (2012). When emotional valence modulates audiovisual integration. *Attention Perception & Psychophysics, 74*(6), 1302-1311.

Mallick, D. B., Magnotti, J. F., & Beauchamp, M. S. (2015). Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin & Review, 22*(5), 1299-1307.

McCormick, D., & Mamassian, P. (2008). What does the illusory-flash look like? *Vision Research, 48*(1), 63-69.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746-748.

Mercier, M. R., Molholm, S., Fiebelkorn, I. C., Butler, J. S., Schwartz, T. H., & Foxe, J. J. (2015). Neuro-oscillatory phase alignment drives speeded multisensory response times: An electro-corticographic investigation. *Journal of Neuroscience, 35*(22), 8546-8557.

Meredith, M. A., Nemitz, J. W., & Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *Journal of Neuroscience, 7*(10), 3215-3229.

Micheyl, C., & Oxenham, A. J. (2010). Objective and subjective psychophysical measures of auditory stream integration and segregation. *Journal of the Association for Research in Otolaryngology, 11*(4), 709-724.

Middlebrooks, J. C., Simon, J. Z., Popper, A. N., & Fay, R. R. (2017). *The Auditory System at the Cocktail Party*. New York: Springer International Publishing.

Mishra, J., Martinez, A., & Hillyard, S. A. (2013). Audition influences color processing in the sound-induced visual flash illusion. *Vision Research, 93*, 74-79.

Mishra, J., Martinez, A., Sejnowski, T. J., & Hillyard, S. A. (2007). Early cross-modal interactions in auditory and visual cortex underlie a sound-induced visual illusion. *Journal of Neuroscience, 27*(15), 4120-4131.

Nahorna, O., Berthommier, F., & Schwartz, J.-L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *The Journal of the Acoustical Society of America, 132*(2), 1061-1077.

Nahorna, O., Berthommier, F., & Schwartz, J.-L. (2015). Audio-visual speech scene analysis: Characterization of the dynamics of unbinding and rebinding the McGurk effect. *The Journal of the Acoustical Society of America, 137*(1), 362-377.

Nath, A. R., & Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *NeuroImage, 59*(1), 781-787.

O'Craven, K. M., Downing, P. E., & Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. *Nature, 401*(6753), 584-587.

Odgaard, E. C., Arieh, Y., & Marks, L. E. (2004). Brighter noise: Sensory enhancement of perceived loudness by concurrent visual stimulation. *Cogn Affect Behav Neurosci, 4*(2), 127-132.

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex, 68*(c), 169-181.

Raposo, D., Sheppard, J. P., Schrater, P. R., & Churchland, A. K. (2012). Multisensory decision-making in rats and humans. *Journal of Neuroscience, 32*(11), 3726-3735.

Rohe, T., & Noppeney, U. (2015). Cortical Hierarchies Perform Bayesian Causal Inference in Multisensory Perception. *PLoS Biol, 13*(2), e1002073.

Shamma, S. A., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in Neuroscience, 34*(3), 114-123.

Shams, L., Kamitani, Y., & Shimojo, S. (2000). Illusions. What you see is what you hear. *Nature, 408*(6814), 788.

Shinn-Cunningham, B. G., Best, V., & Lee, A. K. C. (2017). Auditory Object Formation and Selection. In J. C. Middlebrooks, J. Z. Simon, A. N. Popper & R. R. Fay (Eds.), *The Auditory System at the Cocktail Party* (Vol. 60, pp. 7-40). New York: Springer International Publishing.

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*(1), 31-49.

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

Spence, C., & McDonald, J. (2004). The Crossmodal Consequences of the Exogenous Spatial Orienting of Attention. In G. A. Calvert, C. Spence & B. E. Stein (Eds.), *The Handbook of Multisensory Processing* (pp. 3-25). Cambridge: MIT Press.

Stein, B. E., & Stanford, T. R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience, 9*(4), 255-266.

Stevenson, R. A., Zemtsov, R. K., & Wallace, M. T. (2012). Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. *Journal of Experimental Psychology-Human Perception and Performance, 38*(6), 1517-1529.

Thelen, A., Talsma, D., & Murray, M. M. (2015). Single-trial multisensory memories affect later auditory and visual object discrimination. *Cognition, 138*, 148-160.

Tiippana, K. (2014). What is the McGurk effect? *Frontiers in Psychology, 5*, Article 725.

Tiippana, K., Andersen, T. S., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology, 16*(3), 457-472.

Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society B: Biological Sciences, 353*(1373), 1295-1306.

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and pop: nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology-Human Perception and Performance, 34*(5), 1053-1065.

van Erp, J. B. F., Philippi, T. G., & Werkhoven, P. (2013). Observers can reliably identify illusory flashes in the illusory flash paradigm. *Experimental Brain Research, 226*(1), 73-79.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia, 45*(3), 598-607.

Wallace, M. T., & Stevenson, R. A. (2014). The construct of the multisensory temporal binding window and its dysregulation in developmental disabilities. *Neuropsychologia, 64*, 105-123.

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 80*(3), 638-667.

Figure Legends

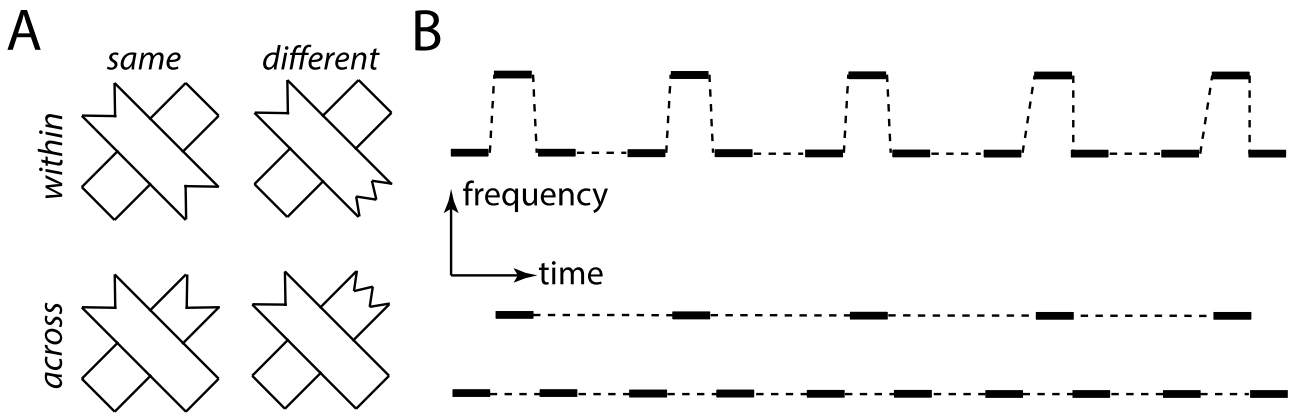

Fig. 1 Visual and auditory examples illustrating object-based attention. **A** Visual example inspired by Behrmann and colleagues (1998) showing two objects intersecting in an X configuration. Subjects are asked to perform a same / different judgment task (whether the "cuttings" at the end of the rectangle(s) are the *same* (left column) or *different* (right column). This task was performed with a faster reaction time when these "cuttings" / features appeared *within* the same object (top row) compared to when they were spread *across* two objects (bottom row), despite having a bigger spatial separation for the displays in the bottom row. **B** Auditory example from Cusack and Roberts (2000). Subjects were asked to detect a change in the isochronous rhythm. Deviation from isochronous (i.e., equally paced) rhythm was much easier to detect when tones were grouped as one object (top) compared to when they were segregated as two objects (bottom).

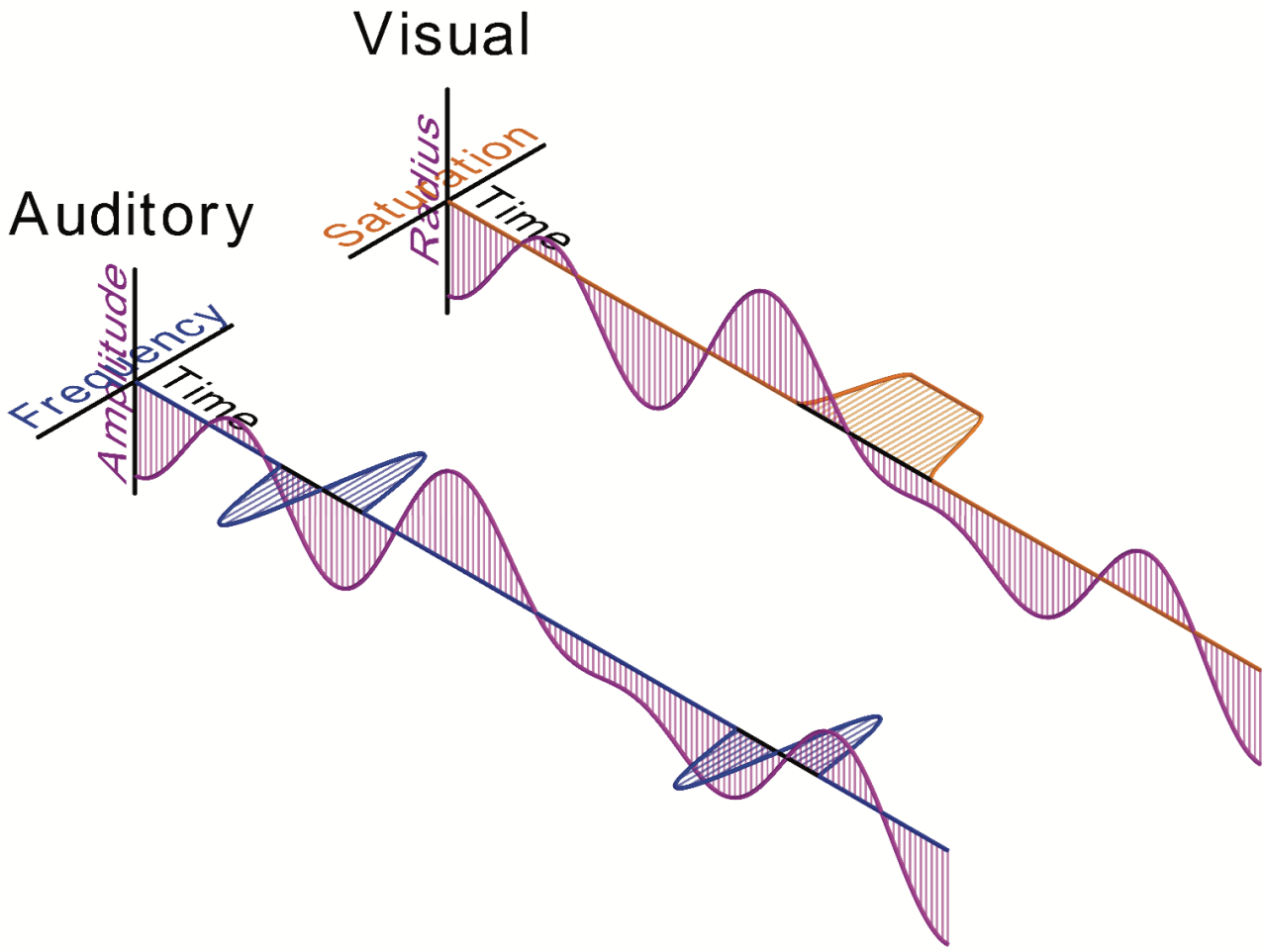

Fig. 2 Auditory and visual stimuli with evolving features over time (in the style of those used by Maddox et al. 2015). In this example, the auditory amplitude and visual radius change coherently (pink features), which facilitates binding into a cross-modal object. The task is based on deviations in the auditory frequency (blue) and visual saturation (orange), which are orthogonal features to those that facilitate binding (amplitude and radius). In other words, the amplitude (radius) provides no task-relevant information about the changes in visual saturation (frequency). Thus, improved perception of these orthogonal feature deviants when the amplitude and radius change coherently (versus when they change independently) demonstrates that this coherence leads to binding.

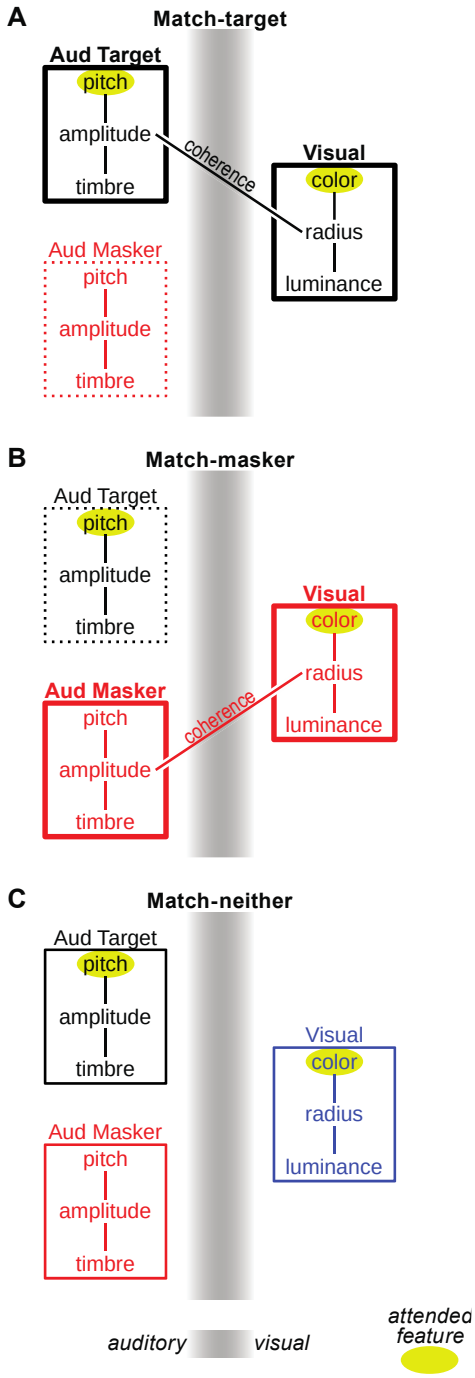

Fig. 3 Conceptual model of binding leading to crossmodal object formation in a task whereby subjects were asked to attend to a pitch change in an auditory target stream (while ignoring an auditory masker stream) and a color perturbation in the visual stimulus (attended

features highlighted in yellow ellipses). Connected sets of features in each sensory stream are shown in a box. Auditory streams are on the left half of the gray sensory boundary and visual on the right. Crossmodal temporal coherence, if present, is shown as a line connecting the coherent features. Specifically, **A** amplitude of the auditory target stream is coherent with the visual size (match-target condition), **B** amplitude of the auditory masker stream is coherent with the visual size (match-masker condition), and **C** no visual features are coherent any features in the auditory streams. Crossmodal binding of the coherent auditory and visual streams enhances each streams' features, resulting in a benefit in the match-target condition (**A**), problematic in the match-masker condition (**B**), and no effect in the match-neither condition (**C**). Enhancement / suppression resulting from object formation is reflected in the strength of the box drawn around each stream's feature (i.e., thick lines denoting enhancement, broken lines, suppression). Reproduced from Maddox et al. (2015).