Audiovisual multisensory integration

Charles Spence*

Crossmodal Research Laboratory, Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford, OX1 3UD, UK

Abstract: Over the last 50 years or so, a large body of empirical research has demonstrated the importance of a variety of low-level spatiotemporal factors in the multisensory integration of auditory and visual stimuli (as, for example, indexed by research on the ventriloquism effect). Here, the evidence highlighting the contribution of both spatial and temporal factors to multisensory integration is briefly reviewed. The role played by the temporal correlation between auditory and visual signals, stimulus motion, intramodal versus crossmodal perceptual grouping, semantic congruency, and the unity assumption in modulating multisensory integration is also discussed. Taken together, the evidence now supports the view that a number of different factors, both structural and cognitive, conjointly contribute to the multisensory integration (or binding) of auditory and visual information.

Keywords: Multisensory integration, Crossmodal, Audiovisual, Binding, Perception

PACS number: 43.10.Mq, 43.66.Mk, 43.66.Qp [doi:10.1250/ast.28.61]

1. INTRODUCTION

When presented with two stimuli, one auditory and the other visual, an observer can either perceive them as referring to the same unitary audiovisual event or else as referring to two separate unimodal events. The binding versus segregation of unimodal stimuli (what Bedford [3] has called the object identity decision; see also [4]) depends on both low-level structural factors, such as the spatial and temporal co-occurrence of the stimuli (e.g., see the chapters in [5]), any temporal correlation between the signals (e.g., [6–8]), and the relative strength of intramodal versus crossmodal perceptual grouping (e.g., [9]), as well as on more cognitive factors, such as whether the stimuli are semantically congruent or not [10,11], and whether or not the observer assumes (for whatever reason) that the two stimuli should 'go together' (e.g., [1,12]). In this review, the various factors (both structural and cognitive; see [2]) that have now been shown to contribute to the multisensory integration of auditory and visual information are highlighted. Given the nature of the journal, the focus is on those studies dealing specifically with crossmodal interactions between auditory and visual stimuli. However, those readers who are particularly interested in the integration of auditory and tactile stimuli are referred to the recent review by Kitagawa and Spence [13].

2. SPATIOTEMPORAL CORRESPONDENCE

Research on the audiovisual binding of sensory information, specifically with regard to the question of stimulus localization, dates back more than 60 years (e.g., see [8,14,15] for early studies). By now, many different studies have shown that auditory stimuli are typically mislocalized toward visual stimuli, provided that they are presented at approximately the same time (see [12,16] for reviews). The magnitude of this spatial ventriloquism effect tends to decrease as the spatial and/or temporal separation between the auditory and visual stimuli is increased (e.g., [4,14,17–20]). Researchers now believe that this form of crossmodal integration (or sensory dominance) may result from the statistically optimal combination of the unimodal inputs (e.g., [21,22]).

Over the last few years, a phenomenon analogous to spatial ventriloquism has been demonstrated in the temporal domain whereby, within a certain temporal window (see [23]), visual stimuli appear to be ‘pulled’ into approximate temporal alignment with the corresponding auditory stimuli (e.g., [24,25]). For example, Morein-Zamir and her colleagues conducted an influential series of experiments in which their participants were presented with pairs of visual stimuli (one from above and the other from below a central fixation point) at a range of different stimulus onset asynchronies (SOAs, see Fig. 1). The participants had to make unspeeded temporal order judgments (TOJs) regarding which of the two lights appeared to have been presented first on each trial. Brief uninformative tones
were also presented from a loudspeaker cone situated directly behind the fixation light. One sound was presented before the onset of the first light, while the other sound was presented after the offset of the second light. Even though the centrally-presented sounds provided no clue as to whether the upper or lower light had been presented first, their presence nevertheless led to a significant improvement in the sensitivity of participants’ TOJ performance when compared to their performance in a condition in which the sounds were presented in synchrony with the lights, or else to a no sound condition. Specifically, the just noticeable differences (JNDs) calculated on the basis of participants’ psychometric data were reduced from 62 ms to 44 ms when the sounds were presented either 75 or 150 ms before/after the lights, rather than at the same time; see [25] for similar results). On the basis of their results, Morein-Zamir et al. [24] argued that the sounds may have temporally ventriloquised the perceived onset of the lights, hence increasing the apparent temporal separation between the two lights, and thus improving participants’ TOJ performance. Findings such as these have been taken to support the view that while visual stimuli typically lead to the spatial ventriloquism of auditory stimuli, auditory stimuli may ventriloquize the time at which visual stimuli are perceived to occur. This pattern of results is entirely consistent with the fact that vision provides more accurate spatial information while audition typically provides more accurate temporal information (see [23,26]).

Auditory stimuli have also been shown to dominate over visual stimuli when it comes to judging the rate at which rapidly-presented streams of auditory and visual stimuli are being presented (e.g., [7,26–28]). Over the years, many different studies have shown that the perceived rate of stimulation in one sensory modality can be modulated by the rate of stimulus presentation in another sensory modality. In particular, researchers have shown that the rate at which a rapidly alternating visual stimulus appears to be flickering can be modulated quite dramatically by changes in the rate at which a simultaneously-presented auditory stimulus is made to flutter (see [9] for a review). For example, participants in a classic study reported by Shipley [27] had to judge the rate at which a sound appeared to flutter or, at other times, to judge the rate at which a light appeared to flicker. Shipley found that changing the physical rate of flutter of an intermittently-presented sound induced a systematic change in the apparent rate at which a flashing light was simultaneously seen to flicker. Indeed, for one of Shipley’s observers, a visual stimulus that was actually flickering at 10 cycles per second was reported at different times to be flickering at anything between 7 and 22 cycles per second depending on the rate of flutter of the simultaneously-presented sound.

3. CORRELATED SENSORY SIGNALS

A number of studies have also highlighted the importance of any temporal correspondence (or correlation) between auditory and visual stimulus streams in facilitating multisensory integration (e.g., [4,6,8,17,19,29]). For example, Radeau and Bertelson [6] followed up on an earlier study by Thomas [8] looking at the effect of any correlation between the temporal properties of pairs of auditory and visual stimuli in the case of the spatial ventriloquism effect. The participants in Radeau and Bertelson’s study were presented with pairs of spatially-misaligned auditory and visual stimuli on each of the bimodal trials. The stimulus in each sensory modality could be presented in one of three different temporal configurations (or patterns): It could either be presented continuously (for 4 seconds), or periodically interrupted at either a fast or slow rate (see Fig. 2). The participants had to decide whether the middle target stimulus had been presented to the left or right of the midline.

Radeau and Bertelson [6] were able to replicate the findings of Thomas’s [8] early study by showing that the visual bias of auditory localization was larger when the temporal configuration of the stimuli in the two modalities matched rather than when the two stimuli were presented with different temporal patterns (see Fig. 2(b)).1 It is also worth noting here that the correlation present between the complex and dynamic time-varying characteristics of
It should however be noted that it is unclear whether it was the matching (or correlation) of the signals presented in each sensory modality, or else simply the increased likelihood of there being simultaneous onsets and offsets in those conditions where the temporal patterns in the two modalities matched that drove the visual biasing effects reported by Radeau and Bertelson [6]. Furthermore, it is also unclear to what extent their results might have been caused simply by the matching of the number of stimuli presented in each sensory modality (which because of the particular temporal patterns used, only occurred in the matching conditions; cf [38–40] on this point). Clearly, further research is needed in order to demonstrate unequivocally that temporal patterning, rather than one of these other factors, is critical for enhancing crossmodal binding.

4. CROSSMODAL DYNAMIC CAPTURE

Stimulus motion (be it real or apparent) has also been shown to play a significant role in multisensory integration (see [31] for a historical overview). Soto-Faraco and his colleagues have conducted a number of experiments over the last 5 years showing that stimulus motion can facilitate audiovisual integration. In their original research, Soto-Faraco et al. [32] demonstrated that the presentation of a visual apparent motion stream consisting of the sequential presentation of two light flashes, one from either side of a central fixation point, has a dramatic effect on the direction in which an auditory apparent motion stream appears to move (either from left-to-right or vice versa) while trying to ignore an irrelevant visual apparent motion stream moving in either the same (i.e., congruent) or opposite (i.e., incongruent or conflicting) direction. The most important result to have emerged

---

1 It should however be noted that it is unclear whether it was the matching (or correlation) of the signals presented in each sensory modality, or else simply the increased likelihood of there being simultaneous onsets and offsets in those conditions where the temporal patterns in the two modalities matched that drove the visual biasing effects reported by Radeau and Bertelson [6]. Furthermore, it is also unclear to what extent their results might have been caused simply by the matching of the number of stimuli presented in each sensory modality (which because of the particular temporal patterns used, only occurred in the matching conditions; cf [38–40] on this point). Clearly, further research is needed in order to demonstrate unequivocally that temporal patterning, rather than one of these other factors, is critical for enhancing crossmodal binding.
have also been ruled out by a number of recent studies (e.g., [34]). Non-perceptual (i.e., decisional or response bias) interpretations of the crossmodal dynamic capture effect [32]).

What's more, crossmodal dynamic capture effects are maximal when the auditory and visual streams are presented asynchronously (separated by 500 ms in this example). The magnitude of the crossmodal dynamic effect is measured in terms of the difference in performance between incongruent and congruent trials (i.e., the difference between the black and white bars in the figure).

from the more that 30 such experiments conducted to date is that participants perform significantly less accurately on incongruent trials than on congruent trials (see Fig. 4), at least when the target and distractor streams are presented at approximately the same time.

Soto-Faraco and his colleagues have argued that this crossmodal dynamic capture effect (defined as the magnitude of the difference in the proportion of correct direction discrimination responses between the incongruent and congruent trials) reflects the mandatory integration of the visual and auditory apparent motion signals. As one might have predicted that this manipulation should stimulate intramodal visual grouping can have on audiovisual integration. Indeed, a large part of ‘the binding problem’ (see [9]) has already been solved for a participant presented with only a single auditory and visual stimulus at around the same point in space and time in an otherwise dark and silent environment. In a way, it is as if the austerity of the experimental situation may in-and-of-itself lead the participants into an ‘assumption of unity’ concerning the likely relation between the two signals (see below). Indeed, the only obvious form of binding that can take place under such conditions is of the crossmodal variety. However, under the more complex (and realistic) multisensory conditions of everyday life there may actually be a constant interplay between those factors promoting the intramodal versus crossmodal grouping of perceptual information ([37]; see [9], for a review).

This issue is particularly important when one comes to consider the putative interplay between the intramodal versus crossmodal perceptual grouping on multisensory integration. Indeed, a large part of ‘the binding problem’ (see [9]) has already been solved for a participant presented with only a single auditory and visual stimulus at around the same point in space and time in an otherwise dark and silent environment. In a way, it is as if the austerity of the experimental situation may in-and-of-itself lead the participants into an ‘assumption of unity’ concerning the likely relation between the two signals (see below). Indeed, the only obvious form of binding that can take place under such conditions is of the crossmodal variety. However, under the more complex (and realistic) multisensory conditions of everyday life there may actually be a constant interplay between those factors promoting the intramodal versus crossmodal grouping of perceptual information ([37]; see [9], for a review).

Sanabria and his colleagues have recently reported a number of experiments in which they have demonstrated the systematic effect that variations in the strength of intramodal visual grouping can have on audiovisual integration [38–40]. For example, Sanabria et al. [40] investigated whether increasing the number of visual stimuli presented in the distractor visual apparent motion stream (from 2 to 6) would affect the magnitude of the crossmodal congruency effect when participants were required to judge the direction in which a sequence of two auditory stimuli moved (either from left-to-right or vice versa; see Fig. 5). Given that increasing the number of stimuli increases the strength of visual apparent motion, one might have predicted that this manipulation should

5. INTRAMODAL VERSUS CROSSMODAL PERCEPTUAL GROUPING

The majority of research on multisensory integration that has been published to date has involved very simplistic conditions of stimulus presentation. In particular, researchers interested in trying to understand the spatial and temporal constraints on crossmodal binding (such as in studies of the spatial and temporal ventriloquism effect outlined above) have tended to present participants with only a single auditory and visual event at any one time (see Fujisaki et al. [35], for one nice recent exception to this generalization). Using this approach, researchers have tried to understand how the presentation of a stimulus in one sensory modality influences people’s perception of (or at least their responses to) a stimulus presented in a different sensory modality at either the same or different point in space or time. It is, however, important to note that the multisensory scenes of everyday life are nearly always far more complex than those found in the austere environment of the psychophysicist’s dark and silent laboratory. In fact, the experimental stimuli used in many of the published studies of multisensory perception have been criticized both on the grounds of their limited ecological validity and contextual saliency (see [18,36]).

Fig. 4 Typical pattern of results from a crossmodal dynamic capture experiment showing that discrimination of the direction of an auditory apparent motion stream can be substantially impaired (i.e., response accuracy is significantly lowered) by the simultaneous presentation of a distracting visual stream moving in the opposite direction. Note that no such performance decrement is reported if the auditory and visual apparent motion streams are presented asynchronously (separated by 500 ms in this example).
have made the distracting visual apparent motion stream harder to ignore, and hence resulted in a larger crossmodal dynamic capture effect. In fact, the crossmodal dynamic capture effect was significantly reduced when the visual distractor stream contained 6-lights (mean capture effect of 21%) rather than just 2-lights (mean capture effect of 34%; see Fig. 6).

Sanabria et al. [40] suggested that increasing the number of stimuli in the visual display may have reduced the magnitude of the crossmodal capture effect because the increase in the number of visual stimuli (relative to the auditory stimuli), should have made it more likely that the auditory and visual streams would segregate from one another and so be treated as separate perceptual events (streams). Any such segregation of the stimuli presented in the two different sensory modalities should have resulted in a reduction in crossmodal grouping (binding) and hence a reduction of multisensory integration. This, in turn, would be expected to lead to a reduction in the influence of the visual stream on participants’ auditory direction-of-motion discrimination responses. Sanabria et al.’s results therefore demonstrate that the nature of the perceptual grouping taking place within the visual modality can influence the extent to which crossmodal binding (in particular, audiovisual integration, as indexed by performance on the crossmodal dynamic capture task) takes place.

Sanabria and his colleagues have now demonstrated the modulatory role of intramodal perceptual grouping over the crossmodal integration of auditory and visual apparent motion stimuli in several different experimental settings (see [38,39]; [9] for a review). Taken together, the results of these studies of the crossmodal dynamic capture effect show that as soon as one moves away from the very simplistic conditions of stimulus presentation traditionally used by many researchers interested in the topic of multisensory integration (e.g., when only a single simple auditory and visual event are presented at any one time, and where the binding problem has effectively been solved for the observer) to more complex conditions in which multiple stimuli (or events) are presented in each sensory modality, then there will be an ongoing battle between the intramodal versus crossmodal grouping of sensory information.

6. SEMANTIC CONGRUENCY

Researchers have recently started to investigate whether the degree of semantic congruency between simultaneously-presented auditory and visual stimuli also modulates multisensory integration (e.g., [10,11,41,42]). For instance, Laurienti et al. [10] reported an experiment in which their participants were presented with target stimuli consisting of either a red or a blue circle and/or the word “red” or “blue” spoken over headphones. On those trials in which a bimodal target was presented, it was always congruent (i.e., a red circle together with the word ‘red’). On many of the trials, a distractor stimulus, consisting of a green circle and/or the word ‘green,’ was also presented. The participants had to make speeded colour (‘red’ vs. ‘blue’) discrimination responses to the target stimuli while at the same time trying to ignore any distractor stimuli that were presented.

The participants in Laurienti et al.’s [10] study responded significantly more rapidly and accurately on the congruent bimodal target trials than on the unimodal
target trials, with the worst performance being observed on those trials where an incongruent distractor was presented. Importantly, the speeding up of participants’ responses observed in the congruent bimodal trials was greater than would have been expected on the basis of probability summation alone (i.e., the race model was violated; see [43]). Laurienti et al. also reported that semantically congruent pairs of visual stimuli did not give rise to a significant facilitation effect, while pairs of semantically incongruent visual stimuli impaired performance just as much as did the incongruent bimodal stimulus pairs. Laurienti et al. therefore concluded that the semantic congruency between pairs of auditory and visual stimuli modulates multisensory integration.

Molholm et al. [11] came to a similar conclusion using a somewhat different experimental paradigm: In their combined behavioral and ERP study, participants were presented with line-drawings and vocalizations from 8 different animals which could either be presented unimodally or else bimodally (once again, in either congruent or incongruent pairings). The participants were able to identify the pre-specified target stimuli (e.g., cows) more rapidly and accurately when presented in a semantically congruent bimodal stimulus pairing (e.g., a drawing of a cow together with a mooing sound) than when presented unimodally or else in a semantically incongruent bimodal pairing (e.g., when a picture of a cow was presented together with the sound of a dog woofing). These results also led Molholm et al. to conclude that the multisensory integration of auditory and visual stimuli is enhanced when auditory and visual stimuli are semantically congruent (as compared to when they are presented unimodally or else together with a semantically incongruent, or irrelevant, distractor; though note that performance in these latter two conditions did not differ in Molholm et al.’s study).2

It is, however, important to note that semantic congruency doesn’t always modulate multisensory integration. For example, Koppen and Spence [41] recently conducted a study of the Colavita visual dominance effect (see [44]) in which the participants were presented with an unpredictable sequence of auditory, visual, and audiovisual stimuli requiring speeded discrimination responses. The auditory stimuli consisted of the sound of a cat meowing or of a dog woofing, while the visual stimuli consisted of the pictures of a cat and of a dog. On the bimodal trials, the auditory and visual stimuli could either be semantically congruent (i.e., the sight and sound of a dog) or else semantically incongruent (i.e., the sound of a cat presented together with the sight of a dog). The participants had to press one response key as rapidly as possible whenever an auditory target was presented and another response key whenever a visual target was presented. On the bimodal trials, the participants were instructed to press both response keys as rapidly as possible. An oft-reported finding from previous research that has used this paradigm with simple stimuli is that while participants experience no problem in responding to the auditory targets on the unimodal trials, they frequently fail to respond to them on the bimodal trials (see [44,45]). In fact, it is as if the simultaneous presentation of the visual stimulus somehow extinguishes the participant’s awareness of, or at least their ability to respond to, the auditory stimulus on a certain proportion of the bimodal trials.

Koppen and Spence [41] recently found that the magnitude of the Colavita effect (i.e., the increased likelihood of missing the auditory targets on the bimodal as compared to the unimodal auditory trials) was completely unaffected by the semantic congruency between the auditory and visual stimuli. Importantly, however, they were able to show that semantic congruency influenced certain other aspects of participants’ performance. Given these mixed results regarding the effects of semantic congruency on audiovisual binding, it is clear that further research will be needed in order to determine the specific conditions under which semantic congruency influences multisensory integration.

7. THE UNITY ASSUMPTION

It has long been argued that whenever two or more sensory inputs are perceived as being highly consistent (i.e., as being related in a way that they appear to ‘go together’), observers will be more likely to treat them as referring to a single audiovisual event (e.g., [12,14]). Consequently, they will be more likely to assume that they have a common spatiotemporal origin, and hence they will be more likely to bind them into a single multisensory perceptual object or event (see [3]). The assumption that a perceiver has as to whether he/she is observing a single multisensory event versus multiple separate unimodal events is a decision that is based, at least in part, on the consistency (or compellingness; see [12,46]) of the information available to each sensory modality, and may

---

2It is important to note that the demonstration of a behavioral facilitation effect attributable to the presentation of semantically congruent stimuli need not necessarily reflect the consequences of multisensory integration. After all, many previous studies have demonstrated Stroop-like interference [58] between pairs of superimposed visual objects, without anyone having claimed that the two visual images were necessarily being bound (or integrated) in any meaningful way (see also [10], p. 412, on this point). This concern over whether or not the auditory and visual stimuli were necessarily bound in these previous studies investigating the effect of semantic congruency is made all the more salient when one considers the fact that the auditory and visual stimuli were presented from different locations in several of these studies (see [59], on this point).
(at least according to [12]) normally take place unconsciously (though see also [2,47]).

The question of whether or not the unity assumption does in fact play a significant role in audiovisual integration has proved to be one of the most contentious issues in multisensory research over the last 50 years ([2,47]; see [1] for a recent review). In 1953, Jackson [14] reported an early study of the ventriloquism effect in which participants’ judgments of the location from which a sound had been presented were biased more by spatially-incongruent visual stimuli that would be predicted to give rise to a stronger assumption of unity (i.e., the sight of a steaming kettle and the sound of a steam whistle) than for auditory and visual stimuli that were paired arbitrarily (such as the sound of a bell and the onset of a light source; i.e., a pair of stimuli which should not necessarily lead to an assumption of unity). However, while the results of this and many other subsequent studies have been taken by some researchers to show that the unity assumption can facilitate multisensory integration, others have criticized such claims on the grounds that the temporal correlation (and informational richness) contained in the auditory and visual stimuli for those stimuli that supposedly promoted an assumption of unity was also higher (see also [46]; and [4] for null results). Thus, it is at present unclear whether the majority of previous research on the unity assumption has demonstrated anything other than the aforementioned importance of temporal correlation in the signals for audiovisual binding. That said, researchers have now started to report evidence in support of the role of the unity assumption in multisensory integration that is seemingly free from any such concerns about possible variations in the richness of the stimuli used [1].

For example, the participants in a series of 4 audiovisual TOJ experiments reported by Vatakis and Spence [1] were presented with pairs of auditory and visual speech stimuli (either single syllables or words) at a range of different SOAs using the method of constant stimuli. The participants had to make unspeeded TOJs regarding whether the auditory or visual speech stream had been presented first on each trial. On half of the trials, the auditory and visual speech stimuli were gender matched (i.e., the sight of a steaming kettle and the sound of a steam whistle) while on the remainder of the trials the auditory and visual speech stimuli were gender mismatched (i.e., a pair of stimuli which should not necessarily lead to an assumption of unity). However, while the results of this and many other subsequent studies have been taken by some researchers to show that the unity assumption can facilitate multisensory integration, others have criticized such claims on the grounds that the temporal correlation (and informational richness) contained in the auditory and visual stimuli for those stimuli that supposedly promoted an assumption of unity was also higher (see also [46]; and [4] for null results). Thus, it is at present unclear whether the majority of previous research on the unity assumption has demonstrated anything other than the aforementioned importance of temporal correlation in the signals for audiovisual binding. That said, researchers have now started to report evidence in support of the role of the unity assumption in multisensory integration that is seemingly free from any such concerns about possible variations in the richness of the stimuli used [1].

For example, the participants in a series of 4 audiovisual TOJ experiments reported by Vatakis and Spence [1] were presented with pairs of auditory and visual speech stimuli (either single syllables or words) at a range of different SOAs using the method of constant stimuli. The participants had to make unspeeded TOJs regarding whether the auditory or visual speech stream had been presented first on each trial. On half of the trials, the auditory and visual speech stimuli were gender matched (i.e., a female face was presented together with the matching female voice), while on the remainder of the trials the auditory and visual speech stimuli were gender mismatched (i.e., the lip-movements of a female face were presented together with a man’s voice; see Fig. 7). The participants in all 4 experiments always found it significantly harder to judge which sensory modality had been presented first (measured in terms of the just noticeable difference; JND) when evaluating the matched speech stimuli (mean JND of 89 ms) than when evaluating the mismatched speech stimuli (mean JND of 68 ms; these values reflect the average JNDs across Vatakis & Spence, Experiments 1–3).

Vatakis and Spence [1] suggested that the presentation of the matched speech stimuli in their study may have resulted in more temporal ventriloquism taking place than was the case for the mismatched videos. Any such ventriloquism of the visual speech event into temporal alignment with the slightly asynchronous auditory speech event would presumably have made it harder for the participants to judge the temporal order of the auditory and visual speech stimuli correctly in the gender matched (as compared to the gender mismatched) videos. Vatakis and Spence’s results therefore provide unequivocal empirical support for the long-standing claim that the unity assumption can influence the multisensory integration of auditory and visual stimuli. It is perhaps also worth noting at this point that the gender matching of auditory and visual speech stimuli has been shown to modulate people’s ability to identify speech stimuli, at least when the observers are familiar with the speakers in the video-clips (see [48,49]; though see also [50] for null results).

8. MULTISENSORY INTEGRATION: THE CONTRIBUTION OF STRUCTURAL VERSUS COGNITIVE FACTORS

Welch [2] has argued that multisensory integration can be modulated by both structural (i.e., stimulus-driven) and cognitive factors (see also [4,12]). Structural factors
include the spatiotemporal correspondence between the stimuli, any temporal correlation between the stimuli, and perhaps even the degree to which the intensity of the stimuli are matched (though no published research has, as yet, addressed this specific question). Cognitive factors refer to the assumption of unity, instructional manipulations, and so on. Welch and Warren [12] have argued that the assumption of unity should be conceptualized as a graded factor (though see [47]), one that is influenced by a number of different variables: from the structural similarity of the stimuli [6], through to an observer’s familiarity with the particular pairing of stimuli (e.g., [14]; though see also [47]); and from the compellingness of the audiovisual stimulus pairing (e.g., [12,46]) through to the instructions given by the experimenter as to whether the auditory and visual stimuli should be considered as having a common distal origin or not (see [46,51]).

It is, however, important to note that it is not always that easy to distinguish between the contribution of structural (or bottom-up) and cognitive (or more top-down) factors to multisensory integration (see also [4,6,12] on this point). This is because many of the structural factors that promote the bottom-up integration of sensory signals also serve to promote the (cognitive) assumption of unity by participants as well. So, for example, while spatial and temporal coincidence are typically thought of as structural factors, the presentation of auditory and visual stimuli from the same (rather than different) spatial location will also presumably increase the likelihood that an observer will assume that those two signals ought to go together (see also [52]), and hence may also facilitate multisensory integration in a more cognitive (i.e., top-down) manner (see [12]).

The same argument also holds true for many of the other putatively structural influences on multisensory integration, such as spatial congruency (or correspondence) between the stimuli (see [6,46], on this point), and the principles of perceptual grouping [9]. Similarly, factors such as semantic congruency, which is sometimes treated as being a cognitive variable, may also, upon inspection, be just as likely to facilitate multisensory integration in a structural manner as well. What’s more, as we have already seen, there has been fierce controversy over whether the enhanced multisensory integration seen in the majority of studies of the unity assumption actually reflects the influence of structural versus more cognitive factors on audiovisual integration (see [1], on this point). In fact, perhaps the only unequivocal demonstration of a purely cognitive influence on multisensory integration comes from an anecdotal finding reported recently by Arnold et al. [51; pp. 1282–1283]. While investigating the phenomenon of audiovisual simultaneity constancy, the authors reported in one experiment that imagining that pairs of auditory and visual stimuli had originated from the same environmental source gave rise to a simultaneity constancy effect which was not observed in the absence of any such assumption of unity by participants (see also [46]).

9. CONCLUDING REMARKS

It should by now be clear that the question of whether or not an observer will bind a particular pair of auditory and visual stimuli is dependent on a number of different factors. In particular, the extant evidence has highlighted the key role that spatiotemporal coincidence plays in many forms of audiovisual integration (e.g., [20,52]; though see [25,29,30] for important exceptions). The temporal patterning (or correlation) between auditory and visual stimuli has also been shown to play a significant role in multisensory integration (as shown by research on the ventriloquism effect; see [6,8,46]), and a growing body of research now points to the critical role that stimulus motion may also play in determining the extent to which the multisensory integration of auditory and visual stimuli will take place (see [31,53] for reviews). Semantic congruency also appears to play an important role in modulating multisensory integration under certain conditions ([10,11]; though see [41]). Finally, the evidence now supports the view that an observer’s beliefs as to whether the auditory and visual stimuli have a common distal source (i.e., the unity assumption) also modulates multisensory integration [1,51].

Given that researchers have by now been reasonably successful in demonstrating the influence of each of these variables on audiovisual integration, one of the most important challenges in the years to come will be to try and understand how the various factors highlighted in this review combine to modulate multisensory integration under more realistic conditions using more ecologically-valid combinations of stimuli. Furthermore, additional research will also be needed in order to more fully understand the neural underpinnings of these perceptual/behavioral effects (e.g., [11,42,53–57]).

REFERENCES

[46] D. H. Warren, R. B. Welch and T. J. McCarthy, “The role of visual-auditory ‘compelliveness’ in the ventriloquism effect: Implications for transitivity among the spatial senses,” Per-


Professor Charles Spence is the head of the Crossmodal Research Laboratory based at the Department of Experimental Psychology, Oxford University (http://www.psy.ox.ac.uk/xmodal/default.htm). Charles is currently a consultant for a number of multinational companies advising on various aspects of multisensory design. He has also conducted research on human-computer interaction issues on the Crew Work Station on the European Space Shuttle, and currently works on problems associated with the design of foods that maximally stimulate the senses, and with the effect of the indoor environment on mood, well-being, and performance. Charles has published more than 200 articles in top-flight scientific journals over the last decade. Charles has been awarded the 10th Experimental Psychology Society Prize, the British Psychology Society: Cognitive Section Award, the Paul Bertelson Award, recognizing him as the young European Cognitive Psychologist of the Year, and, most recently, the prestigious Friedrich Wilhelm Bessel Research Award from the Alexander von Humboldt Foundation in Germany.