Optimality and Limitations of Audio-Visual Integration for Cognitive Systems

Boyce, William Paul; Lindsay, Anthony; Zgonnikov, Arkady; Rañó, Iñaki; Wong-Lin, Kong Fatt

DOI
10.3389/frobt.2020.00094

Publication date
2020

Document Version
Final published version

Published in
Frontiers In Robotics and AI

Citation (APA)
Boyce, W. P., Lindsay, A., Zgonnikov, A., Rañó, I., & Wong-Lin, K. F. (2020). Optimality and Limitations of Audio-Visual Integration for Cognitive Systems. *Frontiers In Robotics and AI*, *7*, Article 94. https://doi.org/10.3389/frobt.2020.00094

frontiers
in Robotics and AI

REVIEW
published: 17 July 2020
doi: 10.3389/frobt.2020.00094

Optimality and Limitations of Audio-Visual Integration for Cognitive Systems

William Paul Boyce[1*†], *Anthony Lindsay*[1], *Arkady Zgonnikov*[2,3], *Iñaki Rañó*[1†] and *KongFatt Wong-Lin*[1]

[1] Intelligent Systems Research Centre, Ulster University, Magee Campus, Derry Londonderry, Northern Ireland, United Kingdom, [2] AiTech, Delft University of Technology, Delft, Netherlands, [3] Department of Cognitive Robotics, Faculty of Mechanical, Maritime, and Materials Engineering, Delft University of Technology, Delft, Netherlands

OPEN ACCESS

Edited by:
German I. Parisi,
University of Hamburg, Germany

Reviewed by:
Christopher C. Berger,
California Institute of Technology,
United States
Cristiano Cuppini,
University of Bologna, Italy

***Correspondence:**
William Paul Boyce
William.Boyce@unsw.edu.au

†Present address:
William Paul Boyce,
School of Psychology, University of New South Wales, Sydney, NSW, Australia
Iñaki Rañó,
SDU Biorobotics, Mærsk Mc Kinney Møller Institut, University of Southern Denmark, Odense, Denmark

Specialty section:
This article was submitted to Sensor Fusion and Machine Perception,
a section of the journal Frontiers in Robotics and AI

Received: 01 December 2019
Accepted: 09 June 2020
Published: 17 July 2020

Citation:
Boyce WP, Lindsay A, Zgonnikov A, Rañó I and Wong-Lin K (2020) Optimality and Limitations of Audio-Visual Integration for Cognitive Systems. Front. Robot. AI 7:94.
doi: 10.3389/frobt.2020.00094

Multimodal integration is an important process in perceptual decision-making. In humans, this process has often been shown to be statistically optimal, or near optimal: sensory information is combined in a fashion that minimizes the average error in perceptual representation of stimuli. However, sometimes there are costs that come with the optimization, manifesting as illusory percepts. We review audio-visual facilitations and illusions that are products of multisensory integration, and the computational models that account for these phenomena. In particular, the same optimal computational model can lead to illusory percepts, and we suggest that more studies should be needed to detect and mitigate these illusions, as artifacts in artificial cognitive systems. We provide cautionary considerations when designing artificial cognitive systems with the view of avoiding such artifacts. Finally, we suggest avenues of research toward solutions to potential pitfalls in system design. We conclude that detailed understanding of multisensory integration and the mechanisms behind audio-visual illusions can benefit the design of artificial cognitive systems.

Keywords: multi-modal processing, multisensory integration, audio-visual illusions, Bayesian integration, optimality, cognitive systems

1. INTRODUCTION

Perception is a coherent conscious representation of stimuli that is arrived at, via processing signals sent from various modalities, by a perceiver: either human or non-human animals (Goldstein, 2008). Humans have evolved multiple sensory modalities, which include not only the classical five (sight, hearing, tactile, taste, olfactory) but also more recently defined ones (for example, time, pain, balance, and temperature, Fitzpatrick and McCloskey, 1994; Fulbright et al., 2001; Rao et al., 2001; Green, 2004). While each modality is capable of resulting in a modality-specific percept, it is often the case that stimulus information gathered by two or more modalities is combined in an attempt to create the most robust representation possible of a given environment in perception (Macaluso et al., 2000; Ramos-Estebanez et al., 2007).

Understanding and mapping just how the human brain combines different types of stimulus information from drastically different modalities is challenging. Behavioral studies have suggested optimal or near-optimal integration of multi-modal information (Alais and Burr, 2004; Shams and Kim, 2010). In the case of Alais and Burr (2004) they examined the classic spatial ventriloquist effect through the lens of near optimal binding. The effect in question describes the apparent "capture" of an auditory stimulus in perceptual space that is then mapped to the perceptual location of a congruent visual stimulus, the famous example being the ventriloquist's voice appearing to emanate from the synchronously animated mouth of the dummy on their knee. Alais and Burr (2004) demonstrated that this process of "binding" the perceived spatial location of an auditory stimulus to the perceived location of a visual stimulus is an example of near optimal audio-visual integration. They achieved this by demonstrating that variations of the effect could be reversed (i.e., a visual stimulus being "captured" and shifted to the same perceptual space as an auditory stimulus) when the auditory signal was less noisy relative to the visual stimulus (when extreme blurring noise was added to the visual stimulus). Additionally, when visual stimuli was blurred, but not extremely so, neither stimulus source captured the other and a mean spatial position was perceived. This in turns hints at a weighting process in audio-visual integration modulated by the level of noise in a given source signal. These findings are consistent with the notion of inverse effectiveness: when a characteristic of a given stimulus has low resolution there tends to be in an increase in "strength" of multisensory integration (Stevenson and James, 2009; de Dieuleveult et al., 2017). See Holmes (2009) for potential issues when measuring multisensory integration "strength" from the perspective of inverse effectiveness.

However, the very existence of audio-visual illusions in these processes highlights that there can be a cost associated with this optimal approach (Shams et al., 2005b); the perceptual illusions here are being considered as unwanted artifacts (costs) that manifest due to optimal integration of signals from multiple modalities. One such well-established audio-visual illusion that combines information from both modalities and arrives at an auditory percept altogether unique is the McGurk-MacDonald effect (McGurk and MacDonald, 1976). When participants watch footage of someone moving their lips, while simultaneously listening to an auditory stimulus (a single syllable repeated in time with the moving lips) that is incongruent to the moving lips, they have a tendency to "hear" a sound that matches neither the mouthed syllable or the auditory stimulus. While not gazing at the moving lips, participants accurately report the auditory stimulus.

The McGurk effect demonstrates that the integration of audio-visual information is an effective process in most natural settings (even when modalities provide competing information), but may occasionally result in an imperfect representation of events. This auditory illusion suggests a "best guess" can sometimes be arrived at when modalities provide contradictory information, where different weightings are given to competing modalities. Crevecoeur et al. (2016) highlighted that the nervous system also considers temporal feedback delays when performing optimal multi-sensory integration (for example, visual input followed by muscle response is slower than proprioceptive input followed by a muscle response with a difference of ~50ms). The faster of the two sensory cues is given a dominant weighting in integration. This shows that temporal characteristics affect optimal integration of information from different modalities, and should be a factor in any models of multi-modal integration.

If artifacts such as illusions can occur in an optimal multi-modal system, these artifacts become a concern when designing artificial cognitive systems. The optimal approach of integrating information from multiple sources may lead to inaccurate representation of an environment (an artifact), which in turn could result in a potentially hazardous outcome. For example, if a autonomous vehicle was trained in a specific environment and then relocated to a novel environment, an artifact manifested via optimal integration of stimuli could compromise the safe navigation through the novel environment and any action decisions taken therein (this scenario is a combination of the "Safe Exploration" and "Robustness to Distributional Shift" accident risks as outlined by Amodei et al., 2016).

The remainder of the paper reviews the processes in audio-visual perception that offers explanations for audio-visual illusions and effects, focusing mainly on how audition affects visual perception, and what this tells us about the audio-visual integration system. We continue by building a case for audio-visual integration as a process of evidence accumulation/discounting, where differing weights are given to different modalities depending on the stimuli information (spatial, temporal, featural etc.) being processed, which follows a hierarchical process (from within-modality discrimination to multi-modal integration). We highlight how some processes are optimal and others suboptimal, and how each have their own drawbacks. Following that, we review cognitive models of multi-modal integration which provide computational accounts for illusions. We then outline the potential implications of the mechanisms behind multisensory illusions for artificial intelligent systems, concluding with our views on future research directions. Additionally, rather than assuming that all attributions of prior entry (discussed below) are accurate, this paper expands on the definition of prior entry to encompass both response bias and undefined non-attentional processes. Doing so circumvents the granular debates surrounding prior entry in favor of better discussing the broader processes on the way to audio-visual integration, of which prior entry is but one. We also consider impletion (discussed below) as a process distinct from prior entry, but one that complements and/or competes with prior entry.

2. AUDIO-VISUAL INTEGRATION

2.1. Visual and Multi-Modal Prior Entry

Prior entry, a term coined by E.B. Titchener in 1908, describes a process whereby a visual stimulus that draws an observer's attention is processed in the visual perceptual system before any unattended stimuli in the perceptual field. This in turn results in the attended stimulus being processed "faster" relative to subsequently attended stimuli (Spence et al., 2001). This suggests that when attention is drawn (usually via a cue) to a specific region of space, a stimulus that is presented to that region is processed at a greater speed than a stimulus presented to unattended space.

Prior entry as a phenomenon is important in multi-modal integration due to the fact that the temporal perception of one modality can be significantly altered by stimuli in another

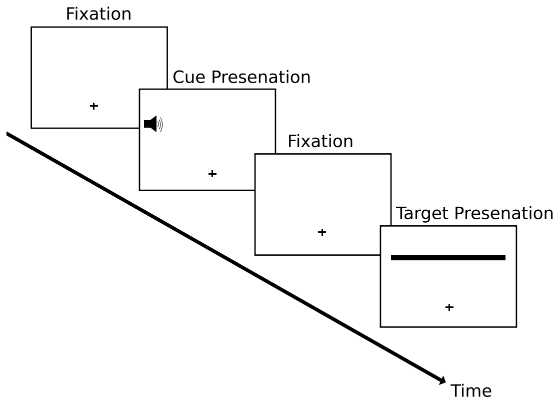

FIGURE 1 | The paradigm of the line-motion illusion similar to that used by Hikosaka et al. (1993a) and Shimojo et al. (1997). Participants were presented with a fixation cross followed by a cue (auditory in this example) to one side of fixation followed by another fixation display. Finally the target stimulus was presented. In this example the left side of the target stimulus was cued via an auditory tone. The resulting perception would be that of the line being drawn from the same side as the cue as opposed to being presented in its entirety at the same time.

modality (as well as within a modality) (Shimojo et al., 1997). The mechanisms underlying prior entry have been the subject of controversy (Cairney, 1975; Downing and Treisman, 1997; Schneider and Bavelier, 2003), but strong evidence has been provided for its existence via orthogonally designed crossmodal experiments (Spence et al., 2001; Zampini et al., 2005). In the case of the orthogonally designed experiments, related information between the attended modality and the subsequent temporal order judgement task was removed, thus ensuring no modality-specific bias.

A classic visual illusion that supports the tenets of prior entry, and demonstrates just how much temporal perception can be affected by it, is the line motion illusion, first demonstrated by Hikosaka et al. (1993a) using visual cues. A cue to one side of fixation prior to the presentation of a whole line to the display can result in the illusion of the line being "drawn" from the cued side (**Figure 1**). Hikosaka et al. (1993b) investigated this effect further and demonstrated illusory temporal order in a similar fashion: namely, cueing one side of fixation in a temporal order judgement task prior to the simultaneous onset of both visual targets. Both these effects were replicated using auditory cues by Shimojo et al. (1997).

Shimojo et al. (1997) demonstrated that the integration of auditory and visual stimuli can cause temporal order perception in one modality to be significantly altered by information in another via audio-visual prior entry. However, Downing and Treisman (1997) suggested that the original line motion illusion was an example of what they termed "impletion": in an ambiguous display multiple stimuli are combined to reflect a single smooth event in perception. For example, when an illusion of apparent motion is created using statically flashed stimuli in different locations (e.g., left space followed by right space), the stimuli can appear to smoothly change from the first stimulus shape (circle) to the second stimulus shape (square) (Kolers and von Grünau, 1976; Downing and Treisman, 1997). It is suggested that a discriminatory process gathers all available stimuli information, combines them, and creates a coherent percept; or the most likely real world outcome where it fills in the gaps on the way to perception. Downing and Treisman (1997) demonstrated that the line motion illusion could in fact be a perception of the visual cue itself streaking across the field of display akin to frames in an animation. Admittedly, when one takes into account the findings of Shimojo et al. (1997) using auditory cues, it may be tempting to dismiss the account of impletion, but illusory visual motion can be induced via auditory stimuli (Hidaka et al., 2009), which demonstrates that auditory stimuli can also induce a perception of motion in visual modality independent of prior-entry. Despite these alternative explanations for phenomena such as the line motion illusion, neuroscience has provided strong evidence for the existence of prior entry: speeded processing when attention was directed to the visual modality rather than the tactile (Vibell et al., 2007), speeded processing associated with attending to an auditory stimulus (Folyi et al., 2012), and speeded processing during a visual task when an auditory tone was presented prior to the onset of the visual stimuli (Seibold and Rolke, 2014).

Evidence thus suggests that prior entry, and indeed audio-visual prior entry, is a robust phenomenon. Whether all effects attributed to prior entry are done so correctly is another matter, but ultimately may be somewhat irrelevant (see Fuller and Carrasco, 2009 where evidence for both prior entry and impletion in the line motion illusion is presented, and suggests prior entry is not requisite). For instance, even if response bias or some non-attentional processes are mistakenly attributed to prior entry, these effects are still predictable, and replicable, and in fact these processes may enhance or exaggerate genuine prior entry effects.

The prior entry and impletion effects discussed above show that shifts in attention, or the combination of separate stimuli into the perception of a single stimulus event, can result in illusory temporal visual perception. It seems likely that evidence gathered from both the audio and visual modalities are combined optimally with some sources of information being given greater weighting in this process. When and how to assign weightings in an artificial system is an important consideration in design in order to avoid artifacts such as those described above. While prior entry and/or impletion can result in an inaccurate representation of temporal events, there exist audio-visual effects that are facilitatory in nature and thus desirable, which we discuss next.

2.2. Temporal Ventriloquism

Illusory visual temporal order, as shown above, can be induced by auditory stimuli. However, auditory stimuli, when integrated with visual stimuli, can also facilitate visual temporal perception: Scheier et al. (1999) discovered an audio-visual effect where spatially non-informative auditory stimuli affected the temporal perception of a visual temporal order judgement task. This effect became known as temporal ventriloquism, analogous to spatial ventriloquism where visual stimuli shifts the perception of auditory localization (Willey et al., 1937; Thomas, 1941; Radeau and Bertelson, 1987). Temporal ventriloquism was further

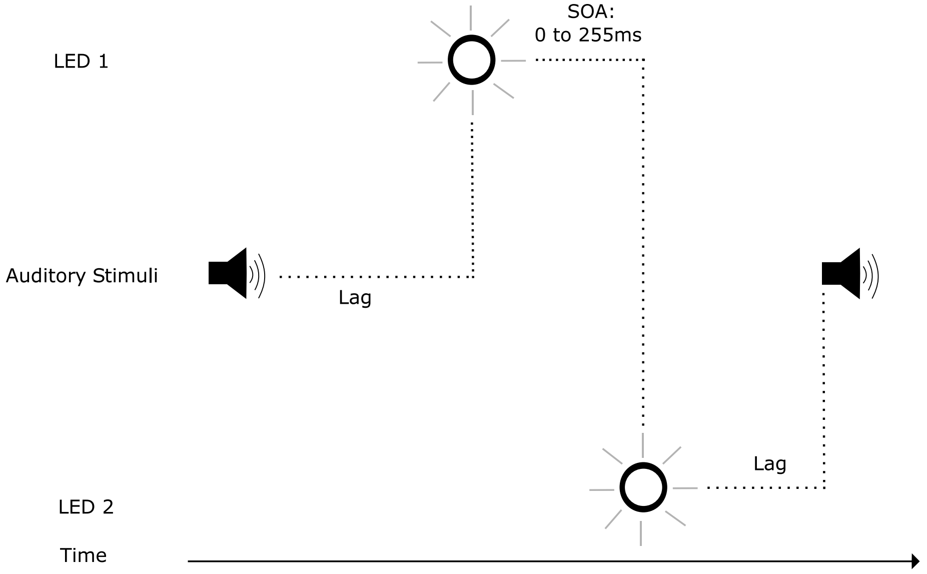

FIGURE 2 | Trial events in the temporal ventriloquism paradigm that resulted in enhanced temporal order judgement performance (Morein-Zamir et al., 2003). The stimulus onset asynchrony (SOA) between illumination of LEDs varied. The first auditory tone was presented before the first LED illumination. The second auditory tone was presented after the second LED illumination. The SOA between tones also varied.

investigated by Morein-Zamir et al. (2003): when accompanied by auditory tones, performance in a visual temporal order judgement task was enhanced (**Figure 2**). This enhancement was abolished when the tones coincided with the visual stimuli in time. When the two tones were presented between the visual stimuli in time (**Figure 3**), a detriment in performance was observed. In both conditions the tones appeared to "pull" the perception of the visual stimuli in time toward the auditory stimuli temporal onsets: further apart in the enhanced performance and closer together when a detriment in performance was observed (see however Hairston et al., 2006 for an argument against the notion of introducing a temporal gap between the stimuli in visual perception). The main driver of this effect was believed to be the temporal relationship between the auditory and visual stimuli, where the higher temporal resolution of the auditory stimuli carried more weight in integration. This is a potent example of how assigning weightings in multi-modal integration can have both positive and negative outcomes in terms of system performance.

Morein-Zamir et al. (2003) hypothesized that the quantity of auditory stimuli must match the quantity of visual stimuli in order for the temporal ventriloquism effects to occur. For example, when a single tone was presented between the presentation of the visual stimuli in time, there was no reported change in performance. Morein-Zamir et al. (2003) refer to the unity assumption: the more physically similar stimuli are to each other across modalities, the greater the likelihood they will be perceived as having originated from the same source (Welch, 1999), we discuss this in more detail later.

However, other research questions the assumption that a matching number of stimuli in both the auditory and visual modality are required to induce temporal ventriloquism. Getzmann (2007) studied an apparent motion paradigm, where participants perceive two sequentially presented visual stimuli behaving as one stimulus moving from one position to another. They found that when a single click was presented between the two visual stimuli, it increased the perception of apparent motion, essentially "pulling" the visual stimuli closer together in time in perception. This casts doubt on the idea that the

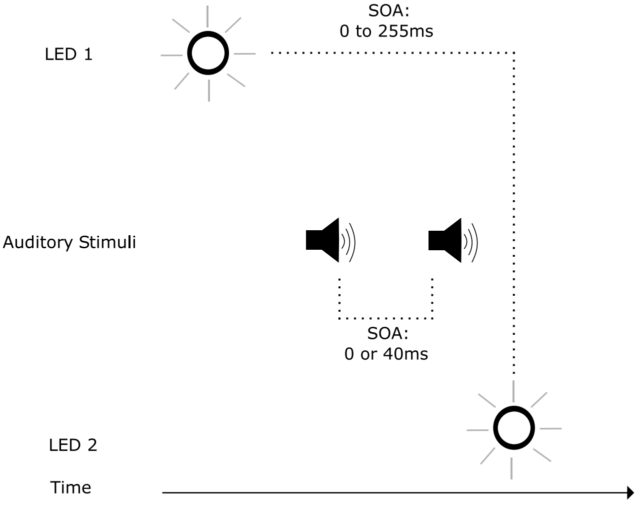

FIGURE 3 | Trial events in the temporal ventriloquism paradigm that resulted in a detriment in temporal order judgement performance (Morein-Zamir et al., 2003). The stimulus onset asynchrony (SOA) between illumination of LEDs varied. The first auditory tone was presented after the first LED illumination. The second auditory tone was presented before the second LED illumination. The SOA between tones also varied.

quantity of stimuli must be equal across modalities in order for, in this instance, an auditory stimulus to affect the perception of visual events.

Indeed, Boyce et al. (2020) demonstrated that a detriment in response bias corresponding to actual visual presentation order can be achieved with the presentation of a single tone in neutral space (different space to the visual stimuli) in a visual ternary response task (temporal order judgement combined with simultaneity judgements where the participant reports if stimuli were presented simultaneously). Importantly, this can be achieved consistently when presenting the single tone *prior* to the onset of the first visual stimulus (similar to the trial in **Figure 2** but without the second auditory stimulus). Often participants were as likely to make a simultaneity judgement report as they were to make a temporal order judgement report corresponding to actual sequential visual stimuli presentation order. This poses a problem for the temporal ventriloquism narrative: a single tone before the sequential presentation of visual stimuli in a ternary task would be expected to "pull" the perception of the first visual stimulus toward it in time, resulting in increased reports corresponding to the sequential order of visual stimuli. Alternatively, it might "pull" the perception of both visual stimuli in time with no observable effect on report bias should the matching quantity rule be abandoned. The repeated demonstrations of a decrease in report bias corresponding to the sequential order of visual stimuli suggest that the processes underlying temporal ventriloquism may be more flexible than previously suggested, and may have impletion-esque elements. Specifically, characteristics of stimuli such as their spatial and temporal relationship, and the featural similarity of stimuli within a single modality, may be weighted to arrive at the most likely real world outcome in perception regardless of whether the number of auditory stimuli equal the number of visual stimuli or not. Indeed, the number of auditory stimuli relative to visual in this example appears to modulate the type of temporal ventriloquist effect that might be expected to be observed.

Clearly not all conditions support the idea that the temporal relationship of an auditory stimulus to a visual stimulus drives temporal ventriloquism and similar effects. However, while there are no easy explanations for the audio-visual integration in temporal ventriloquism, efforts have been made to show that auditory stimuli do indeed "pull" visual stimuli in temporal perception. Freeman and Driver (2008) created an innovative paradigm that tested the idea that temporal ventriloquism is driven by auditory capture (in a similar fashion to that of Getzmann, 2007), that is to say there is a "pulling" of visual stimuli toward auditory stimuli in temporal perception. They began by determining the relative timings of visual stimuli that resulted in illusory apparent visual motion (**Figure 4**). Once visual stimulus onset asynchronies (SOAs) were established for the effect, Freeman and Driver (2008) adjusted the timings to remove bias in the illusion (**Figure 4B**). Following that, they introduced auditory stimuli (**Figure 4C**) with the same SOAs used to induce the effect in the visual-only condition (**Figure 4A**). In the presence of the auditory stimuli, both visual stimuli were "pulled" toward each other in time perception, and participants perceived a bias in the illusion.

This demonstrated that auditory stimuli had the ability to "pull" the respective visual stimuli in perceptual time toward the respective auditory onsets. In doing so, the visual stimuli now appeared in perception to have the same SOA as the auditory stimuli. This introduced a perceptual bias consistent with that observed for the visual stimuli SOA (in the absence of auditory stimuli) necessary to induce the same bias in illusory apparent visual motion. This meant predictable manipulation of the effect, and more specifically, demonstrated auditory capture of visual events in time.

Freeman and Driver (2008) suggest that the timings of the flanker stimuli (the stimuli used to induce temporal ventriloquism effects) in relation to the visual are the main drivers of temporal ventriloquism. Roseboom et al. (2013a) demonstrated that, in fact, the featural similarity of the flanker stimuli used to induce the effects described by Freeman and Driver (2008) have arguably as important a role to play at these time scales. Specifically, Roseboom et al. (2013a) replicated the findings of Freeman and Driver (2008) using auditory flankers. When flankers were featurally distinct (e.g., a sine wave and a white-noise burst) or flanker types were mixed via audio-tactile flankers, a mitigated effect was observed. It was significantly weaker compared to featurally identical audio-only or tactile-only flankers. This suggests that temporal capture in-and-of-itself is not sufficient when describing the underlying mechanisms that account for this effect, or temporal ventriloquism in general at the reported time scales. Roseboom et al. (2013a) also demonstrated that the reported illusory apparent visual motion could be induced when the flanker stimuli was presented synchronously with the target visual stimuli. This suggests that temporal ansynchrony is not a requisite to induce this illusion in a directionally ambiguous display. Keetels et al. (2007) further highlighted the importance of featural characteristics when inducing the temporal ventriloquism effect. However, at shorter time scales, featural similarity appears not to play as large a role where timing is reasserted as the main driver (Kafaligonul and Stoner, 2010, 2012; Klimova et al., 2017).

The above research is consistent with the unity assumption, where an observer makes an assumption about two sensory signals, such as auditory and visual (or indeed, signals from the same modality), originating from a single source or event (Vatakis and Spence, 2007, 2008; Chen and Spence, 2017). Vatakis and Spence (2007) demonstrated that when auditory and visual stimuli were mismatched (for example, speech presentation where the voice did not match the gender of the speaker) participants found it easier to judge which stimulus was presented first; auditory or visual. The task difficulty increased when the stimuli were matched suggesting an increased likelihood of perceiving the auditory and visual stimuli occurring at the same time. This finding provides support for the unity assumption in audio-visual temporal integration of speech via the process of temporal ventriloquism. See Vatakis and Spence (2008) for limitations of the unity assumption's influence over audio-visual temporal integration of complex non-speech stimuli. See also Chen and Spence (2017) for a thorough review of the unity assumption and the myriad debates surrounding it, and how it relates to Bayesian causal inference.

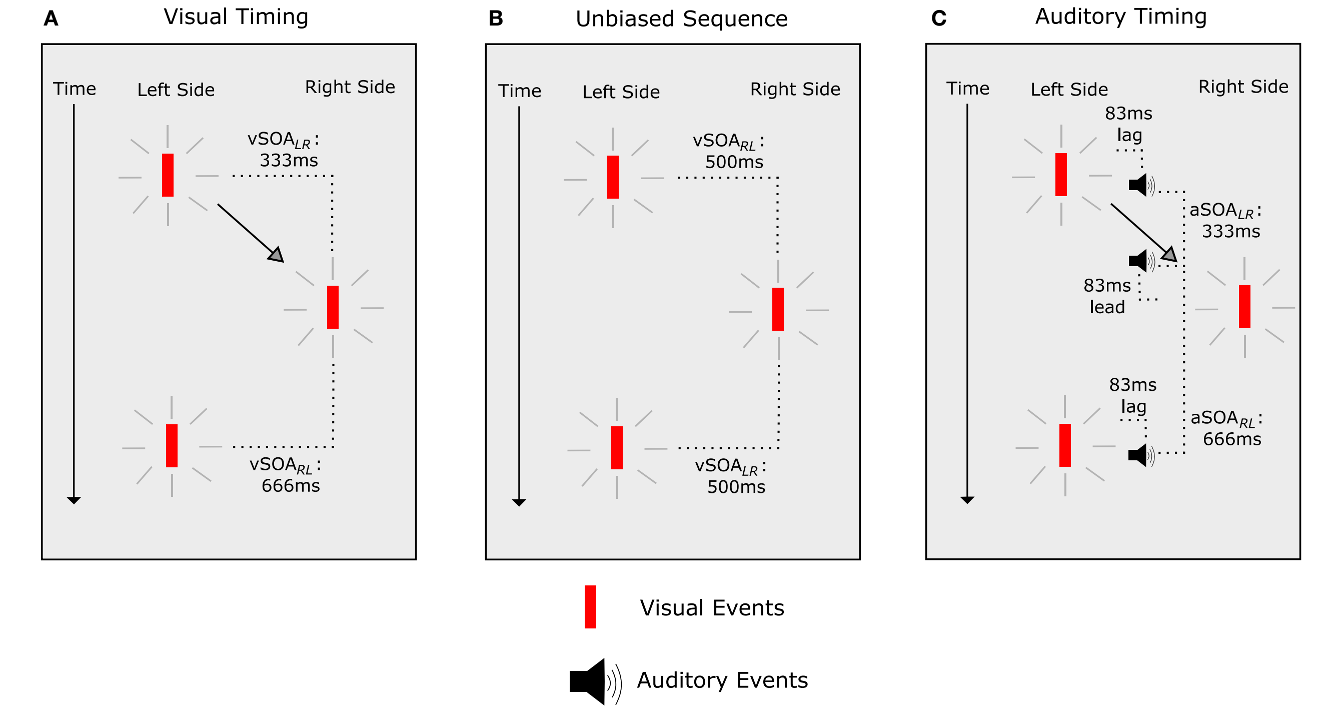

FIGURE 4 | Illusory apparent visual motion paradigm (Freeman and Driver, 2008). **(A)** The visual SOAs (vSOA) between stimuli. "L" denotes the left stimulus, and "R" the right stimulus. When the vSOA was 333 ms, apparent motion in the direction of the second stimulus was perceived. When there was a vSOA of 666 ms, no apparent motion was perceived. **(B)** When vSOAs were 500 ms, there was no bias in apparent motion. **(C)** When vSOAs were 500 ms but auditory stimuli were presented with an SOA (aSOA) of 333 ms, participants perceived apparent motion in the same direction as would be expected with a vSOA of 333 ms. When the aSOA was 666 ms, illusory apparent visual motion was abolished.

The findings by Roseboom et al. (2013a) and Keetels et al. (2007) show that there is often a much more complex process of integration than simply auditory stimuli (or other stimuli of high temporal resolution) capturing visual stimuli in perception. There would appear to be a process of evidence accumulation and evidence discounting: when two auditory events are featurally similar, and their temporal relationship with visual stimuli is close, the auditory and visual stimuli are integrated. However, when two auditory stimuli meet the temporal criteria for integration with visual stimuli, but these auditory stimuli are featurally distinct and therefore deemed to be from unique sources, they are not wholly integrated with the visual stimuli. In the second example, some of the accumulated temporal evidence is discounted due to evidence of unique sources being present.

Taken together, this evidence builds a more complicated picture of temporal ventriloquism in general, and the modulated illusory apparent visual motion direction effect (Freeman and Driver, 2008) in particular. Indeed, the unity assumption potentially plays a role here when one considers the effect featural similarity has on the apparent grouping of flankers.

2.3. Additional Audio-Visual Effects

Most of the research discussed thus far has focused on the effects of audio-visual integration on space and time perception in the visual modality. As highlighted by the McGurk effect, audio-visual integration can also have other surprising outcomes in perception. Shams et al. (2002) demonstrated that when a single flash of a uniform disk was accompanied by two or more tones, participants tended to perceive multiple flashes of the disk (**Figure 5**). When multiple physical flashes were presented and accompanied by a single tone, participants tended to perceive a single presentation of the disk (Andersen et al., 2004). These effects were labeled as *fission* in the case of illusory flashes, and *fusion* in the case of illusory single presentation of the disk. Fission and fusion differ from the likes of temporal ventriloquism and prior entry in that they increase or decrease the quantity of perceived stimuli. After training, or when there was a monetary incentive, qualitative differences were detectable between illusory and physical flashes (Rosenthal et al., 2009; vanErp et al., 2013). However, the illusion persisted despite the ability to differentiate. Similarities may be drawn between the effect reported by Shams et al. (2002) and Shipley (1964) where, when the flutter rate of an auditory signal was increased, participants perceived an increased flicker frequency of a visual signal. However, there was a relatively small quantitative change in flicker frequency, whereas fission is a pronounced change in the visual percept (a single stimulus perceived as multiple stimuli).

Neuroimaging evidence provided further insights into fission/fusion effects. Specifically, in the presence of auditory stimuli the BOLD response in the retinotopic visual cortex increased whether fission was perceived or not (Watkins et al., 2006). The inverse was true when the fusion illusion was

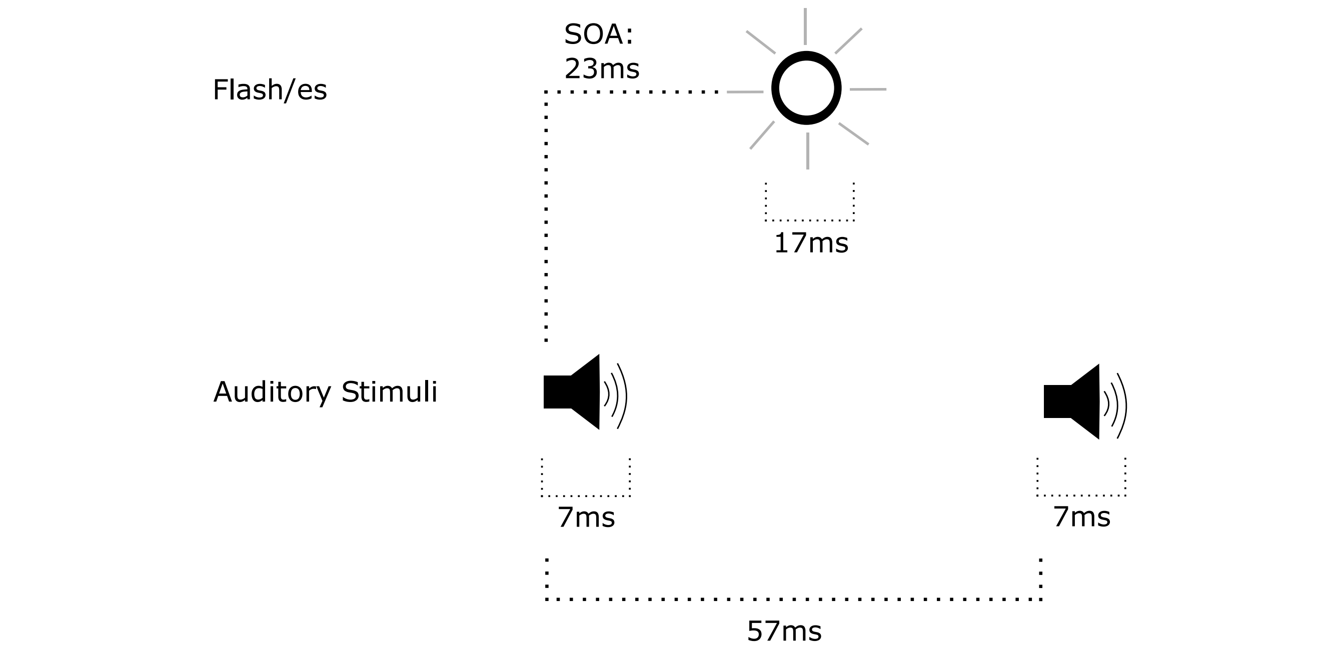

FIGURE 5 | Trial events in the multiple flash illusion paradigm (Shams et al., 2002). A tone was presented before visual stimulus onset, and a tone was presented after visual stimulus onset. The SOA between tones was 57 ms. This paradigm resulted in the perception of multiple flashes when in fact the visual stimulus was presented only once for 17 ms.

perceived. This suggests that the auditory and visual perceptual systems are intrinsically linked, and reflects the additive nature of the fission illusion and the suppressive nature of the fusion illusion that "removes" information from visual perception (auditory stimuli has also been shown to have suppressive effects on visual perception Hidaka and Ide, 2015).

Shams et al. (2002) proposed the discontinuity hypothesis as an underlying explanation for the fission effect: discontinuous stimuli must be present in one modality in order to "dominate" another modality during integration. However, as alluded to above, Andersen et al. (2004) demonstrated that this was not the case via the fusion illusion. Fission and fusion once again align with the ideas of impletion and the unity assumption. Consistent with the influence featural similarity of flankers had on illusory apparent visual motion, the fission effect was completely abolished when the tones used were distinct from each other: one a sine wave, the other a white noise burst; or both featurally distinct sine wave tones (a 300 Hz sine wave and a 3,500 Hz) (Roseboom et al., 2013b; Boyce, 2016).

Another effect that seems to be governed by the featural similarity of auditory stimuli is the stream bounce illusion. In this illusion, two uniform circles move toward each other from opposite space, and when a tone presented at the point of overlap of the circles differed featurally to other presented tones, the circles are perceived to "bounce" off each other. When multiple tones were featurally identical, the circles often appeared to cross paths and continue on their original trajectory (Sekuler et al., 1997; Watanabe and Shimojo, 2001). Taken with the above and similar research (Keetels et al., 2007; Cook and Van Valkenburg, 2009), this suggests that auditory streaming (where a sequence of auditory stimuli are assigned the same or differing origins) processes play an integral role in audio-visual illusions and integration in general.

Auditory motion can also have a profound effect on visual perception where a static flashing visual target is perceived to move in the same direction as auditory stimuli (**Figure 6**, see also Hidaka et al., 2011; Fracasso et al., 2013). Perceived location of apparent motion visual stimuli can also be modulated by auditory stimuli (Teramoto et al., 2012). The visual motion direction selective brain region MT/V5 is activated in the presence of moving auditory stimuli, suggesting processing for auditory motion occurs there (Poirier et al., 2005), which in turn hints at an intrinsic link between auditory and visual perceptual systems. These effects taken together again point to evidence accumulation in audio-visual integration as an optimal process, where different weight is given to auditory and visual inputs.

Consideration should be lent to how and when multiple stimuli in single modality should be grouped together as originating from a single source, or not, before pairing with another modality. As demonstrated above from neuropsychological evidences, and also from recent systems neuroscience evidences (Ghazanfar and Schroeder, 2006; Meijer et al., 2019), the human audio-visual integration system appears to operate in rather complex processing steps, in addition to the traditional thinking of hierarchical processing from single modality. Hence, there is a need for modeling these cognitive processes. When designing artificial cognitive systems, efforts should be made to isolate sources of auditory and visual stimuli,

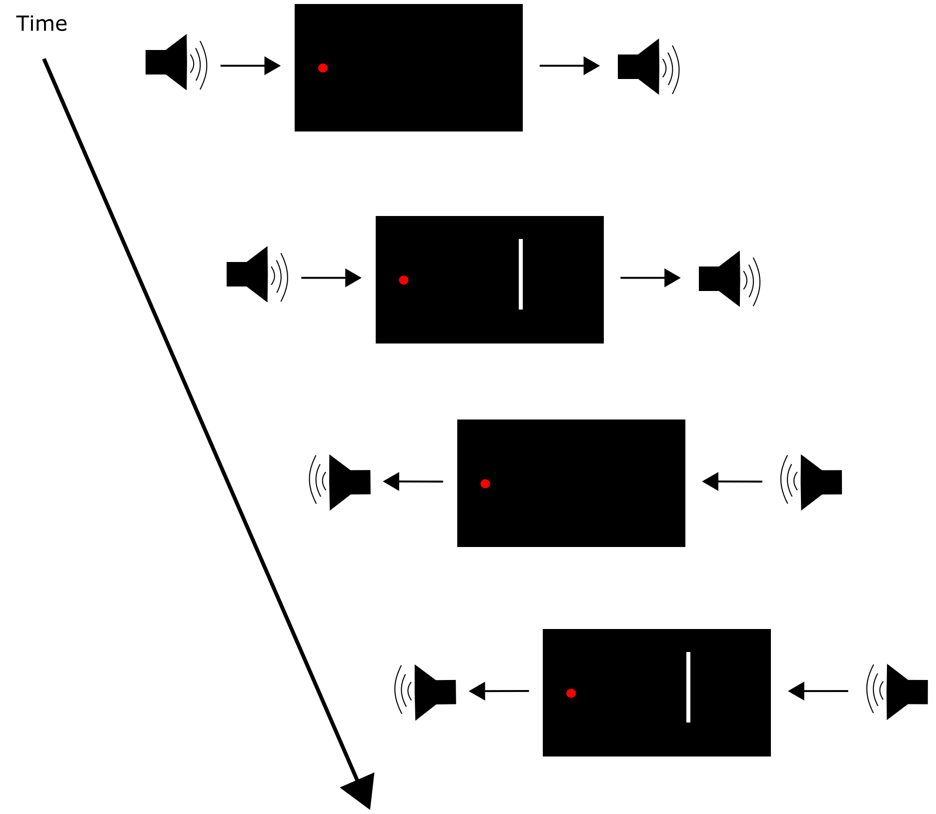

FIGURE 6 | Trial events in the sound induced illusory visual motion paradigm (Hidaka et al., 2011). The flashing visual stimulus (white bar above) was presented with varying eccentricities from fixation (red dot) depending on the trial condition. The auditory stimulus in the illusory condition was first presented to one ear and panned to the opposing ear. This paradigm resulted in the illusion of motion often in the same direction as that of the auditory stimulus motion.

and identify characteristics that would suggest they are related events. As shown above, relying on similar temporal signatures alone is not a robust approach when integrating signals across modalities. The illusions discussed above are summarized in **Table 1**.

3. COMPUTATIONAL COGNITIVE MODELS

We have discussed how auditory stimuli can have a pronounced effect on the perception of visual events, and vice versa, be it temporal or qualitative in nature. For auditory stimuli affecting the perception of visual signals, some effects were additive, facilitatory, and others suppressive. Regardless of the outcome, the influence of auditory stimuli on visual perception provides evidence of complexity of audio-visual integration processes on the way to visual perception.

These complex mechanisms of how and when auditory stimuli alter visual perception have been clarified through computational modeling. Chandrasekaran (2017) presents a review of computational models of multisensory integration, categorizing the computational models into accumulator models, probabilistic models, or neural network models. These types of models are also typically used in single-modal perceptual decision-making (e.g., Ratcliff, 1978; Wang, 2002; Wong and Wang, 2006). In this review, we will only focus on the accumulator and probabilistic models; the neural network (connectionist) models provide finer-grained, more biologically plausible description of neural processes, but on the behavioral level are mostly similar to the models reviewed here (Bogacz et al., 2006; Ma et al., 2006; Wong and Wang, 2006; Ma and Pouget, 2008; Roxin and Ledberg, 2008; Liu et al., 2009; Pouget et al., 2013; Ursino et al., 2014, 2019; Zhang et al., 2016; Meijer et al., 2019).

3.1. Accumulator Models

The race model is a simple model that accounts for choice distribution and reaction time phenomena, e.g., faster reaction times of multisensory than unisensory stimuli (Raaja, 1962; Gondan and Minakata, 2016; Miller, 2016). More formally, the multisensory processing time D_{AV} is the winner of two channel's processing times D_A and D_V for audio and visual signals:

$$D_{AV} = min(D_A, D_V) \tag{1}$$

Another type of accumulator model, the coactivation model (Schwarz, 1994; Diederich, 1995), is based on the classic accumulator-type drift diffusion model (DDM) of decision-making (Stone, 1960; Link, 1975; Ratcliff, 1978; Ratcliff and

TABLE 1 | Illusions summary.

Illusion	Description
The line-motion illusion:	When one side of space is cued prior to the presentation of the entire physical line it results in the perception of the line being drawn from that cued side of space
Illusory temporal order I:	When a tone is presented to one side of congruent space prior to the simultaneous presentation of both targets in a simultaneity judgement task, illusory sequential order is perceived
Illusory temporal order II:	When a tone is presented in neutral space prior to the simultaneous presentation of both targets in a ternary judgement task, and a tone is presented in neutral space after the onset of both targets, illusory sequential order is perceived
Temporal ventriloquism - performance enhancement:	When a tone is presented before the first visual stimulus in a temporal order judgement sequence and a tone is presented after the second visual stimulus, performance is improved
Temporal ventriloquism - performance detriment I:	When a tone is presented after the first visual stimulus in a temporal order judgement sequence and a tone is presented before the second visual stimulus, performance is worsened
Temporal ventriloquism - performance detriment II:	When a single tone is presented before the first visual stimulus in a temporal order judgement sequence, response bias matching the presentation order of visual stimuli is reduced
Temporal ventriloquism - illusory apparent visual motion:	When auditory stimuli are presented in neutral space but with specific SOAs in relation to visual stimuli the perception of apparent motion can be modulated
Multiple flash illusion:	When two or more tones are presented either side of a single presentation of a circle in time, multiple flashes of the circle are perceived
Single flash illusion:	When a single tone is presented with multiple flashes of a circle a single presentation of the circle is perceived
Sound-induced illusory apparent visual motion:	When auditory stimuli are presented panning from one ear to the other in time with a static flashing visual target, visual apparent motion is perceived

Rouder, 1998). The DDM is a continuous analog of a random walk model (Bogacz et al., 2006), using a drift particle with state X at any moment in time to represent a decision variable (relating in favor of one over another choice). This is obtained through integrating noisy sensory evidence over time in the form of a stochastic differential equation, a biased Brownian motion equation:

$$dX = Adt + cdW, \tag{2}$$

where A is the stimulus signal (i.e., the drift rate), c is the noise level, and W represents the stochastic Wiener process. Integration of the sensory evidence begins from an initial point (usually origin point 0), and is bounded by the lower and upper decision thresholds, $-z$ and z, respectively. Each threshold corresponds to a decision in favor of one of the two choices. Integration of the sensory information continues until the drift particle encounters either the upper or the lower threshold, at which stage a decision is made in favor of the corresponding option. The drift particle is then reset to the origin point to allow the next decision to be processed. The DDM response time (RT) is calculated as the time taken for the drift particle to move from its origin point to the either of the decision thresholds and can include a brief, fixed non-decision latency. For the simplest DDM, the RT has a closed form analytical solution (Ratcliff, 1978; Bogacz et al., 2006):

$$RT = \frac{z}{A}\tanh(\frac{Az}{c^2}) \tag{3}$$

Similarly, the corresponding analytical solution for the DDM's error rate (ER) is:

$$ER = \frac{1}{1 + \exp(\frac{2Az}{c^2})} \tag{4}$$

A simple unweighted coactivation model would combine evidence from two modalities and integrate it over time using the DDM (Schwarz, 1994; Diederich, 1995). For example, with unimodal sensory evidence X_1 and X_2 (e.g., auditory and visual information), the combined evidence is just a simple summation over time using (2):

$$X_c = X_1 + X_2 \tag{5}$$

3.2. Bayesian Models

In contrast to these models, the Bayesian modeling framework offers an elegant approach to modeling multisensory integration (Angelaki et al., 2009), although they share some similar characteristics with drift-diffusion models (Bitzer et al., 2014; Fard et al., 2017). This approach can provide optimal or near optimal integration of multimodal sensory cues by weighting the incoming evidences from each modality. For example, if the modalities follow a Gaussian distribution, the optimal integration estimate X_c is (Bülthoff and Yuille, 1996):

$$X_c = \frac{k_1^2}{k_1^2 + k_2^2}\widehat{X_1} + \frac{k_2^2}{k_1^2 + k_2^2}\widehat{X_2} \tag{6}$$

where $k_1 = \frac{1}{\sigma_1^2}$ and $k_2 = \frac{1}{\sigma_2^2}$ and $\widehat{X_1}$, $\widehat{X_2}$, σ_1 and σ_2 are the means and standard deviations for modality 1 and 2, respectively. Interestingly, the model can show that the combined variance (noise) will always be less than their individual estimates when the latter are statistically independent (Alais and Burr, 2004). This justifies why combining the signals help reduce the overall noise. In fact, Beck et al. (2012) makes a strong case for suboptimal inference, that the larger variability is due to deterministic, but suboptimal computation, and that the latter, not internal or external noise, is the major cause of variability in behavior.

A more complex model, TWIN (time window of integration), involves a combination of the race model and the coactivation model (Colonius and Diederich, 2004). Specifically, whichever modality is first registered (as in winning a "race"), the size of the window is dynamically adapted to the level of reliability of the sensory modality. This would ensure, for instance, that if the less reliable modality wins the race, the window would be increased to give the more reliable modality a relatively higher contribution in multisensory integration. This model accounts for the illusory temporal order induced via a tone after visual stimuli onset, where the more reliable temporal information (auditory stimulus) dictates the perceptual outcome—illusory temporal order (Boyce et al., 2020).

The fission and fusion in audio-visual integration were suggested to result from statistically optimal computational strategy (Shams and Kim, 2010), similar to Bayesian inference where audio-visual integration implies decisions about weightings assigned to signals and decisions whether to integrate these signals. Battaglia et al. (2003) applied this Bayesian approach to reconcile two seemingly separate audio-visual integration theories. The first theory, called visual capture, is a "winner-take-all" model where the most reliable signal (least variance) dominates, while the second theory used a maximum-likelihood estimation to identify the weight average of the sensory input. The visual signal was shown to be dominant because of the subject perceptual bias, but the weighting given to auditory signals increased as visual reliability decreased. Battaglia et al. (2003) showed that Equation (6) can naturally account for both theories by having the weights to vary based on the signals' variances.

To study how children and adults differ in audio-visual integration, Adams (2016) also used the same Bayesian approach in addition to two other models of audio-visual integration: a focal switching model, and a modality-switching model. The focal switching model stochastically sampled either auditory or visual cues based on subjects' reports of the observed stimulus. For the modality-switching model, the stochastically sampled cues were probabilistically biased toward the likelihood of the stimulus being observed. Adams (2016) found that the sub-optimal switching models modeled sensory integration in the youngest study groups best. However, the older participants followed the partial integration of an optimal Bayesian model.

3.2.1. Illusions as a By-product of Optimal Bayesian Integration

A variety of perceptual illusions have been shown to result from optimal Bayesian integration of information coming from multiple sensory modalities. In the context of sound-induced flash illusion (**Figure 5**), given independent auditory and visual sensory signals A, V, the ideal Bayesian observer estimates posterior probabilities of the number of source signals as a normalized product of single-modality likelihoods $P(A|Z_A)$ and $P(V|Z_V)$ and joint priors $P(Z_A, Z_V)$

$$P(Z_A, Z_V|A, V) = \frac{P(A|Z_A)P(V|Z_V)P(Z_A, Z_V)}{P(A, V)}. \tag{7}$$

Regardless of the degree of consistency between auditory and visual stimuli, the optimal observer (7) have been shown to be consistent with the performance of human observers (Shams et al., 2005b). Specifically, when the discrepancy between the auditory and visual source signals is large, human observers rarely integrate the corresponding percepts. However, when the source signals overlap to a large degree, the two modalities are partially combined; in these cases the more reliable auditory modality shifts the visual percepts, thereby leading to sound-induced flash illusion.

Existence of different causes for signals of different modalities is the key assumption of the optimal observer model developed in Shams et al. (2005b), which allowed it to capture both full and partial integration of multisensory stimuli, with the latter resulting in illusions. Körding et al. (2007) suggested that in addition to integration of sensory percepts, optimal Bayesian estimation is also used to infer the causal relationship between the signals; this was consistent with spatial ventriloquist illusion found in human participants. Alternative Bayesian accounts developed by Alais and Burr (2004) [using Equation (6)] and Sato et al. (2007) also suggest that the spatial ventriloquist illusion stems from the near optimal integration of spatial and auditory signals.

Evidence for optimal Bayesian integration as the primary mechanism behind perceptual illusions comes from the paradigms involving not only audio-visual, but also other types of information. Wozny et al. (2008) applied the model of Shams et al. (2005b) to trimodal, audio-visuo-tactile perception, through simple extension of Equation (7). This Bayesian integration model accounted for cross-modal interactions observed in human participants, including touch-induced auditory fission, and flash- and sound-induced tactile fission (Wozny et al., 2008). Further evidence for Bayesian integration of visual, tactile, and proprioceptive information is provided by the *rubber hand illusion* (Botvinick and Cohen, 1998), in which a feeling of ownership of a dummy hand emerges soon after simultaneous tactile stimulation of both the concealed own hand of the participant and the visible dummy hand (see Lush, 2020 for a critique of control methods used in the "rubber hand" illusion). The optimal causal inference model (Körding et al., 2007) adapted for this scenario accounted for this illusion (Samad et al., 2015). Moreover, the model predicted that if the distance between the real hand and the rubber hand is small, the illusion would not require any tactile stimulation, which was also confirmed experimentally (Samad et al., 2015).

Finally, Bayesian integration has recently been shown to account even for those illusions which were previously striking researchers as "anti-Bayesian", for the reason that the empirically observed effects had the direction opposite to the effects predicted by optimal integration. Such "anti-Bayesian" effects are the size-weight illusion (Peters et al., 2016) (of the two objects with same mass but different size, the larger object is perceived to be lighter), and the material-weight illusion (Peters et al., 2018) (of the two objects with the same mass and size, the denser-looking object is perceived to be lighter). In both cases, the models explaining these two illusions involved optimal Bayesian estimation of latent variables (e.g., density), which affected the final estimation of weight.

Altogether, the reviewed evidence from diverse perceptual tasks illustrates the ubiquity of optimal Bayesian integration and its role in emergence of perceptual illusions.

3.2.2. Temporal Dimension in Bayesian Integration

Basic Bayesian modeling framework often does not come with a temporal component, unlike dynamical models such as accumulator. However, a recent study shows that when optimal Bayesian model is combined with the DDM, it can provide optimal and dynamic weightings to the individual sensory modalities. In the case of visual and vestibular integration, using an experimental setup similar to that of Fetsch et al. (2009) and Drugowitsch et al. (2014) found a Bayes-optimal DDM to integrate vestibular and visual stimuli in a heading discrimination task. It allowed the incorporation of time-variant features of the vestibular motion, i.e., motion acceleration, and visual motion velocity. The Bayesian framework allowed the calculation of a combined sensitivity profile $d(t)$ from the individual stimulus sensitivities.

$$d(t) = \sqrt{\frac{k_{vis}^2(c)}{k_{comb}^2(c)}v^2(t) + \frac{k_{vest}^2(c)}{k_{comb}^2(c)}a^2(t)} \tag{8}$$

where $k_{vis}(c)$, $k_{vest}(c)$, and $k_{comb}(c)$ are the visual, vestibular and combined stimulus sensitivities, and $v(t)$ and $a(t)$ are the temporal sensitivities of the visual and vestibular stimuli, respectively. Drugowitsch et al. (2014) found that Bayes-optimal DDM led to suboptimal integration of stimuli when subject response times were ignored. However, when response times were considered, the decision-making process took longer but resulted in more accurate responses. That said, a significant limitation of the study by Drugowitsch et al. (2014) and related work is that it does not incorporate delays in information processing. More generally, current Bayesian models do not consider how temporal delays impact sensory reliability. Delays are particularly relevant for feedback control in the motor system and processes like audio-visual speech because different sensory systems are affected by different temporal delays (McGrath and Summerfield, 1985; Jain et al., 2015; Crevecoeur et al., 2016).

So far, the modeling approaches do not generally take into account the effects of attention, motivation, emotion, and other "top-down" or cognitive control factors that could potentially affect multimodal integration. However, there are experimental studies of top-down influences, mainly attention (Talsma et al., 2010). More recently, Maiworm et al. (2012) showed that aversive stimuli could reduce the ventriloquism effect. Bruns et al. (2014) designed a task paradigm in which rewards were differentially allocated to different spatial locations (hemifields), creating a conflict between reward maximization and perceptual reliance. The auditory stimuli were accompanied by task-irrelevant, spatially misaligned visual stimuli. They showed that the hemifield with higher reward had a smaller ventriloquism effect. Hence, reward expectation could modulate multimodal integration and illusion, possibly through some cognitive control mechanisms. Future computational studies, e.g., using reward rate analysis (Bogacz et al., 2006; Niyogi and Wong-Lin, 2013), should address how reward and punishment are associated with such effects.

4. AUDIO-VISUAL SYSTEMS IN THE ARTIFICIAL

Multimodal integration and sensor fusion in artificial systems have been an active research field for decades (Luo and Kay, 1989, 2002), since using multiple sources of information can improve artificial systems in many application areas, including smart environments, automation systems and robotics, intelligent transportation systems. Integration of sensory modalities to generate a percept can occur at different stages, from low (feature) to high (semantic) level. The integration of several sources of unimodal information at middle and high level representations (Wu et al., 1999; Gómez-Eguíluz et al., 2016) has clear advantages: interpretability, simplicity of system design, and avoiding the problem of increasing dimensionality of the resulting integrated feature. Although model dependent, lower dimensionality of the feature space typically leads to better estimates of parametric models and computationally faster non-parametric models for a fixed amount of training data, which in turn can reduce the number of judgement failures. However, percept integration at the representation level lacks robustness and does not account for the way humans integrate multisensory information (Calvert et al., 2001; Shams et al., 2005a; Watkins et al., 2006; Stein et al., 2014) to create these percepts (Cohen, 2001). Temporal ventriloquism and the McGurk effect are just two examples of the result of the lower-level integration of sensory modalities in humans to create percepts, yet the differences with artificial systems go even further. While human perceptual decision-making is based on a dynamic process of evidence accumulation of noisy sensory information over time (see above), artificial systems typically follow a snapshot approach, i.e., percepts are created on the basis of instantaneous information, and only from data over time-windows when the perception mainly unfolds over time. Therefore, we can distinguish between decisions made over accumulated evidence, i.e., decision-making, and decisions made following the snapshot approach, i.e., classification, even though sometimes these two approaches are combined.

Audio-visual information integration is one of the multi-sensory mechanisms that has increasingly attracted research interests in the design of artificial intelligent systems. This is mainly due to the fact that humans heavily rely on these sensing modalities, and advances in this area have been facilitated by the high level of maturity of the individual areas involved, for instance signal processing, speech recognition, machine learning, and computer vision. See Parisi et al. (2017) and Parisi et al. (2018) for examples of how human multisensory integration in spatial ventriloquism has been used to model human-like spatial localization responses in artificial systems in which—given a scenario where sensor uncertainty exists in audio-visual information streams—they propose artificial neural architectures for multisensory integration. An interesting characteristic of audio-visual processing compared to other multimodal systems is the fact that the information unfolds over time for audio signals, but also for visual systems when video is considered instead of still images. However, most of the research in artificial visual systems follow the snapshot approach mentioned above to build percepts, while video processing mainly focuses on integrating and updating of these instantaneous percepts over time, which can be seen as evidence accumulation. Like for other multimodal integration modalities, audio-visual integration in artificial systems can be performed at different levels, although is generally used for classification purposes, while decision-making, when performed, is based on the accumulation of classification results. Optimal temporal integration of visual evidence together with audio information can be prone to the sort of illusory effects on percepts illustrated above in humans. However, because artificial systems are designed with very specific objectives, an emergent deviation of the measurable targets of the system would be considered as a failure or bug of the system. Therefore, although artificial systems can display features that could be the emergent results of the multimodal integration, they will be regarded as failures to be avoided, and most likely not reported in the literature. A close example related to reinforcement learning is the reward hacking effect (Amodei et al., 2016), where a learning agent finds an unexpected (maybe undesired) optimal policy for a given learning problem.

As stated earlier, multimodal integration is typically performed at high level, as low-level integration generates higher dimensional data, thereby increasing the difficulty of processing and analysis. Moreover, the low-level integration of raw data can have the additional problem of combining data of very different nature. The dimensionality problem is magnified by the massive amount of data visual perception produces, therefore most approaches to audio-visual processing in intelligent systems also address the problem of integration at a middle and higher levels across diverse applications: object and person tracking (Nakadai et al., 2002; Beal et al., 2003), speaker localization and identification (Gatica-Perez et al., 2007), multimodal biometrics (Chibelushi et al., 2002), lip reading and speech recognition (Sumby and Pollack, 1954; Luettin and Thacker, 1997; Chen, 2001; Guitarte-Pérez et al., 2005) and video annotation (Wang et al., 2000; Li et al., 2004), and others. The computational models described above can be identified with these techniques for artificial systems, as they generate percepts and perform decision-making on the basis of middle-level fusion of evidence. However, some work in artificial systems deals with the challenging problem of combining data at the signal level (Fisher et al., 2000; Fisher and Darrell, 2004). While artificial visual systems were dominated by feature definition, extraction, and learning (Li and Allinson, 2008), the success of deep learning and convolutional neural networks in particular has shifted the focus of computer vision research. Likewise, speech processing is adopting this new learning paradigm, yet audio-visual speech processing with deep learning is still based mainly on high-level integration (Deng et al., 2013; Noda et al., 2015). Although the human audio-visual processing is not fully understood, our knowledge of the brain strongly inspires (and biases) the design of artificial systems. Besides the well-known (yet not widely reported) reward hacking in reinforcement learning and optimization (Amodei et al., 2016), to the best of our knowledge no illusory percepts have been reported in specific-purpose artificial systems, as they are typically situations to be avoided.

5. DISCUSSION

The multi-modal integration processes and related illusions outlined above are closely related in terms of how audition affects visual perception. Untangling whether prior entry (whatever form it may take), impletion, temporal ventriloquism, or featural similarity of auditory stimuli are the drivers of audio-visual effects can be a challenge, and may be missing the bigger picture when trying to understand how perception is arrived at in a noisy world. The most likely explanation of the discussed effects is one of an overarching unified process of evidence accumulation and evidence discounting. This perspective would state that evidence is gathered via multiple modalities and is filtered through multiple sub-processes: prior entry, auditory streaming, impletion, and temporal ventriloquism. Two or more of these sub-processes will often interact, with various weightings given to each process. For example, prior entry using a single auditory cue can induce an illusion of temporal order, but with the addition of a cue in the unattended side of space after the presentation of both target visual stimuli, extra information in favor of temporal order can be accumulated, which would increase the strength of the illusion (Boyce et al., 2020). Similarly, illusory temporal order can be induced via spatially neutral tones (an orthogonal design), as demonstrated by Boyce et al. (2020), which appears to combine prior entry and temporal ventriloquism, and impletion-like processes generally. The illusory temporal order induced via spatially neutral tones is significantly weaker when compared to the spatially congruent audio and visual stimuli equivalent, highlighting the relative weight given to spatial congruency. Additionally, when featurally distinct tones are used for both of these effects, the prior entry illusory order is preserved while the illusory order induced by spatially neutral tones is completely abolished (Boyce et al., 2020). This highlights how spatial information carries greater weight than the featural information of the tones used when the auditory and visual stimuli are spatially congruent. Conversely, it also highlights how featural characteristics carry greater weight in the absence of audio-visual spatial congruency.

As outlined above, temporal ventriloquism effects can also interact with auditory streaming where the features of auditory stimuli undergo a process of grouping, and the outcome can dictate whether the stimuli is paired or not with visual stimuli. The mere fact that the temporal signature of stimuli is not in-and-of-itself enough to induce an effect (at the times discussed here) suggests that sub-processes interact across modalities. Specifically, when auditory stimuli are not grouped in the streaming process, there is less evidence that they belong to the same source, and in turn it is less likely both auditory stimuli belong to the same source as the visual events. These types of interactions taken with different outcomes in visual perception, depending on the number of auditory stimuli used, point toward an overarching process that fits an expanded version of impletion, or a unifying account of impletion (Boyce et al., 2020) (aligning with Bayesian inference), where the most likely real world outcome is reflected in perception. The observer weighs evidence from both modalities in multi-modal perception and also weighs evidence within a single modality. This suggests an inherent weighted hierarchy, where spatial, featural, and temporal information are all taken into account.

The discussed integration processes are often statistically optimal in nature (Alais and Burr, 2004; Shams and Kim, 2010). This has implications for designing artificial cognitive systems. An optimal approach may be an intuitive one: minimizing the average error in perceptual representation of stimuli. However, as discussed, this approach can come with costs in terms of illusions, or artifacts, despite a reduction in the average error. Of course, some systems will not rely wholly on mimicking human integration of modalities, and indeed will supersede human abilities: for example, a system may be designed to perform multiple tasks simultaneously, something a human cannot do. However, future research should aim to identify when an optimal approach is not suitable in multi-modal integration.

Using illusion research in human perception as a guide, researchers could identify and model when artifacts occur in multi-modal integration, and apply these findings to system design. This might take the shape of modulating the optimality of integration depending on conditions via increasing or decreasing weightings as deemed appropriate. This approach could contribute to a database of "prior knowledge" where specific conditions that can result in artifacts are cataloged and can inform the degree of integration between sources in order to avoid undesired outcomes. For instance, Roach et al. (2006) examines audio-visual integration from just such a perspective using a Bayesian model of integration, where prior knowledge of events are taken into account and a balance between benefits and costs (optimal integration and potential erroneous perception) of integration is reached. They examined interactions between auditory and visual rate perception (where a judgement is made in a single modality and the other modality is "ignored") and found that there is a gradual transition between partial cue integration and complete cue segregation as inter-modal discrepancy increases. The Bayesian model they implemented took into account prior knowledge of the correspondence between audio and visual rate signals, when arriving at an appropriate degree of integration.

Similarly, a comparison between unimodal information and the final multi-modal integration might offer a strategy for identifying artifacts. This strategy might be akin to the study of Sekiyama (1994), which demonstrated that Japanese participants, in contrast to their American counterparts, have a different audio-visual strategy in the McGurk paradigm: less weight was given to discrepant visual information, which in turn affected the integration with auditory stimuli, ultimately resulting in a smaller McGurk effect. The inverse was shown in the participants with cochlear implants who demonstrated a larger McGurk effect: more weight was given to visual stimuli in general (Rouger et al., 2008). Magnotti and Beauchamp (2017) suggested that a causal inference (determining if audio and visual stimuli have the same source) "type" calculation is a step in multisensory speech perception, where some, but not all, incongruent audio-visual speech stimuli are integrated based on the likelihood of a shared, or separate, sources. Should that be the case, and this step is part of a near optimal strategy, a suboptimal process—such as a comparison of unimodal information and final multi-modal integration, or adjusting relative stimulus feature weightings when estimating likelihoods of source—could ensure that a McGurk-like effect is avoided. Additionally, it is worth noting the research by Driver (1996) who demonstrated that when there are competing auditory speech stimuli ostensibly from the same source and a matching visual speech stimuli from a different spatial location this has the effect of "pulling" the matching auditory stimuli in perceptual space toward the visual stimuli improving separation of the auditory streams reflected in report accuracy.

Dynamic adjustment of prior expectations is a vital consideration when designing local "prior knowledge" databases for artificial cognitive systems. This is illuminated by the fact that dynamically updated prior expectations can increase the likelihood of audio-visual integration: When congruent audio-visual stimuli is interspersed with incongruent McGurk audio-visual stimuli, the illusory McGurk effect emerges (Gau and Noppeney, 2016). Essentially, when there is a high instance of audio-visual integration due to congruent stimuli, incongruent stimuli have a greater chance of being deemed as originating from the same source and therefore being integrated. These behavioral results were supported by fMRI recordings that showed the left inferior frontal sulcus arbitrates between multisensory integration and segregation by combining top-down prior congruent/incongruent expectations with bottom-up congruent/incongruent cues (Gau and Noppeney, 2016). This suggests that in artificial cognitive systems, even though the prior knowledge databases should cater for updates, it should not be done live and "on-the-fly." If the probability of audio and visual stimuli originating from the same source was calculated near-optimally in the manner described by Gau and Noppeney (2016) dynamically, it could result in artifacts in an artificial cognitive system, where, for example, unrelated audio-visual events could be classified as being characteristics of the same event. If a dynamic approach is required, an optimal strategy should be avoided for these reasons.

In addition to the approaches suggested above, it is important that temporal characteristics, such as processing differences across artificial modalities are also taken into account. For instance, even though light is many hundreds of thousands of times faster than sound, the human perceptual system processes sound stimuli faster than visual stimuli (Recanzone, 2009). Indeed, it has been suggested that characteristics such as processing speeds of auditory and visual stimuli changing as a person ages (for example, visual processing slowing) may be responsible for increasing audio-visual integration in older participants where auditory tones had a greater influence on the perceived number of flashes in the sound-induced flash illusion compared to younger participants (DeLoss et al., 2013; McGovern et al., 2014). Similar considerations should be made for artificial systems. Regardless of how sensitive or fast at processing a given artificial sensor is, light will always reach a sensor before sound if the respective stimuli originate from the same distance/location. Setting aside the physical attribute of the speed of light vs. the speed of sound, there is an additional level of complexity even in an artificial system where it presumably would require a lot more computational power, and thus time, to process and separate the stimulus of interest in a given visual scene (with other factors such as feature resolution playing a role). Indeed, as mentioned previously, temporal feedback delay in the nervous system is a factor in optimal multi-sensory integration (Crevecoeur et al., 2016). Additionally, a unimodal auditory strategy for separating sources of auditory stimuli in a noisy environment via extracting and segregating temporally coherent features into separate streams has been developed by Krishnan et al. (2014). These considerations taken with the multi-modal audio-visual strategies deployed in speech (where temporal relationships of mouth movement and auditory onset play a role, specifically the voice onsets between 100 and 300ms before the mouth visibly moves Chandrasekaran et al., 2009) highlight the importance of temporal characteristics, correlations, and strategies when designing artificial cognitive systems. Finally, to handle noisy sensory information, artificial cognitive systems should perhaps consider incorporating temporal integration of sensory evidence (Rañó et al., 2017; Yang et al., 2017; Mi et al., 2019) instead of employing snapshot decision processing.

In summary, we highlight a wide range of audio-visual illusory percepts from the psychological and neuroscience literature, and discussed how computational cognitive models can account for some of these illusions—through seemingly optimal multimodal integration. We provide cautions regarding the naïve adoption of these human multimodal integration computations for artificial cognitive systems, which may lead to unwanted artifacts. Further investigations of the mechanisms of multimodal integration in humans and machines can lead to efficient approaches for mitigating and avoiding unwanted artifacts in artificial cognitive systems.

AUTHOR CONTRIBUTIONS

WB wrote sections 1, 2, and 5. AL, AZ, and KW-L wrote section 3. IR wrote section 4. All authors contributed to structuring, revising, and proofreading of the manuscript.

FUNDING

This research was supported by the Northern Ireland Functional Brain Mapping Project (1303/101154803), funded by Invest NI and the University of Ulster (KW-L), The Royal Society (KW-L and IR), and ASUR (1014-C4-Ph1-071) (KW-L, IR, and AL).

ACKNOWLEDGMENTS

The authors are grateful to Shufan Yang for helpful discussions on the computational modeling.

REFERENCES

Adams, W. J. (2016). The development of audio-visual integration for temporal judgements. *PLoS Comput. Biol.* 12:e1004865. doi: 10.1371/journal.pcbi.1004865

Alais, D., and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Curr. Biol.* 14, 257–262. doi: 10.1016/j.cub.2004.01.029

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:160606565*.

Andersen, T. S., Tiippana, K., and Sams, M. (2004). Factors influencing audiovisual fission and fusion illusions. *Cogn. Brain Res.* 21, 301–308. doi: 10.1016/j.cogbrainres.2004.06.004

Angelaki, D. E., Gu, Y., and DeAngelis, G. C. (2009). Multisensory integration: psychophysics, neurophysiology, and computation. *Curr. Opin. Neurobiol.* 19, 452–458. doi: 10.1016/j.conb.2009.06.008

Battaglia, P. W., Jacobs, R. A., and Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *J. Opt. Soc. Am. A* 20, 1391–1397. doi: 10.1364/JOSAA.20.001391

Beal, M. J., Jojic, N., and Attias, H. (2003). A graphical model for audiovisual object tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* 25, 828–836. doi: 10.1109/TPAMI.2003.1206512

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., and Pouget, A. (2012). Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron.* 74, 30–39. doi: 10.1016/j.neuron.2012.03.016

Bitzer, S., Park, H., Blankenburg, F., and Kiebel, S. J. (2014). Perceptual decision making: drift-diffusion model is equivalent to a Bayesian model. *Front. Hum. Neurosci.* 8:102. doi: 10.3389/fnhum.2014.00102

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., and Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.* 113:700. doi: 10.1037/0033-295X.113.4.700

Botvinick, M., and Cohen, J. (1998). Rubber hands 'feel? touch that eyes see. *Nature.* 391:756. doi: 10.1038/35784

Boyce, W. P. (2016). *Perception of space and time across modalities* (Doctoral dissertation). Swansea University, Swansea, Wales.

Boyce, W. P., Whiteford, S., Curran, W., Freegard, G., and Weidemann, C. T. (2020). Splitting time: sound-induced illusory visual temporal fission and fusion. *J. Exp. Psychol. Hum. Percept. Perform.* 46, 172–201. doi: 10.1037/xhp0000703

Bruns, P., Maiworm, M., and Röder, B. (2014). Reward expectation influences audiovisual spatial integration. *Attent. Percept. Psychophys.* 76, 1815–1827. doi: 10.3758/s13414-014-0699-y

Bülthoff, H. H., and Yuille, A. L. (1996). "A Bayesian framework for the integration of visual modules," in *Attention and Performance Vol. XVI: Information Integration in Perception and Communication*, eds J. L. McClelland and T. Inui (Cambridge, MA: MIT Press), 49–70.

Cairney, P. T. (1975). The complication experiment uncomplicated. *Perception* 4, 255–265. doi: 10.1068/p040255

Calvert, G. A., Hansen, P. C., Iversen, S. D., and Brammer, M. J. (2001). Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. *Neuroimage*. 14, 427–438. doi: 10.1006/nimg.2001.0812

Chandrasekaran, C. (2017). Computational principles and models of multisensory integration. *Curr. Opin. Neurobiol*. 43, 25–34. doi: 10.1016/j.conb.2016.11.002

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., and Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Comput. Biol*. 5:e1000436. doi: 10.1371/journal.pcbi.1000436

Chen, T. (2001). Audiovisual speech processing. lip reading and lip synchronization. *IEEE Signal Process. Mag*. 18, 9–21. doi: 10.1109/79.911195

Chen, Y. C., and Spence, C. (2017). Assessing the role of the "unity assumption" on multisensory integration: a review. *Front. Psychol*. 8:445. doi: 10.3389/fpsyg.2017.00445

Chibelushi, C. C., Deravi, F., and Mason, J. S. D. (2002). A review of speech-based bimodal recognition. *IEEE Trans. Multimedia* 4, 23–37. doi: 10.1109/6046.985551

Cohen, M. H. (2001). "Multimodal integration - a biological view," in: *Proceedings of the* 15^th *International Conference on Artificial Intelligence* (La Palma), 1417–1424.

Colonius, H., and Diederich, A. (2004). Multisensory interaction in saccadic reaction time: a time-window-of-integration model. *J. Cogn. Neurosci*. 16, 1000–1009. doi: 10.1162/0898929041502733

Cook, L. A., and Van Valkenburg, D. L. (2009). Audio-visual organisation and the temporal ventriloquism effect between grouped sequences: evidence that unimodal grouping precedes cross-modal integration. *Perception* 38, 1220–1233. doi: 10.1068/p6344

Crevecoeur, F., Munoz, D. P., and Scott, S. H. (2016). Dynamic multisensory integration: somatosensory speed trumps visual accuracy during feedback control. *J. Neurosci*. 36, 8598–8611. doi: 10.1523/JNEUROSCI.0184-16.2016

de Dieuleveult, A. L., Siemonsma, P. C., van Erp, J. B., and Brouwer, A. M. (2017). Effects of aging in multisensory integration: a systematic review. *Front. Aging Neurosci*. 9:80. doi: 10.3389/fnagi.2017.00080

DeLoss, D. J., Pierce, R. S., and Andersen, G. J. (2013). Multisensory integration, aging, and the sound-induced flash illusion. *Psychol. Aging*. 28:802. doi: 10.1037/a0033289

Deng, L., Hinton, G., and Kingsbury, B. (2013). "New types of deep neural network learning for speech recognition and related applications: an overview," in *IEEE International Conference on Acoustics, Speech and Signal Processing* (Vancouver, CA), 8599–8603. doi: 10.1109/ICASSP.2013.6639344

Diederich, A. (1995). Intersensory facilitation of reaction time: evaluation of counter and diffusion coactivation models. *J. Math. Psychol*. 39, 197–215. doi: 10.1006/jmps.1995.1020

Downing, P. E., and Treisman, A. M. (1997). The line-motion illusion: attention or impletion? *J. Exp. Psychol. Hum. Percept. Perform*. 23, 768–779. doi: 10.1037/0096-1523.23.3.768

Driver, J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature*. 381, 66–68. doi: 10.1038/381066a0

Drugowitsch, J., DeAngelis, G. C., Klier, E. M., Angelaki, D. E., and Pouget, A. (2014). Optimal multisensory decision-making in a reaction-time task. *eLife* 3:e03005. doi: 10.7554/eLife.03005

Fard, P. R., Park, H., Warkentin, A., Kiebel, S. J., and Bitzer, S. (2017). A Bayesian reformulation of the extended drift-diffusion model in perceptual decision making. *Front. Comput. Neurosci*. 11:29. doi: 10.3389/fncom.2017.00029

Fetsch, C. R., Turner, A. H., DeAngelis, G. C., and Angelaki, D. E. (2009). Dynamic reweighting of visual and vestibular cues during self-motion perception. *J. Neurosci*. 29, 15601–15612. doi: 10.1523/JNEUROSCI.2574-09.2009

Fisher, J. W., and Darrell, T. (2004). Speaker association with signal-level audiovisual fusion. *IEEE Trans. Multimedia* 6, 406–413. doi: 10.1109/TMM.2004.827503

Fisher, J. W., Darrell, T., Freeman, W. T., and Viola, P. A. (2000). "Learning joint statistical models for audio-visual fusion and segregation," in *Neural Information Processing Systems* (Denver, CO), 772–778.

Fitzpatrick, R., and McCloskey, D. (1994). Proprioceptive, visual and vestibular thresholds for the perception of sway during standing in humans. *J. Physiol*. 478(Pt 1):173. doi: 10.1113/jphysiol.1994.sp020240

Folyi, T., Fehér, B., and Horváth, J. (2012). Stimulus-focused attention speeds up auditory processing. *Int. J. Psychophysiol*. 84, 155–163. doi: 10.1016/j.ijpsycho.2012.02.001

Fracasso, A., Targher, S., Zampini, M., and Melcher, D. (2013). Fooling the eyes: the influence of a sound-induced visual motion illusion on eye movements. *PLOS ONE* 8:e62131. doi: 10.1371/journal.pone.0062131

Freeman, E., and Driver, J. (2008). Direction of visual apparent motion driven solely by timing of a static sound. *Curr. Biol*. 18, 1262–1266. doi: 10.1016/j.cub.2008.07.066

Fulbright, R. K., Troche, C. J., Skudlarski, P., Gore, J. C., and Wexler, B. E. (2001). Functional MR imaging of regional brain activation associated with the affective experience of pain. *Am. J. Roentgenol*. 177, 1205–1210. doi: 10.2214/ajr.177.5.1771205

Fuller, S., and Carrasco, M. (2009). Perceptual consequences of visual performance fields: the case of the line motion illusion. *J. Vis*. 9:13. doi: 10.1167/9.4.13

Gatica-Perez, D., Lathoud, G., Odobez, J. M., and McCowan, I. (2007). Audiovisual probabilistic tracking of multiple speakers in meetings. *IEEE Trans. Audio Speech Lang. Process*. 15, 601–616. doi: 10.1109/TASL.2006.881678

Gau, R., and Noppeney, U. (2016). How prior expectations shape multisensory perception. *Neuroimage*. 124, 876–886. doi: 10.1016/j.neuroimage.2015.09.045

Getzmann, S. (2007). The effect of brief auditory stimuli on visual apparent motion. *Perception* 36:1089. doi: 10.1068/p5741

Ghazanfar, A. A., and Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends Cogn. Sci*. 10, 278–285. doi: 10.1016/j.tics.2006.04.008

Goldstein, E. B. (2008). *Sensation and Perception, 8th Edn*. Belmont, CA: Thomson Wadsworth.

Gómez-Eguíluz, A., Rañó, I., Coleman, S. A., and McGinnity, T. M. (2016). "A multi-modal approach to continuous material identification through tactile sensing," in *IEEE/RSJ International Conference on Intelligent Robots and Systems* (Daejeon).

Gondan, M., and Minakata, K. (2016). A tutorial on testing the race model inequality. *Attent. Percept. Psychophys*. 78, 723–735. doi: 10.3758/s13414-015-1018-y

Green, B. G. (2004). Temperature perception and nociception. *J. Neurobiol*. 61, 13–29. doi: 10.1002/neu.20081

Guitarte-Pérez, J. F., Frangi, A. F., Lleida-Solano, E., and Lukas, K. (2005). "Lip reading for robust speech recognition on embedded devices," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (Philadelphia, PA), 473–476.

Hairston, W. D., Hodges, D. A., Burdette, J. H., and Wallace, M. T. (2006). Auditory enhancement of visual temporal order judgment. *Neuroreport* 17, 791–795. doi: 10.1097/01.wnr.0000220141.29413.b4

Hidaka, S., and Ide, M. (2015). Sound can suppress visual perception. *Sci. Rep*. 5:10483. doi: 10.1038/srep10483

Hidaka, S., Manaka, Y., Teramoto, W., Sugita, Y., Miyauchi, R., Gyoba, J., et al. (2009). The alternation of sound location induces visual motion perception of a static object. *PLoS ONE* 4:e8188. doi: 10.1371/journal.pone.0008188

Hidaka, S., Teramoto, W., Sugita, Y., Manaka, Y., Sakamoto, S., and Suzuki, Y. (2011). Auditory motion information drives visual motion perception. *PLoS ONE* 6:e17499. doi: 10.1371/journal.pone.0017499

Hikosaka, O., Miyauchi, S., and Shimojo, S. (1993a). Voluntary and stimulus-induced attention detected as motion sensation. *Perceptio* 22, 517–526. doi: 10.1068/p220517

Hikosaka, O., Miyauchi, S., and Shimojo, S. (1993b). Focal visual attention produces illusory temporal order and motion sensation. *Vis. Res*. 33, 1219–1240. doi: 10.1016/0042-6989(93)90210-N

Holmes, N. P. (2009). The principle of inverse effectiveness in multisensory integration: some statistical considerations. *Brain Topogr*. 21, 168–176. doi: 10.1007/s10548-009-0097-2

Jain, A., Bansal, R., Kumar, A., and Singh, K. (2015). A comparative study of visual and auditory reaction times on the basis of gender and physical activity levels of medical first year students. *Int. J. Appl. Basic Med. Res*. 5:124. doi: 10.4103/2229-516X.157168

Kafaligonul, H., and Stoner, G. R. (2010). Auditory modulation of visual apparent motion with short spatial and temporal intervals. *J. Vis*. 10:31. doi: 10.1167/10.12.31

Kafaligonul, H., and Stoner, G. R. (2012). Static sound timing alters sensitivity to low-level visual motion. *J. Vis*. 12:2. doi: 10.1167/12.11.2

Keetels, M., Stekelenburg, J., and Vroomen, J. (2007). Auditory grouping occurs prior to intersensory pairing: evidence from temporal ventriloquism. *Exp. Brain Res*. 180, 449–456. doi: 10.1007/s00221-007-0881-8

Klimova, M., Nishida, S., and Roseboom, W. (2017). Grouping by feature of cross-modal flankers in temporal ventriloquism. *Sci. Rep*. 7:7615. doi: 10.1038/s41598-017-06550-z

Kolers, P. A., and von Grünau, M. (1976). Shape and color in apparent motion. *Vis. Res*. 16, 329–335. doi: 10.1016/0042-6989(76)90192-9

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., and Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE* 2:e943. doi: 10.1371/journal.pone.0000943

Krishnan, L., Elhilali, M., and Shamma, S. (2014). Segregating complex sound sources through temporal coherence. *PLoS Comput. Biol*. 10:e1003985. doi: 10.1371/journal.pcbi.1003985

Li, J., and Allinson, N. A. (2008). A comprehensive review of current local features for computer vision. *Neurocomputing* 71, 1771–1787. doi: 10.1016/j.neucom.2007.11.032

Li, Y., Narayanan, S., and Kuo, C. C. J. (2004). Content-based movie analysis and indexing based on audiovisual cues. *IEEE Trans. Circ. Syst. Video Technol*. 14, 1073–1085. doi: 10.1109/TCSVT.2004.831968

Link, S. W. (1975). The relative judgment theory of two choice response time. *J. Math. Psychol*. 12, 114–135. doi: 10.1016/0022-2496(75)90053-X

Liu, Y. S., Yu, A., and Holmes, P. (2009). Dynamical analysis of Bayesian inference models for the Eriksen task. *Neural Comput*. 21, 1520–1553. doi: 10.1162/neco.2009.03-07-495

Luettin, J., and Thacker, N. A. (1997). Speechreading using probabilistic models. *Comput. Vis. Image Understand*. 65, 163–178. doi: 10.1006/cviu.1996.0570

Luo, R. C., and Kay, M. G. (1989). Multisensor integration and fusion in intelligent systems. *IEEE Trans. Syst. Man Cybernet*. 19, 901–931. doi: 10.1109/21.44007

Luo, R. C., and Kay, M. G. (2002). Multisensor fusion and integration: approaches, applications, and future research directions. *IEEE Sens. J*. 2, 107–119. doi: 10.1109/JSEN.2002.1000251

Lush, P. (2020). Demand characteristics confound the rubber hand illusion. *Collabra Psychol*. 6:22. doi: 10.1525/collabra.325

Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nat. Neurosci*. 9:1432. doi: 10.1038/nn1790

Ma, W. J., and Pouget, A. (2008). Linking neurons to behavior in multisensory perception: a computational review. *Brain Res*. 1242, 4–12. doi: 10.1016/j.brainres.2008.04.082

Macaluso, E., Frith, C. D., and Driver, J. (2000). Modulation of human visual cortex by crossmodal spatial attention. *Science* 289, 1206–1208. doi: 10.1126/science.289.5482.1206

Maiworm, M., Bellantoni, M., Spence, C., and Röder, B. (2012). When emotional valence modulates audiovisual integration. *Attent. Percept. Psychophys*. 74, 1302–1311. doi: 10.3758/s13414-012-0310-3

MaMagnotti, J. F., and Beauchamp, M. S. (2017). A causal inference model explains perception of the McGurk effect and other incongruent audiovisual speech. *PLoS Comput. Biol*. 13:e1005229. doi: 10.1371/journal.pcbi.1005229

McGovern, D. P., Roudaia, E., Stapleton, J., McGinnity, T. M., and Newell, F. N. (2014). The sound-induced flash illusion reveals dissociable age-related effects in multisensory integration. *Front. Aging Neurosci*. 6:250. doi: 10.3389/fnagi.2014.00250

McGrath, M., and Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *J. Acous. Soc. Am*. 77, 678–685. doi: 10.1121/1.392336

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Meijer, G. T., Mertens, P. E., Pennartz, C. M., Olcese, U., and Lansink, C. S. (2019). The circuit architecture of cortical multisensory processing: distinct functions jointly operating within a common anatomical network. *Prog. Neurobiol*. 174, 1–15. doi: 10.1016/j.pneurobio.2019.01.004

Mi, Y., Lin, X., Zou, X., Ji, Z., Huang, T., and Wu, S. (2019). Spatiotemporal information processing with a reservoir decision-making network. *arXiv preprint arXiv:190712071*.

Miller, J. (2016). Statistical facilitation and the redundant signals effect: what are race and coactivation models? *Attent. Percept. Psychophys*. 78, 516–519. doi: 10.3758/s13414-015-1017-z

Morein-Zamir, S., Soto-Faraco, S., and Kingstone, A. (2003). Auditory capture of vision: examining temporal ventriloquism. *Cogn. Brain Res*. 17, 154–163. doi: 10.1016/S0926-6410(03)00089-2

Nakadai, K., Hidai, K., and Okuno, H. G. (2002). "Real-time speaker localization and speech separation by audio-visual integration," in *Proceedings of the IEEE International Conference on Robotics and Automation*, *Vol. 1* (Washington, DC).

Niyogi, R. K., and Wong-Lin, K. (2013). Dynamic excitatory and inhibitory gain modulation can produce flexible, robust and optimal decision-making. *PLoS Comput. Biol*. 9:e1003099. doi: 10.1371/journal.pcbi.1003099

Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., and Ogata, T. (2015). Audio-visual speech recognition using deep learning. *Appl. Intell*. 42, 722–737. doi: 10.1007/s10489-014-0629-7

Parisi, G. I., Barros, P., Fu, D., Magg, S., Wu, H., Liu, X., et al. (2018). "A neurorobotic experiment for crossmodal conflict resolution in complex environments," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Madrid: IEEE), 2330–2335. doi: 10.1109/IROS.2018.8594036

Parisi, G. I., Barros, P., Kerzel, M., Wu, H., Yang, G., Li, Z., et al. (2017). "A computational model of crossmodal processing for conflict resolution," in *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (Lisbon: IEEE), 33–38. doi: 10.1109/DEVLRN.2017.8329784

Peters, M. A. K., Ma, W. J., and Shams, L. (2016). The size-weight illusion is not anti-Bayesian after all: a unifying Bayesian account. *PeerJ* 4:e2124. doi: 10.7717/peerj.2124

Peters, M. A. K., Zhang, L. Q., and Shams, L. (2018). The material-weight illusion is a Bayes-optimal percept under competing density priors. *PeerJ* 6:e5760. doi: 10.7717/peerj.5760

Poirier, C., Collignon, O., DeVolder, A. G., Renier, L., Vanlierde, A., Tranduy, D., et al. Specific activation of the V5 brain area by auditory motion processing: an fMRI study. *Cogn. Brain Res*. (2005) 25, 650–658. doi: 10.1016/j.cogbrainres.2005.08.015

Pouget, A., Beck, J. M., Ma, W. J., and Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nat. Neurosci*. 16:1170. doi: 10.1038/nn.3495

Raaja, D. (1962). Statistical facilitation of simple reaction time. *Trans. N. Y. Acad. Sci*. 24, 574–590. doi: 10.1111/j.2164-0947.1962.tb01433.x

Radeau, M., and Bertelson, P. (1987). Auditory-visual interaction and the timing of inputs. *Psychol. Res*. 49, 17–22. doi: 10.1007/BF00309198

Ramos-Estebanez, C., Merabet, L. B., Machii, K., Fregni, F., Thut, G., Wagner, T. A., et al. (2007). Visual phosphene perception modulated by subthreshold crossmodal sensory stimulation. *J. Neurosci*. 27, 4178–4181. doi: 10.1523/JNEUROSCI.5468-06.2007

Rañó, I., Khamassi, M., and Wong-Lin, K. (2017). "A drift diffusion model of biological source seeking for mobile robots," in *2017 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE), 3525–3531.

Rao, S. M., Mayer, A. R., and Harrington, D. L. (2001). The evolution of brain activation during temporal processing. *Nat. Neurosci*. 4, 317–323. doi: 10.1038/85191

Ratcliff, R. (1978). A theory of memory retrieval. *Psychol. Rev*. 85:59. doi: 10.1037/0033-295X.85.2.59

Ratcliff, R., and Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychol. Sci*. 9, 347–356. doi: 10.1111/1467-9280.00067

Recanzone, G. H. (2009). Interactions of auditory and visual stimuli in space and time. *Hear. Res*. 258, 89–99. doi: 10.1016/j.heares.2009.04.009

Roach, N. W., Heron, J., and McGraw, P. V. (2006). Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration. *Proc. R. Soc. Lond. B Biol. Sci*. 273, 2159–2168. doi: 10.1098/rspb.2006.3578

Roseboom, W., Kawabe, T., and Nishida, S. Y. (2013a). Direction of visual apparent motion driven by perceptual organization of cross-modal signals. *J. Vis*. 13, 1–13. doi: 10.1167/13.1.6

Roseboom, W., Kawabe, T., Nishida, S. Y. (2013b). The cross-modal double flash illusion depends on featural similarity between cross-modal inducers. *Sci. Rep*. 3:3437. doi: 10.1038/srep03437

Rosenthal, O., Shimojo, S., and Shams, L. (2009). Sound-induced flash illusion is resistant to feedback training. *Brain Topogr*. 21, 185–192. doi: 10.1007/s10548-009-0090-9

Rouger, J., Fraysse, B., Deguine, O., and Barone, P. (2008). McGurk effects in cochlear-implanted deaf subjects. *Brain Res*. 1188, 87–99. doi: 10.1016/j.brainres.2007.10.049

Roxin, A., and Ledberg, A. (2008). Neurobiological models of two-choice decision making can be reduced to a one-dimensional nonlinear diffusion equation. *PLoS Comput. Biol*. 4:e1000046. doi: 10.1371/journal.pcbi.1000046

Samad, M., Chung, A. J., and Shams, L. (2015). Perception of body ownership is driven by Bayesian sensory inference. *PLoS ONE*. 10:e0117178. doi: 10.1371/journal.pone.0117178

Sato, Y., Toyoizumi, T., and Aihara, K. (2007). Bayesian inference explains perception of unity and ventriloquism aftereffect: identification of common sources of audiovisual stimuli. *Neural Comput*. 19, 3335–3355. doi: 10.1162/neco.2007.19.12.3335

Scheier, C., Nijhawan, R., and Shimojo, S. (1999). Sound alters visual temporal resolution. *Invest. Ophthalmol. Vis. Sci.* 40:792.

Schneider, K. A., and Bavelier, D. (2003). Components of visual prior entry. *Cogn. Psychol*. 47, 333–366. doi: 10.1016/S0010-0285(03)00035-5

Schwarz, W. (1994). Diffusion, superposition, and the redundant-targets effect. *J. Math. Psychol*. 38, 504–520. doi: 10.1006/jmps.1994.1036

Seibold, V. C., and Rolke, B. (2014). Does temporal preparation speed up visual processing? Evidence from the N2pc. *Psychophysiology* 51, 529–538. doi: 10.1111/psyp.12196

Sekiyama, K. (1994). Differences in auditory-visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility. *J. Acous. Soc. Japan*. 15, 143–158. doi: 10.1250/ast.15.143

Sekuler, R., Sekuler, A. B., and Lau, R. (1997). Sound alters visual motion perception. *Nature* 385:308. doi: 10.1038/385308a0

Shams, L., Iwaki, S., Chawla, A., and Bhattacharya, J. (2005a). Early modulation of visual cortex by sound: an MEG study. *Neurosci. Lett*. 378, 76–81. doi: 10.1016/j.neulet.2004.12.035

Shams, L., Kamitani, Y., and Shimojo, S. (2002). Visual illusion induced by sound. *Cogn. Brain Res*. 14, 147–152. doi: 10.1016/S0926-6410(02)00069-1

Shams, L., and Kim, R. (2010). Crossmodal influences on visual perception. *Phys. Life Rev*. 7, 269–284. doi: 10.1016/j.plrev.2010.04.006

Shams, L., Ma, W. J., and Beierholm, U. (2005b). Sound-induced flash illusion as an optimal percept. *Neuroreport* 16, 1923–1927. doi: 10.1097/01.wnr.0000187634.68504.bb

Shimojo, S., Miyauchi, S., and Hikosaka, O. (1997). Visual motion sensation yielded by non-visually driven attention. *Vis. Res*. 37, 1575–1580. doi: 10.1016/S0042-6989(96)00313-6

Shipley, T. (1964). Auditory flutter-driving of visual flicker. *Science*. 145, 1328–1330. doi: 10.1126/science.145.3638.1328

Spence, C., Shore, D. I., and Klein, R. M. (2001). Multisensory prior entry. *J. Exp. Psychol*. 130, 799–831. doi: 10.1037/0096-3445.130.4.799

Stein, B. E., Stanford, T. R., and Rowland, B. A. (2014). Development of multisensory integration from the perspective of the individual neuron. *Nat. Rev. Neurosci*. 15, 520–535. doi: 10.1038/nrn3742

Stevenson, R. A., and James, T. W. (2009). Audiovisual integration in human superior temporal sulcus: inverse effectiveness and the neural processing of speech and object recognition. *Neuroimage* 44, 1210–1223. doi: 10.1016/j.neuroimage.2008.09.034

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*. 25, 251–260. doi: 10.1007/BF02289729

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acous. Soc. Am*. 26, 212–215. doi: 10.1121/1.1907309

Talsma, D., Senkowski, D., Soto-Faraco, S., and Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends Cogn. Sci*. 14, 400–410. doi: 10.1016/j.tics.2010.06.008

Teramotoa, W., Hidaka, S., Sugita, Y., Sakamoto, S., Gyoba, J., Iwaya, Y., et al. Sounds can alter the perceived direction of a moving visual object. *J. Vis*. (2012) 12, 1–12. doi: 10.1167/12.3.11

Thomas, G. J. (1941). Experimental study of the influence of vision on sound localization. *J. Exp. Psychol*. 28:163. doi: 10.1037/h0055183

Ursino, M., Cuppini, C., and Magosso, E. (2014). Neurocomputational approaches to modelling multisensory integration in the brain: a review. *Neural Netw*. 60, 141–165. doi: 10.1016/j.neunet.2014.08.003

Ursino, M., Cuppini, C., Magosso, E., Beierholm, U., and Shams, L. (2019). Explaining the effect of likelihood manipulation and prior through a neural network of the audiovisual perception of space. *Multisens. Res*. 32, 111–144. doi: 10.1163/22134808-20191324

van Erp JBF, Philippi, T. G., and Werkhoven, P. (2013). Observers can reliably identify illusory flashes in the illusory flash paradigm. *Exp. Brain Res*. 226, 73–79. doi: 10.1007/s00221-013-3413-8

Vatakis, A., and Spence, C. (2007). Crossmodal binding: evaluating the "unity assumption" using audiovisual speech stimuli. *Percept. Psychophys*. 69, 744–756. doi: 10.3758/BF03193776

Vatakis, A., and Spence, C. (2008). Evaluating the influence of the "unity assumption" on the temporal perception of realistic audiovisual stimuli. *Acta Psychol*. 127, 12–23. doi: 10.1016/j.actpsy.2006.12.002

Vibell, J., Klinge, C., Zampini, M., Spence, C., and Nobre, A. C. (2007). Temporal order is coded temporally in the brain: early event-related potential latency shifts underlying prior entry in a cross-modal temporal order judgment task. *J. Cogn. Neurosci*. 19, 109–120. doi: 10.1162/jocn.2007.19.1.109

Wang, X. J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron* 36, 955–968. doi: 10.1016/S0896-6273(02)01092-9

Wang, Y., Liu, Z., and Huang, J. C. (2000). Multimedia content analysis-using both audio and visual clues. *IEEE Signal Process. Mag*. 17, 12–36. doi: 10.1109/79.888862

Watanabe, K., and Shimojo, S. (2001). When sound affects vision: effects of auditory grouping on visual motion perception. *Psychol. Sci*. 12, 109–116. doi: 10.1111/1467-9280.00319

Watkins, S., Shams, L., Tanaka, S., Haynes, J. D., and Rees, G. (2006). Sound alters activity in human V1 in association with illusory visual perception. *Neuroimage*. 31, 1247–1256. doi: 10.1016/j.neuroimage.2006.01.016

Welch, R. B. (1999). Meaning, attention, and the unity assumption in the intersensory bias of spatial and temporal perceptions. *Adv. Psychol*. 129, 371–387. doi: 10.1016/S0166-4115(99)80036-3

Willey, C. F., Inglis, E., and Pearce, C. (1937). Reversal of auditory localization. *J. Exp. Psychol*. 20:114. doi: 10.1037/h0056793

Wong, K. F., and Wang, X. J. (2006). A recurrent network mechanism of time integration in perceptual decisions. *J. Neurosci*. 26, 1314–1328. doi: 10.1523/JNEUROSCI.3733-05.2006

Wozny, D. R., Beierholm, U. R., and Shams, L. (2008). Human trimodal perception follows optimal statistical inference. *J. Vis*. 8:24. doi: 10.1167/8.3.24

Wu, L., Oviatt, S. L., and Cohen, P. R. (1999). Multimodal integration - a statistical view. *Trans. Multimedia* 1, 334–341. doi: 10.1109/6046.807953

Yang, S., Wong-Lin, K., Rano, I., and Lindsay, A. (2017). "A single chip system for sensor data fusion based on a Drift-diffusion model," in *2017 Intelligent Systems Conference (IntelliSys)* (IEEE), 198–201. doi: 10.1109/IntelliSys.2017.8324291

Zampini, M., Shore, D. I., and Spence, C. (2005). Audiovisual prior entry. *Neurosci. Lett*. 381, 217–222. doi: 10.1016/j.neulet.2005.01.085

Zhang, W. H., Chen, A., Rasch, M. J., and Wu, S. (2016). Decentralized multisensory information integration in neural systems. *J. Neurosci*. 36, 532–547. doi: 10.1523/JNEUROSCI.0578-15.2016

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.