Psychonomic Bulletin & Review (2026) 33:113
https://doi.org/10.3758/s13423-026-02877-9

THEORETICAL/REVIEW

What role does temporal synchrony play in mid-level audiovisual crossmodal correspondences?

Charles Spence[1] · Nicola Di Stefano[2]

Received: 8 July 2025 / Accepted: 5 February 2026

Abstract

Temporal synchrony is widely recognized as one of the key factors facilitating the emergence of crossmodal correspondences and affecting their crossmodal effects. However, several issues regarding the definition of temporal synchrony and the mechanisms underlying its crossmodal effects remain open, depending on the specific experimental/perceptual context/stimuli used, as well as the influence of crossmodal congruency and structural (including isomorphic) crossmodal correspondences. In this review, we take a closer look at the literature that has been published in this area over recent decades in order to critically evaluate what is currently known concerning the crossmodal effects that are mediated by temporal synchrony. We focus especially on mid-level audiovisual crossmodal correspondences, defined as those that involve multi-element, or dynamic, auditory and visual stimuli. We examine the different experimental methodologies used and their limitations as well as the theoretical frameworks that have been proposed to account for the viewer's impression of (and the meaning/affect that is associated with) such experimental audiovisual displays, including those that are based on the 'Congruency-Associationist Model', Gestalt perceptual grouping, as well as the phenomenon of multisensory emergence. Finally, we outline several directions for future research on temporal synchrony in the context of audiovisual crossmodal correspondences.

Keywords Synchrony · Gestalt perceptual grouping · Dynamic crossmodal interactions · Crossmodal correspondences

Introduction

In recent years, there has been an explosive growth of interest in the topic of crossmodal correspondences (e.g., for reviews, see Motoki, Velasco, & Marks, 2023; Spence, 2011, 2018a). Crossmodal correspondences are the sometimes surprising crossmodal associations between features, attributes, or dimensions of experience, either directly perceived or else merely imagined, in different sensory modalities. For instance, people often intuitively choose to associate high-pitched sounds with bright, small, visual objects that are positioned high in space, while associating low-pitched sounds with darker, larger, visual objects that happen to be positioned lower in space (Evans & Treisman, 2010; Parise & Spence, 2012). Similarly, certain tastes, such as sweetness, are frequently matched with round shapes or soft textures, whereas bitterness tends to be linked with angular forms or low-pitched sounds and rough textures instead (for reviews, see Deroy, Crisinel, & Spence, 2013; Spence, 2023). These intuitive pairings are not only robust across individuals but have also been shown to influence perception, preference, and behaviour across a range of experimental and real-world contexts.

Recently, Spence and Di Stefano (2025a) introduced an important distinction between simple, mid-level, and complex crossmodal correspondences. According to this classificatory framework, simple crossmodal correspondences involve individual sensory attributes or dimensions, such as pitch, hue, or intensity (for reviews, see Marks, 2004; Spence, 2011); mid-level crossmodal correspondences involve combinations of (possibly dynamic) unisensory stimuli, such as short sequences of sounds (or video animations); and complex crossmodal correspondences involve semantically rich and/or emotionally meaningful complex

✉ Charles Spence
charles.spence@psy.ox.ac.uk

[1] Crossmodal Research Laboratory, Department of Experimental Psychology, Oxford University, Oxford OX1 3PS, UK

[2] Institute of Cognitive Sciences and Technologies, National Research Council, Rome, Italy

Published online: 17 March 2026

stimuli, such as paintings, music, film clips, and other works of art (for reviews, see Spence, 2020; Spence & Di Stefano, 2025b). While the majority of research that has been published to date in this area has studied crossmodal correspondences occurring between stimuli at the same level of complexity, such as pitch-hue correspondences (though see Spence & Di Stefano, 2024), or crossmodal associations between music and paintings/drawings (e.g., Di Stefano et al., 2025; Di Stefano et al., 2024; for review, see Spence, 2020), a few researchers have also studied cross-level crossmodal correspondences; for example, consider here only the crossmodal correspondences that have been documented between classical music selections and colour patches by Palmer, Schloss, Xu, and Prado-León (2013; see also Hauck, von Castell, & Hecht, 2022; McDonald et al., 2022).

Empirical findings and theoretical reflections highlight the key role of temporal synchrony in binding the unisensory contents that are actually associated crossmodally (for reviews, see Vatakis & Spence, 2010; Vroomen & Keetels, 2010). In its narrowest sense, the term refers to the precise co-occurrence of sensory events, such as an auditory click and a flash of light, which is believed to promote multisensory integration. This millisecond-level simultaneity underpins many controlled experiments on audiovisual integration and appears to benefit from transient stimulus onsets (e.g., Andersen & Mamassian, 2008; Cook, Van Valkenburg, & Badcock, 2011; Raij et al., 2010; Van der Burg et al., 2010). However, it should be noted that such a definition is most readily applicable to those contexts involving discrete or intermittent events, where temporal alignment can be clearly marked and easily measured, such as in laboratory experiments using bouncing balls or flashing lights.

In everyday contexts, however, the perception of synchrony can be more flexible (or ambiguous): that is, visual and auditory stimuli can appear synchronized when they share a rhythmic pattern or common temporal structure, even if their individual elements do not exactly align (Di Stefano & Spence, 2025). For instance, dancers may align with the beat of a soundtrack only intermittently, and yet still appear 'in sync' due to the presence of periodic or expressive correspondences (Heins et al., 2021; Keller & Repp, 2008; Wakiyama, Tsubaki, Kuno-Mizumura, & Sakaguchi, 2025). Complicating matters further, synchrony can be attributed to or recognized in expressive or structural patterns in the stimuli across modalities. For instance, auditory and visual events might match in terms of their dynamic changes in tension, trajectory, or salience over time (Munárriz Ortiz, 2017). These occurrences of higher-order synchrony that involve looser, but nevertheless still perceptually compelling, mappings raise questions regarding the very notion of temporal synchrony and the extent to which precision in time is required to generate the desired crossmodal effect. Indeed, separately, other researchers have demonstrated how it may be the correlation between unimodal temporal stimulus patterns, rather than the exact temporal synchrony of the elements in each stream, that is key to multisensory integration (see Parise, Harrar, Ernst, & Spence, 2013; Parise, Spence, & Ernst, 2012).

In this review, we focus on the role of temporal synchrony in facilitating crossmodal associations and in mediating their crossmodal effects. We start by reviewing the key literature that has been published on mid-level crossmodal correspondences, with a particular focus on the temporal dimension of stimuli matching (Section "Critical review of research involving mid-level crossmodal correspondences"). Thereafter, we examine crossmodal grouping at a perceptual level, with the aim of identifying a number of organizing principles that recur across the empirical literature and constrain how audiovisual integration can occur (Section "Crossmodal grouping and perceptual organization"). In the "Discussion" section, we discuss the key methodological issues of the reviewed literature as well as their implications for theoretical reflections before outlining directions for future research in the area of audiovisual correspondences and temporal synchrony. Overall, this review suggests that a more nuanced account of temporal relationships in audiovisual perception is required, one that is capable of accommodating both low-level synchrony and higher-order structural and expressive relations underlying reported crossmodal effects.

Critical review of research involving mid-level crossmodal correspondences

The majority of the studies targeting mid-level crossmodal correspondences that have been published to date involve studying the effect, if any, of adding different kinds of auditory and musical stimuli to visual stimuli. Several, at times overlapping, experimental literatures can be distinguished here: A number of researchers have, for instance, chosen to investigate the impact of (primarily classical) music on people's ratings of static visual stimuli (often consisting of pictures of paintings; for a review, see Spence, 2020). Meanwhile, a separate empirical literature has emerged documenting the impact of music on people's perception of dynamic visual film clips (for a review, see Spence & Di Stefano, 2025c). Meanwhile, a third body of research has investigated more synaesthesia-like connections between auditory and visual stimuli, often in the context of what might be called multimedia art (e.g., Daniels, Naumann, &

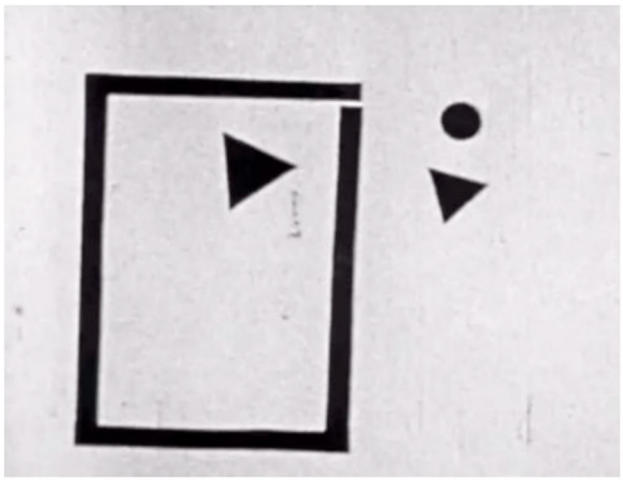

Fig. 1 Screenshot of the film figures (Heider & Simmel, 1944)

Thoben, 2010; Deutsch, 2012; Kargon, 2011; Toccafondi, 2025; Zika, 2013).[1]

In one early study, Marshall and Cohen (1988) presented their participants with a 2-min abstract animation showing a large and a small triangle and a small circle in motion (see Fig. 1). Typically, when people view this short animation, they tend to interpret the large triangle as a bully who is victimizing the two smaller geometric figures (the small triangle and small circle). Marshall and Cohen studied the effect of two contrasting pieces of background music played while participants viewed this film clip on the latter's attitudes toward the figures shown in the animation. Baseline information about the meaning of the music and visual stimuli was first obtained on 12 bipolar adjective rating scales (e.g., powerful/powerless). For the visual stimuli, the participants had to judge each of the three stimuli individually, and then provide an overall evaluation of the film as a whole. Next, a new group of participants was simultaneously presented with one combination of music plus film, and once again had to judge the film 'characters' and the film overall on the same 12 bipolar adjective ratings. In total, there were five groups of five participants in the preliminary study, with each group being presented with the same stimulus twice. The first time round the participants had to describe what happened in the film and supply adjectives to describe the three shapes or the music. The participants were also asked whether the music reminded them of anything. The second time through, the participants were instructed to assess each figure and/or the music on 13 rating scales.

[1] Synaesthetes are those rare individuals who experience a consistent vivid and automatic concurrent percept in the same or different sensory modality in response to a specific inducer (Deroy & Spence, 2013). For instance, colour-music is a form of synaesthesia in which people see idiosyncratically-matched colours (that are not physically present) in response to specific musical notes and/or timbres.

In Marshall and Cohen's (1988) main study, two originally composed contrasting but structurally similar musical accompaniments were created. They incorporated three main musical 'themes' used at specific points in the visual animation. One theme accompanied the introduction of the large triangle, the second theme coincided with the large triangle contacting the small triangle, and the third auditory theme was introduced when the small shapes are 'chased' around the rectangular enclosure by the large triangle. One version of the music was in major mode, the other in minor mode. The former was classified as 'weak', the other as 'strong'. A no-music baseline condition was also presented. In the main study, Evaluation, Potency, and Activity, were once again assessed using the Semantic Differential technique (with four bipolar adjective scales for each dimension). Intriguingly, judgments of the individual shapes differed for a given musical accompaniment. So, for example, the activity ratings for the three geometric characters were shown to differ significantly under the two music backgrounds, although the overall activity rating of the two pieces of music did not differ significantly.

Marshall and Cohen (1988) attempted to explain these results by suggesting that an interaction had taken place between the temporal structure of the film and music that influenced the participants' visual attention, such that its focus differed under the two musical conditions (that said, no objective measure of participants' fixation was provided to back up such post hoc claims, though note that covert shifts of attention would not be observable using such techniques either) (see Table 1 for a summary of results). The sample size (namely five participants per condition) in Marshall and Cohen's preliminary study was likely statistically underpowered by today's standards; such a limitation is all the more important given the recent replication crisis in (psychological) science (see Bohannon, 2015), and the absence of any attempts at replication in this area that we are aware of. Nevertheless, Marshall and Cohen suggested that congruence in terms of the three principal dimensions of the semantic differential (Evaluation (good/bad), Potency (strong/weak), and Activity (active/passive)) likely influenced the meaning of the stimuli in this study and were computed simultaneously in each modality. In particular, different ratings on the calm-agitation dimension of the film figure as a function of the sound that was paired with the animation suggest that sound influenced the meaning of the film characters for the participants.

According to Marshall and Cohen (1988), if music, through 'structural similarity', were to have resulted in participants' attention being directed to a particular feature of the animation while, at the same time, providing particular connotative meaning, then this particular connotation may have become associated with the attended visual feature. Marshall and Cohen (1988, p. 95) also talk of "congruent

Table 1 Chronological summary of studies involving audivisual mid-level crossmodal correspondences (typically involving visual abstract animimations paired with either dynamic configurations of sounds or pre-composed music clips)

Study	N (participants)	Audiovisual Stimuli	Results
Marshall & Cohen (1988)	25 psychology students; 5 per condition; Stimuli presented twice	2-min Heider & Simmel (1944) abstract animation; Prelminary study: *Adagio* (weak) & *Allegro Marcato* (strong) section from Prokofiev's Symphony No.5. Main study: Major (Weak) or Minor (Strong) compositions.	Activity ratings for three characters in animation changed by the presence of music. E.g., activity rating of small triangle increases with strong music as compared to weak music or no music.
Cohen (1993)	E1: 12 participants	Video of bouncing ball varying in height and speed of bounce; Repeating melodic tone varying in pitch height and tempo.	Pitch height and tempo of music influenced participants' ratings of the happiness/sadness of the visual animations.
Sirius & Clarke (1994)	10 students x 4 films, 8 rate visual only, & 9 rate music only	Abstract video animation; 4 pieces of music composed in Disco, Spanish, Thriller, & Epic music (i.e., 'Spaghetti Western') styles.	Music exerted a consistent, additive effect on the evaluation of visual stimuli, buit no interaction effects observed.
Iwamiya (1994)	E1: 9 students	40 video disc excerpts, 20 matching & 20 mismatching to some degree (either desynchronized or mismatching).	Audition affected vision for higher-level factors ('cleanliness' & 'uniqueness') for matching stimuli; Brightness and evaluation also affected.
Lipscomb & Kim (2004)	28 participants	Rate degree of 'match' for 48 audivisual composites, varying in e.g., tempo, pitch, duration, & size.	Pitch associaed with vertical location, loudness with size, & timbre with shape.
Kim & Iwamiya (2008)	E1: 13 students	Nonsense letter strings projected on screen ('Telops'); Range of sound effects from broadcast programs in Japan (e.g., 'noise burst with gradual attack and delay; two consecutive Japanese drum sounds).	Formal audiovisual congruency led to higher ratings of audivisual stimuli: In particular, synchronous audiovisual onset and/or matching of changing patterns.
Kendall (2010)	E1: 16 students E2: 3 music majors	56 audiovisual composite stimuli (7 auditory x 8 visual); 7 audiovisual combinations	Participants sensitive to similar structures presented in the auditory and visual modalities.
Millet et al. (2021)	60 students; 12 per condition	70 s abstract animation from Heider & Simmel (1944); *Adagio* (weak) & *Allegro Marcato* (strong) from Prokofiev's Symphony No.5; Stimuli presented twice.	Presence of music influenced speed of first fixation on film objects & emotion associated with the characters in the abstract animation, but didn't influence attitudes to filmed events.

auditory and visual structure". However, and importantly, it remains unclear what exactly the authors mean with the phrase "crossmodal (audiovisual) temporal structure": Are they referring to synchronized elements, correlated temporal, isomorphic inputs, or perhaps something else entirely? Finally, it is worth noting that the animation used in the study had originally been developed by Heider and Simmel (1944) in the context of social psychology research, with participants being asked to describe what was going on by providing some sort of narrative, meaning that this narrative was not generated spontaneously by the participants.

Reflections on the joint process of the emergence of congruent music-visual temporal structure followed by the ascription of meaning (associations) to the visual stimuli resulted in the formulation of the Congruence-Associationist Framework (sometimes referred to as the Congruency-Associationist Model, or CAM for short) for understanding the effects of film music in film and video presentation (albeit with a film showing the spatiotemporal relations between simple elements). Marshall and Cohen (1988) also raise the possibility that this binding of semantic meaning associated with sound onto visual objects presented in the animated film could be considered as a kind of 'illusory conjunction' (cf. Treisman & Schmidt, 1982). As a post hoc explanation of the findings obtained, Marshall and Cohen's suggestions would appear plausible enough (cf. Kahneman & Henik, 1981). That being said, within the CAM framework, the causal relationship between association and congruency remains unclear, that is, whether percepts are associated because they are congruent, or whether they are judged as being congruent because they are associated. Moreover, it is uncertain whether similar occurrences should be considered as a genuine case of multisensory integration or not. Finally, the concept of congruency itself also appears rather ambiguous, unless it is clearly specified in relation to which particular features or dimensions the comparison is being made.

Certainly, the dichotomy between the structure (organization) and (semantic) meaning of the stimuli, no matter whether they happen to be auditory or visual, has some appeal. However, it is unclear whether Marshall and Cohen (1988) ever followed up on their suggestion in order to provide robust support for the claimed mechanism despite the model being re-presented and refined in many subsequent articles. As an aside, one might also note the visual bias that is inherent in Marshall and Cohen's experimental approach (cf. Hutmacher, 2019) in that one could presumably equally well have asked how visual stimuli affect people's interpretation of sound, or how audition and vision merge to deliver what Audissino (2017) calls a 'macro-configuration'. However, that said, the majority of the researchers working in this area to date have tended to focus on people's interpretation of the visual stimuli.

Cohen (1993) reported a study in which melodies lasting several seconds were sometimes presented at the same time as computerized visual animations showing a single object moving up and down on a screen (looking very much like a bouncing ball). The repeating single melodic notes varied in terms of their tempo (slow, moderate, or fast) and pitch (low, middle, or high), while the tempo (slow, medium, and fast) and height of the bouncing ball (low, medium, and high) were also varied. The participants used a single five-point scale to rate the apparent happiness/sadness of the music and video examples when presented individually. As might have been expected, given the extensive literature on musical emotion (see Hevner, 1936, 1937; Juslin, 2011; Riggs, 1964), the background melody was judged as happier when the tempo was faster and when the pitch was higher. Meanwhile, the rated 'happiness' of the bouncing ball increased for faster tempos of bouncing as well as for those bouncing videos where the 'ball' bounced higher.

The findings do not appear to support the conclusion subsequently drawn by Cohen (2005, pp. 20–21), namely that: "the meaning of the music and visual materials was systematically related to physical characteristics of the sound and light patterns." That the small number of participants in a forced-choice task consensually rate physical stimuli along some 'arbitrary' response scale, in-and-of-itself tells you absolutely nothing about what those stimuli 'mean'. At least not if by 'meaning', one considers the associations that are spontaneously brought to the top of mind. Certainly, one might worry about the possibility of halo-dumping (Clark & Lawless, 1994); this the name given for the tendency of participants to dump their feelings and experience onto whatever response scale they have been presented with, no matter whether those scales capture their experience or not.

Cohen's (2005) study also demonstrated a significant crossmodal influence of the music on participants' rating of the dynamic visual stimuli. So, for example, a ball that bounced high and fast was judged as looking very happy when the accompanying background music happened to be high pitched and fast as well, but was judged as looking less happy when paired with low-pitched and slow background music instead. Taken together, these results therefore appear to show that the happiness attributed to the accompanying music (mediated by variations in tempo and pitch height) influenced the judged happiness of the bouncing ball. What remains unclear, at least from more of a philosophical angle, is whether this should be considered as an example of one audiovisual object/event, or as separate visual and auditory objects/events. However, as a starting point in this debate, one might point to the absence of precisely synchronized or correlated inputs as likely reducing the likelihood of any kind of perceptual unification (see also Spence & Di Stefano, 2025a).

Some researchers have explicitly manipulated the synchrony of the auditory and visual elements in audiovisual film clips (e.g., Iwamiya, 1992, 1994, 2013; see also Bruce Nauman's (1969) Lip Sync video installation: https://www.moma.org/collection/works/107669). So, for example, Iwamiya (1994) presented short audiovisual film clips that were either synchronized or else had been deliberately desynchronized by 500 ms (note the shift from mid-level, or structural, correspondences to more complex correspondences). The participants had to respond on 22 different bipolar adjective pairs that could be applied to auditory and visual elements of the commercial films selected for this study. It is, however, worth bearing in mind the very large number of semantic differential scales that each participant had to complete in Iwamiya's Experiment 1 – namely, 40 film clips × 22 SD scales (for four of five conditions) (auditory rating without video; video rating without sound; sound and video ratings in the presence of the other modality) and then rating the degree of matching of sound and video for all clips on a 7-step scale) = 3,560 ratings in total per participant. Factor analysis of participants' semantic differential responses was interpreted as revealing five dimensions of meaning: Tightness, evaluation, brightness, cleanness, and uniqueness. The results revealed that the participants rated the original synchronized audiovisual clips as matching better than the clips that had been deliberately desynchronized. That said, it should be noted how people rapidly adapt to asynchronous audiovisual speech stimuli, even when the magnitude of the asynchrony is quite large (i.e., several hundred ms; e.g., Dixon & Spitz, 1980; Macaluso et al., 2004; Nahrstedt, 2024; Steinmetz, 1996).

Sirius and Clarke (1994) used a modified version of the approach first presented by Marshall and Cohen (1988), once again using the semantic differential technique to measure people's ratings of various combinations of auditory and visual stimuli. In this case, computer-generated moving images and music that had been specially composed were used as stimuli. Twenty-seven participants were divided into three groups, with each group exposed to audiovisual (10), visual-only (8), or auditory-only stimuli (9). The results indicated that music exerted a consistent, additive effect on the evaluation of visual stimuli, but no significant interaction effects were documented between specific musical styles and visual sequences. In this case, the absence of any synergistic audiovisual meanings was attributed to the simplicity of the visual material that was presented. Sirius and Clarke proposed an interpretative framework grounded in principles of ecological social perception to account for the observed patterns in crossmodal evaluation. However, with the benefit of hindsight, this study once again appears to be statistically underpowered, especially given the between-participants nature of the experimental design (see Brysbaert, 2019).

Lipscomb and Kim (2004) investigated the relationship between the auditory and visual components of an audiovisual composite. The 28 participants in this study rated the perceived degree of 'match' between audio/video components in a series of randomly presented audiovisual composites. (It is, though, unclear whether such a 'matching' task can provide meaningful insights into the 'congruency' between the stimuli.) The audio parameters that were manipulated included pitch, loudness, timbre, and duration, while the visual parameters that were manipulated included colour, vertical location, shape, and size. Audiovisual composites were created by combining all possible pairs of audio and visual stimuli using three different magnitudes of change (small, moderate, large), giving rise to 48 stimuli in total. The participants' mean response ratings revealed the following primary relationships: pitch with vertical location, loudness with size, and timbre with shape. Note that the colour was equally matched by the participants with both pitch and loudness.

Kim and Iwamiya (2008) studied people's perception of simple auditory stimuli and moving letter patterns (referred to as 'Telops'; that is, animated text on a display as might be seen in television commercials). The studies, which involved participants completing 22 semantic differential ratings scales for each of 64 stimuli (eight auditory patterns crossed with eight Telops patterns), revealed a sensitivity to similar patterns of motion across the visual and audio modalities (see also Kendall, 2010; Lipscomb, 2005). Once again, the semantic differential approach, based on the three dimensions of Evaluation, Potency, and Activity, was used. Two types of formal congruency were found to be effective in creating subjective congruency: Specifically, the synchronization of temporal structures and the matching of changing patterns of auditory and visual events. The former relies on correspondence of the onsets of sound and Telops pattern (i.e., temporal synchrony), the latter, typically, on the combinations of the gradually rising of some acoustic feature, for example, loudness or pitch of sounds, and the expanding (or approaching) of the Telops patterns. The combinations of the gradually rising loudness, or increasing pitch, of sounds and the expanding (or approaching) of the Telops pattern were rated as matching. Formal congruency also contributed to enhancing the evaluation of the audiovisual productions by the participants.

Kendall (2010) conducted a couple of experiments to assess the impact of music in and the perception of visual animation. In a first experiment, seven auditory structures were combined with eight visual patterns (i.e., a total of 56 patterns were presented). The results demonstrated that participants were sensitive to similar structures being presented in the two modalities, such as an arch, ramp, and undulation. However, there appeared to be a preference for left-to-right interpretation of stimuli. In a second study, seven audiovisual contours were presented and participants rated the match between the arch structures in both senses as very good (see also Karwoski, Odbert, & Osgood, 1942).

Finally, Millet, Chattah, and Ahn (2021) conducted a mixed-design experiment in which 60 participants viewed a 70-s dynamic visual animation from Heider and Simmel (1944) while listening to different pieces of music. They chose the same two movements from Prokofiev's Symphony No. 5, as has first been used in Marshall and Cohen's (1988) preliminary study. The participants listened to each piece of music by itself; they viewed the animation in silence; and they watched the video when paired with each piece of music. These five within-participant conditions were presented to participants in a randomized order. Participants' eye position was monitored, as was electrodermal activity, their self-reported emotional state, perception of the film characters, and the three dimensions of the semantic differential scale. Millet et al. assessed the effect of music on participants' visual attention, their emotional responses, and their attitude towards film objects, and the continuation of narratives. The authors assessed the impact of the latter on participants' visual attention and their affective responses to the animation. As had originally been suggested by Marshall and Cohen (1988), the presentation of the music while watching the video led to faster first fixations on the visual objects depicted in the animation and supplied emotional content, increasing positive sentiment for the animation's characters. As such, these results can be taken as providing support for crossmodal spatial attention, driven by structural correspondence between the auditory and visual channels, playing some role in influencing people's perception of the visual animation (see Ansani et al., 2020, for a similar approach involving a short video made up of static images; and Clemente, Friberg, & Holzapfel, 2023).

Interim synthesis: Mechanisms underlying reported effects

Taken together, the studies reviewed above point to a number of shared explanatory accounts that extend beyond their original framing in terms of film music or audiovisual matching (e.g., Cohen, 2005), and instead invite interpretation in terms of crossmodal perceptual organization (see Spence & Di Stefano, 2025b). What is noticeable from this review of the literature is how widely Osgood, Suci, and Tannenbaum's (1957) semantic differential technique has been used by the researchers working in this area (e.g., Iwamiya, 1994; Marshall & Cohen, 1988; Sirius & Clarke, 1994). At the same time, however, it is also worth noting that some film study researchers have criticized this approach, arguing that forcing people to respond along a small (or even large) set of bipolar adjective scales may fail to capture the richness of any emergent properties or sensations that may result from combining the senses (e.g., Audissino, 2017; Wells, 1980; see also Spence & Di Stefano, 2025b).

In addition to the above-mentioned 'Congruency-Associationist Model' (Cohen, 2005), and a more conventional cognitive, or experimental, psychology approach (often using the semantic differential technique), various additional explanations for the crossmodal effects have been observed in the reviewed literature. First, any affect (or emotional response) that happens to be associated with, or triggered by, the music may carry over (perhaps as a result of sensory transfer) to influence the viewer's interpretation of visuals (as in the case of film music, see Spence & Di Stefano, 2025c). Second, there might be some form of crossmodal influence such that visual stimuli appear different as a result of crossmodal perceptual interactions (crossmodal perceptual effects, e.g., Armontrout, Schutz, & Kubovy, 2009; Olivers & Van der Burg, 2008; for a review, see Spence, 2018b). Relevant here, crossmodal research has revealed evidence of extensive crossmodal influences on perceptual grouping (for reviews, see Spence, 2015; Spence & Di Stefano, 2025a; Spence, Sanabria, & Soto-Faraco, 2007), albeit in the absence of any genuine 'inter-sensory Gestalten' (see Gilbert, 1938). This can be considered as similar to what Kubovy and Yu (2012, p. 963) term 'trans-modal gestalts'. In such cases, note, it might be the difference in the perception of the visual stimuli that leads to the change in participants' ratings. Third, synchronized crossmodal inputs may capture a viewer's visual attention (see Millet et al., 2021, for some preliminary answers here). While synchronous transients have sometimes been shown to be important, a looser definition of synchrony has been shown to be sufficient to give rise to crossmodal effects at the level of temporally extended stimulus sequences (including correlated accent structures across the senses).

Crossmodal grouping and perceptual organization

A central question emerging from the reviewed literature on audiovisual integration concerns how and under what conditions signals from different sensory modalities are grouped into a single perceptual event. This section focuses on crossmodal grouping at a perceptual level, with the aim of identifying a number of organizing principles that recur across the empirical literature and constrain how audiovisual integration can occur. Rather than treating synchrony as a sufficient condition for audiovisual binding, it will be clear that temporal alignment between auditory and visual streams does not exert a uniform effect across emotions, but instead interacts with the expressive qualities of the stimuli. This pattern supports a Gestalt interpretation of synchrony as a context-sensitive cue the

perceptual consequences of which depend on higher-order structure rather than on simultaneity alone.

Temporal synchrony as a basic grouping cue

A series of studies on the so-called 'pip-and-pop' effect in the field of cognitive psychology have demonstrated that the mere audiovisual synchrony of sound and visual stimuli can lead to the 'pop-out' of the latter in complex dynamic visual displays (e.g., Klapetek, Ngo, & Spence, 2012; Van der Burg et al., 2010; Van der Burg et al., 2008). This effect refers to the phenomenon whereby the presentation of a sudden-onset sound (or other transient, such as a tactile stimulus) leads to the pop-out of a synchronized change in a complex visual array of changing items, thus making the synchronized visual stimulus appear more perceptually salient. Such contemporary crossmodal research findings, and many others like them, demonstrate how synchronized audiovisual stimuli can give rise to measurable crossmodal effects on visual perception (see also Grassi & Casco, 2010; Ryan, 1940; Shams, Kamitani, & Shimojo, 2000; Staal & Donderi, 1983).

Moving to real-world examples involving both mid-level and complex stimuli, it is interesting to observe, with Strachan (2006), the synchrony of the elements in Daft Punk's 1997 house music track *Around the World*. In the video that accompanies the song, the appearance of four sets of four dancers are synchronized to coincide with (and thus represent) different parts of the music. According to Strachan (2006, p. 200), each musical unit "is choreographed to a particular group of dancers and dance moves in the video in a kinetic representation of musical structure, melody, and harmony." Indeed, music videos may more frequently capture the temporal aspects of the music by synchronized visuals than other forms of audiovisual media.

While these examples help to illustrate how synchrony is one of the most effective ways to induce a sense of objective unity in the observer or to establish a causal link between multiple sensory inputs, the effectiveness of synchrony is strongest for discrete, punctate events having a clearly defined onset. Under these conditions, synchrony can operate as a powerful bottom-up cue that captures attention and facilitates perceptual binding. At the same time, this strength also highlights a limitation: temporal synchrony alone does not scale well to extended or structurally complex stimuli, such as sequences, rhythms, or continuous audiovisual streams. In these cases, simultaneity at isolated time points is insufficient to account for stable perceptual grouping.

Beyond simultaneity: Correlated temporal structure and isomorphism

In many mid-level audiovisual displays, especially those involving extended sequences, binding does not depend on millisecond-level simultaneity. Instead, correlated temporal structures (e.g., shared rhythm, accent patterns, or dynamic contours) are sufficient to support perceptual grouping (see Iwamiya, Sugano, & Kouda, 2000; Lipscomb, 2013; Müller, 2010). From a Gestalt perspective, the same temporal form or rhythm can be instantiated in different sensory modalities, giving rise to what has often been described as structural or isomorphic correspondence. As observed by Pratt:

"An auditory rhythm is auditory, and that's that; but the same rhythm – a Gestalt – may also be visual or tactual, and the graceful lilt, let us say of a waltz rhythm […] will be present in all three modalities. Gestalten […] reveal innumerable iconic relations and resemblances across modalities. Therein lies the great power of art, for the moods and feelings of mankind are capable of iconic *presentation* in visual and auditory patterns a mode obviously far more direct and effective than symbolic *representation*." (Pratt, 1969, pp. 25–26 [italics in original]; see also Collopy, 2000; Kulezic-Wilson, 2004).

Recently, Di Stefano and Spence (2025) reviewed the literature supporting the existence of similar temporal grouping principles in both audition and vision (e.g., Aksentijević, Elliott, & Barber, 2001; Allen, Walker, Symonds, & Marcell, 1977; Kang, Lancelin, & Pressnitzer, 2018; Marks, 1987). This body of empirical research demonstrates that temporal patterns that are perceived in one sensory modality – such as audition – can often be recognized when presented via another sensory modality, such as vision. While this crossmodal recognition is most robust between audition and vision, it may sometimes extend to the sense of touch, though it is far less evident in the chemical senses. However, when considering the perception of temporal structure distributed across modalities (one form of intersensory Gestalten) – where some portions of the temporal information are presented to one sensory modality and the remainder to another – the evidence becomes sparse and less conclusive (though see Huang et al., 2012; for a review, see Spence, 2015).

At one level more of abstraction, the relations between multi-element (or dynamically changing) auditory and visual stimuli may be isomorphic (cf. Omwake, 1940; Ravignani & Sonnweber, 2017; Thornley Head, 2006). Here, for example, we might perhaps think of transposing a temporal auditory pattern into a spatial visual arrangement (though what constitutes the most appropriate transformation between the auditory and visual modalities undoubtedly remains something of a contentious issue: see Handel, 1988a, b; Julesz & Hirsh, 1972; Kubovy, 1988). Crucially, isomorphism involves a formal structural correspondence between elements, and thus requires a clearly defined criterion or mapping rule – something that makes it different from mere perceptual similarity based on the sharing of phenomenological properties (see also Schöffer, 1985; Vanel, 2009). As Di Stefano and Spence (2024) note, while judgements of perceptual similarity are

relatively easy to understand within a single sensory modality, they become considerably more difficult to define and apply crossmodally. This is likely because different senses often rely on distinct representational codes and processing mechanisms, and lack a common metric, thus making direct comparisons ambiguous and highly dependent on context, task demands, and perhaps also individual experience. One might question the relation between isomorphism and metaphor (Dahl & Adachi, 2013; Liu & Kennedy, 1997; Ramachandran et al., 2020; Wagner, Winner, Cicchetti, & Gardner, 1981). However, it should be noted that getting to grips with this thorny question lies beyond the scope of the present article.

In an intriguing early article, Julesz and Hirsh (1972) looked for analogies between auditory and visual pattern perception (see also Harris, 1950; Stevens, 1958). They consider the various Gestalt grouping principles and consider the extent to which they may be present in both the visual and auditory modalities. They also suggested that grouping by proximity and similarity likely operate in both senses, and that good continuation (or smooth transformation) also seems to apply to both senses, as does figure/ground segregation. Numerosity and matching spatiotemporal patterns (perhaps involving melody) could be present in both senses. Set effects too can presumably be extended across the senses (Liu, 1976). By contrast, symmetry and closure appear more relevant to the visual rather than the auditory modality, and emergence has been documented in the visual modality but is not so obviously apparent in audition.

There is also an intriguing literature emerging on morphodynamics, connecting time-varying sounds/music with complex visual forms. For instance, Wanke et al. (2025) recently investigated the audiovisual associations between contemporary experimental music and abstract shapes. The authors chose sonic compositions belonging to specific styles of experimental music, namely spectralism and electronic-glitch music. These compositions are typically perceived as a sequence of sound patterns which can prompt a phenomenological and sensory engagement (Wanke, 2021) rooted in processing mechanisms associated with the early stages of perception, such as primal scene segregation (Bregman, 1990) and the formation of auditory Gestalten (Koelsch & Siebel, 2005). Overall, Wanke et al.'s results revealed that participants consistently selected images that shared a large number of morphological features with the corresponding auditory stimuli. This not only reveals a general preference for linear spectrotemporal visual representations of auditory experiences (Wanke, 2023), but also supports the notion that listeners spontaneously transpose auditory patterns into spatial visual forms.

Emergence of multisensory perceptual unit in the context of structured audiovisual stimuli

In some cases, the combination of structured audiovisual stimuli has been interpreted in terms of a multisensory perceptual unit (e.g., Kim & Iwamiya, 2008). This outcome is believed to occur especially in the case of functionally congruent stimuli, such as music and images in movies, where neither perfect synchrony nor any alternative structural congruence seem to apply. Beyond temporal congruency (Bolivar, Cohen, & Fentress, 1994), emotional congruency has also been evoked as an effective means to induce the emergence of a multisensory unit. This occurs when two sensory inputs, taken separately, are associated with the same or similar emotional meaning. Note that, for the binding of the two sensory inputs to occur, they have to be temporally aligned and/or the inputs available to each sensory modality should be temporally correlated with one another (Parise et al., 2012, 2013; see also Jertberg et al., 2024).

Importantly, the kind of multisensory emergence just discussed cannot be explained solely by temporal synchrony (§3.1) or structural isomorphism (§3.2). Although synchrony may play a facilitatory role, what distinguishes emergent audiovisual experiences is the superadditive unity and dynamic interplay of the various elements. That is, the audiovisual whole expresses something that neither sensory modality conveys by itself, and which is seemingly more than the mere sum of the parts (i.e., of the two modalities when considered individually; though see Spence, 2025). This may involve affective or conceptual layering, narrative construction, or spatial-temporal abstraction, often guided more by expressive fit than by precise temporal alignment.

Multiple outcomes of audiovisual co-presentation

Figure 2 highlights some of the various outcomes that may result when independent auditory and visual stimulus streams are presented together. Taken together, the literature reviewed here suggests that audiovisual co-presentation can lead to several distinct perceptual outcomes, including perceptual binding, in which signals are grouped into a single multisensory event; attentional modulation, where one modality enhances processing in another without full integration; partial interaction, such as biasing or priming effects across modalities; perceptual independence, where signals remain segregated despite temporal alignment; competition or interference, particularly when crossmodal cues conflict. This underscores the fact that occurrences of temporal synchrony and structural isomorphisms in the context of mid-level audiovisual correspondences are best understood as constraints on possible outcomes, rather than as mechanisms that uniquely determine integration.

Notice how a number of these possible outcomes have been reviewed in depth elsewhere (e.g., see Spence & Di Stefano,

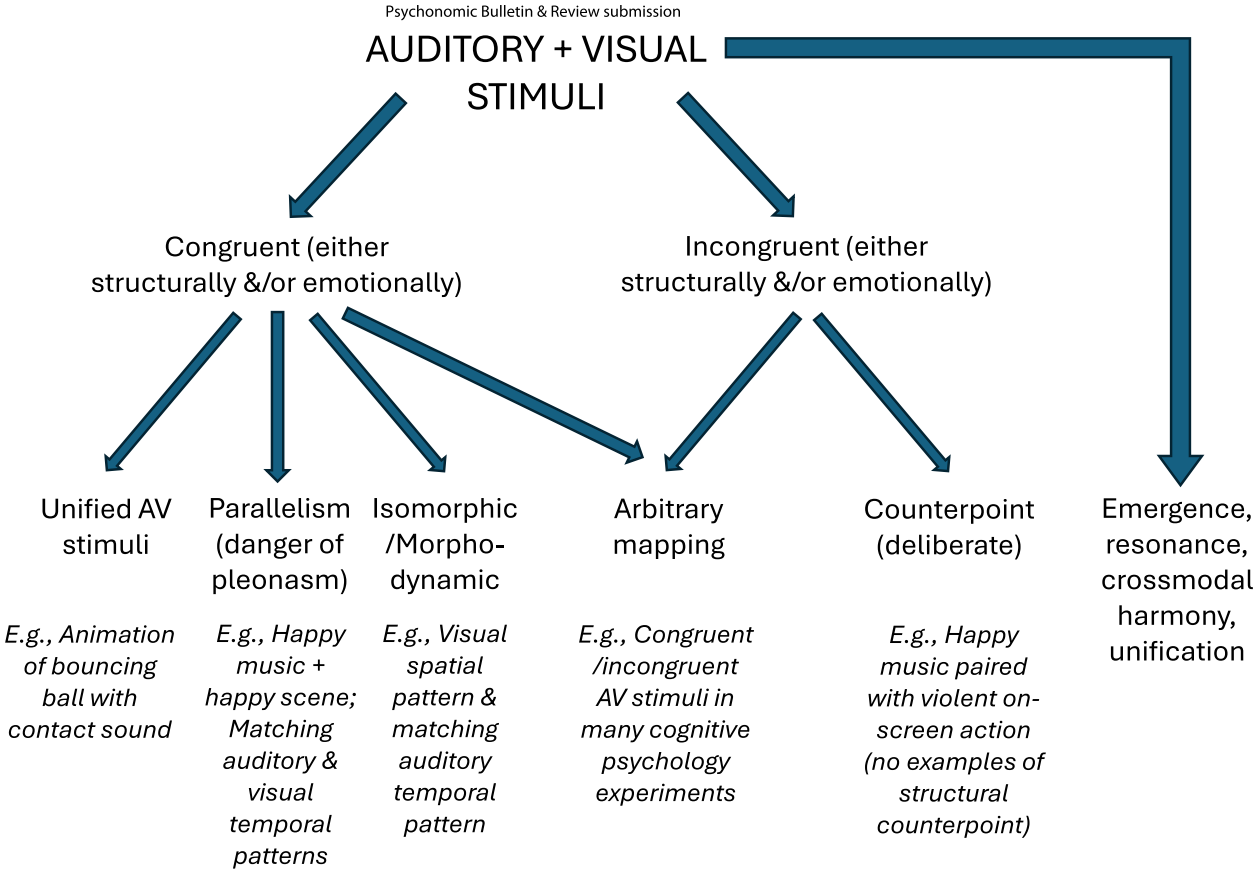

Fig. 2 Possible outcomes when mid-level auditory and visual stimuli are presented

2025a, b, c). Synchrony is clearly an important cue to binding (working best with transients occurring simultaneously in both modalities). However, it is important to recognize how beat/accent structure also provides grounds for the synchronization of sequences of unimodal sensory signals (e.g., Kendall, 2010; Lipscomb, 2013). There is, however, also a separate more recent experimental literature on correlated inputs (Parise et al., 2012, 2013) that provides an even more powerful cue to crossmodal binding. Subliminal audiovisual temporal congruency in music videos enhances perceptual pleasure (Lin, Yeh, & Shams, 2022). One might also wonder with regard to the context of emotive music, whether precise sensory synchronization matters, or whether instead it is the simultaneity of the emotions evoked by the music that are key to biasing the interpretation of any visual stimuli (cf. Su, 2014).

Discussion

The research that has been reviewed in this narrative historical review primarily concerns those crossmodal interactions involving mid-level audiovisual crossmodal correspondences (see Spence & Di Stefano, 2025b), as well as a few cross-level examples involving mid-level stimuli in one modality and complex stimuli in the other. While some of the early findings in this area were taken to inform film music studies (e.g., Cohen, 2005), our view is that they may actually be more informative with regard to (and/or better explained in terms of) crossmodal Gestalt perceptual grouping (see also Daurer, 2010; Staal & Donderi, 1983; Welch, DuttonHurt, & Warren, 1986; for review, see Spence & Di Stefano, 2025a). Certainly, stimulus structure (namely synchronized audiovisual inputs and correlated auditory and visual signals) appears to play an important role in the case of crossmodal perceptual grouping (Parise et al., 2012, 2013; see also Müller, 2010).

The literature on mid-level (or structural) crossmodal correspondences shares much with the cognitive psychology literature investigating the effects of pairing music with paintings (for a review, see Spence, 2020), or else with short movie clips (Spence & Di Stefano, 2025c), in that those participants taking part in these studies are rarely, if ever, told anything about why exactly they are being presented with the particular combinations of auditory and visual stimuli that the experimenter(s) has/have settled on. Given that there is no obvious intentionality behind combining mismatching auditory and visual stimuli, any possibility of meaningful

counterpoint is presumably lost (see Spence & Di Stefano, 2025b). However, one difference that comes out in this particular literature is the importance of the various different kinds of temporal alignment that may be experienced (see Fig. 3). From this perspective, temporal alignment should not be treated as a unitary construct, but rather as a family of relations that constrain, without determining, the outcome of audiovisual co-presentation.

Vision typically dominates over audition in laboratory-based experimental psychology research. This is thought to be due to the 'greater cortical real estate', information processing bandwidth, and attention that are apparently available to the visual modality (Gallace et al., 2012; Heilig, 1992; Posner, Nissen, & Klein, 1976). The dominant approach in this area involves assessing how music affects visual perception, rather than vice versa. This asymmetry reflects both theoretical assumptions and practical methodological choices, rather than a principled claim about multisensory perception. Such an asymmetry is, however, consistent with the focus on vision that has been highlighted in multisensory research more generally (Hutmacher, 2019). While the visual stimuli in the research reviewed here have mostly been mid-level (except Iwamiya, 1994), i.e., involving multiple, possibly dynamically changing simple stimuli, the auditory stimuli have included both mid-level and more complex semantically/emotionally meaningful music clips. It may be for this reason that auditory stimuli have typically been found to dominate visual stimuli in this literature (see Vines et al., 2011, for evidence that semantically meaningful visuals can sometimes impact people's auditory perception too). As Embler (1974) noted more than half a century ago: "music and film each depend upon the phenomena of movement and are thereby allied aesthetically…sound movement reinforces visual movement."

One other important methodological point to consider here relates to the widespread use of the semantic differential technique (e.g., Iwamiya, 1994; Iwamiya et al., 2000; Marshall & Cohen, 1988; Millet et al., 2021; Sirius & Clarke, 1994), though, as has been mentioned already, some commentators have criticized how this popular experimental approach, imported from cognitive psychology, may ultimately fail to capture important aspects of meaning, such as any emergent macro-structural or super-additive responses (Audissino, 2017). As such, there may be occasions in which a more open (and hence less constrained) response format is more appropriate for those wanting to know how people's responses to dynamic visual displays (such as animations) may be changed by the simultaneous presentation of sound (as can be found in a subset of the studies of mood music in the context of film studies; for a review, see Spence & Di Stefano, 2025b).

A related methodological issue concerns the frequent conflation of stimulus co-presentation with stimulus coherence. In many experimental designs, auditory and visual stimuli are presented simultaneously (or within a limited temporal window) precisely because the goal is to assess crossmodal effects. However, such designs make it difficult to disentangle the effects of mere co-presentation from those arising from genuine structural, temporal, or expressive relations between the component stimuli. As a result, observed crossmodal influences may reflect the presence of concurrent multisensory input rather than the specific coherence or congruency of the audiovisual pairings.

Taken together, these considerations highlight the need for a more nuanced account of temporal alignment and congruency in audiovisual perception in order to accommodate both low-level synchrony and higher-order structural and expressive relations.

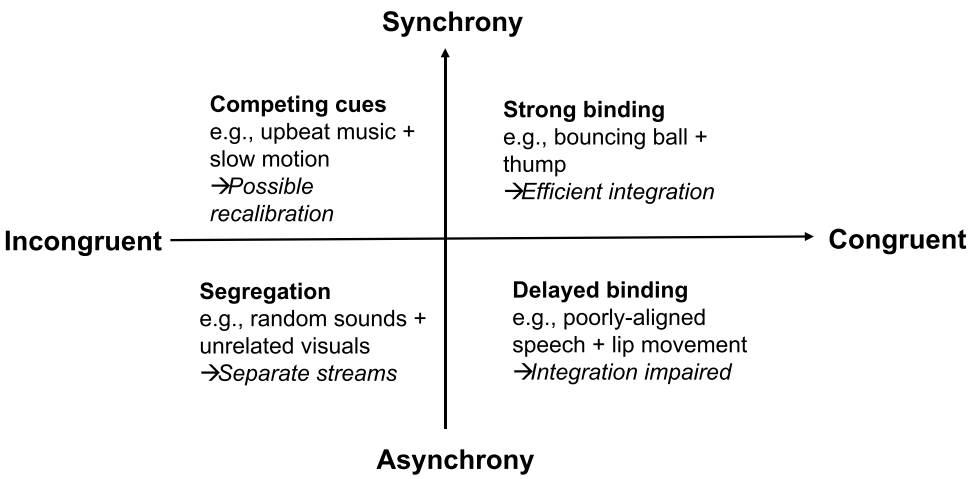

Fig. 3 Possible outcomes when mid-level auditory and visual stimuli (varying in terms of their synchrony and congruency) are presented

Future directions

While the research that has been published to date highlights the importance of synchrony in establishing mid-level crossmodal correspondences, future research should clarify the specific contribution of precise temporal alignment relative to other factors, such as structural isomorphism, temporally correlated inputs, and semantic relatedness, in the perception of audiovisual correspondences. Moreover, it is crucial to examine how specific musical features (e.g., tempo, pitch, rhythm) interact with dynamic visual elements (e.g., motion, speed, direction) to influence people's perception and any emotional response that might be evoked by the stimuli that happen to be presented.

An important direction for future research concerns the systematic disentangling of stimulus co-presentation from crossmodal coherence. While many existing studies have chosen to focus on the simultaneous or temporally proximate presentation of auditory and visual stimuli in order to elicit crossmodal effects, future experimental designs could benefit from independently manipulating the presence of concurrent multisensory input and the degree of structural, temporal, or expressive correspondence between modalities. Such approaches would make it possible to assess not only whether audiovisual co-presentation influences perception, but under which conditions coherence constrains perceptual grouping and interpretation.

In the era of the replication crisis in psychological and other sciences, the very small sample sizes used in many of the early between-participants studies relevant to the study of mid-level crossmodal correspondences would not be accepted today. Thus, another important direction looks back to published literature with the aim of replicating some of the foundational observations in this particular area of research with adequate sample sizes (e.g., Marshall & Cohen, 1988; Sirius & Clarke, 1994).

In parallel, there is a need to determine the cognitive mechanism(s) by which music influences people's perception of, and responses to, dynamic visual content, likely involving some combination of subjective report and psychophysiological measures (cf. Thaiwong & Fukumoto, 2024, 2025). Subjective reports, including continuous ratings and semantic differential scales, can capture the nuances of conscious experience, while techniques like eye-tracking, electrodermal activity, and heart-rate variability can provide insights into attentional and emotional responses. Furthermore, neuroimaging techniques could potentially be used to investigate the neural mechanisms underlying the integration of synchronous versus asynchronous, and congruent versus incongruent, audiovisual information at different levels of processing (e.g., expecting higher gamma band oscillations in response to congruent/synchronous audiovisual stimuli, based on the available evidence on separate senses: see Knief et al., 2000; Müller et al., 1997). Understanding how the brain differentiates and integrates these dimensions will be crucial for those wishing to develop any kind of comprehensive model of crossmodal perception involving the more complex configurations of stimuli that we tend to be presented with in everyday life.

Finally, future research should aim to extend the current findings to more ecologically valid scenarios. While controlled laboratory studies have undoubtedly provided researchers with essential insights, examining how synchrony, temporal correlation, and congruency operate in complex real-world contexts, such as film viewing, human-computer interaction, and artistic performances (see Daniels et al., 2010, for many such examples), will be crucial for understanding their broader significance. That said, the latter examples would appear to fall more in the realm of complex rather than mid-level correspondences. This investigation can benefit from the use of portable and low-invasive devices for monitoring physiological parameters, such as cardiac or electrodermal activity.

Conclusions

This review has examined empirical research on mid-level audiovisual crossmodal correspondences, identifying a set of recurring patterns across otherwise heterogeneous paradigms. Crossmodal effects are most consistently observed when auditory and visual streams exhibit some form of temporal alignment or structural correspondence, whereas simple co-occurrence or semantic association alone often proves insufficient to account for the reported outcomes. Rather than interpreting these findings primarily in terms of film music or modality-specific meanings, we have argued that many such effects are more parsimoniously understood within a framework of crossmodal perceptual organization. From this perspective, temporal alignment, structural similarity, and expressive coherence act as partially independent constraints on perceptual grouping, shaping how multisensory information is integrated over time rather than functioning as independent mechanisms.

By synthesizing results across diverse experimental approaches, this review highlights the importance of distinguishing between the effects of stimulus co-presentation and those that emerge from crossmodal coherence. Note here how making this distinction explicit has implications both for the interpretation of existing findings and for the design of future studies aimed at isolating the conditions under which mid-level audiovisual correspondences emerge. More broadly, an organizational view of audiovisual interaction may help unify research across domains while supporting experimental paradigms that better capture the dynamic and emergent nature of multisensory perception.

Authors' contributions CS and NDiS: Original concept, draft, and editing.

Funding CS would like to thank the AHRC grant entitled 'Rethinking the senses' (AH/L007053/1) for supporting this research.

Data availability Not applicable.

Code availability Not applicable.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Open practices statement Not applicable.

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

References

Aksentijević, A., Elliott, M. A., & Barber, P. J. (2001). Dynamics of perceptual grouping: Similarities in the organization of visual and auditory groups. *Visual Cognition, 8*(3–5), 349–358. https://doi.org/10.1080/13506280143000043

Allen, T. W., Walker, K., Symonds, L., & Marcell, M. (1977). Intrasensory and intersensory perception of temporal sequences during infancy. *Developmental Psychology, 13*(3), 225–229. https://doi.org/10.1037/0012-1649.13.3.225

Andersen, T. S., & Mamassian, P. (2008). Audiovisual integration of stimulus transients. *Vision Research, 48*(25), 2537–2544. https://doi.org/10.1016/j.visres.2008.08.018

Ansani, A., Marini, M., D'Errico, F., & Poggi, I. (2020). How soundtracks shape what we see: Analyzing the influence of music on visual scenes through self-assessment, eye tracking, and pupillometry. *Frontiers in Psychology, 11*, Article 556697. https://doi.org/10.3389/fpsyg.2020.02242

Armontrout, J. A., Schutz, M., & Kubovy, M. (2009). Visual determinants of a cross-modal illusion. *Attention, Perception & Psychophysics, 71*, 1618–1627. https://doi.org/10.3758/APP.71.7.1618

Audissino, E. (2017). A gestalt approach to the analysis of music in films. *Musicology Research, 2*(1), 69–88.

Bohannon, J. (2015). Many psychology papers fail replication test. *Science, 349*(6251), 910–911. https://doi.org/10.1126/science.349.6251.910

Bolivar, V. J., Cohen, A. J., & Fentress, J. C. (1994). Semantic and formal congruency in music and motion pictures: Effects on the interpretation of visual action. *Psychomusicology : a Journal of Research in Music Cognition, 13*, 28–59. https://doi.org/10.1037/h0094102

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. MIT Press.

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition, 2*(1), Article 16. https://doi.org/10.5334/joc.72

Clark, C. C., & Lawless, H. T. (1994). Limiting response alternatives in time-intensity scaling: An examination of the halo dumping effect. *Chemical Senses, 19*(6), 583–594. https://doi.org/10.1093/chemse/19.6.583

Clemente, A., Friberg, A., & Holzapfel, A. (2023). Relations between perceived affect and liking for melodies and visual designs. *Emotion, 23*(6), 1584–1605. https://doi.org/10.1037/emo0001141

Cohen, A. J. (1993). Associationism and musical soundtrack phenomena. *Contemporary Music Review, 9*(1–2), 163–178. https://doi.org/10.1080/07494469300640421

Cohen, A. J. (2005). How music influences the interpretation of film and video: Approaches from experimental psychology. *Selected Reports in Ethnomusicology, 12*, 15–40.

Collopy, F. (2000). Colour, form, and motion – Dimensions of a musical art of light. *Leonardo, 33*(5), 355–360. https://doi.org/10.1162/002409400552829

Cook, L. A., Van Valkenburg, D. L., & Badcock, D. R. (2011). Predictability affects the perception of audiovisual synchrony in complex sequences. *Attention, Perception & Psychophysics, 73*(7), 2286–2297. https://doi.org/10.3758/s13414-011-0185-8

Dahl, C. D., & Adachi, I. (2013). Conceptual metaphorical mapping in chimpanzees (*Pan troglodytes*). *eLife, 2*, Article e00932. https://doi.org/10.7554/eLife.00932

Daniels, D., Naumann, S., & Thoben, J. (2010). *See this sound: Audiovisuology. An interdisciplinary survey of audiovisual culture*. Verlag der Buchhandlung Walter König.

Daurer, G. (2010). Audiovisual perception. In D. Daniels, S. Naumann, & J. Thoben (Eds.), *See this sound: Audiovisuology. An interdisciplinary survey of audiovisual culture* (pp. 328–347). Verlag der Buchhandlung Walter König.

Deroy, O., Crisinel, A.-S., & Spence, C. (2013). Crossmodal correspondences between odors and contingent features: Odors, musical notes, and geometrical shapes. *Psychonomic Bulletin & Review, 20*, 878–896. https://doi.org/10.3758/s13423-013-0397-0

Deroy, O., & Spence, C. (2013). Why we are not all synesthetes (not even weakly so). *Psychonomic Bulletin & Review, 20*, 643–664. https://doi.org/10.3758/s13423-013-0387-2

Deutsch, J. (2012). Synaesthesia and synergy in art. Gustav Mahler's "Symphony No. 2 in C minor" as an example of interactive music visualization. In F. G. Barth, P. Giampieri-Deutsch, & H.-D. Klein (Eds.), *Sensory perception—Mind and matter* (pp. 215–235). Springer.

Dixon, N. F., & Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception, 9*(6), 719–721. https://doi.org/10.1068/p090719

Di Stefano, N., Ansani, A., Schiavio, A., Saarikallio, S., & Spence, C. (2025). Audiovisual associations in Saint-Saëns' *Carnival of the Animals*: A cross-cultural investigation on the role of timbre. *Empirical Studies of the Arts, 43*(2), 1162–1180. https://doi.org/10.1177/02762374241308810

Di Stefano, N., Ansani, A., Schiavio, A., & Spence, C. (2024). Prokofiev was (almost) right: A cross-cultural exploration of auditory-conceptual associations in *Peter and the Wolf*. *Psychonomic Bulletin & Review, 31*, 1735–1744. https://doi.org/10.3758/s13423-023-02435-7

Di Stefano, N., & Spence, C. (2024). Perceptual similarity: Insights from crossmodal correspondences. *Review of Philosophy and Psychology, 15*(3), 997–1026. https://doi.org/10.3758/s13414-025-03045-2

Di Stefano, N., & Spence, C. (2025). Perceiving temporal structure within and between the senses: A multisensory/crossmodal perspective. *Attention, Perception, & Psychophysics, 87*, 1811–1838. https://doi.org/10.3758/s13414-025-03045-2

Embler, J. (1974). The structure of film music. In J. L. Limbacher (Ed.), *Film music: From violin to video* (pp. 61–65). Scarecrow.

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision, 10*(1), Article 6. https://doi.org/10.1167/10.1.6

Gallace, A., Ngo, M. K., Sulaitis, J., & Spence, C. (2012). Multisensory presence in virtual reality: Possibilities & limitations. In G. Ghinea, F. Andres, & S. Gulliver (Eds.), *Multiple sensorial media advances and applications: New developments in MulSeMedia* (pp. 1–38). IGI Global.

Gilbert, G. M. (1938). A study in inter-sensory Gestalten. *Psychological Bulletin, 35*(9), 698. https://doi.org/10.1037/h0055433

Grassi, M., & Casco, C. (2010). Audiovisual bounce-inducing effect: When sound congruence affects grouping in vision. *Attention, Perception, & Psychophysics, 72*(2), 378–386. https://doi.org/10.3758/APP.72.2.378

Handel, S. (1988a). Space is to time as vision is to audition: Seductive but misleading. *Journal of Experimental Psychology. Human Perception and Performance, 14*(2), 315–317. https://doi.org/10.1037/0096-1523.14.2.315

Handel, S. (1988b). No one analogy is sufficient: Rejoinder to Kubovy. *Journal of Experimental Psychology. Human Perception and Performance, 14*(2), 321. https://doi.org/10.1037/h0092865

Harris, J. D. (1950). *Some relations between vision and audition*. Charles C. Thomas.

Hauck, P., von Castell, C., & Hecht, H. (2022). Crossmodal correspondence between music and ambient color is mediated by emotion. *Multisensory Research, 35*(5), 407–446. https://doi.org/10.1163/22134808-bja10077

Heider, F., & Simmel, M. (1944). An experimental study of apparent behaviour. *American Journal of Psychology, 57*, 243–259.

Heilig, M. L. (1992). El cine del futuro: The cinema of the future. *Presence (Cambridge, Mass.), 1*(3), 279–294. https://doi.org/10.1162/pres.1992.1.3.279

Heins, N., Pomp, J., Kluger, D. S., Vinbrüx, S., Trempler, I., Kohler, A., et al. (2021). Surmising synchrony of sound and sight: Factors explaining variance of audiovisual integration in hurdling, tap dancing and drumming. *PLoS ONE, 16*(7), Article e0253130. https://doi.org/10.1371/journal.pone.0253130

Hevner, K. (1936). Experimental studies of the elements of expression in music. *American Journal of Psychology, 48*, 246–268. https://doi.org/10.2307/1415746

Hevner, K. (1937). The affective value of pitch and tempo in music. *American Journal of Psychology, 49*, 621–630. https://doi.org/10.2307/1416385

Huang, J., Gamble, D., Sarnlertsophon, K., Wang, X., & Hsiao, S. (2012). Feeling music: Integration of auditory and tactile inputs in musical meter perception. *PLoS ONE, 7*(10), Article e48496. https://doi.org/10.1371/journal.pone.0048496

Hutmacher, F. (2019). Why is there so much more research on vision than on any other sensory modality? *Frontiers in Psychology, 10*, Article 2246. https://doi.org/10.3389/fpsyg.2019.02246

Iwamiya, S. (1992). The interaction between auditory and visual processing when listening to music via audio-visual media. *Journal of the Acoustical Society of Japan, 48*, 146–153.

Iwamiya, S. (1994). Interaction between auditory and visual processing when listening to music in an audiovisual context: 1. Matching 2. Audio quality. *Psychomusicology, 13*(1–2), 133–154. https://doi.org/10.1037/H0094098

Iwamiya, S. (2013). Perceived congruence between auditory and visual elements in multimedia. In S.-L. Tan, A. J. Cohen, S. D. Lipscomb, & R. A. Kendall (Eds.), *The psychology of music in multimedia* (pp. 141–164). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199608157.003.0007

Iwamiya, S., Sugano, Y., & Kouda, K. (2000). The effects of synchronization of temporal structures of sound and motion picture on the impression of audio-visual contents. *Systems, Man, and Cybernetics, 2000 IEEE International Conference*, 2, 1222-1225.

Jertberg, R. M., Begeer, S., Geurts, H. M., Chakrabarti, B., & Van der Burg, E. (2024). Perception of temporal synchrony not a prerequisite for multisensory integration. *Scientific Reports, 14*, Article 4982. https://doi.org/10.1038/s41598-024-55572-x

Julesz, B., & Hirsh, I. J. (1972). Visual and auditory perception - An essay of comparison. In E. E. David & P. B. Denes (Eds.), *Human communication: A unified view* (pp. 283–340). McGraw-Hill.

Juslin, P. N. (2011). Music and emotion: Seven questions, seven answers. In I. Deliège & J. Davidson (Eds.), *Music and the mind: Essays in honour of John Sloboda* (pp. 113–135). Oxford University Press.

Kahneman, D., & Henik, A. (1981). Perceptual organization and attention. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 181–211). Lawrence Erlbaum Associates.

Kang, H., Lancelin, D., & Pressnitzer, D. (2018). Memory for random time patterns in audition, touch, and vision. *Neuroscience, 389*, 118–132. https://doi.org/10.1016/j.neuroscience.2018.03.017

Kargon, J. (2011). Harmonizing these two arts: Edmund Lind's *The music of color*. *Journal of Design History, 24*(1), 1–14. https://doi.org/10.1093/jdh/epq042

Karwoski, T. F., Odbert, H. S., & Osgood, C. E. (1942). Studies in synesthetic thinking. II. The rôle of form in visual responses to music. *Journal of General Psychology, 26*(2), 199–222. https://doi.org/10.1080/00221309.1942.10545166

Keller, P. E., & Repp, B. H. (2008). Multilevel coordination stability: Integrated goal representations in simultaneous intra-personal and inter-agent coordination. *Acta Psychologica, 128*(2), 378–386. https://doi.org/10.1016/j.actpsy.2008.03.012

Kendall, R. A. (2010). Music in film and animation: Experimental semiotics applied to visual, sound and musical structures. In B. E. Rogowitz & T. N. Pappas (Eds.), *Proceedings of SPIE, 7525, sponsored by IS & T Electronic Imaging and SPIE: Human Vision and Electronic Imaging XV*. San Jose, CA: 1-13. https://doi.org/10.1117/12.849097.

Kim, K.-H., & Iwamiya, S.-I. (2008). Formal congruency between Telop patterns and sounds effects. *Music Perception, 25*(5), 429–448. https://doi.org/10.1525/mp.2008.25.5.429

Klapetek, A., Ngo, M. K., & Spence, C. (2012). Does crossmodal correspondence modulate the facilitatory effect of auditory cues on visual search? *Attention, Perception & Psychophysics, 74*, 1154–1167. https://doi.org/10.3758/s13414-012-0317-9

Knief, A., Schulte, M., Bertrand, O., & Pantev, C. (2000). The perception of coherent and non-coherent auditory objects: A signature in gamma frequency band. *Hearing Research, 145*(1–2), 161–168. https://doi.org/10.1016/S0378-5955(00)00091-5

Koelsch, S., & Siebel, W. A. (2005). Towards a neural basis of music perception. *Trends in Cognitive Sciences, 9*(12), 578–584. https://doi.org/10.1016/j.tics.2005.10.001

Kubovy, M. (1988). Should we resist the seductiveness of the space: Time: : Vision: Audition analogy? *Journal of Experimental Psychology. Human Perception and Performance, 14*(2), 318–320. https://doi.org/10.1037/0096-1523.14.2.318

Kubovy, M., & Yu, M. (2012). Multistability, cross-modal binding and the additivity of conjoint grouping principles. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 367*(1591), 954–964. https://doi.org/10.1098/rstb.2011.0365

Kulezic-Wilson, D. (2004). The musicality of film rhythm. In K. Rockett & J. Hill (Eds.), *National cinema and beyond: Studies in Irish film I* (pp. 115–119). Four Courts Press.

Lin, C., Yeh, M., & Shams, L. (2022). Subliminal audio visual temporal congruency in music videos enhances perceptual pleasure. *Neuroscience Letters, 779*, Article 136623. https://doi.org/10.1016/j.neulet.2022.136623

Lipscomb, S. D. (2005). The perception of audio-visual composites: Accent structure alignment of simple stimuli. *Selected Reports in Ethnomusicology, 12*, 37–67.

Lipscomb, S. D. (2013). Cross-modal alignment of accent structures in multimedia. The psychology of music in multimedia. In S. Tan, A. J. Cohen, S. D. Lipscomb, & R. A. Kendall (Eds.), *The psychology of music in multimedia* (pp. 192–215). Oxford University Press.

Lipscomb, S. D., & Kim, E. M. (2004). Perceived match between visual parameters and auditory correlates: An experimental multimedia investigation. *Proceedings of the 8th International Conference on Music Perception and Cognition*, 72-75.

Liu, A. (1976). Cross-modality set effect on the perception of ambiguous pictures. *Bulletin of the Psychonomic Society, 7*(3), 331–333. https://doi.org/10.3758/BF03337206

Liu, C. H., & Kennedy, J. M. (1997). Form symbolism, analogy, and metaphor. *Psychonomic Bulletin & Review, 4*, 546–551. https://doi.org/10.3758/BF03214347

Macaluso, E., George, N., Dolan, R., Spence, C., & Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech perception: A PET study. *NeuroImage, 21*(2), 725–732. https://doi.org/10.1016/j.neuroimage.2003.09.049

Marks, L. E. (1987). On cross-modal similarity: Perceiving temporal patterns by hearing, touch, and vision. *Perception & Psychophysics, 42*(3), 250–256. https://doi.org/10.3758/BF03203076

Marks, L. E. (2004). Cross-modal interactions in speeded classification. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of multisensory processes* (pp. 85–105). MIT Press.

Marshall, S. K., & Cohen, A. J. (1988). Effects of musical soundtracks on attitudes toward animated geometric figures. *Music Perception, 6*(1), 95–112. https://doi.org/10.2307/40285417

McDonald, J., Canazza, S., Chmiel, A., De Poli, G., Houbert, E., Murari, M., Rodà, A., Schubert, E., & Zhang, J. D. (2022). Illuminating music: Impact of color hue for background lighting on emotional arousal in piano performance videos. *Frontiers in Psychology, 13*, Article 828699. https://doi.org/10.3389/fpsyg.2022.828699

Millet, B., Chattah, J., & Ahn, S. (2021). Soundtrack design: The impact of music on visual attention and affective responses. *Applied Ergonomics, 93*, Article 103301. https://doi.org/10.1016/j.apergo.2020.103301

Motoki, K., Marks, L. E., & Velasco, C. (2023). Reflections on cross-modal correspondences: Current understanding and issues for future research. *Multisensory Research, 37*(1), 1–23. https://doi.org/10.1163/22134808-bja10114

Müller, J. P. (2010). Synchronization as a sound-image relationship. In D. Daniels, S. Naumann, & J. Thoben (Eds.), *See this sound: Audiovisuology. An interdisciplinary survey of audiovisual culture* (pp. 401–413). Verlag der Buchhandlung Walter König.

Müller, M. M., Junghöfer, M., Elbert, T., & Rochstroh, B. (1997). Visually induced gamma-band responses to coherent and incoherent motion: A replication study. *Neuroreport, 8*(11), 2575–2579. https://doi.org/10.1097/00001756-199707280-00031

Munárriz Ortiz, J. (2017). Sonic landscapes, visual environments. Interaction and synchronicity in compositions and live performance. In *Synchresis audio vision tales* (pp. 1-16).

Nahrstedt, K. (2024). Impact of Steinmetz' synchronization work on multimedia community. In S. Schulte & B. Koldehofe (Eds.), *From multimedia communications to the future internet. Lecture Notes in Computer Science*, 15200, 20-30. Springer, Cham. https://doi.org/10.1007/978-3-031-71874-8_2.

Olivers, C. N. L., & Van der Burg, E. (2008). Bleeping you out of the blink: Sound saves vision from oblivion. *Brain Research, 1242*, 191–199. https://doi.org/10.1016/j.brainres.2008.01.070

Omwake, L. (1940). Visual responses to auditory stimuli. *Journal of Applied Psychology, 24*(4), 468–481. https://doi.org/10.1037/h0057603

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. University of Illinois Press.

Palmer, S. E., Schloss, K. B., Xu, Z., & Prado-León, L. R. (2013). Music-color associations are mediated by emotion. *Proceedings of the National Academy of Sciences, 110*(22), 8836–8841. https://doi.org/10.1073/pnas.1212562110

Parise, C. V., Harrar, V., Ernst, M. O., & Spence, C. (2013). Cross-correlation between auditory and visual signals promotes multisensory integration. *Multisensory Research, 26*(3), 307–316. https://doi.org/10.1163/22134808-00002417

Parise, C. V., & Spence, C. (2012). Audiovisual crossmodal correspondences and sound symbolism: A study using the implicit association test. *Experimental Brain Research, 220*, 319–333. https://doi.org/10.1007/s00221-012-3140-6

Parise, C. V., Spence, C., & Ernst, M. (2012). When correlation implies causation in multisensory integration. *Current Biology, 22*, 46–49. https://doi.org/10.1016/j.cub.2011.11.039

Posner, M. I., Nissen, M. J., & Klein, R. M. (1976). Visual dominance: An information-processing account of its origins and significance. *Psychological Review, 83*(2), 157–171. https://doi.org/10.1037/0033-295X.83.2.157

Pratt, C. C. (1969). Wolfgang Köhler 1887–1967. In W. Köhler (Ed.), *The task of Gestalt Psychology* (pp. 3–29). Princeton University Press.

Raij, T., Ahveninen, J., Lin, F. H., Witzel, T., Jääskeläinen, I. P., Letham, B., Israeli, E., Sahyoun, C., Vasios, C., Stufflebeam, S., Hämäläinen, M., & Belliveau, J. W. (2010). Onset timing of cross-sensory activations and multisensory interactions in auditory and visual sensory cortices. *European Journal of Neuroscience, 31*(10), 1772–1782. https://doi.org/10.1111/j.1460-9568.2010.07213.x

Ramachandran, V. S., Marcus, Z., & Chunharas, C. (2020). Bouba-Kiki: Cross-domain resonance and the origins of synesthesia, metaphor, and words in the human mind. In K. Sathian & V. S. Ramachandran (Eds.), *Multisensory perception* (pp. 3–40). Academic Press. https://doi.org/10.1016/B978-0-12-812492-5.00001-2

Ravignani, A., & Sonnweber, R. (2017). Chimpanzees process structural isomorphisms across sensory modalities. *Cognition, 161*, 74–79. https://doi.org/10.1016/j.cognition.2017.01.005

Riggs, M. G. (1964). The mood effect of music: A comparison of data from four investigators. *Journal of Psychology, 58*(2), 427–438. https://doi.org/10.2190/8Y6G-KTM8-VDX4-UHRW

Ryan, T. A. (1940). Interrelations of the sensory systems in perception. *Psychological Bulletin, 37*(9), 659–698. https://doi.org/10.1037/h0060252

Schöffer, N. (1985). Sonic and visual structures. *Leonardo, 18*(2), 59–68. https://doi.org/10.2307/1577872

Shams, L., Kamitani, Y., & Shimojo, S. (2000). What you see is what you hear: Sound induced visual flashing. *Nature, 408*, 788. https://doi.org/10.1038/35048669

Sirius, G., & Clarke, E. F. (1994). The perception of audiovisual relationships: A preliminary study. *Psychomusicology: A Journal of Research in Music Cognition, 13*(1–2), 119–132. https://doi.org/10.1037/h0094099

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics, 73*(4), 971–995. https://doi.org/10.3758/s13414-010-0073-7

Spence, C. (2015). Cross-modal perceptual organization. In J. Wagemans (Ed.), *The Oxford handbook of perceptual organization* (pp. 649–664). Oxford University Press.

Spence, C. (2018a). Crossmodal correspondences: A synopsis. In D. Howes (Ed.), *Senses and sensation: Critical and primary sources* (Vol. Vol. III, pp. 91–125). Bloomsbury Academic.

Spence, C. (2018b). Multisensory perception. In J. Wixted (Ed.-in-Chief), J. Serences (Vol. Ed.), *The Stevens' handbook of experimental psychology and cognitive neuroscience* (4th Ed, Vol. 2, pp. 1-56). Hoboken, NJ: John Wiley & Sons.

Spence, C. (2020). Assessing the role of emotional mediation in explaining crossmodal correspondences involving musical stimuli. *Multisensory Research, 33*(1), 1–29. https://doi.org/10.1163/22134808-20191469

Spence, C. (2023). Explaining visual shape-taste crossmodal correspondences. *Multisensory Research, 36*(4), 313–345. https://doi.org/10.1163/22134808-bja10096

Spence, C. (2025). Reflecting on *the merging of the senses*: A cognitive psychology perspective. *Multisensory Research, 38*(4–5), 231–253. https://doi.org/10.1163/22134808-bja10139

Spence, C., & Di Stefano, N. (2024). Sensory translation between audition and vision. *Psychonomic Bulletin & Review, 31*, 599–626. https://doi.org/10.3758/s13423-023-02343-w

Spence, C., & Di Stefano, N. (2025a). Gestalt perceptual grouping and crossmodal art. In W. Coppola (Ed.), *Handbook of Gestalt-theoretic psychology of art* (pp. 202–230). Routledge. https://doi.org/10.4324/9781032694467-11

Spence, C., & Di Stefano, N. (2025b). Augmenting art crossmodally: Possibilities and pitfalls. *Frontiers in Psychology, 16*, Article 1605110. https://doi.org/10.3389/fpsyg.2025.1605110

Spence, C., & Di Stefano, N. (2025c). Mood music: Studying the impact of background music on film. *Multisensory Research, 39*(1), 1–45. https://doi.org/10.1163/22134808-bja10172

Spence, C., Sanabria, D., & Soto-Faraco, S. (2007). Intersensory Gestalten and crossmodal scene perception. In K. Noguchi (Ed.), *Psychology of beauty and Kansei: New horizons of Gestalt perception* (pp. 519–579). Fuzanbo International.

Staal, H. E., & Donderi, D. C. (1983). The effect of sound on visual apparent movement. *American Journal of Psychology, 96*(1), 95–105. https://doi.org/10.2307/1422212

Steinmetz, R. (1996). Human perception of jitter and media synchronization. *IEEE Journal on Selected Areas in Communications, 14*(1), 61–72.

Stevens, S. S. (1958). Some similarities between hearing and seeing. *Laryngoscope, 68*(3), 508–527. https://doi.org/10.1002/lary.5540680332

Strachan, R. (2006). Music video and genre. In S. Brown & U. Volgsten (Eds.), *Music and manipulation: On the social uses and social control of music* (pp. 187–206). Berghahn Books.

Su, Y. H. (2014). Content congruency and its interplay with temporal synchrony modulate integration between rhythmic audiovisual streams. *Frontiers in Integrative Neuroscience, 8*(8), Article 92. https://doi.org/10.3389/fnint.2014.00092

Thaiwong, T., & Fukumoto, M. (2024). The effects of selected preferred music on perceived emotions through audiovisual stimuli. Paper presented at *2024 IEEE/ACIS 9th International Conference on Big Data, Cloud Computing, and Data Science (BCD)* (pp. 175-180). https://doi.org/10.1109/BCD61269.2024.10743095.

Thaiwong, T., & Fukumoto, M. (2025). A study on the effects of combining music and animation on emotional induction. *International Journal of Affective Engineering, 24*(2), 181–191. https://doi.org/10.5057/ijae.IJAE-D-24-00007

Thornley Head, P. D. (2006). Synaesthesia: Pitch-colour isomorphism in RGB-space? *Cortex, 42*(2), 164–174. https://doi.org/10.1016/s0010-9452(08)70341-1

Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology, 14*(1), 107–141. https://doi.org/10.1016/0010-0285(82)90006-8

Toccafondi, F. (2025). History of intersensoriality and art. In W. Coppola (Ed.), *Handbook of Gestalt-theoretic psychology of art* (pp. 41–55). Routledge. https://doi.org/10.4324/9781032694467-2

Van der Burg, E., Cass, J., Olivers, C. N. L., Theeuwes, J., & Alais, D. (2010). Efficient visual search from synchronized auditory signals requires transient audiovisual events. *PLoS ONE, 5*, Article e10664. https://doi.org/10.1371/journal.pone.001066

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and pop: Non-spatial auditory signals improve spatial visual search. *Journal of Experimental Psychology. Human Perception and Performance, 34*(5), 1053–1065. https://doi.org/10.1037/0096-1523.34.5.1053

Vanel, H. (2009). Visual muzak and the regulation of the senses: Notes on Nicolas Schöffer. In C. Lund & H. Lund (Eds.), *Audio • visual - On visual music and related media* (pp. 58–75). Arnoldsche Verlagsanstalt.

Vatakis, A., & Spence, C. (2010). Audiovisual temporal integration for complex speech, object-action, animal call, and musical stimuli. In M. J. Naumer & J. Kaiser (Eds.), *Multisensory object perception in the primate brain* (pp. 95–121). Springer. https://doi.org/10.1007/978-1-4419-5615-6_7

Vines, B. W., Krumhansl, C. L., Wanderley, M. M., Dalca, I. M., & Levitin, D. J. (2011). Music to my eyes: Cross-modal interactions in the perception of emotions in musical performance. *Cognition, 118*(2), 157–170. https://doi.org/10.1016/j.cognition.2010.11.010

Vroomen, J., & Keetels, M. (2010). Perception of intersensory synchrony: A tutorial review. *Attention, Perception & Psychophysics, 72*, 871–884. https://doi.org/10.3758/APP.72.4.871

Wagner, S., Winner, E., Cicchetti, D., & Gardner, H. (1981). "Metaphorical" mapping in human infants. *Child Development, 52*(2), 728–731. https://doi.org/10.2307/1129200

Wakiyama, T., Tsubaki, Y., Kuno-Mizumura, M., & Sakaguchi, Y. (2025). Temporal relationship between dancer's body movements and music beats in classical ballet. *Scientific Reports, 15*, Article 29976. https://doi.org/10.1038/s41598-025-15571-y

Wanke, R. (2021). *Sound in the ecstatic-materialist perspective on experimental music*. Routledge, Taylor & Francis.

Wanke, R. (2023). Listening to contemporary art music: A morphodynamic model of cognition. *Journal of Cognition, 6*(1), Article 32. https://doi.org/10.5334/joc.280

Wanke, R., Ansani, A., Di Stefano, N., & Spence, C. (2025). Exploring auditory morphodynamics: Audio-visual associations in sound-based music. *I-Perception, 16*(0), 1–21. https://doi.org/10.1177/20416695251338718

Welch, R. B., DuttonHurt, L. D., & Warren, D. H. (1986). Contributions of audition and vision to temporal rate perception. *Perception & Psychophysics, 39*(4), 294–300. https://doi.org/10.3758/BF03204939

Wells, A. (1980). Music and visual color: A proposed correlation. *Leonardo, 13*, 101–107. https://doi.org/10.2307/1577978

Zika, F. (2013). Color and sound: Transcending the limits of the senses. In M. Blassnigg, G. Deutsch, & H. Schimek (Eds.), *Light, image, imagination* (pp. 29–46). Amsterdam University Press. https://doi.org/10.2307/j.ctt45kdxn.6