Perceiving Tempo in Incongruent Audiovisual Presentations of Human Motion: Evidence for a Visual Driving Effect

Xinyue Wang1,*, Clemens Wöllner1,*** and Zhuanghua Shi2,****
1Institute for Systematic Musicology, Universität Hamburg, Hamburg, Germany
2Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany

Received 31 August 2020; accepted 30 March 2021

Abstract
Compared to vision, audition has been considered to be the dominant sensory modality for tempo-ral processing. Nevertheless, recent research suggests the opposite, such that the apparent inferiority of visual information in tempo judgements might be due to the lack of ecological validity of experimental stimuli, and reliable visual movements may have the potential to alter the temporal location of perceived auditory inputs. To explore the role of audition and vision in overall time perception, audiovisual stimuli with various degrees of temporal congruence were developed in the current study. We investigated which sensory modality weighs more in holistic tempo judgements with conflicting audiovisual information, and whether biological motion (point-light displays of dancers) rather than auditory cues (rhythmic beats) dominate judgements of tempo. A bisection experiment found that participants relied more on visual tempo compared to auditory tempo in overall tempo judgements. For fast tempi (150 to 180 BPM), participants judged ‘fast’ significantly more often with visual cues regardless of the auditory tempo, whereas for slow tempi (60 to 90 BPM), they did so significantly less often. Our results support the notion that visual stimuli with higher ecological validity have the potential to drive up or down the holistic perception of tempo.

Keywords
Tempo judgement, audiovisual timing, visual driving, bisection paradigm
1. Introduction

Perceiving inconsistent audiovisual information is common in daily life. In many cases, conflicting inputs of one modality are able to alter the percept of another. The McGurk effect (McGurk & MacDonald, 1976), for example, is a famous example that lip movements not corresponding to the speech alternate the perceived sounds. Different pianists’ performances coupled with the same soundtrack have been perceived to be different in a number of dimensions (Behne & Wöllner, 2011). It is of interests whether similar observations can be extended to the perception of timing and tempo. The question of how audiovisual asynchrony affects temporal processing has also attracted much attention. The dominant role of audition has been long recognised in temporal processing in the sense that it provides higher accuracy and precision than vision (Grondin, 2010). However, this view is challenged by emerging evidence of the superiority of meaningful visual movements in time perception with both abstract (Grahn, 2012) and real-life stimuli (Hove & Keller, 2010; London et al., 2016). Thus, it remains controversial whether audition or vision dominates our perception of tempo.

There has been a long debate in research about whether timing is based on a central or on a distributed system (Occelli et al., 2011; Penney, 2003; Van Wassenhove et al., 2008; for an overview, see Wang & Wöllner, 2019). Some argue that the timing mechanism is distributed, which can explain the discrepancies in timing performance between, for instance, vision and audition (Grondin et al., 2008); while others favour the notion that the modality difference in time perception comes from the interaction between different sensory modalities in the central timing (e.g., Levitan et al., 2015). The distributed account is supported by findings that audition has an advantage over vision and other sensory modalities in terms of duration discrimination (Grondin et al., 2008), reproduction (Gamache & Grondin, 2010), and estimation (Kanai et al., 2011). However, it should be noted that the domination of audition in temporal processing does not apply to all cases, particularly with biological trajectories (Hove et al., 2010) or movements (Allingham et al., 2020).

Evidence of temporal entrainment, which specifies the synchronisation between two rhythms, has been observed with musical (Hammerschmidt & Wöllner, 2020), visual (Iversen et al., 2015) as well as tactile stimuli (Occelli et al., 2011). For tempo judgements, the temporal ventriloquism effect (Burr et al., 2009) and the auditory driving effect (Shipley, 1964) have both suggested the dominance of audition over visual displays by ‘dragging’ the temporal location of the latter to that of the former. Even when the auditory stimuli were not attended to, or reduced in salience, duration judgements clearly leaned towards that of the perceived tone rather than of the visual circle in this case, suggesting that the processing of auditory temporal cues was possibly autonomous and occupied minimal cognitive resources (Ortega et al., 2014). Alternatively, the reliability
of the sensory inputs was crucial when perceiving durations: whichever channel provided the least noise was assigned the most weight in duration estimation (Hartcher-O’Brien et al., 2014; Shi et al., 2010). It is therefore likely that auditory temporal inputs were more reliable in studies where audition outweighed vision, given that multisensory temporal information is integrated in an optimal fashion (Shi et al., 2013).

Relatively few studies have explored tempo judgement in the context of audiovisual stimuli with high ecological validity. Among those who adopted naturalistic visual stimuli, a strong influence of the visual over the auditory inputs has been found. For example, videos of musicians playing long notes on the marimba, coupled with long and short corresponding sounds, shifted the perceived note length towards ‘long’ (Schutz & Lipscomb, 2007). The effect of human point-light displays on unimodal (auditory or visual) and bimodal (audiovisual) stimuli of varying musical tempi indicated that movements of high energy led to faster perceived tempo of auditory stimuli (London et al., 2016). However, it remains unclear how tempo is holistically perceived when participants are asked to judge it based on both sensory information channels. Furthermore, to our knowledge, no study investigated a combination of audition and vision in a controlled variation of audiovisually inconsistent stimuli.

Evidence has suggested that vision might not always be less accurate than audition in duration and tempo judgements. Past research believes that vision dominates spatial rather than temporal localisation (e.g., Burr et al., 2009; Repp, 2003), and has a lower temporal sensitivity than audition in low-level information processing (Ortega et al., 2014). However, it should be noted that the evidence for high precision of audition comes from simple and controlled laboratory stimuli, and the arrangements of their presence in the task, such as simple visual flickers (Shipley, 1964; Treisman et al., 1990), coloured squares (Grahn et al., 2011), looming or receding discs or dots (Van Wassenhove et al., 2008) rather than naturalistic stimuli. Naturalistic stimuli, such as biological motion derived from human behaviours and social activities (Boltz, 2005), often yield movements with high compatibility (Hove et al., 2010). The point-light technique for human motion has first been used in experimental research by Johansson (1973). Point-light displays (PLD) of biological motion are derived from natural human (or animal) movements such that visual details including facial expressions or clothes are not shown, yet the naturalness in terms of movement kinematics is preserved. Abstract visual stimuli such as flashes, on the other hand, provide fewer cues, no biological movement trajectories, and thus a less rich temporal structure. An earlier attempt by Boltz (2005) compared the effects of naturalistic scenes when they were presented in either auditory, visual, or audiovisual channels. Results suggested that modality differences were not observed in terms of duration reproduction or estimation accuracy. This led to a string of studies that adopted the naturalistic approach, in other words using visual movements that were common
in everyday life to achieve the same salience and temporal discriminability as of the auditory stimuli presented in past research (e.g., Grahn, 2012; Hove et al., 2010; Hove et al., 2013; London et al., 2016). Extraction of rhythmic patterns from visual movements was not only viable, but also independent from auditory interferences, indicating a robust temporal representation in the visual module (Su & Salazar-López, 2016).

The above evidence calls for further exploration in multisensory timing. The current study intends to advance research as follows: (1) Provide a scenario where auditory and visual information is equally important to tempo judgements. (2) Explore the potential interaction of audition and vision in tempo judgement. In light of this, the current study examined the effects of competing auditory and visual information on tempo judgement with biological motion and drumbeats, taking into account the ecological validity of both. We hypothesised that, with meaningful visual information such as point-light displays (PLDs), the visual tempo relative to auditory tempo would contribute more to the overall tempo judgement. A given unit change in visual tempo should thus lead to larger changes in the tempo judgement ratio (fast/slow). Accordingly, more participants should rely on visual rather than auditory information when judging tempo. Finally, we hypothesised that perceived naturalness would decrease as the audiovisual tempo discrepancy increases.

2. Methods

2.1. Participants

Twenty-four participants were recruited for the study (12 female; aged \( M = 24.21 \) years, SD = 4.68). Participants had a mean of 10.04 years (SD = 7.09) of regular practice with musical instruments (including voice), and a mean of 7.65 years (SD = 7.18) of lessons on their instrument. Thus, the current sample represents a population that has moderate to advanced musical training. The sample size had been calculated a priori for a \( 3 \times 3 \) design (\( \alpha = 0.05, \text{Cohen's} \ f = 0.25, \text{power} = 0.8 \)), requiring at least 15 participants (using G*Power; Faul et al., 2009). For a conservative approach, we recruited 24 participants. We also followed the guidelines of the Ethics Committee of the Faculty of Humanities, Universität Hamburg, and each participant was compensated 10 Euro for taking part.

2.2. Material

Participants were presented with audiovisual stimuli synthesised from isochronous drumbeats of nine tempi (60, 75, 90, 105, 120, 135, 150, 165, 180 beats per minute [BPM]), and visuals of the same tempo spectrum. The visual stimulus showed a PLD of a person jumping from left to right with the hands moving up and down (Fig. 1). This movement pattern was recorded with an eleven-camera motion-capture system (Qualisys Oqus, Qualisys AB, Göteborg, Sweden) at a
framerate of 200 frames per second. Thirty-one markers were attached to the performer. The movement pattern was intended to be neither towards an action-based nor to a habitual (highly automatised) outcome, in order to avoid familiarity with the movement (Calvo-Merino et al., 2006). The movement was originally recorded at a speed of 120 BPM. The motion was presented from a 30-degree angle, in frontal view. The MATLAB Motion Capture (MoCap) Toolbox (Burger & Toiviainen, 2013) was adopted to speed up and slow down the PLD to the eight further tempi as specified above, while ensuring that the visual resolution and the number of data points per second were unchanged. The auditory stimuli of nine tempi (60 to 180 BPM, 15 BPM per step), on the other hand, were directly synthesized from a bass drum on the online drumbeat generator Drumbit (https://drumbit.app). Drum beats can be found in real life scenarios such as listening to techno music, thus providing higher naturalness than abstract auditory stimuli adopted in past studies such as sine waves. The PLDs and the drumbeat soundtracks of all tempi were then combined in Adobe Premiere Pro CC 2017 (Adobe Systems, San Jose, CA, USA) to create a total of 81 stimuli with all audiovisual tempo combinations. That is, all stimuli were bimodal videos varying in tempi.

The experiment was conducted in the SloMo laboratory at Universität Hamburg on a Dell U2414Hb monitor (Dell Technologies Inc., Round Rock, TX, USA), controlled by the software OpenSesame (Mathôt et al., 2012). A Sennheiser HD600 headphone set (Sennheiser GmbH, Hanover, Germany) was provided for the soundtrack. Participants responded to the experimental task by pressing the leftward or rightward button on the keyboard.

2.3. Design and Procedure

The current study introduced a $3 \times 3$ design where three auditory tempi ranges (slow: 60, 75, 90 BPM; medium: 105, 120, 135 BPM; fast: 150, 165, 180 BPM) and
three visual tempi of the same spectrum acted as the independent variables, while taking the temporal bisection, a two-alternative forced-choice task (2AFC), to examine the ratio of ‘fast’ judgement at different tempo and modality conditions. The 2AFC method has been used in various studies of audiovisual integration (Chen et al., 2018; Gori et al., 2012; Shi et al., 2010). The bisection task has been widely used in cue combination research for both spatial stimuli and time in audiovisual integration processes (Gori et al., 2012). In an audiovisual Ternus apparent motion study, Shi et al. (2010) used the bisection task to measure the audiovisual duration integration. Roach et al. (2006) adopted the 2AFC temporal discrimination (higher or lower than 10 Hz) to estimate the threshold for the audiovisual temporal integration. Similar to other direct measures of duration or tempi, such as reproduction tasks, the bisection task is able to probe the audiovisual temporal integration as well as decisions in the tempo judgements. One benefit of using the bisection task, compared to the direct tempo reproduction or other motor-related tasks, is that the task is not influenced by motor noise. In a similar manner, here we applied the temporal bisection point to measure the holistic tempo judgements, that is whether observers shifted their judgements towards a fast or a slow tempo.

In the current study, participants were first presented with two audiovisual anchors (a fast tempo and a slow tempo) at the beginning of the experiment and were asked to judge the tempo of a given stimulus as close to the fast or the slow tempo holistically. In other words, they should focus on both auditory and visual information in the video stimuli. The slow anchor was a bimodal video with an audiovisually consistent tempo at 60 BPM, and a fast anchor at 180 BPM. They were then shown randomised trials of 81 bimodal videos generated from nine auditory stimuli (60, 75, 90, 105, 120, 135, 150, 165, 180 BPM) and nine visual stimuli of the same tempo spectrum, repeated three times. Each auditory tempo was combined with each visual tempo. The bimodal stimuli include both tempo-consistent and -inconsistent presentations. A total of 243 trials were presented to each participant.

Participants were seated in a quiet room approximately 80 cm in front of the monitor. Instructions were given by an experimenter who was trained to follow fixed protocols to ensure a standardised procedure. Each trial started with a fixation point for 100ms, followed by a PLD presentation of 5 s while drum sounds were simultaneously played through the headphones. After the presentation, a ‘?’ was shown in the centre of the screen, prompting participants to judge if the tempo of the presented stimulus was closer to the slow or the fast anchor tempo. They were asked to press the leftward arrow key for the slow and the rightward arrow key for the fast anchor. To refresh participants’ memory of the two anchors, a text reminder ‘anchors’ popped up after every nine trials, and the anchors were played once each time. Participants pressed any key to proceed and to watch the
fast and slow videos, with no time limit imposed. They were not required to make any response. In addition, an optional short break was offered every 40 or 41 trials. After completing the bisection task, participants were asked to rate the naturalness (‘how natural does this video feel?’) of all 81 conditions in randomised order. A trial started with a fixation point for 100 ms, followed by a 5-s bimodal video stimulus. A visual instruction of the naturalness question ‘Please rate how natural the video feels’ was presented. On a horizontal gauge bar from 1 (marked as ‘least natural’) to 100 (marked as ‘most natural’), participants placed the cursor in a relative position to give a response.

2.4. Data Analyses

All statistical analyses were conducted with R (Version 3.5.3; R Core Team, 2019). Given that the distributions of individual participants’ tempo judgements were heavily skewed, we used nonparametric analyses, more specifically a series of chi-square tests, to compare differences between the numbers of ‘fast’ versus ‘slow’ judgements for auditory or visual tempo conditions. In addition, we fitted the response (‘fast’ versus ‘slow’) as a logistic function of the auditory and visual tempi to obtain a 2D psychometric function, such that we can obtain the points of subjective equality (PSEs). A separation boundary (auditory and visual tempi) by the PSE, yielded by the logistic model when the likelihood of ‘fast’ judgement was 0.5, was then estimated on the individual and group levels. Pearson’s correlations were conducted to explore the relationship between perceived naturalness and audiovisual discrepancy.

3. Results

3.1. Visual Versus Auditory Tempo

We first examined the effect of modality on tempo judgement at each tempo condition. First, we grouped the tempo of the presentation either by (a) tempo of drumbeats or (b) tempo of visual PLD, by slow (60, 75, and 90 BPM), medium (105, 120, and 135 BPM), and fast (150, 165, and 180 BPM) tempo ranges. Figure 2 shows the mean proportion of ‘fast’ responses as a function of tempo, separated by modality. A chi-square test shows that participants were more likely to judge ‘fast’ with visual rather than auditory cues, $\chi^2 (1, N = 3878) = 81.96, p < 0.001$, with a small to medium effect size ($\phi = 0.15$). Correspondingly, for slow stimuli the proportion of ‘fast’ responses was significantly higher when participants relied on auditory cues than visual ones, $\chi^2 (1, N = 3907) = 59.16, p < 0.001$. The effect size was also small to medium ($\phi = 0.12$), according to Cohen (1988). There was no significant difference between modalities when the stimuli were presented at intermediate tempo, $\chi^2 (1, N = 3879) = 0.07, p = 0.80, \phi = 0.004$. These results suggest that visual information plays a more important role than the auditory
information in tempo judgement at both ends of the tempo spectrum: When the PLDs were shown at a fast (150, 165, 180 BPM) or a slow (60, 75, 90 BPM) tempo, participants judged stimuli overall to be fast or slow, regardless of the auditory tempo.

To get a detailed picture of individual contributions of auditory and visual tempo in temporal judgements, we plotted the average response heatmap in Fig. 3. Both auditory and visual tempo contribute to judgements. In general, the faster the tempo, the more likely a participant would judge ‘fast’. Consistent with the analysis shown above, the change of response is more sensitive in the ‘vision’ direction than in the ‘audition’ direction, as evinced by the response contour changes along the visual rather than the auditory modality. To further quantify this, we applied a two-dimensional logistic regression, which is an extension of the one-dimensional psychometric function. We assume participants’ bisection response proportion $p$ depends on both auditory and visual tempi $(T_A, T_V)$ with a logistic relation:

$$\log \frac{p}{1-p} = \alpha + \beta_A T_A + \beta_V T_V,$$

where $\alpha, \beta_A, \beta_V$ are coefficients of the logistic function. When $p = 0.5$, the above equation indicates the boundary separation between the fast and slow responses (see the red curve in Fig. 3).
Figure 3. Heat map of the proportion of 'fast' responses distributed over the audio and video tempo spectrum. Yellow stands for high proportions, and blue for low proportions. Note that the red line marks the corresponding audio and video conditions at which the average ratio of fast response versus all responses equals 0.5, or the line of subjective equality.
The tempi in both conditions were first standardised by dividing each value by the median (120 BPM). The logistic model suggested a significant relationship among tempo judgement and the auditory as well as visual tempo, \( \chi^2(5827) = 2757.21, p < 0.001 \). McFadden's \( R^2 = 0.34 \), which fell between 0.2 and 0.4, indicating a good fit. The estimated coefficients are shown in Table 1. The coefficients for the visual and auditory tempi were \( \beta_V = 5.04 \) and \( \beta_A = 3.52 \), respectively. These reflect the degree of sensitivity in change of responses (according to the model, a unit change in the relative tempo contributes a change of log likelihood of two responses). Furthermore, \( \beta_V \) was significantly larger than \( \beta_A \) (based on the non-overlapping 95% confidence interval, see Table 1), which confirms that in general the visual tempo contributed more than the auditory tempo.

3.2. Individual Modality Reliance

Based on the separation boundary, we then categorized participants into one the following types: vision-, audition-, or bimodal-reliant types. We used the log-ratio between the auditory and visual coefficients (\( \log \left| \frac{\beta_A}{\beta_V} \right| \)) as an indicator of modality reliance. A \( \log \left| \frac{\beta_A}{\beta_V} \right| \) of 0 is an ideal case of equal reliance for the visual and auditory modality. Considering random fluctuation, \( \log \left| \frac{\beta_A}{\beta_V} \right| \) between −0.05 and 0.05 was regarded as equal reliance. A ratio higher than 0.05 suggests auditory reliance, while a ratio lower than −0.05 indicates visual reliance. Figure 4 shows examples of participants for the three types of modality reliance.

According to the categorisation, 16 participants were vision-reliant, seven audition-reliant, and one bimodal-reliant. A chi-square test of independence indicated a significant difference among the three groups, \( \chi^2(2, N = 24) = 14.25, p < 0.001 \). That is to say, a larger proportion of the sample favoured visual information when it came to tempo judgement, regardless of the auditory tempo (Fig. 5). This finding again supports our hypothesis that the visual tempo, when presented as natural human movements, has higher priority than auditory tempo.

3.3. Naturalness and Audiovisual Discrepancy

A Pearson’s correlation between the overall naturalness rating, ranging from 0 (least natural) to 100 (very natural), and the absolute values of the audiovisual tempo discrepancy suggested that the smaller the discrepancy between the audio and video tempo, the more natural a stimulus was perceived \( r(81) = -0.56, p < 0.01 \) (see Fig. 6).

A two-way ANOVA was conducted to examine the effect of auditory and visual tempo on perceived naturalness. Again, tempo ranges were categorised into slow (60 to 90 BPM), medium (105 to 135 BPM), and fast (150 to 180 BPM) for the analysis. Simple main effects suggested that fast visual tempo led to significantly
higher naturalness \( (F_{2,1935} = 100.38, p < 0.001) \). A statistically significant interaction between auditory and visual tempo on perceived naturalness was found, \( F_{4,1935} = 37.74, p < 0.001 \). Tukey’s HSD post-hoc tests revealed that, for auditory tempo, no statistically significant differences were observed between different tempi. However, for visual tempo, fast stimuli were associated with higher naturalness ratings than the medium \( (p < 0.001, d = 0.24) \) and the slow ones \( (p < 0.001, d = 0.74) \). The medium-speed visuals were rated more natural than the slow ones \( (p < 0.001, d = 0.50) \), regardless of the auditory tempo.

### 4. Discussion

The present study examined the role of audition and vision in tempo judgements of naturalistic stimuli of biological motion, when the tempi of the two modalities are not consistent. First, the tempo of visual information (here the PLD stimuli) affected overall tempo judgements to a greater extent than that of the auditory information (drumbeats). Secondly, a higher proportion of the participants relied on visual rather than on auditory information for tempo judgement. Different modality weightings exhibited by individual participants again support our hypothesis that visual information, when presented as biological motion PLDs, should possess high ecological validity and consequently serves as the dominant tempo reference. Finally, a larger audiovisual tempo discrepancy led to lower perceived naturalness.

The results are consistent with our main hypothesis in the sense that naturalistic visual input dominated overall tempo judgement. Past studies with visual movements of varying complexities have observed similar effects where ecological validity could be derived from the stimuli. For abstract movements, the ‘visual driving effect’ has been found for both rhythm perception (Su & Jonikaitis, 2011) and duration estimation (Van Wassenhove et al., 2008). Su and Jonikaitis (2011) revealed that changes in the moving speed of dots or in luminance provided
Figure 4. Examples of participants of typical audition-reliant judgement type (a), vision-reliant judgement type (b), and bimodal-reliant judgement type (c). The x-axis stands for auditory tempo, and the y-axis for visual tempo. The heatmap represents the proportion of ‘fast’ judgements, with lighter colour for higher proportion of ‘fast’ judgements. The yellow lines stand for the audiovisual tempi at which participants were equally likely to respond ‘fast’ or ‘slow’.
Figure 5. The distribution of $\log|\frac{\beta_L}{\beta_V}|$ against the three modality types among participants: visual, auditory, or bimodal reliance.

Figure 6. Perceived naturalness and audiovisual tempo discrepancies (absolute values), varying from small to large. Bottom and top of the boxes represent the first and third quartiles, and the centre line the median. The red curve represents the trend for perceived naturalness, as the audiovisual tempo discrepancy increases.
velocity cues that biased the perceived auditory tempo. Other studies (e.g., van Wassenhove et al., 2008) have shown that looming and receding discs alter the perceived duration of pure tones. Those studies have pointed out the importance of vision in audiovisual temporal perception, which was previously believed to be dominated by audition.

Other experiments that have adopted more complex audiovisual stimuli than abstract rhythmic sequences have successfully replicated the visual driving effect too. For example, participants were equally accurate in their discrimination of complex rhythmic patterns for both auditory and visual presentation (Grahn, 2012). Further attempts have taken biological movements into account. When watching vigorous dance movements, the music tempo was perceived as faster compared to the conditions where only music, or music and relaxed dance movements were presented (London et al., 2016). Our study also used point-light dance-like movements, which appeared to entrain the overall perceived tempo toward the visual tempo. This suggests that the preferred modality for tempo judgement may not entirely depend on the precision of the modality (i.e., the modality precision account). Rather, it might depend on how well information of this modality could assist with action prediction (i.e., the modality appropriateness account).

As discussed earlier in the Introduction, the reliability of the sensory modality determines its contribution weight in the overall judgement (Hartcher-O’Brien et al., 2014; Shi et al., 2013). The more reliable the prediction from the signal, the higher the weight that would be assigned to that modality. Hence, our findings may suggest that predictable biological motions relative to drumbeats may offer reliable cues for temporal judgement.

This in particular is in line with human action prediction. Various studies have suggested that visual attention driven by the action goal was accompanied by higher processing efficiency (Decroix & Kalénine, 2019; Loucks & Pechey, 2016). Loucks and Nagel (2018) found more accurate tempo discrimination performance with human actions compared to non-human actions. In this vein, higher temporal sensitivity with goal-directed biological motions may also be reflected in the current study where repetitive dance-like movements became highly predictable, and thus endowed with more weights in tempo judgement. Compared to drumbeats, the biological movements provide more timing information than discreet bursts of sound. In other words, the continuous nature of the visual information may have provided a more reliable source of tempo information. When both modalities possess information of similar continuities, there might not be an advantageous modality in tempo judgement. In this regard, one of the earlier studies by Boltz (2005) found no modality effect on duration reproduction performances with continuous, natural human behaviours including sports or conversations, presented in either the auditory, visual, or audiovisual channel. However, it is as yet inconclusive whether the continuity or the biological plausibility affects the role of vision in temporal processing. To disentangle the effects of the two features,
Hove et al. (2010) examined the efficiency of facilitating audiomotor synchronisation with continuous and/or direction-compatible motions that were either abstract or biological trajectories. Higher synchronisation rates were observed with continuous, direction-compatible, but not necessarily biological motions. It can be speculated that continuous stimuli contribute more temporal information than discontinuous stimuli, regardless of forms (abstract vs biological) and modalities. In the same vein, the continuity of visual and auditory rhythms has direct effects on participants’ timing performances: The sensory modality with continuous inputs was assigned more weight than the discontinuous one (Varlet et al., 2012). Similarly, compared with the discrete bursts of sounds, the continuous biological motions in our study might provide more reliable information in the overall tempo judgements.

It should be noted that the audiovisual source locations might also contribute to the weight in audiovisual judgements. In a study of multisensory simultaneity (Di Luca, Machulla, & Ernst, 2009), it has been shown that in the headphone-based relative to the co-location audiovisual presentation, the auditory estimates are likely to be biased as they are trusted less. However, the contribution of this spatial discrepancy to the visual-dominant temporal judgements, if any, is likely very mild, given that the headphone-based presentation potentially reduced the interference of other external sounds, which would potentially boost the reliability of the auditory modality. In a similar vein, visual and tactile stimuli appearing in different spatial location were associated with less accurate discrimination responses than those in the same location (Spence et al., 2008), indicating that the interference from one modality might pose a threat to the credibility of the other. In the current study, the auditory source that was closer to the participants should have provided a more reliable source of temporal information than the visual displays, yet failed to do so.

The discrepancy between auditory and visual tempi in our study was reflected by the naturalness ratings. The current results suggest that high perceived naturalness could be particularly derived from a fast visual tempo, as well as from high audiovisual temporal congruence. Not surprisingly, the smaller the audiovisual temporal discrepancy, the more natural the stimulus was perceived to be. Stimuli with small or no discrepancies presumably posed the least difficulty in binding multisensory inputs to one (Vatakis & Spence, 2008). Our results align well with past findings in that meaningful visual motion, especially when following an expected movement direction, has a strong impact on timing and, as shown in other research, sensorimotor synchronisation (Hove et al., 2010). This finding was supported by research comparing the effect of biological movements (finger tapping) with abstract visual stimuli (flashes) on timing accuracy, which found higher stability when synchronising with finger movements than with flashes (Hove & Keller, 2010).
There is scarce neurobiological evidence supporting the role of visual input compared to auditory input in temporal processing. Evidence for the latter including beat detection and time estimation, however, can be found mostly in studies using abstract unimodal stimuli (for a review, see Buonomano & Maass, 2009). An fMRI study where participants were asked to discriminate multisensory inputs (visual, auditory, and tactile) revealed that the auditory dorsal pathway was partly specific to beat processing and functioned as a supra-modal network (Araneda et al., 2017). In Kanai and colleagues’ (2011) study, Transcranial Magnetic Stimulation (TMS) disrupted the activities in the auditory cortex and consequently impeded the participants’ performances in a two-alternative-force choice task where two durations, presented either in pure tones or visual flickers, were compared. By contrast, disrupting the activities in the primary visual cortex only affected the performances of duration judgements with visual stimuli, suggesting the dominant role of the auditory cortex in temporal processing. The evidence above, nevertheless, may not generalise to neural mechanisms of timing with naturalistic stimuli. Biological motion carries spatiotemporal information that helps in forming action predictions, along with the timing of the action. The findings with behavioural data in the current study call for future neurobiological research on the visual dominance with meaningful visual stimuli such as continuous movements.

Furthermore, attentional processes might also contribute the ‘visual driving’ effect observed in our study. Visual dominance in spatial attention has been supported by ample studies. In an audiovisual context, the visual modality was associated with faster reaction time and less response errors in modality-switching, spatially-incongruent tasks (Lukas, Philipp, & Koch, 2010) and detected with greater sensitivity (Spence et al., 2012). The Colavita effect, more specifically, referred to a phenomenon where visual stimuli were associated with higher salience than auditory stimuli when both appeared simultaneously (Colavita, 1974). The perception of human biological motions in point-light displays led to even higher visual salience than abstract visual displays (Johansson, 1973). Selective attention oriented towards biological movement, particularly motion with a purpose compared to scrambled motion, has been shown to activate the part of the motor cortex associated with action mirroring (Gao et al., 2014). The ‘imagined’ imitation led to determination of action intention (Knoblich & Sebanz, 2008), in this case prediction of motion trajectories and their spatiotemporal information. The allocation of attention to natural motions (the PLD in the current study) could then explain participants’ reliance for visual tempo. According to the Dynamic Attending Theory (Jones & Boltz, 1989), the pace of the internal clock is subject to the environmental rhythm when the limited attentional resources orient towards the rhythm of exogenous stimuli. Synchronising one’s attentional ‘pulses’ with environmental rhythms is also known as the entrainment effect — in this case, participants were predominantly under the
influence of the visual tempo. The biological motion in the PLD was presumably attended to more often than the drumbeats, therefore dominating the perceiver’s tempo judgement.

Both naturalistic biological motion and spatial attention towards to biological motion contribute to the visual driving effect as observed in the current study. Yet a few questions remain: firstly, by modifying the ecological validity of auditory stimuli (e.g., beats vs more complex music), it is possible to observe changes in overall tempo judgement. Secondly, instead of multisensory gain, it could be investigated whether there is also a multisensory loss, or more specifically, whether the presence of inconsistent multisensory information could impede temporal processing such as timing or duration judgement. Lastly, studies may explore whether other senses such as tactile perception are capable of multisensory integration and what their weights are in the timing process. A few studies have attempted to explore the potential of tactile-assisted metre judgement when auditory or visual stimuli were presented (Araneda et al., 2017; Huang et al., 2012). In a rhythm pattern identification task, for example, congruent vibrations (tactile) and tones (auditory) raised the correct rhythm discrimination rate to 90%, while incongruent inputs resulted in a decline (Huang et al., 2012). Interestingly, the correct rate was significantly higher when the dominant (correct) pattern was presented with sounds than with vibrations, again confirming the dominant role of audition.

There are a number of limitations that should be addressed in future studies. First, a response bias in the decision process, which might be reflected by participants’ preference towards one modality under certain tempo conditions, cannot be fully ruled out. According to the causal inference model (Körding et al., 2007), response bias towards one source of information can be observed when the key feature (tempo) differs between two sources (sensory modalities) to an extreme extent. However, if such a response bias widely existed among the sample population, the stimuli with a large audiovisual gap, regardless of the modality of the fast tempo, should be equally often judged to be ‘fast’, which would be reflected by a (inverted) U-shaped threshold of equality in Fig. 3. Secondly, considering the essential role of naturalistic stimuli, the perceived predictability of the auditory beats and the PLD has not been quantitatively pre-evaluated. For visual stimuli, the predictability should take into account direction compatibility as in Hove and colleagues’ (2010) work. In future studies, a baseline test prior to the experiment collecting familiarity and naturalness ratings, as well as eye fixation concerning the compatibility of motion, should be considered. The imbalance between the two modalities can be minimised by collecting the perceived naturalness of both auditory and visual stimuli respectively from multiple independent raters before the experiment commences. As for auditory stimuli, the predictability is frequently measured by sensorimotor synchronisation accuracy in tapping tasks (e.g., Stupacher et al., 2017). Furthermore, past studies tended to require the participants to judge the temporal information of one modality only (e.g., Klink, et
al., 2011). The advantages and disadvantages of an experimental paradigm that allows participants the liberty to exhibit their modality reliance should be further examined. In addition, the current study did not systematically evaluate differences between auditory and visual attention. Future studies should seek to disentangle the effects of multisensory attention, especially visual attention, from the effect of naturalistic stimuli on temporal judgements. To verify the link between a visual driving effect and direction compatibility, future studies should also consider a control condition in which inverted biological movements are presented.

Taken together, the current study provides evidence for a visual driving effect in multisensory tempo judgements with meaningful movements. On the group level, visual tempo contributed more to the overall tempo judgement than the auditory tempo. On the individual level, in addition, when presented with tempo-inconsistent audiovisual stimuli, more participants relied on the visual tempo to make the overall tempo judgements. The modality reliance provided insights into tempo judgement strategies adopted by different individuals. Future studies should further investigate the apparent dominance of visual information in timing with real-life audiovisual scenes as well as the factors influencing individuals’ modality reliance in temporal judgements.

Acknowledgements

This research was supported by a Consolidator Grant from the European Research Council to the second author. The research is part of the five-year project: “Slow motion: Transformations of musical time in perception and performance” (SloMo; Grant No. 725319). We are grateful to Emma Allingham for helpful comments on a previous version of the manuscript.

References

Allingham, E., Hammerschmidt, D., & Wöllner, C. (2020). Time perception in human movement: Effects of speed and agency on duration estimation, Q. J. Exp. Psychol. 74, 559–572. doi:10.1177/1747021820979518.

Araneda, R., Renier, L., Ebner-Karestinos, D., Dricot, L., & De Volder, A. G. (2017). Hearing, feeling or seeing a beat recruits a supramodal network in the auditory dorsal stream, Eur. J. Neurosci. 45, 1439–1450. doi: 0.1111/ejn.13349.

Behne, K.-E., & Wöllner, C. (2011). Seeing or hearing the pianists? A synopsis of an early audiovisual perception experiment and a replication, Musicae Sci. 15, 324–342. doi:10.1177/1029864911410955.

Boltz, M. G. (2005). Duration judgments of naturalistic events in the auditory and visual modalities, Percept. Psychophys. 67, 1362–1375. doi:10.3758/BF03193641.

Buonomano, D. V., & Maass, W. (2009). State-dependent computations: spatiotemporal processing in cortical networks, Nat. Rev. Neurosci. 10, 113–125. doi:10.1038/nrn2558.
Burr, D., Banks, M. S., & Morrone, M. C. (2009). Auditory dominance over vision in the perception of interval duration, Exp. Brain Res. 198, 49. doi:10.1007/s00221-009-1933-z.

Burger, B., & Toiviainen, P. (2013). MoCap Toolbox — A Matlab toolbox for computational analysis of movement data. In R. Bresin (Ed.), Proceedings of the sound and music computing conference 2013 (pp. 172–178). Berlin, Germany: Logos Verlag.

Calvo-Merino, B., Grèzes, J., Glaser, D. E., Passingham, R. E., & Haggard, P. (2006). Seeing or doing? Influence of Visual and motor familiarity in action observation, Curr. Biol. 16, 1905–1910. doi:10.1016/j.cub.2006.07.065.

Chen, L., Zhou, X., Müller, H. J., & Shi, Z. (2018). What you see depends on what you hear: Temporal averaging and crossmodal integration, J. Exp. Psychol. Gen. 147, 1851–1864. doi:10.1037/xge0000487.

Cohen, J. (1988). Statistical power analysis for the behavioural sciences. Hillside, NJ, USA: Lawrence Earlbaum Associates.

Colavita, F. B. (1974). Human sensory dominance, Percept. Psychophys. 16, 409–412. doi:10.3758/BF03203962.

Decroix, J., & Kalénine, S. (2019). What first drives visual attention during the recognition of object-directed actions? The role of kinematics and goal information, Atten. Percept. Psychophys. 81, 2400–2409. doi:10.3758/s13414-019-01784-7.

Di Luca, M., Machulla, T.-K., & Ernst, M. O. (2009). Recalibration of multisensory simultaneity: Crossmodal transfer coincides with a change in perceptual latency, J. Vis. 9, 7. doi:10.1167/9.12.7.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses, Behav. Res. Methods 41, 1149–1160. doi:10.3758/BRM.41.4.1149.

Gamache, P.-L., & Grondin, S. (2010). Sensory-specific clock components and memory mechanisms: investigation with parallel timing, Eur. J. Neurosci. 31, 1908–1914. doi:10.1111/j.1460-9568.2010.07197.x.

Gao, Z., Lu, X., Shen, M., Shui, R., & Chen, S. (2014). Rehearsing biological motion in working memory: an fMRI study, J. Vis. 14, 1009. doi:10.1167/14.10.1009.

Gori, M., Sandini, G., & Burr, D. (2012). Development of visuo-auditory integration in space and time, Front. Integr. Neurosci. 6, 77. doi:10.3389/fnint.2012.00077.

Grahn, J. A. (2012). See what I hear? Beat perception in auditory and visual rhythms, Exp. Brain Res. 220, 51–61. doi:10.1007/s00221-012-3114-8.

Grahn, J. A., Henry, M. J., & McAuley, J. D. (2011). FMRI investigation of cross-modal interactions in beat perception: Audition primes vision, but not vice versa, NeuroImage, 54, 1231–1243. doi:10.1016/j.neuroimage.2010.09.033.

Grondin, S. (2010). Timing and time perception: A review of recent behavioral and neuroscience findings and theoretical directions, Atten. Percept. Psychophys. 72, 561–582. doi:10.3758/APP.72.3.561.

Grondin, S., Gamache, P.-L., Tobin, S., Bisson, N., & Hawke, L. (2008). Categorization of brief temporal intervals: An auditory processing context may impair visual performances, Acoust. Sci. Technol. 29, 338–340. doi:10.1250/ast.29.338.

Hammerschmidt, D., & Wöllner, C. (2020). Sensorimotor synchronization with higher metrical levels in music shortens perceived time, Music Percept. 37, 263–277. doi:10.1525/mp.2020.37.4.263.
Hartcher-O’Brien, J., Luca, M. Di, & Ernst, M. O. (2014). The duration of uncertain times: audiovisual information about intervals is integrated in a statistically optimal fashion, *PLoS ONE* 9, e89339. doi:10.1371/journal.pone.0089339.

Hove, M. J., & Keller, P. E. (2010). Spatiotemporal relations and movement trajectories in visuomotor synchronization, *Music Percept.* 28, 15–26. doi:10.1525/mp.2010.28.1.15.

Hove, M. J., Spivey, M. J., & Krumhansl, C. L. (2010). Compatibility of motion facilitates visuomotor synchronization, *J. Exp. Psychol. Hum. Percept. Perform.* 36, 1525–1534. doi:10.1037/a0019059.

Hove, M. J., Iversen, J. R., Zhang, A., & Repp, B. H. (2013). Synchronization with competing visual and auditory rhythms: bouncing ball meets metronome, *Psychol. Res.* 77, 388–398. doi:10.1007/s00426-012-0441-0.

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, e48496. doi:10.1371/journal.pone.0048496.

Iversen, J. R., Patel, A. D., Nicodemus, B., & Emmorey, K. (2015). Synchronization to auditory and visual rhythms in hearing and deaf individuals, *Cognition* 134, 232–244. doi:10.1016/j.cognition.2014.10.018.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis, *Percept. Psychophys.* 14, 201–211. doi:10.3758/BF03212378.

Jones, M. R., & Boltz, M. (1989). Dynamic attending and responses to time, *Psychol. Rev.* 96, 459–491. doi:10.1037/0033-295X.96.3.459.

Kanai, R., Lloyd, H., Buetti, D., & Walsh, V. (2011). Modality-independent role of the primary auditory cortex in time estimation, *Exp. Brain Res.* 209, 465–471. doi:10.1007/s00221-011-2577-3.

Klink, P. C., Montijn, J. S., & van Wezel, R. J. A. (2011). Crossmodal duration perception involves perceptual grouping, temporal ventriloquism, and variable internal clock rates, *Attent. Percept. Psychophys.* 73, 219–236. doi:10.3758/s13414-010-0010-9.

Knoblich, G., & Sebanz, N. (2008). Evolving intentions for social interaction: from entrainment to joint action, *Phil. Trans. R. Soc. B Biol. Sci.* 363, 2021–2031. doi:10.1098/rstb.2008.0006.

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

Levitan, C. A., Ban, Y.-H. A., Stiles, N. R. B., & Shimojo, S. (2015). Rate perception adapts across the senses: evidence for a unified timing mechanism, *Sci. Rep.* 5, 8857. doi:10.1038/srep08857.

London, J., Burger, B., Thompson, M., & Toiviainen, P. (2016). Speed on the dance floor: Auditory and visual cues for musical tempo, *Acta Psychol.* 164, 70–80. doi:10.1016/j.actpsy.2015.12.005.

Loucks, J., & Nagel, N. (2018). Temporal perception is enhanced for goal-directed biological actions, *Vis. Cogn.* 26, 530–544. doi:10.1080/13506285.2018.1516708.

Loucks, J., & Pechey, M. (2016). Human action perception is consistent, flexible, and orientation dependent, *Perception* 45, 1222–1239. doi:10.1177/0301006616652054.

Lukas, S., Philipp, A. M., & Koch, I. (2010). Switching attention between modalities: further evidence for visual dominance, *Psychol. Res.* 74, 255–267. doi:10.1007/s00426-009-0246-y.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences, *Behav. Res. Methods* 44, 314–324. doi:10.3758/s13428-011-0168-7.

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

Occelli, V., Spence, C., & Zampini, M. (2011). Audiotactile interactions in temporal perception, *Psychon. Bull. Rev.* 18, 429–454. doi:10.3758/s13423-011-0070-4.
Ortega, L., Guzman-Martinez, E., Grabowecky, M., & Suzuki, S. (2014). Audition dominates vision in duration perception irrespective of salience, attention, and temporal discriminability, *Atten. Percept. Psychophys.* 76, 1485–1502. doi:10.3758/s13414-014-0663-x.

Penney, T. (2003). Modality differences in interval timing: Attention, clock speed, and memory. In W. H. Meck (Ed.), *Functional and neural mechanisms of interval timing* (pp. 209–233). Boca Raton, FL, USA: CRC Press/Routledge/Taylor & Francis Group. doi:10.1201/9780203009574.

Repp, B. H. (2003). Rate limits in sensorimotor synchronization with auditory and visual sequences: the synchronization threshold and the benefits and costs of interval subdivision, *J. Mot. Behav.* 35, 355–370. doi:10.1080/00222890309603156.

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

Schutz, M., & Lipscomb, S. (2007). Hearing gestures, seeing music: vision influences perceived tone duration, *Perception* 36, 888–897. doi:10.1068/p5635.

Shi, Z., Chen, L., & Müller, H. J. (2010). Auditory temporal modulation of the visual Ternus effect: The influence of time interval, *Exp. Brain Res.* 203, 723–735. doi:10.1007/s00221-010-2286-3.

Shi, Z., Church, R. M., & Meck, W. H. (2013). Bayesian optimization of time perception, *Trends Cogn. Sci.* 17, 556–564. doi:10.1016/j.tics.2013.09.009.

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

Spence, C., F. Pavan, A. Maravita & N.P. Holmes. (2008). Multi-sensory interactions. In M.C. Lin & M.A. Otaduy (Eds), *Haptic rendering: foundations, algorithms, and applications* (pp. 21–52). Wellesley, MA, USA: AK Peters/CRC Press. doi:10.1201/b10636.

Spence, C., Parise, C., & Chen, Y. C. (2012). The Colavita visual dominance effect. In M. M. Murray & M. T. Wallace (Eds), *The neural bases of multisensory process* (pp. 523–550). Taylor & Francis Group. doi:10.1201/b11092.

Stupacher, J., Witte, M., & Wood, G. (2017). Go with the flow: Subjective fluency of performance is associated with sensorimotor synchronization accuracy and stability, Proc. 25th Anniv. Conf. Eur. Soc. Cogn. Sci. Mus., Ghent, Belgium, 163–166.

Su, Y.-H., & Jonikaitis, D. (2011). Hearing the speed: Visual motion biases the perception of auditory tempo, *Exp. Brain Res.* 214, 357–371. doi:10.1007/s00221-011-2835-4.

Su, Y. H., & Salazar-López, E. (2016). Visual timing of structured dance movements resembles auditory rhythm perception, *Neural Plast.* 2016, 1678390. doi:10.1155/2016/1678390.

Treisman, M., Faulkner, A., Naish, P. L. N., & Brogan, D. (1990). The internal clock: evidence for a temporal oscillator underlying time perception with some estimates of its characteristic frequency, *Perception* 19, 705–743. doi:10.1068/p190705.

Van Wassenhove, V., Buonomano, D. V., Shimojo, S., & Shams, L. (2008). Distortions of subjective time perception within and across senses, *PLoS ONE* 3, e1437. doi:10.1371/journal.pone.0001437.

Varlet, M., Marin, L., Issartel, J., Schmidt, R. C., & Bardy, B. G. (2012). Continuity of visual and auditory rhythms influences sensorimotor coordination, *PLoS ONE* 7, e44082. doi:10.1371/journal.pone.0044082.

Vatakis, A., & 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.

Wang, X., & Wöllner, C. (2019). Time as the ink that music is written with: a review of internal clock models and their explanatory power in audiovisual perception, *Jahrb. Musikpsychol.* 29, e67. doi:10.5964/jbdgm.2019v29.67.