# Multimodal perception of interpersonal synchrony: Evidence from global and continuous ratings of improvised musical duo performances

Kelly Jakubowski[1], Tuomas Eerola[1], Arwen Blackwood Ximenes[2], Wai K. Ma[2], Martin Clayton[1], and Peter E. Keller[2]

[1]Durham University, [2]Western Sydney University

Author Note: This research was funded by the Arts and Humanities Research Council (grant no. AH/N00308X/1).

Corresponding Author: Kelly Jakubowski, Durham University, Department of Music, Palace Green, Durham, DH1 3RL, UK, +44(0)1913341546, kelly.jakubowski@durham.ac.uk

## Abstract

Investigating cues that underpin perceptual judgments of interpersonal coordination has important implications for understanding socio-cognitive evaluations of the quality of human interactions. With a focus on musical interpersonal coordination, we conducted two experiments investigating the impact of music style, modality of stimulus presentation, rater expertise, and audio/visual stimulus features on ratings of perceived synchrony in improvised duo performances. In the first experiment participants made synchrony ratings following 10-second excerpts of musical performances, while in the second experiment participants rated longer (up to 1 minute) excerpts continuously as the music unfolded. Several consistent results emerged across the two experiments, including that participants perceived standard jazz improvisations featuring a regular beat as significantly more synchronous than free improvisations that aimed to eschew the induction of such a beat. However, ratings of perceived synchrony were more similar across these two styles when only the visual information from the performance was available, suggesting that performers' bodily cues functioned similarly to communicate and coordinate musical intentions. Computational analysis of the audio and visual aspects of the performances indicated that synchrony ratings increased with increases in audio event density and when co-performers engaged in periodic movements at similar frequencies, while the salience of visual information increased when synchrony ratings were made continuously over longer timescales. These studies reveal new insights about the correspondence between objective and subjective measures of synchrony, and contribute methodological advances indicating both parallels and divergences between the results obtained in paradigms utilizing global versus continuous ratings of musical synchrony.

*Keywords:* entrainment, synchrony, music performance, audiovisual perception, interpersonal coordination

# Multimodal perception of interpersonal synchrony: Evidence from global and continuous ratings of improvised musical duo performances

Successful interpersonal interactions are contingent on one's ability to anticipate, perceive, and respond to time-sensitive, multimodal cues (Garrod & Pickering, 2004; Shockley, Richardson, & Dale, 2009; van der Steen & Keller, 2013). Research on conversation has shown that verbal exchanges are precisely and rapidly coordinated (Levinson, 2016), with turn-taking gaps between interlocutors often averaging 200 ms or less (Stivers et al., 2009). In addition, non-verbal aspects of communication (e.g., gestures, movement, and other visual cues that accompany conversations) are often spontaneously coordinated between interlocutors (e.g., Richardson, Dale, & Kirkham, 2007; Shockley, Richardson, & Dale, 2009; Shockley, Santana, & Fowler, 2003), which can in turn facilitate cognitive processing of the conversation and increase social cohesion (Bernieri, 1988; Chartrand & Bargh, 1999; Latif, Barbosa, Vatiokiotis-Bateson, Castelhano, & Munhall, 2014; Macrae, Duffy, Miles, & Lawrence, 2008). These observations indicate that humans are highly sensitive, at least implicitly, to fine-grained temporal cues during communication in both the auditory and visual domains. Investigating such sensitivities is therefore important for understanding subjective judgments of the success of an interaction and its social consequences.

Temporal sensitivities to multimodal, communicative cues have also been explored in the domain of music performance (e.g., Arrighi, Alais, & Burr, 2006; Goebl & Parncutt, 2001; Moran, Hadley, Bader, & Keller, 2015). Music performance provides a useful forum for investigating temporal aspects of nonverbal multimodal communication, as different styles of music vary naturally in terms of their timing structure, from metrical frameworks with isochronous beat subdivisions to those with asymmetrical subdivision structures to nonmetrical styles (i.e. avoidance of regularly patterned beat groupings), to give just a few examples. In ensemble performance, musicians must not only carefully execute the timing of their own parts in relation to the prescribed metrical framework, but also need to constantly monitor and temporally adapt to the behaviors of their co-performers (Keller, 2014; Repp & Keller, 2008; van der Steen & Keller, 2013), which is achieved through the use of both auditory and visual cues (Goebl & Palmer, 2009; Williamon & Davidson, 2002). The term ‘interpersonal entrainment’ has been used to describe a broad range of behaviors in which musical co-performers’ actions—and resultant produced sounds—become temporally coupled, either deliberately or spontaneously (see Clayton, Sager, & Will, 2005, for an overview).

The present study investigated audience perception of entrainment between co-performers in musical duos, and how such perceptions varied as a function of stylistic and audiovisual features of a performance. Specifically, participants were asked to make synchrony ratings of performances, with ‘synchrony’ used as a short, comprehensible term intended to direct their attention toward temporal aspects of the performance (rather than pitch or extramusical factors, for instance); we also specified that we were using this term to describe the general temporal coordination/coherence between two performers, rather than whether they specifically produced sounds at the same instant in time. This definition of ‘synchrony’ (which is broader than that typically implicated in experimental research on sensorimotor synchronization, e.g. Repp & Su, 2013) was employed to reflect the fact that real musical performances often comprise extensive passages where co-performers play in counterpoint or assume disparate roles, such as melody player and accompanist, yet still demonstrate systematic temporal coupling between the sounds they produce. This definition was also intended to encompass several different dimensions and timescales of interpersonal entrainment between co-performers (Clayton et al., 2005; Clayton et al., in review), which may be more or less salient depending on the audio and/or visual cues available to the listener (MacRitchie, Varlet, & Keller, 2017).

## Auditory Aspects of Musical Synchronization

Previous research on musical synchronization in the auditory modality has typically focused on analyzing note-to-note asynchronies between sounds produced by co-performers. Analyses of audio-recorded musical performances have revealed that synchronization often varies throughout the course of a piece (e.g., Shaffer, 1984; Wing, Endo, Bradbury, & Vorberg, 2014) and can depend on the particular style or musical tradition. For instance, recent comparative corpus analyses have demonstrated that Afrogenic drum ensembles (Malian jembe and Uruguayan candombe groups) play in tighter synchrony on average than styles implicating plucked or bowed strings (e.g., North Indian instrumental ragas and string quartets), which may be related to the sharper acoustic onsets produced by percussion instruments (Clayton, et al., in review). A number of additional factors may contribute to this variety in synchronization across different types of music, from aesthetic preferences and expressive intentions (Iyer, 2002; Keller, 2014) to motor or musical constraints on the performers (Wing, 1993). Ensemble roles can also affect the degree or directionality of asynchronies between co-performers, as evidenced, for instance, by the tendency for the melodic voice to play slightly before the accompaniment in Western classical music ('melody lead'; Keller & Appel, 2010; Palmer, 1989; Rasch, 1988; see also Clayton et al., in review, for examples of 'melody lead' in non-Western music styles).

In terms of perceptual sensitivity to these auditory timing deviations between co-performers, previous research has revealed that humans can perceive asynchronies of up to around 2 ms from steady-state synthesized tones (Wallach, Newman, & Rosenzweig, 1949; Zera & Green, 1993), whilst temporal order of two sounds can be determined from asynchronies as little as 20 ms for pure tones (Hirsh, 1959; Rosen & Howell, 1987) or 30 ms for musical tones (Goebl & Parncutt, 2001). However, perceptual ratings of synchronization can also be prone to various biases, including the tendency to perceive melody lead when attention is focused on the melody line even when no asynchrony is present (Ragert, Fairhurst, & Keller, 2014) and possible attentional biases towards instruments which are more rhythmic or produce sharper onsets (see Butterfield, 2010, in which participants tended to report a drum lead over bass, even when no asynchrony was present). In addition, several studies by Repp (1992, 1995, 1999) have indicated that the detection of timing deviations in solo piano music varies systematically based on the structural position within the music (e.g., participants tend to perceive lengthening of notes at the end of phrases, in line with common practice in Western classical music, even when such lengthening is not actually present), suggesting that asynchrony detection may also be affected by aspects of the musical structure.

**Visual Aspects of Musical Synchronization**

As in conversation, visual information also plays a key role in facilitating cohesion in musical performances. In particular, musicians use cues from co-performers' instrumental movements (i.e. sound-producing movements, such as drawing a violin bow across the strings) and ancillary movements (i.e. sound-facilitating movements, such as head nods) in order to produce music that is temporally and expressively coordinated (Bishop, Cancino-Chacón, & Goebl, 2019; Bishop & Goebl, 2018; Chang, Kragness, Livingstone, Bosnyak, & Trainor, 2019; Glowinski et al., 2013; Kawase, 2014; King & Ginsborg, 2011; Williamon & Davidson, 2002). As the visual system exhibits lower temporal resolution than the auditory system (Holcombe, 2009), visual information tends to be better suited for facilitating coordination of high-level aspects of the musical structure (e.g., phrase boundaries and expressive intentions), whilst auditory information may be more vital at the level of note-to-note synchronization (MacRitchie, Varlet, & Keller, 2017). Thus, musicians are likely to rely on cues in the auditory domain for achieving temporal coordination over short timescales and visual cues for relatively long-term coordination. The efficacy with which such visual cues can be utilized for temporal coordination is also modulated by factors such as motor expertise; for instance, it has been found that string players are more effective than both non-musicians and non-string playing musicians at predicting timing from visual cues from a first violinist of a string quartet (Wöllner & Cañal-Bruland, 2010).

Audiences also rely substantially on visual cues from musical performers in order to interpret expressive intentions and judge the quality of performances (Dahl & Friberg, 2007; Platz & Kopiez, 2012; Wanderley, Vines, Middleton, McKay, & Hatch, 2005). In some cases, visual information has even been found to be a more effective indicator of expressive information (Davidson, 1993) and had a more prominent effect on judgments of performance quality (Tsay, 2013) than auditory information. The perception of temporal contingencies between co-performers' movements has been investigated by Moran et al. (2015), who asked participants to make real/fake judgments of pairings of musicians presented via point-light displays in which one performer was soloing and the other was a silent 'back-channeler'. It was found that participants were able to distinguish real from fake pairings for non-pulsed free improvisations (characterized by the avoidance of a regular pulse) but not for standard jazz performances with a regular underlying pulse. However, the lack of effect for the standard jazz excerpts may be due to the fact that most of the duos performed at similar tempi, resulting in potentially imperceptible discrepancies between movement and audio tempi within the fake pairings. This is supported by the additional finding that participants with better rhythm perception skills were more accurate in distinguishing the real from fake pairings for the standard jazz duos, suggesting a more refined level of temporal sensitivity may have been required to successfully complete this task. Positive effects of musical expertise on the ability to detect asynchronies between audio and visual information have also been reported in studies of multisensory integration with rhythmic drumming stimuli (Petrini et al., 2009a; Petrini, Russell, & Pollick, 2009b). Finally, in work that bears some parallels to that of Moran et al. (2015), Wöllner (2018) investigated bodily interactions between soloing versus silent partners in free improvisation duo performances. Correspondences were found in the movement features (movement variability and cumulative distance of head motion) between duos, which also varied systematically as a function of the intended emotion of the performance (happy or sad).

## The Present Research

The research reviewed above has revealed that observers are sensitive to fine-grained temporal contingencies between both sounds and movements produced by musical co-performers, although perceptual judgments of the level of coordination within musical ensembles can be influenced or biased by a variety of factors, from elements of the musical structure and ensemble roles to the modality of stimulus presentation and the expertise of the listener. As subjective judgments of the synchronicity of a performance can also influence audience reactions to music, from the appraisal of performance quality to emotional and motor responses (Engel, Hoefle, Monteiro, Bramati, et al., 2014; Engel, Hoefle, Monteiro, Moll, & Keller, 2014; Labbé & Grandjean, 2014; Trost, Labbé, & Grandjean, 2017), understanding the factors that potentially underlie these judgments has key implications for research on both nonverbal interpersonal communication and music performance evaluation.

In the present work we investigated the perception of musical synchrony in two types of natural video-recorded musical performances that varied in their temporal structure: pulsed and non-pulsed duo improvisations (a subset of the performances from Moran et al., 2015).

Specifically, the pulsed improvisations consisted of duos performing and improvising over a jazz standard that has a regular underlying beat and simple metrical structure, whereas the non-pulsed performances were free improvisations that are characterized by the avoidance of both a regular, predictable beat and hierarchical metrical structure. Participants were exposed to the performances in audio only, visual only, and audiovisual conditions, to test how information from the different sensory channels is weighted and combined during the process of judging ensemble synchrony. Two groups of participants were recruited (high and low training) on the basis of years of previous formal musical training.

A second aim of this work was to conduct a novel investigation into the effects of audio and visual features of the performances on perceptual ratings of ensemble synchrony. Feature selection was based on previous literature on musical rhythm and coordination. Audio features included timing-related properties of the music—event density and pulse clarity—as well as measures of RMS energy and spectral flux—the extent to which the frequency spectrum changes over time. Spectral flux has previously been shown, particularly in lower frequency bands, to be positively related to perceptual ratings of rhythmic strength, propensity to move, and musical groove (Burger, Ahokas, Keipi, & Toiviainen, 2013; Stupacher, Hove, Novembre, Schütz-Bosbach, & Keller, 2013). In terms of visual features, we used computer vision methods to track (primarily ancillary) upper body movements of performers from the video recordings (Jakubowski et al., 2017) and quantified the level of temporal coordination between co-performers' movements using cross wavelet transform (CWT) analysis. In previous research, the temporal coupling of such movements as quantified by CWT analysis been shown to be a significant predictor of audience judgments of visual bouts of communicative interaction between musical co-performers (Eerola et al., 2018). We also included an aggregate measure of the quantity of motion of both performers in each duo and a summed measure of periodic movement from both performers.

A final aim was to investigate the time course of musical synchrony judgments, in particular via continuous ratings of synchrony by participants. Continuous ratings have proven to be informative in studies of emotional response to music (e.g., Dean & Bailes, 2016; Dean, Bailes, & Dunsmuir, 2014; Schubert, 2004) and performance quality more broadly (Thompson, Williamon, & Valentine, 2007), but have been rare in research on interpersonal musical coordination (see Vicary, Sperling, Zimmermann, Richardson, & Orgs, 2017, for an example from the dance research domain). Such an approach is clearly warranted given that temporal relationships between co-performers unfold and evolve over time. Here, we present the results of two experiments: Experiment 1 required participants to give a single synchrony rating at the end of short excerpts of the performances, while Experiment 2 required participants to provide continuous ratings of moment-to-moment synchrony for longer excerpts. Comparing the results of these two experiments allowed us to probe whether single, global ratings can approximate the observations obtained using a more complex, continuous rating paradigm, and explore whether different audio and visual features play a more or less prominent role depending on the timescale of the rating.

In sum, we investigated how synchrony ratings of musical duo performances varied in relation to the temporal structure (pulsed/non-pulsed) of the music, modality of stimulus presentation, and musical expertise of the rater. We also investigated how audio and visual (movement) features of the performances might predict perceptual ratings of synchrony, and probed how such relationships between these features and synchrony ratings varied depending on the timescale of the rating. The results of this work provide insights into the cues that listeners make use of when assessing multimodal aspects of temporal coordination in music performance.

## Experiment 1: Global ratings of perceived synchrony

## Method

**Design.** The experiment used a mixed 2 x (2 x 3) design, with one between-subjects independent variable of musical training (high/low groups, based on years of formal training in music), and two within-subjects independent variables: music style (pulsed/non-pulsed music) and modality (audiovisual (AV)/audio only (AO)/visual only (VO)) of stimulus display. The dependent variable was a subjective rating of synchrony between the two performers for each stimulus (on a rating scale from 1 to 9). The experiment was granted ethical approval by the local institutional review board.

**Participants.** Fifty-two participants were recruited for the experiment. These participants were recruited for low/high musical training groups on the basis of the following criteria: For the high musical training group we required participants with more than four consecutive years of training on a musical instrument in a private setting and for the low musical training group we recruited individuals who had received no more than three years of such training. Participants in the high musical training group were recruited via contacts in one of the authors' music community, and received reimbursement for travel costs, while participants in the low musical training group were studying Introductory Psychology at Western Sydney University and received course credit for participation. In total, we tested 26 participants (14 female and 12 male; mean age = 25 years; *SD* = 5.51; range = 18 – 42 years) in the high musical training group (*M* = 10.98 years of training, *SD* = 3.21, range = 6 – 17 years). The low musical training group comprised another 26 participants (21 female and 5 male; mean age = 24 years; *SD* = 9.41; range = 18 – 54 years) who reported, on average, 0.43 years of musical training (SD = 0.83, range = 0 – 3 years). It should be noted that years of musical training was the sole measure used to differentiate the two groups, and we did not require the musically trained participants to have expertise in the specific styles used in the present experiment (only 1 high training and 1 low training participant reported experience in performing jazz music, with no participants reporting experience in performing free improvisation). None of the participants reported having a hearing impairment and any participants requiring vision correction (e.g., glasses/contact lenses) were asked to wear their corrective lenses throughout the experiment.

**Stimuli.** The stimuli were selected from the Improvising Duos video corpus recorded at the Max Planck Institute for Human Cognitive and Brain Sciences in Leipzig, Germany (first reported in Moran et al., 2015). The musicians in these duos were filmed facing each other, and their whole bodies could be seen in the video frame (see Moran, Jakubowski, & Keller, 2017 to view the video corpus). The videos were recorded using a SONY HDR-HC9 camera in AVI format at a frame rate of 25 frames per second and a frame size of 720 × 576 pixels.

Two different groups of instrumental musicians performed the pulsed and non-pulsed improvisations. Performers in these duos were recruited on the basis of public performance experience of around 10 years in their respective styles. Some performers had played together before, but this was not a primary recruitment criterion and some duos were only introduced to each other on the day of the recording. No performer played in more than one duo. The pulsed music comprised improvisations over the jazz standard *Autumn Leaves*, while the non-pulsed music comprised free improvisations, in which a regular beat is purposely avoided. All performances contained sections of both joint and solo playing; however, only sections of joint playing were included in the present stimulus set since the primary experimental task was to rate interpersonal synchrony between the two musicians. The final stimulus set comprised five duos from each of the two music styles (see Table 1). The different number of clips per duo that was used here simply reflects the fact that different amounts of footage had been captured for each duo in the original corpus, as well as the fact that different duos varied in the number and duration of joint playing sections.

The original recordings were edited using Adobe Premiere Pro 6.0.0 to produce video clips of 10 s each, with audio fade in/out of 500 ms and visual fade in/out of 1 second. The 10 s clips were cut from the longer video recordings at random (i.e. not in line with phrase boundaries), with the precondition that the clip was taken from a section of joint (rather than solo) playing. In addition to the audiovisual (AV) modality condition, the 10 s clips were further processed to create two additional modalities for presentation within the experiment: 1) audio only (AO), in which the video channel was replaced with a still image of the duo, and 2) visual only (VO), in which the audio channel was deleted.

Table 1

*Instrumentation and Number of Clips in the Stimulus Set for the Non-Pulsed and Pulsed Duos (Experiment 1)*

| Music style | Duo | Instrument 1 | Instrument 2 | Number of clips |
|---|---|---|---|---|
| Non-pulsed | 1 | Flute | Double bass | 5 |
| | 2 | Soprano saxophone | Drums | 5 |
| | 3 | Drums | Tenor saxophone | 8 |
| | 4 | Clarinet | Alto saxophone | 7 |
| | 5 | Cello | Soprano saxophone | 3 |
| Pulsed | 6 | Piano | Tenor saxophone | 4 |
| | 7 | Trumpet | Electric guitar | 6 |
| | 8 | Tenor saxophone | Electric bass guitar | 4 |
| | 9 | Violin | Piano | 9 |
| | 10 | Double bass | Acoustic guitar | 5 |

**Apparatus.** Testing took place in sound treated rooms or quiet conditions. All stimuli were presented on a Sony Trinitron screen by a Macintosh 15.4-inch Macbook Pro running OS X 10.9.5 with screen display resolution of 1680 × 1050 pixels, using OpenSesame experimental software (version Jazzy James 3.1; Mathôt, Schreij, & Theeuwes, 2012). Auditory stimuli were presented through Beyerdynamic DT 770 PRO or KOSS UR20 headphones. A questionnaire was used to collect participants' demographic information, music and dance experience (18 musical experience questions from the Ollen Musical Sophistication Index and eight dance experience questions), as well as strategies used in the judgment of synchrony.

**Procedure.** The experiment began with a practice session, which included both verbal and written instructions and an opportunity to ask questions about the procedures. Participants were informed that the performers in the videos were professional musicians who were highly regarded in their respective styles of performance. Participants were told that the music clips would be of different styles of music, but were not given detailed information about what a typical standard jazz or free improvisation performance ‘should’ comprise, and were simply exposed to the two styles through the several practice clips that were presented at the beginning of the session and preceding each block of the experiment. The initial practice session consisted of one trial from each music style, followed by three trials of the same clip presented in all three possible modalities of stimulus presentation. Participants were asked to “*Rate the synchrony of the performers*” on a scale from 1 (“very poor” synchrony) to 9 (“excellent” synchrony). To ensure all participants had the same understanding of the term “synchrony” for this task, they were asked to evaluate the quality of coordination between the performers and were told that synchrony in this particular context did not mean that both performers were “*playing the same note exactly at the same time*” but rather referred to “*how well they were playing together to produce a coherent piece of music*”.

The main experiment was presented in 6 blocks, each of which comprised one music style (pulsed/non-pulsed) and one modality (AV/AO/VO) of stimulus presentation. Participants were exposed to all possible music style x modality pairings across these 6 blocks, which were presented in a counterbalanced order across participants, and the order of presentation of the individual stimuli within each block was randomized. Each of the 6 blocks included 16 trials of the rating task (96 trials in total: 12 practice trials, 84 experimental trials). The initial 2 trials of each block were designated as practice trials for participants to familiarize themselves with the new modality and/or music style. These practice trials (as well as the practice clips that were presented before the main experiment, as described above) were excerpts from the same corpus as those used in the experimental trials, therefore they included some of the same performers as the experimental trials, but different sections of music than were seen in the experimental trials. In order to avoid carry-over effects between the AV block and the other two modalities, the experimental stimuli were split into two file pools, such that participants who were presented with stimuli from file pool 1 for the AV modality would be presented with file pool 2 stimuli for the AO modality and VO modality, and vice versa. As such, two versions of the experiment were created and the version order was counterbalanced between subjects.

Each participant completed the synchrony rating experiment twice (with identical order of block presentation, separated by a two-minute break) so that two ratings of each stimulus were obtained from each participant. The demographic/ musical background questionnaire was completed following the main experiment.

**Extraction of audio and visual features.** Audio and visual features (aggregated across each clip) were extracted from each stimulus in order to examine their relationship with the perceived synchrony ratings. Audio features of each stimulus were extracted using the MIR Toolbox for MATLAB (Lartillot, Toiviainen, & Eerola, 2008). The audio features considered in this study were mean event density, mean pulse clarity (Lartillot, Eerola, Toiviainen, & Fornari, 2008), mean RMS energy, and mean spectral flux in one low-frequency (Sub-Band 2: 50-100 Hz) and one high-frequency sub-band (Sub-Band 9: 6400-12800 Hz; see Alluri & Toiviainen, 2010). The low-level features (spectral flux, RMS energy) were extracted over 25 ms frames with 50% overlap. Event density and pulse clarity, which both rely on onset extraction using a filterbank decomposition of the half-wave rectified envelope, were computed over 1000 ms frames with 50% overlap.

To obtain visual (movement) features from the video clips, the upper body movements of each performer were extracted using automated computer vision tools in EyesWeb (http://www.infomus.org/eyesweb_ita.php). Specifically, a region of interest (ROI) was manually set around each performer's upper body and two-dimensional movement within that region was then tracked using dense optical flow (Farnebäck, 2003; see Jakubowski et al., 2017 for details of this implementation) at a sampling rate of 25 Hz. The aim of this movement extraction procedure was to capture many of the ancillary, communicative movements commonly made by performers (e.g. head nods, body sway), although it should be noted that some performative, sound-producing movements (e.g. bowing of a violin) may still be captured, the extent of which may vary from one instrument to another (though see Jakubowski et al., 2017, in which this method was able to capture a significant proportion of head and upper torso movements across a range of instruments including piano, strings, and woodwinds). Cross wavelet transform (CWT) analysis was then applied to obtain a measure of joint periodic movement across the two performers' individual movement profiles, following the implementation used by Eerola et al. (2018). Specifically, we computed the mean CWT Energy in sub-bands of 0.6 seconds in width centered around 0.4 and 2.0 Hz, which correspond to a slower (hereafter *CWT 0.4 Hz*) and faster (hereafter *CWT 2.0 Hz*) band of joint periodic movement respectively. We also computed a measure of co-occurring periodic movement at periodicities that are not necessarily related; this was done by computing the Energy of the Wavelet Transform (WT) for each performer across time for each clip (at 25 Hz), summing these values across the two performers at each timepoint in this series, and then taking the mean of this summed time series for each clip (hereafter *Summed WT Energy*). In addition, we applied a frame differencing technique

(Wren, Azarbayejani, Darrell, & Pentland, 1997) to the ROIs to obtain a measure of the Quantity of Motion (QoM) of each performer on a frame-by-frame basis (sampling rate of 25 Hz; see Jakubowski et al., 2017 for further details). The QoM values for both performers were then summed across time, and the mean value of this time series was taken as a measure of combined QoM across both performers (hereafter *Summed QoM*).

## Results

**Effects of modality, music style, and musical training on perceived synchrony ratings**. A 3-way mixed ANOVA was run to investigate the effects of modality (AV/AO/VO) of stimulus presentation, music style (pulsed/non-pulsed), and musical training (low/high) on perceived synchrony ratings. Synchrony ratings from the two trials completed for each stimulus by each participant were averaged. Pearson correlations computed between the synchrony ratings across the two trials of each stimulus for each participant showed good agreement between trial 1 and 2 (mean $r = .65$, SD = .15, range = .26 to .90). We corrected for multiple comparisons within this 3-way ANOVA using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995), as described by Cramer et al. (2016), for controlling the false discovery rate (FDR).

A statistically significant main effect of music style ($F(1, 50) = 206.44$, $p < .001$, $\eta_p^2 = .81$) revealed that overall ratings of synchrony were higher for the pulsed ($M = 6.77$, $SD = 1.76$) than non-pulsed music ($M = 4.15$, $SD = 2.12$). Thus, the presence of a regular, predictable beat appears to lead to higher judgments of overall temporal cohesion. No main effects of modality ($F(2, 100) = 0.90$, $p = .41$, $\eta_p^2 = .02$) or musical training ($F(1, 50) = 3.30$, $p = .08$, $\eta_p^2 = .06$) were found. Secondary analyses using years of musical training as a continuous variable revealed a small, positive correlation between musical training and mean synchrony ratings (Kendall's $\tau$ (50)= .20, $p = .05$), which appears to be driven primarily by responses to the non-pulsed music style (non-pulsed: Kendall's τ (50)= .23, *p* = .02, pulsed: Kendall's τ (50)= .04, *p* = .66).

Significant two-way interactions were present between modality and music style (*F*(2, 100) = 37.53, *p* < .001, $\eta_p^2$ = .43) and modality and musical training (*F*(2, 100) = 3.84, *p* = .02, $\eta_p^2$ = .07), while the interaction of music style and musical training was not statistically significant (*F*(1, 50) = 3.30, *p* = .08, $\eta_p^2$ = .06). Mean synchrony ratings across all three independent variables are visualized in Figure 1. In regard to the interaction of modality and music style, for the pulsed music style, synchrony ratings were significantly lower in the VO modality (*M* = 6.33, *SD* = 1.78) than both the AV (*M* = 7.00, *SD* = 1.74; *p* < .001) and AO modalities (*M* = 6.97, *SD* = 1.68; *p* < .001), with no statistically significant difference between the AV and AO modalities (*p* > 1) in Bonferroni-corrected, paired-samples t-tests. The same analysis revealed the opposite pattern of results for the non-pulsed music style, with *higher* ratings in the VO modality (*M* = 4.54, *SD* = 2.09) than both the AV (*M* = 4.09, *SD* = 2.27) and AO modalities (*M* = 3.82, *SD* = 2.11), with a statistically significant difference only between the VO versus AO modalities (*p* = .002; for VO vs. AV *p* = .05; for AV vs. AO *p* = .09). This suggests the auditory cues may have played a more substantial role than the visual cues in leading to the overall difference in synchrony ratings between pulsed versus non-pulsed music. In terms of the interaction between modality and musical training, it was found in Bonferroni-corrected, independent-samples t-tests that participants in the high musical training group gave significantly higher synchrony ratings in the AO modality (*M* = 5.69, *SD* = 0.86) than the low musical training group (*M* = 5.11, *SD* = 0.69; *p* = .03), but that the groups did not significantly differ in ratings in the AV or VO modality (*p* = .13 and p > 1, respectively). The corresponding three-way interaction between modality, music style, and musical training was also statistically significant (*F*(2, 100) = 3.41, *p* = .04, $\eta_p^2$ = .06), but did not survive FDR correction (critical α = .029).

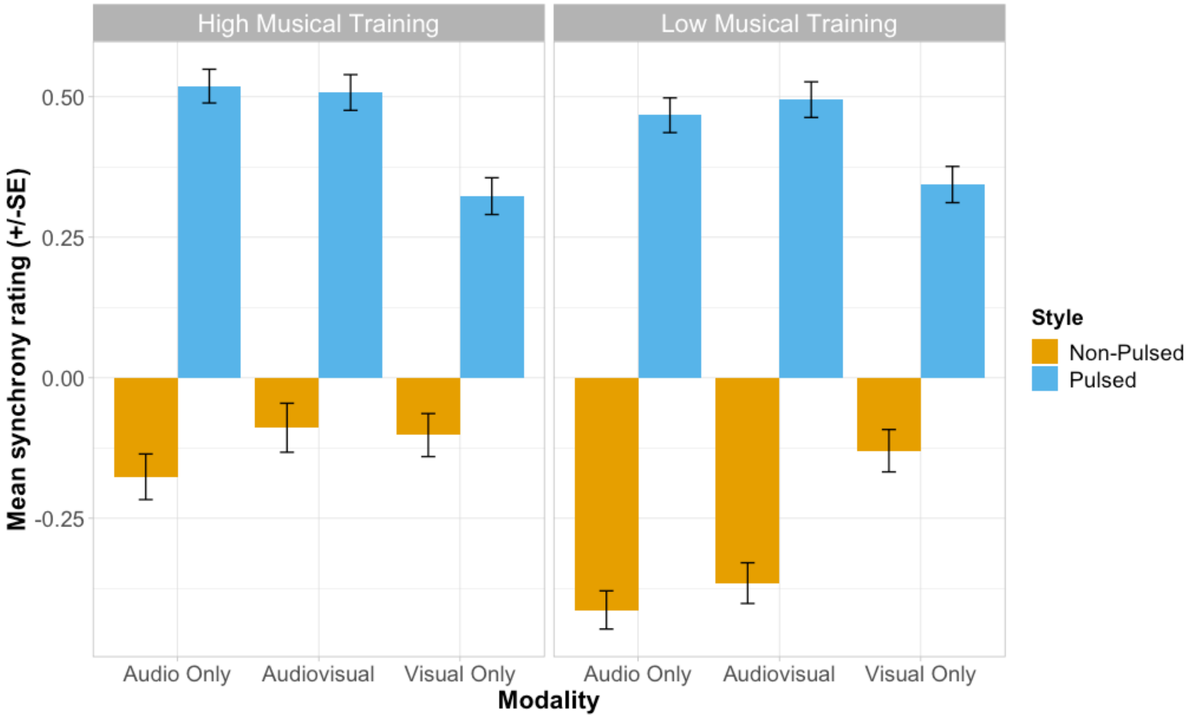

*Figure 1.* Mean synchrony ratings (+/-SE) by modality, music style, and musical training. The original ratings (on a scale of 1 to 9) have been normalized in this visualization to a scale ranging from -1 to +1 for ease of comparison with the results of Experiment 2 (see Figure 4).

Mean ratings of each individual stimulus clip were highly correlated across all modalities (AV & AO: $r(54) = .98$, $p < .001$, AV & VO: $r(54) = .85$, $p < .001$, AO & VO: $r(54) = .84$, $p < .001$). Thus, stimuli that were rated high in synchrony in the AV modality were also typically rated high in synchrony in both the AO and VO modalities. This is despite the fact that the stimuli were separated into two file pools and no participant ever rated the same stimulus in the AO or VO modality that (s)he would have already seen/heard in the AV modality (or vice versa). In addition, a regression model was fit to predict mean AV ratings using the mean AO and VO ratings for the same stimulus. This revealed that AV ratings were significantly predicted by AO ratings, but not VO ratings (see Table 2), suggesting a greater reliance on auditory than visual cues in the AV condition. Consistent with this asymmetry, the semi-partial correlation of AO and AV ratings (when controlling for VO ratings) was moderate and statistically significant (*r*(54) = .50, *p* < .001), while the semi-partial correlation of VO and AV ratings (controlling for AO ratings) was not significant (*r*(54) = .04, *p* = .76).

Table 2

*Linear Regression to Predict Audiovisual (AV) Synchrony Ratings from Audio Only (AO) and Visual Only (VO) Ratings of the Same Stimulus*

| Predictor | ß | SE | z-value | p-value |
|---|---|---|---|---|
| Intercept | 0.214 | 0.250 | 0.854 | .397 |
| AO | 0.870 | 0.047 | 18.563 | < .001*** |
| VO | 0.118 | 0.077 | 1.540 | .130 |

Note: *** *p* < .001; Adjusted $R^2$ = 0.96

**Relationship between audio/visual stimulus features and perceived synchrony ratings.** Our next aim was to investigate whether certain audio and visual features of the clips used in the experiment might help to explain differences in ratings of perceived synchrony between stimuli. Due to the relatively low number of musical stimuli (56) in comparison to the number of audio (5) and visual features (4), this was an exploratory analysis. The main aim was to identify trends that could be further investigated in Experiment 2 via continuous ratings, which yield a substantially larger number of data points for analysis (as both participant ratings and stimulus features can be sampled on a moment-to-moment basis over time).

Figure 2 displays the relationship between the mean ratings of synchrony for each clip (from the AV Modality condition) and the audio features. The effect of music style—that is, lower synchrony ratings for non-pulsed than pulsed performances as reported in the previous analysis—is clearly visible throughout. However, this descriptive analysis also indicates that synchrony ratings may be influenced by specific audio features, such as RMS energy and pulse clarity (although the pulse clarity result does not survive Bonferroni correction for 5 tests, with a critical α of .01). In this case it is not possible to disentangle the effects of these features from the effects of music style (pulsed music tends to be higher on these features than non-pulsed music). Therefore, these features will be investigated more systematically in terms of their moment-to-moment effects on synchrony ratings, separately for each music style, in Experiment 2.

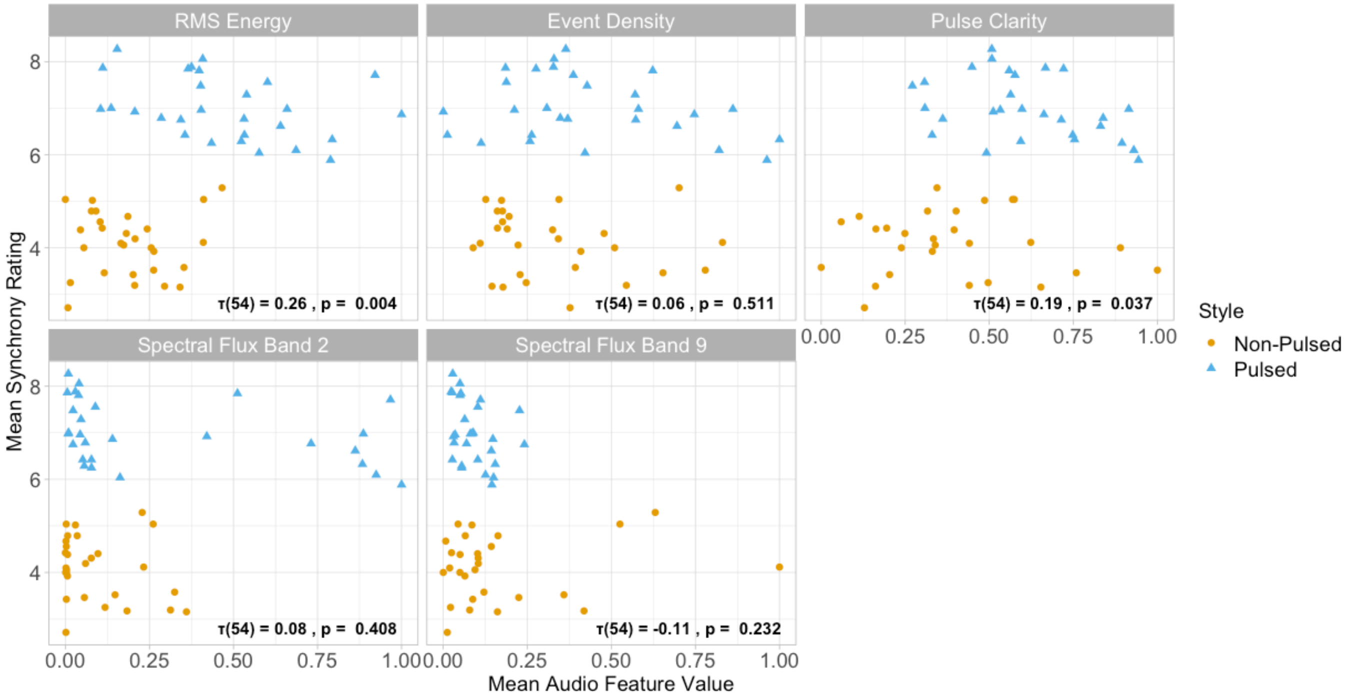

*Figure 2*. Mean audio features (normalized on a scale from 0 to 1) by mean synchrony ratings in the audiovisual (AV) modality for each stimulus. Pairwise correlations (non-parametric Kendall's tau) and their significance values are reported at the bottom of each subplot.

The visual features showed fewer relationships to the synchrony ratings (Figure 3), with the exception of Summed QoM, which was negatively related to synchrony ratings.

Again, this may be due to differences between the music styles, as the non-pulsed music clips exhibited significantly greater Summed QoM than the pulsed music clips ($t(52) = 3.69$, $p < .001$).

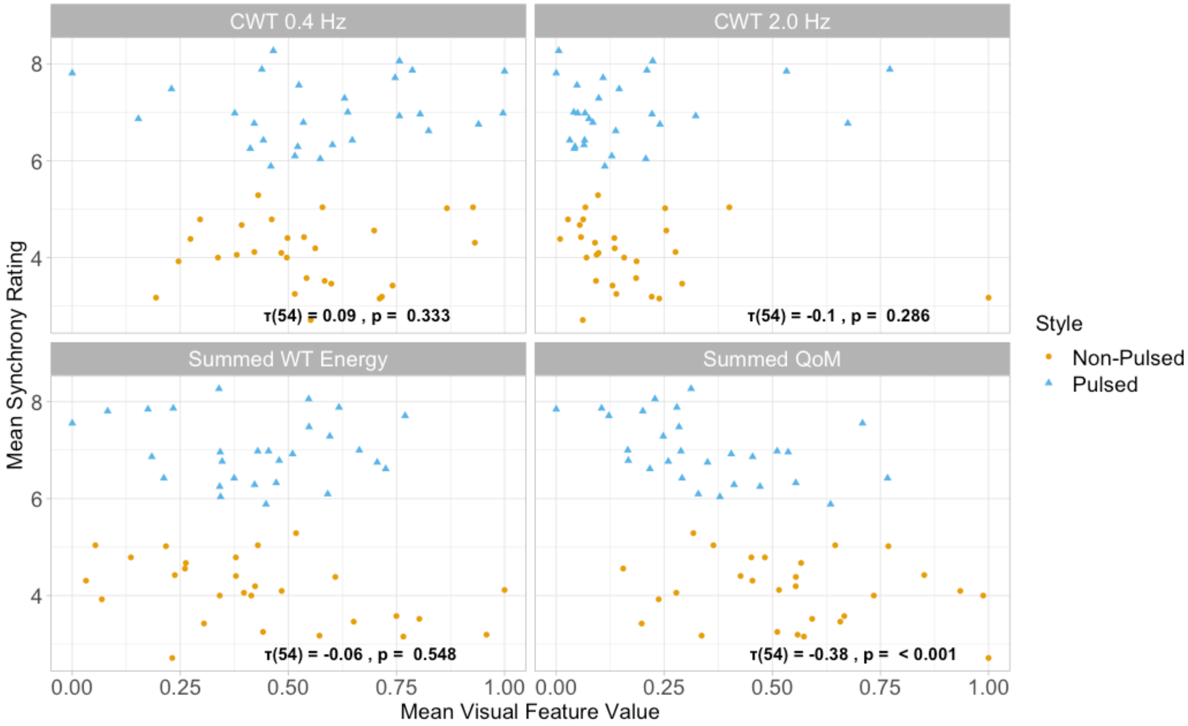

*Figure 3.* Mean visual features (normalized on a scale from 0 to 1) by mean synchrony ratings in the audiovisual (AV) modality for each stimulus. Pairwise correlations (non-parametric Kendall's tau) and their significance values are reported at the bottom of each subplot.

## Discussion

The results of Experiment 1 indicate that ratings of perceived synchrony between musical performers were positively influenced by the presence of a regular temporal framework, as evidenced by the strong main effect of music style, with pulsed music receiving consistently higher synchrony ratings than non-pulsed music. Judgments of the synchronicity of the same stimuli were similar across all three modalities of presentation, although ratings from the AO condition were a better predictor of AV ratings than VO ratings, suggesting that judgments of audiovisual stimuli were more influenced by the auditory than visual aspects of the performance. In addition, an interaction between music style and modality indicated that synchrony ratings became more similar for pulsed vs. non-pulsed music in the VO modality. This suggests that the types of visual cues for coordination that are used between co-performers may be similar between these two styles. Just as body gestures that accompany conversational speech have universal aspects in addition to features that vary across languages (Abner, Cooperrider, & Goldin-Meadow, 2015), communicative body motion during ensemble performances might generalize to some degree across musical styles and instruments (Clayton et al., in review). This is more likely to be the case for ancillary movements yoked to phrase boundaries (which are similar for pulsed and non-pulsed performances; Eerola et al., 2018) than for instrumental movements that trigger specific sounds.

There was no main effect of musical training on synchrony ratings, although an interaction between musical training and modality indicated that the high musical training group gave higher synchrony ratings than the low musical training group in the AO modality; this is particularly apparent in the non-pulsed music (see Figure 1). In addition, supplementary analysis revealed a positive correlation between years of musical training and mean synchrony ratings for the non-pulsed (but not the pulsed) duos. A possible explanation for these results is that more highly trained musicians may be more lenient in judging the synchronicity of non-pulsed music, perhaps due to a greater familiarity with this style than the less musically trained participants. An alternative explanation is that musical training might heighten sensitivity to auditory cues indicating complementary roles and turn-taking between co-performers (Phillips-Silver & Keller, 2012), lessening reliance on cues related to the simultaneity of sound onsets. In the VO modality the judgments of the two groups of participants were more similar, which again may be related to the two music styles being more similar in their visual than auditory features.

Some relationship between the audio features of the stimuli and synchrony ratings was evidenced; however, this relationship may be confounded by naturally occurring differences in these features between the two music styles used in Experiment 1. We thus aimed to rectify this potential limitation in Experiment 2, by investigating the effects of these audio features on synchrony ratings separately for each music style using a substantially larger number of data points (by sampling participant ratings and stimulus features more continuously over time). Nevertheless, this initial, descriptive analysis gave some indication that perceived synchrony ratings may be positively related to audio features such as pulse clarity and RMS energy. In terms of visual features of the stimuli, weaker relationships were revealed, with the exception of a negative relationship between Summed QoM and synchrony ratings that may also be an artefact of natural differences between the two music styles. Higher QoM in non-pulsed duos may reflect a heightened need for 'coordination smoothers'—specifically, exaggerating movements to make them easier to predict—seen in various forms of joint action (Vesper, Butterfill, Knoblich, & Sebanz, 2010).

Many of the movement cues that are captured by our visual features (e.g. ancillary body sway and head movements) evolve over a relatively slower timescale than auditory events (i.e. seconds rather than milliseconds; Eerola et al., 2018). Therefore, it may be especially informative to examine the possible influence of such cues over time within longer clips, as will be done in Experiment 2. In addition, both perceived synchrony and these audio/visual features can change over the time course of a performance. Thus, we sought to investigate the influence of these features more thoroughly by instructing participants to rate synchrony of longer clips from this corpus continuously over time.

## Experiment 2: Continuous ratings of perceived synchrony

Experiment 2 followed a similar design to Experiment 1, but with a focus on comparing and predicting moment-to-moment judgments of synchrony in musical performances collected via a continuous rating paradigm. This approach allowed us to test the degree of correspondence between continuous judgments and global judgments, as well as to account for potential factors such as memory demands on global judgments. Listener responses to musical stimuli are not only a function of musical features but also memory processes (Schubert, 2004). The general dependence on memory may have implications for the reliability of global ratings collected following stimulus presentation. Memory is more robust for tonal music than atonal music (see Halpern & Bartlett, 2010), which could have contributed to the differences observed between the pulsed and non-pulsed conditions, as the two styles also varied in terms of tonal content (Hadley, Sturt, Moran, & Pickering, 2018). With continuous ratings, memory demands are reduced as responses are made while the performances unfold, placing the two musical styles on a more level playing field in this respect. In addition, previous research comparing global ratings of performance quality for 20 s versus 60 s excerpts of the same solo piano performances found that the longer excerpts typically elicited higher and more consistent mean ratings, indicating that listener attitudes often change with increased exposure to a particular performance (Wapnick et al., 2005). In Experiment 2, synchrony ratings were again compared in relation to pulsed/non-pulsed musical style, modality of stimulus presentation, musical expertise of the rater, and audio/visual features of each stimulus. Assessing the relationship between these audio/visual features and continuous ratings allowed us to examine whether the effects observed across stimulus items in Experiment 1 also occur in response to feature-specific variation within stimuli.

## Method

**Design**. Experiment 2 used the same mixed 2 x (2 x 3) design as Experiment 1 to investigate the effects of musical training, music style, and modality of stimulus display on ratings of synchrony between the two performers. However, this experiment synchrony ratings were made continuously while the stimulus was displayed, and measured by the position of a joystick on a rating scale from "less in sync" to "more in sync". The experiment was granted ethical approval by the local institutional review board.

**Participants**. Participants (*N* = 49 total) were recruited and assigned to the high and low musical training groups following the same criteria as Experiment 1. The vast majority of participants had not taken part in Experiment 1, although 2 participants (both in the high musical training group) took part in both experiments. The high musical training group included 25 participants (18 female and 7 male; mean age = 33 years; *SD* = 10.2; range = 18 – 55 years) and the low musical training group comprised 24 participants (15 female and 9 male; mean age = 26 years; *SD* = 8.5; range = 18 – 46 years). Participants in the high training group had 4 to 18 years of musical training (*M* = 10.48 years of training, SD = 3.84) and the low musical training group had 0 to 3 years of training (*M* = 0.49 years, SD = 0.98). Two participants in the high training group reported experience in performing jazz, and one reported experience in performing experimental music; none of the low training participants reported experience in performing either of the two styles used in this experiment. None of the participants reported having a hearing impairment and participants requiring vision correction were asked to wear their corrective lenses throughout the experiment.

**Stimuli.** Stimuli were selected from the same video corpus as Experiment 1. Excerpts of joint playing were extracted from five duos from each of the two music styles (see Table 3). To maximize the duration of the clips, we selected entire sections of joint (as opposed to solo) playing, which meant the different clips varied naturally in duration. The duration of the pulsed music clips (*M* = 54.5 s, *SD* = 24.5, *range* = 37- 102 s) was longer on average than the non-pulsed clips (*M* = 41.9 s, *SD* = 6.5, *range* = 35- 53 s). As such, 6 pulsed and 8 non-pulsed clips were used, in order to equate the overall duration of the clips used across the two styles. The original recordings were edited using Adobe Premiere Pro 6.0.0 to produce the stimuli with audio fade in/out of 500 ms and visual fade in/out of 1 second. Audiovisual, audio only, and visual only versions of each stimulus were created using the same method as Experiment 1.

Table 3.

*Instrumentation and Number of Clips in the Stimulus Set for the Non-Pulsed and Pulsed Duos (Experiment 2)*

| Music style | Duo | Instrument 1 | Instrument 2 | Number of clips |
|---|---|---|---|---|
| Non-pulsed | 1 | Flute | Double bass | 1 |
| | 2 | Soprano saxophone | Drums | 2 |
| | 3 | Drums | Tenor saxophone | 2 |
| | 4 | Clarinet | Alto saxophone | 2 |
| | 5 | Cello | Soprano saxophone | 1 |
| Pulsed | 6 | Piano | Tenor saxophone | 1 |
| | 7 | Trumpet | Electric guitar | 1 |
| | 8 | Tenor saxophone | Electric bass guitar | 1 |
| | 9 | Violin | Piano | 2 |
| | 10 | Double bass | Electric guitar | 1 |

**Apparatus**. Testing took place in sound treated rooms or quiet conditions using the same set-up as Experiment 1. The main experiment was programmed in OpenSesame and participant ratings of synchrony were collected by recording the position of a Thrustmaster USB joystick every 200 ms. A casing was placed around the joystick to limit its movement to one dimension. Demographic information, music and dance experience, and responses to an open question on the strategies used[1] in the judgment of synchrony were collected via a questionnaire in an analogous fashion to Experiment 1.

**Procedure.** The experiment began with a practice session that followed the same structure and procedures as detailed in Experiment 1. Participants were asked to "*give a continuous rating of the synchrony of the performers throughout the clip*" using a joystick. The levels of synchrony ratings to be given ranged from "less in sync" (position on the joystick closest to the participant) to "more in sync" (position farthest away from the participant), and were accompanied by the written instructions: "*The position on the extreme end farthest away from you indicates an EXCELLENT level of synchrony between the performers. Likewise, the position on the extreme end closest to you indicates a VERY POOR level of synchrony*." The same definition of synchrony was provided to participants as used in Experiment 1.

The blocking and randomization of stimuli followed the same format as used in Experiment 1, with one practice trial presented at the start of each block. In total, the experiment comprised 27 trials of the rating task (6 practice trials, 21 experimental trials). Blocks containing pulsed stimuli included 3 trials, while blocks with non-pulsed stimuli included 4 trials. As in Experiment 1, to avoid carry-over effects between the AV block and the other two modalities, the experimental stimuli were split into two file pools; as such, two versions of the experiment were created and the version order was counterbalanced between subjects. The summed duration all trials in version 1 was 1233 s (20.55 min) and version 2 was 1209 s (20.15 min).

[1] In general, the self-reported strategies for judging synchrony in the AO condition made frequent reference to the rhythm, beat, and tempo of the music, although infrequent references to harmony and pitch indicated that these other (non-timing related) musical parameters were not always possible to disregard. In the VO condition frequent references were made to movements, eye contact, and body language, while in the AV condition references to both the sound and movement were made, although several participants explicitly stated that they relied more heavily on the audio than the visual cues when both were available.

Each participant completed the synchrony rating experiment twice (with identical order of block presentation, separated by a two-minute break) so that two continuous rating series of each stimulus were obtained from each participant. The demographic/ musical background questionnaire was completed following the main experiment.

**Extraction of audio and visual features.** Four audio features were extracted in MATLAB using MIR Toolbox: event density, pulse clarity, RMS energy, and spectral flux in Sub-Band 2 (50-100 Hz), using the same extraction parameters outlined in Experiment 1.[2] The same four visual (movement) features were extracted as in Experiment 1 (CWT 0.4 Hz; CWT 2.0 Hz, Summed WT Energy, Summed QoM). The primary difference here is that we retained each of these features as a time series spanning the duration of each stimulus (rather than computing a single, mean value of each feature across the whole clip, as in Experiment 1).

**Data analysis**. In a first stage of analysis, and to provide a comparison to the results of Experiment 1, we tested the effects of modality of stimulus presentation, music style, and musical training on the *mean* synchrony ratings for these longer musical extracts (in essence, converting the continuous ratings to a single, global rating for each trial) in a 3-way mixed ANOVA. In a second strand of analysis, we focused on using the continuous audio and visual features to predict *continuous* ratings of synchrony over time. For this analysis, the synchrony rating time series were first adjusted slightly via a linear interpolation function. This was necessary as the OpenSesame interface was subject to an occasional, small amount of delay in sampling the joystick position; across the dataset, 38% of the data points were sampled at

[2] Spectral flux in Sub-Band 9 (6400-12800 Hz) was excluded in Experiment 2 due to the fact that both Experiment 1 and previous studies (e.g., Burger et al., 2013; Stupacher, et al., 2013) have confirmed that low-frequency spectral flux is more relevant to perceptual judgments of rhythmic properties of music than high-frequency spectral flux. In addition, the results of Experiment 1 indicated that these particular musical stimuli do not show a high degree of variation in Sub-Band 9 spectral flux (see Figure 2).

the desired time interval of exactly 200 ms, 28% were sampled at a time interval of 201 ms, and the remaining 34% of data points were collected at time intervals of 202 to 222 ms. The interpolation function was therefore applied to impose a constant sampling rate of 5 Hz across the dataset. The data were then scaled on a participant-wise basis, to ensure that all participants' responses ranged from -1 to 1 across the experiment. We then examined the response styles of individual participants for each trial. We excluded from this continuous analysis (but not from first analysis using the mean ratings) the data from 7 participants (1 high musical training, 6 low musical training) who moved the joystick less than 85% of the time across all trials, as these participants exhibited a strategy that resembled making static ratings for all or a large portion of each clip, rather than continuous ratings throughout.

In preparation for the analysis, the audio and visual features and synchrony ratings were all resampled to a common sampling rate of 5 Hz via symmetric low-pass filtering (including both reverse forward and backward filtering, to avoid introducing time delays to the signal) with a Butterworth filter. This operation smoothed the occasional 'jerky' synchrony ratings and brought all features to a similar rate of change (for similar low-pass filtering, see Metallinou, Katsamani & Narayanan, 2013; Mauss, Levenson, McCarter, Wilhelm & Gross, 2005). Finally, the first and the last 4 seconds of each trial were excluded to prevent the initial identical synchrony values from impacting the results (all trials started with the joystick at the midpoint position of 0) and to avoid constraints in the magnitude of the continuous ratings that are introduced by the low-pass filtering procedure at the very beginning and end of each trial.

For the main continuous ratings analysis, we first considered calculating an aggregated response series for each stimulus—the mean synchrony rating across all participants—to be predicted using the audio and visual features. However, continuous rating data for responses to music is often highly idiosyncratic (see for instance Dean, Bailes, &

Dunsmuir, 2014, where inter-individual differences were greater within than between groups of different musical expertise levels), and the mean rating necessarily covers a smaller range of the response scale than the range employed by individual responses (Upham & McAdams, 2018). In addition, the average standard deviation between ratings across all timepoints in our dataset was 0.50 (25% of the range of the rating scale), which indicated a consistently large amount of variation between participants. We also considered employing a method based on identifying significant moments of coordination of responses between participants such as Activity Analysis (Upham & McAdams, 2018). However, the stimuli we employed were all shorter than the recommend two-minute duration for applying Activity Analysis, and hence would have required modification of the thresholds in a relatively arbitrary fashion to obtain promising results. We therefore decided upon a multilevel Bayesian regression approach for testing the effects of the audio and visual features on synchrony ratings. The motivation for this approach was to incorporate group-level effects embedded in the data (i.e. individual participants and stimuli, which we considered as random effects in linear mixed models). In addition, Bayesian regression allowed us to estimate the posterior probability distribution of each coefficient ($\beta$). A Bayesian approach also minimizes the problems created by non-independent observations often present in time-series data since the analysis does not rely on p-values that require independent observations. These analyses were carried out in the Stan computational framework (http://mc-stan.org/) accessed via the brms package (Bürkner, 2017) in R.

## Results

**Effects of modality, music style, and musical training on mean synchrony ratings.** A 3-way mixed ANOVA was run to investigate the effects of modality (AV/AO/VO) of stimulus presentation, music style (pulsed/non-pulsed), and musical training (low/high) on mean synchrony ratings, with the Benjamini-Hochberg FDR correction applied for multiple comparisons. Pearson correlations computed between the mean synchrony ratings across the two trials of each stimulus for each participant showed good agreement between trial 1 and 2 for most participants (mean $r$ = .66, SD = .22, range = - .06 to .94).

As in Experiment 1, a main effect of music style ($F$(1, 47) = 124.92, $p$ < .001, $\eta_p^2$ = .73) revealed that overall ratings of synchrony were higher for the pulsed ($M$ = 0.47, $SD$ = 0.47) than non-pulsed music ($M$ = -0.06, $SD$ = 0.52). Unlike in Experiment 1, a main effect of modality also emerged ($F$(2, 94) = 7.15, $p$ = .001, $\eta_p^2$ = .13), in which the VO stimuli ($M$ = 0.11, $SD$ = 0.55) were rated significantly lower in synchrony than the AO ($M$ = 0.20, $SD$ = 0.57) and AV stimuli ($M$ = 0.19, $SD$ = 0.56) in Bonferroni-corrected, paired-samples t-tests (VO vs. AO $p$ = .009, VO vs. AV $p$ = .03, AO vs. AV $p$ > 1). As before, no main effect of musical training ($F$(1, 47) = 3.22, $p$ = .08, $\eta_p^2$ = .06) was found. In follow-up analyses, the number of years of musical training reported by each participant was found to be positively correlated with mean synchrony ratings (Kendall's τ (47)= .25, $p$ = .02). As in Experiment 1, this difference appears to be driven primarily by responses to the non-pulsed music style (non-pulsed: Kendall's τ (47)= .29, $p$ = .01, pulsed: Kendall's τ (47)= .07, $p$ = .51).

As in Experiment 1, a significant two-way interaction was present between modality and music style ($F$(2, 94) = 25.65, $p$ < .001, $\eta_p^2$ = .35). For the pulsed music style, synchrony ratings were significantly lower in the VO modality ($M$ = 0.31, $SD$ = 0.51) than both the AV ($M$ = 0.53, $SD$ = 0.46; $p$ < .001) and AO modalities ($M$ = 0.59, $SD$ = 0.39; $p$ < .001), with no statistically significant difference between the AV and AO modalities ($p$ = .82) in Bonferroni-corrected, paired-samples t-tests. There were no significant differences in analogous tests comparing synchrony ratings for the non-pulsed music style between the VO ($M$ = -0.04, $SD$ = 0.54), AV ($M$ = -0.06, $SD$ = 0.50) and AO modalities ($M$ = -0.09, $SD$ = 0.52; $p$s > 1, see Figure 4). The two-way interactions between modality and musical training

($F(2, 94) = 0.18$, $p = .84$, $\eta_p^2 = .004$) and between music style and musical training ($F(1, 47) = 0.25$, $p = .62$, $\eta_p^2 = .005$) were not statistically significant, nor was the three-way interaction of all the independent variables ($F(2, 94) = 0.96$, $p = .39$, $\eta_p^2 = .02$).

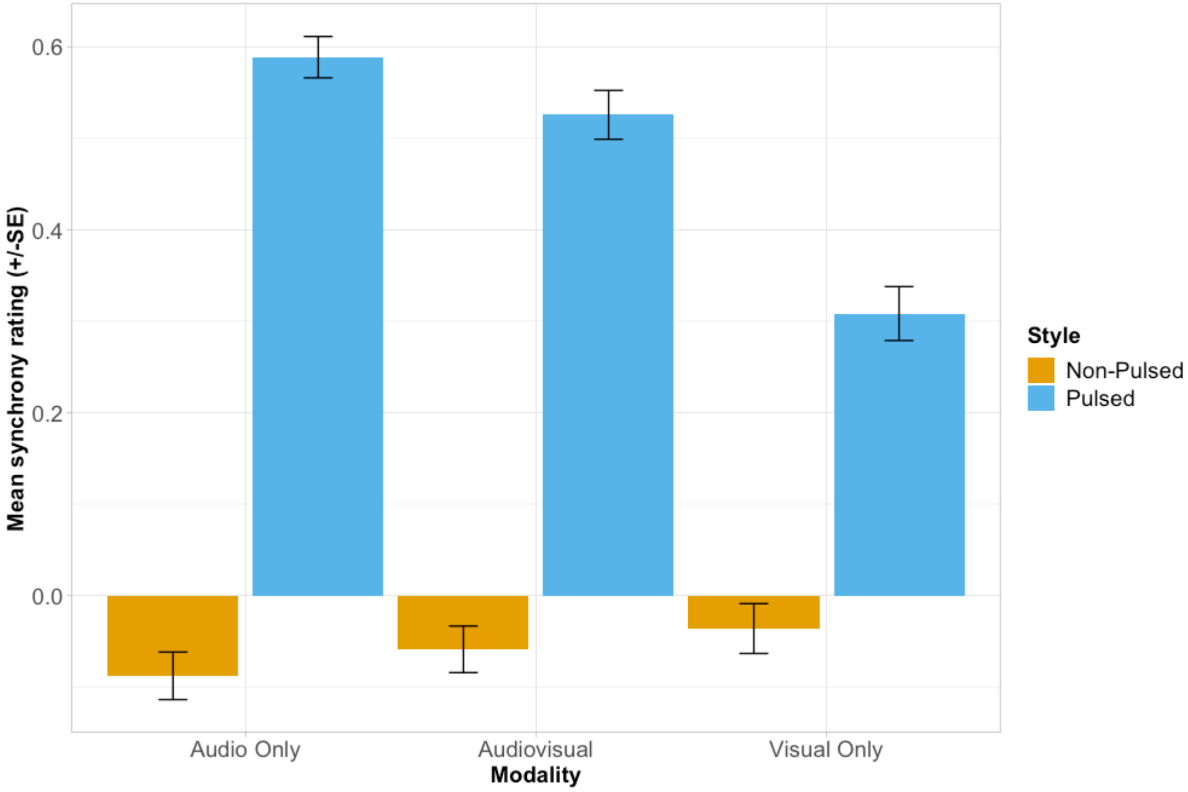

Figure 4. Mean synchrony ratings (+/-SE) by modality and music style. Synchrony ratings were measured on a scale of -1 (“less in sync”) to 1 (“more in sync”).

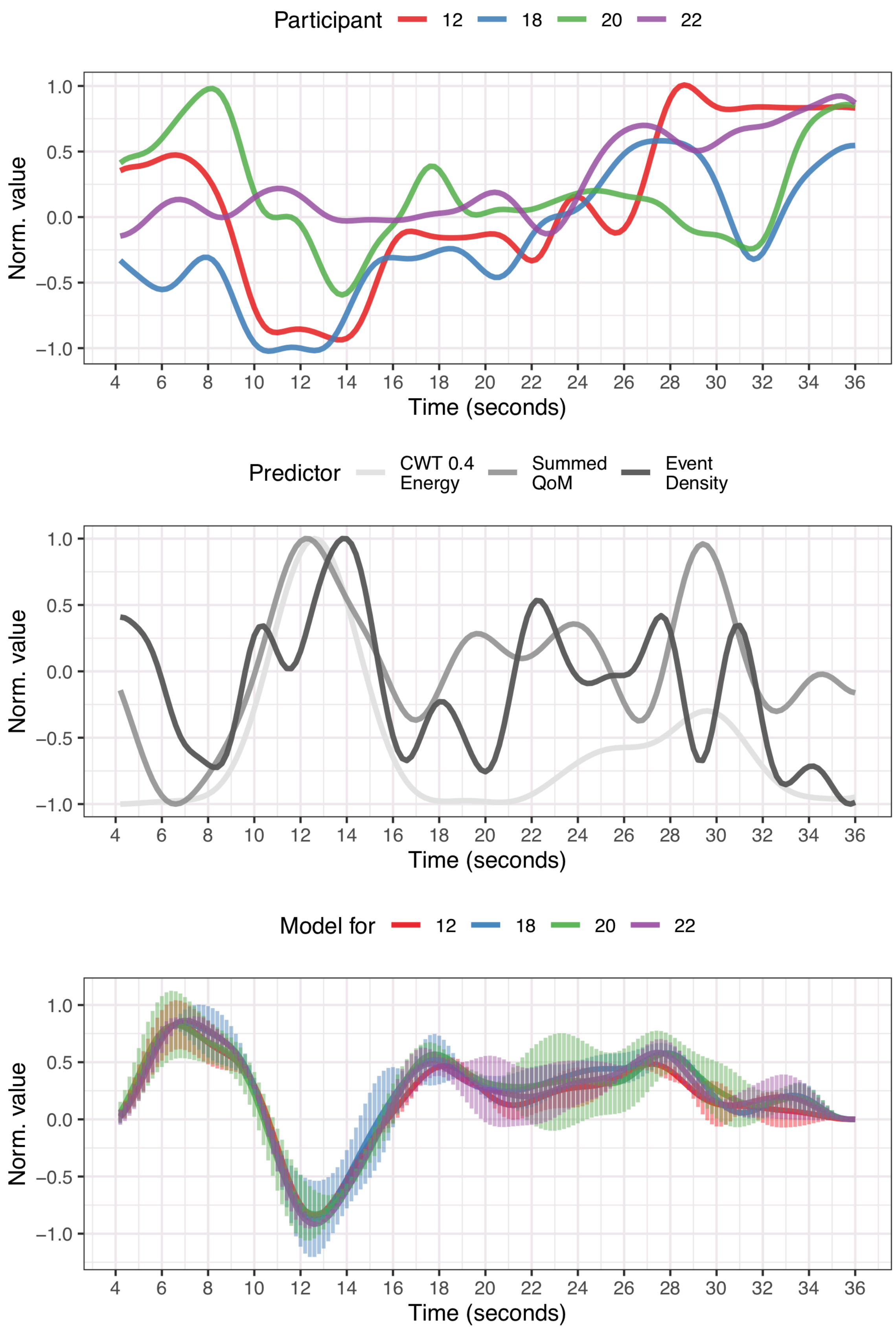
Participant 12 18 20 22
Norm. value
1.0
0.5
0.0
−0.5
−1.0
4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36
Time (seconds)
Predictor CWT 0.4 Energy Summed QoM Event Density
Norm. value
1.0
0.5
0.0
−0.5
−1.0
4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36
Time (seconds)
Model for 12 18 20 22
Norm. value
1.0
0.5
0.0
−0.5
−1.0
4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36
Time (seconds)

*Figure 5*. One example trial (Pulsed Music Style; Musicians11&12-Take21) showing synchrony ratings from four participants displaying a typical response profile (i.e. similar to the overall response pattern across the sample) for this trial in the Audiovisual modality (top panel), three of the visual and audio features (middle panel), and model predictions from the multilevel Bayesian regression analysis for each participant using these three features (bottom panel; bars represent standard deviation, model prediction series have been smoothed using the same method applied to the rating and feature data).

**Predicting continuous ratings of synchrony with audio and visual stimulus features.** In line with our prior analyses and due to the significant differences in music style revealed previously, we built separate multilevel Bayesian regression models for pulsed and non-pulsed music styles. Figure 5 displays an example trial for one stimulus, showing the resampled and smoothed synchrony rating data for four participants, examples of the audio and visual features, and the resultant Bayesian regression model predictions for these features and participants. For predicting synchrony ratings in the AV modality, we entered the four audio and four visual features into the model. For the VO and AO modalities, only the modality-relevant (visual or audio) features were used. All models were constructed without interaction terms between the predictors since preliminary analyses indicated that such interactions added only minimal predictive power while increasing the complexity of the models.

Overall, the models achieved moderately good levels of prediction across the genres, $R^2 = 0.356$ (CI$_{95\%}$ 0.351–0.361) in the non-pulsed music style and $R^2 = 0.406$ (CI$_{95\%}$ 0.400–0.411) in pulsed music using all features to predict the synchrony ratings in the AV modality. Looking at the individual features (see Figure 6), it is clear that the models are driven primarily by the visual features, particularly Summed WT Energy, CWT 0.4 Hz, and Summed QoM, where the distributions of the estimates (betas) are furthest away from zero, although event density also makes some positive contribution in both music styles. In Figure 7, we used the same approach to predict synchrony ratings in the AO modality using the audio features (top panel) and synchrony ratings in the VO modality using the visual features (bottom panel). The model prediction rates using the four audio features in the AO Modality were similar across music styles, $R^2 = 0.400$ (CI$_{95\%}$ 0.395–0.406) for pulsed music and $R^2 = 0.400$ (CI$_{95\%}$ 0.396–0.406) for non-pulsed music. In the VO modality, the models achieved a similar level of moderate success, although this varied across the styles, $R^2 = 0.411$ (CI$_{95\%}$ 0.405–0.415) for pulsed music and $R^2 = 0.362$ (CI$_{95\%}$ 0.357–0.367) for non-pulsed music. The same features possessed the largest beta estimates as in the AV modality (Summed WT Energy, Summed QoM). The visual features operated consistently across the two music styles (with the exception of Summed QoM), while effects of the audio features varied more by music style (i.e. spectral flux and RMS energy had effects in opposite directions for pulsed vs. non-pulsed music, see Figure 7).

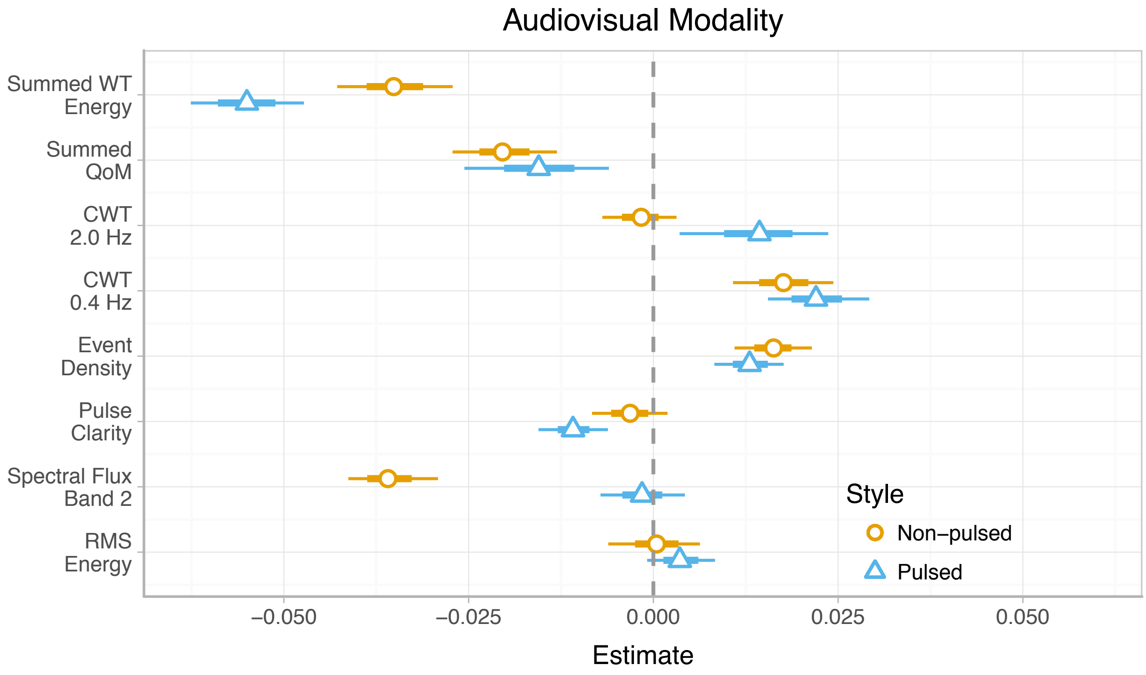

*Figure 6*. Mean, 66% (thick line) and 95% (thin line) quantile intervals of the beta ($\beta$) estimates from the models using all audio and visual features to predict synchrony ratings in the AV modality.

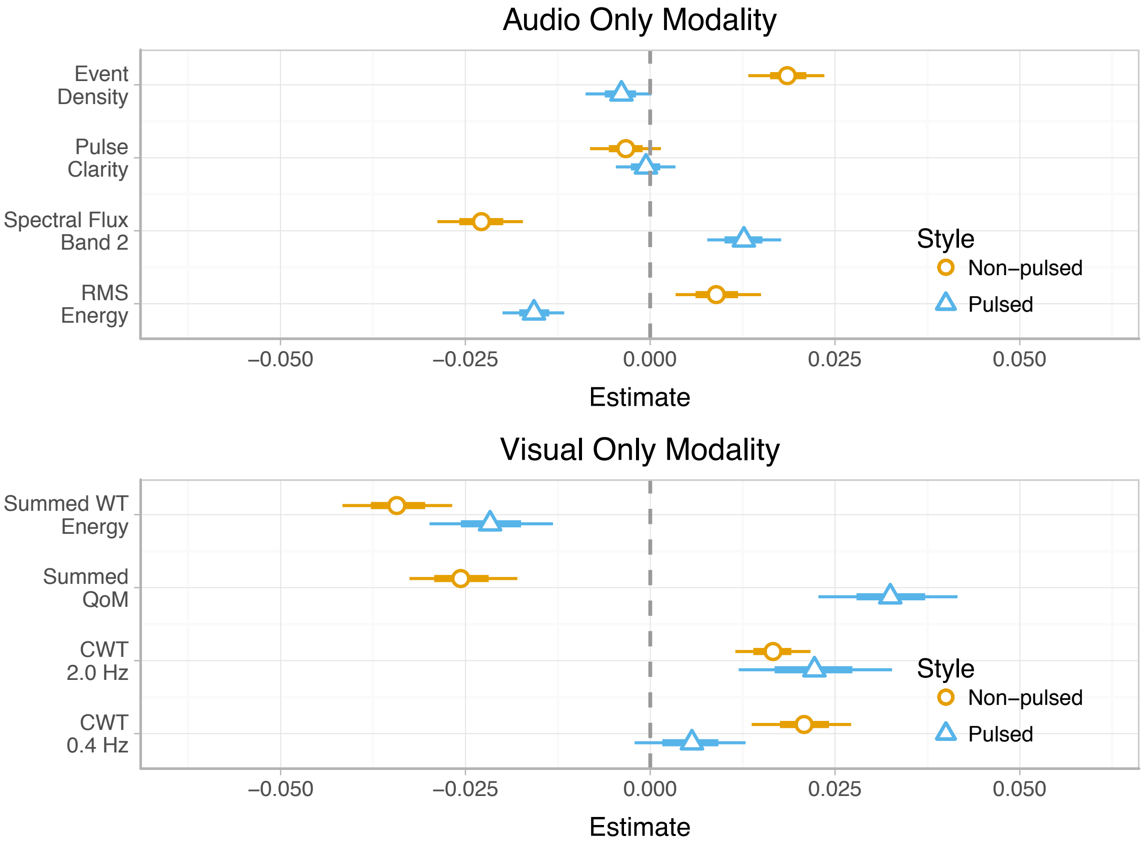

*Figure 7*. Mean, 66% and 95% quantile interval of the beta ($\beta$) estimates from the models using audio features to predict synchrony ratings in the Audio Only modality (top panel) and visual features to predict synchrony ratings in the Visual Only modality (bottom panel).

## Discussion

Analysis of the mean synchrony ratings across these longer clips primarily revealed results that parallel those of Experiment 1. In particular, pulsed music was perceived as significantly more synchronous than non-pulsed music and the differences in synchrony ratings between these music styles were more extreme when auditory information was present than when only visual cues were present. In addition, a main effect of modality was found in Experiment 2 that was not present in Experiment 1. Mean synchrony ratings were higher when auditory information was present, and this effect seems to be primarily driven by the pulsed music stimuli (see Figure 4). It may be that participants found it more difficult to judge musical synchrony in the absence of auditory cues, and were thus inclined to use the middle of the continuous rating scale more than the extremes. These initial results indicate that participants were relying on similar cues for making continuous ratings of synchrony as the participants in Experiment 1 who were asked to make an overall (global) rating at the end of a short clip of music.

The methodology of Experiment 2 also allowed us to investigate the influence of audio and visual features on a moment-to-moment basis. In terms of visual features, the results indicate that, in general, higher perceived synchrony was associated with a lower amount of overall movement (Summed QoM), lower summed periodic movement across the two performers in any frequency band (Summed WT Energy), and increased mutual movement in the same frequency band (CWT 0.4 Hz and CWT 2.0 Hz). These results follow a quite similar pattern regardless of whether the musical stimuli were presented in an audiovisual (see Figure 6) or visual only (see Figure 7) version, with perhaps the most notable exception being the difference in the effect of Summed QoM on ratings of pulsed vs. non-pulsed music in the VO modality; pulsed music with more overall movement was rated as more synchronous, whereas the opposite pattern emerged in the non-pulsed music. Future research is needed to explore the extent to which this result could be related to baseline differences in Summed QoM between the two styles (Experiment 1 showed non-pulsed music to be significantly higher in Summed QoM overall) or subtle differences in the way larger versus smaller movements are employed across the styles.

When considering results of the visual features as a whole (across both AV and VO modalities), however, it appears that movements that are unrelated (in terms of quantity or period) across the performers decrease ratings of perceived synchrony, whereas similar periodic movements between the performers increase the synchrony ratings. Since the Summed QoM measure simply takes account of the amount of motion summed across both performers, this measure often picks up on large gestures made by a single performer, which does not necessarily entail interpersonal synchrony (and indeed could signal the opposite, if such gestures are being used to get the performers back together or navigate a difficult transition point, for example). Similarly, Summed WT Energy captures moments when performers exhibit clear periodic movements but the actual periodicities are not necessarily related between the two performers. In contrast, when both performers were moving broadly at the same or related frequencies, using either relatively slow or fast movements (as indexed by the CWT 0.4 Hz and CWT 2.0 Hz features, respectively), the participants perceived the performers to exhibit higher synchrony.

The audio features used here were generally less successful in predicting synchrony ratings across the trials and participants than the visual features. Greater event density was related to higher synchrony ratings in both the AV and AO modalities (with one exception for pulsed music in AO). This could indicate that participants were rating passages as more synchronous when both instruments were playing more often, or that perceived synchrony is related to tempo. This finding aligns with results of a large corpus study of several musical styles indicating that synchronization between instrument pairings typically increases with greater event density (Clayton et al., in review). In addition, previous research indicates that asynchronies tend to be larger and more variable at slow than at fast tempi (Rasch, 1988) and, while such general scaling effects might not be readily noticeable due to similar constraints operating on perceptual processes (Repp, 2006a), the likelihood of occasional, atypically large and perceptually salient asynchronies may increase at slower tempi due to increased temporal uncertainty (Repp, 2006b).

In the pulsed music style, low-frequency (band 2) spectral flux showed the predicted pattern of increasing synchrony ratings in the AO modality, with no significant effect in the AV modality. In non-pulsed music, for both AO and AV modalities, greater low-frequency spectral flux was related to *lower* synchrony ratings. Although this result was not predicted, previous studies demonstrating the importance of low-frequency spectral flux in perception of rhythmicity have all used strongly beat-based music (e.g., Burger et al., 2013; Stupacher, et al., 2013), thus instances of increased low-frequency spectral flux in the non-pulsed music used here may be indexing something different than a strong beat or rhythmic drive. Pulse clarity and RMS energy showed fewer pronounced effects across the datasets, although RMS energy made some contribution in predicting the AO ratings, showing a positive relationship to synchrony ratings in non-pulsed music and negative relationship to synchrony ratings in pulsed music. This may be simply an artefact of this relatively specific dataset (e.g., these 10 pulsed music performers happened to play more quietly during more synchronized sections and vice versa for the 10 non-pulsed performers) and may not be generalizable to these music genres as a whole, but should be explored further in future research.

## General Discussion

In this paper we have presented two experiments investigating global ratings (Experiment 1) and continuous ratings (Experiment 2) of perceived synchrony in natural musical duo performances. The results of these experiments have revealed that synchrony judgments are affected by the reliance of the music on a regular pulse, the modality of stimulus presentation, and specific visual and auditory features of the stimuli. These results offer insights into the factors that influence the perception of interpersonal synchrony in musical ensembles, as well as methodological insights comparing global to continuous ratings of perceived musical synchrony.

The two styles of music utilized here are similar in terms of their improvisatory nature; both styles also comprise instrumental duos that were filmed in the same location under the same conditions. A primary feature on which these performances vary is their reliance on (in the case of the pulsed music) or avoidance of (in the case of the non-pulsed music) a regular, predictable beat. Pronounced main effects of music style across both experiments indicate that participants perceived music with a regular beat as significantly more synchronous than music that aimed to eschew the induction of such a beat. This bears some similarity to previous evidence indicating that the presence of a regular rhythmic framework can aid perceptual judgments about the timing of tone pairs (e.g., Jones, Kidd, & Wetzel, 1981) and general theories on how regular rhythmic contexts can guide attention to specific points in time (e.g., Large & Jones, 1999). Interestingly, the interaction of music style with modality of stimulus presentation that was found in both experiments indicated that participants' judgments were more similar across the two music styles in the condition in which only visual information was available. This shows some parallels to the results reported in Experiment 2 in which the visual features predicted synchrony ratings in a more similar way across music styles than the audio features (see Figures 6 and 7). Therefore, it seems that the visual cues used across both music styles communicate synchrony in a relatively similar way, whereas the auditory cues affect synchrony judgments somewhat differently across the styles. The generalizability of visual cues to synchrony across styles suggests that co-performer body motion during ensemble performance, like body gestures during spoken conversation (Abner et al., 2015), contains spatiotemporal features that signal the effectiveness of social communication in a reliable manner.

Musical training (primarily defined here in terms of low/high musical training groups) did not have a substantial influence on synchrony ratings. For global ratings in Experiment 1, the interaction of musical training with modality indicated that participants more highly trained in music exhibited greater tolerance to the relatively unpredictable auditory conditions of the non-pulsed music style, and were thus more likely to give higher synchrony ratings than less musically trained participants (see Figure 1), although this interaction was not replicated by continuous ratings in Experiment 2. Similarly, correlational analyses showed that participants with more years of musical training gave higher synchrony ratings on average, in particular for the non-pulsed music. Future research should investigate such effects in groups with more extreme differences in their levels of musical expertise (including groups with expertise in performing the specific music styles in question), as well as how such groups make use of different auditory and visual cue combinations in judging musical synchrony.

We also explored the effects of auditory and visual features of the individual stimuli on ratings of perceived synchrony. The overall results across all modalities of stimulus presentation in Experiment 2 indicated that visual information provided more salient cues for synchrony ratings than the audio features that we utilized. While this visual dominance is generally consistent with studies of aesthetic judgments for solo performance (Davidson, 1993; Tsay, 2013), the results of Experiment 2 contrast somewhat with the initial correlational results reported in Experiment 1 (see Figures 2 and 3). This could relate to the fact that the music styles were analyzed separately in Experiment 2, allowing different effects to be detected than in the relatively small dataset utilized in Experiment 1. Or this could relate to the fact that Experiment 1 employed short (10 s) sound clips (for the global rating task), which may have increased the salience of auditory cues over the visual cues. In particular, the visual cues used in our study were based on movements that unfold relatively slowly in comparison to the auditory cues. Therefore, the longer clips used for continuous ratings in Experiment 2 may have afforded the opportunity for the visual cues to exhibit a more pronounced effect than that initially seen in Experiment 1. This result bears some parallels to the work of Wapnick et al. (2005), who found differences in ratings of performance quality for shorter (20 s) versus longer (60 s) clips of the same solo piano performances. In the timing domain more specifically, a distinction can be made between entrainment mechanisms at different time-scales: sensorimotor synchronization which enables sound event onsets to be closely aligned and depends for its precision primarily on auditory information, and coordination at section boundaries (points of change in the music) which may be detectable in coordinated body movement (Clayton et al., in review). The greater salience of visual information in the longer clips may be related to this distinction.

In Experiment 2, coarse movement features that captured simply the summed amount of upper body movement across the two performers (Summed QoM) or the summed amount of periodic movement at frequencies that were not necessarily related (Summed WT Energy) were inversely related to synchrony ratings, whereas more specific measures of co-occurring periodic movement of the two performers at related slow (CWT 0.4 Hz) or fast (CWT 2.0 Hz) frequencies were positive predictors of synchrony ratings. This finding extends previous results from Eerola et al. (2018), who found that CWT analysis of performances from the same corpus used here could effectively predict ratings of ‘bouts of interaction’ between co-performers, whereas a measure of summed QoM was not an effective predictor of these interactions. Thus, it appears that the perception of visual synchrony in musical performances is not increased simply by the amount of movement seen, but rather by the detection of co-occurring movements of the performers at similar frequencies. Future research should extend this approach to further differentiate between ancillary movements and sound-producing movements (e.g., bow strokes, key presses) using additional technologies such as motion capture.

Although the audio features used here were less consistent predictors of synchrony ratings, we were able to obtain similar prediction rates for the AO rating data as in the VO rating data (Figure 7). Some of the effects were as predicted, including a generally positive association between synchrony ratings and event density and a positive relationship between low-frequency spectral flux in the pulsed music in the AO condition. The other audio features, RMS energy and pulse clarity, were less consistent predictors of synchrony ratings across the datasets. This is despite the fact that these two features demonstrated initial positive correlations with synchrony ratings in Experiment 1 (see Figure 2), which disappeared when the music styles were treated separately in Experiment 2. Future research should include additional audio features in an attempt to increase these prediction rates; in particular, our study was limited by the lack of separate audio tracks for the two instruments from which onsets of each instrument could be extracted and compared.

This study also represents, to our knowledge, the first published comparison between global ratings of musical synchronization made retrospectively and continuous ratings of synchronization made as a performance unfolded. Although we made use of different stimuli in each experiment to account for the demands of the different tasks (i.e., shorter files for the global ratings to decrease memory demands), the fact that these stimuli were drawn from the same corpus allowed us to compare across the two experiments using an otherwise identical design for each. An important methodological finding here is that the results of Experiment 2 closely paralleled those of Experiment 1 when the mean synchrony ratings across the clips from Experiment 2 were used as the dependent variable. This suggests that studies employing an overall synchrony rating at the end of a short clip of music, despite memory demands, can provide a reliable estimate of what can be achieved using a more complicated paradigm of rating synchrony continuously over the course of longer clips. However, a main advantage of the continuous rating method was that it allowed us to examine how changes in features of the stimuli covaried with changes in perceived synchrony over time. Given that fluctuating patterns of tension and relaxation evoked by musical features influence aesthetic appreciation and enjoyment (Huron, 2006; Lehne & Koelsch, 2015; Meyer, 1956), tracking the dynamics of the relationship between objective and subjective measures of musical properties is a necessary step towards understanding the mechanisms that underlie musical communication in general.

Although the recordings used as stimuli here are relatively high in ecological validity, the disadvantage of using natural performances in such a design is that it entails a correlational approach rather than an experimental approach in which the controlled manipulation of features of interest allows causal relations between stimuli and responses to be discerned. Future experimental research could take this work as a starting point to introduce more pronounced manipulations in the stimuli (e.g., making recordings in which performers are asked to vary their movements or sounds in specific ways to affect movement coordination, event density, loudness, etc.). Such research could help to further disentangle effects that were conflicting or non-significant in the present work, which may have been limited by the range of variation in certain features in these particular performances. In addition, although the different instruments featured within the duo recordings in our study reflect the natural diversity of instruments used for performing these music styles, more work is needed to understand the extent to which instrument-specific factors (e.g. movement patterns and playing styles that are idiosyncratic to a particular instrument) may also impact on synchrony judgments.

In conclusion, this work has revealed new insights on the perception of synchrony in musical ensemble performance and made methodological progress in exploring the possibilities afforded by collecting musical synchrony rating data via a continuous response paradigm. We have demonstrated that the perception of musical synchrony in duos is affected by various features of stimulus, including the presence of a regular beat and co-occurring periodic movements of the performers, which vary somewhat depending on the degree to which auditory and/or visual information is available to the rater. Charting the relationship between objective and subjective measures of ensemble synchrony is informative about social communicative processes that mediate interactions between groups of co-performers and audiences. More broadly, exploring this relationship may hold the key to understanding the adaptive functions of collective musical behavior in evolution, for example, with reference to the role of coordinated group activity in signaling coalition strength to potential allies or competitors (Hagen & Bryant, 2003). In terms of methodological advancement, our study revealed several parallels between a method requiring a single rating of synchrony across a clip and one implicating a rating of changes in synchrony over time. These results open new directions for further exploration into the factors that affect the perception of musical synchrony, which is an important area of research that can be used to inform our understanding of higher-level, socio-cognitive processes such as judgments of performance quality and musical preferences.

# References

Abner, N., Cooperrider, K., & Goldin-Meadow, S. (2015). Gesture for linguists: A handy primer. *Language and Linguistics Compass, 9*(11), 437-451. doi:10.1111/lnc3.12168

Alluri, V., & Toiviainen, P. (2010). Exploring perceptual and acoustical correlates of polyphonic timbre. *Music Perception: An Interdisciplinary Journal*, *27*(3), 223-242.

Arrighi, R., Alais, D., & Burr, D. (2006). Perceptual synchrony of audiovisual streams for natural and artificial motion sequences. *Journal of Vision*, *6*(3), 6-6.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B, 57,* 289–300.

Bernieri, F. J. (1988). Coordinated movement and rapport in teacher-student interactions. *Journal of Nonverbal Behavior*, *12*(2), 120-138.

Bishop, L., Cancino-Chacón, C., & Goebl, W. (2019). *Moving to communicate, moving to interact: Patterns of body motion in musical duo performance.* Music Perception, 37(1), 1-25. doi:10.1525/mp.2019.37.1.1

Bishop, L., & Goebl, W. (2018). Communication for coordination: gesture kinematics and conventionality affect synchronization success in piano duos. *Psychological Research, 82*(6), 1177–1194. doi:10.1007/s00426-017-0893-3

Burger, B., Ahokas, R., Keipi, A., & Toiviainen, P. (2013). Relationships between spectral flux, perceived rhythmic strength, and the propensity to move. In *Proceedings of the Sound and Music Computing Conference 2013, SMC 2013, Stockholm, Sweden*. Logos Verlag Berlin.

Butterfield, M. (2010). Participatory discrepancies and the perception of beats in jazz. *Music Perception: An Interdisciplinary Journal*, *27*(3), 157-176.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80(1)*, 1–28.

Carpenter, B., Gelman, Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software* 76(1). doi: 10.18637/jss.v076.i01

Chang, A., Kragness, H. E., Livingstone, S. R., Bosnyak, D. J., & Trainor, L. J. (2019). Body sway reflects joint emotional expression in music ensemble performance. *Scientific Reports, 9*(1), 205. doi:10.1038/s41598-018-36358-4

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*(6), 893-910.

Clayton, M., Jakubowski, K., Eerola, T., Keller, P.E., Camurri, A., Volpe, G., & Alborno, P. (in review). Interpersonal entrainment in music performance: Theory, method and model. *Music Perception*.

Clayton, M., Sager, R. & Will, U. (2005). In time with the music: The concept of entrainment and its significance for ethnomusicology. *European Meetings in Ethnomusicology 11* (ESEM Counterpoint 1): 1-82.

Cramer, A. O., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P., ... & Wagenmakers, E. J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review*, *23*(2), 640-647.

Dahl, S., & Friberg, A. (2007). Visual perception of expressiveness in musicians' body movements. *Music Perception: An Interdisciplinary Journal, 24*(5), 433-454. doi: 10.1525/mp.2007.24.5.433

Davidson, J. W. (1993). Visual perception of performance manner in the movements of solo musicians. *Psychology of Music, 21*(2), 103-113. doi: 10.1177/030573569302100201

Dean, R. T., & Bailes, F. (2016). Modeling perceptions of valence in diverse music. *Roles of Acoustic Features, Agency, and Individual Variation, 34*(1), 104-117. doi:10.1525/mp.2016.34.1.104

Dean, R. T., Bailes, F., & Dunsmuir, W. T. (2014). Time series analysis of real-time music perception: Approaches to the assessment of individual and expertise differences in perception of expressed affect. *Journal of Mathematics and Music, 8*(3), 183-205.

Eerola, T., Jakubowski, K., Moran, N., Keller, P. E., & Clayton, M. (2018). Shared periodic performer movements coordinate interactions in duo improvisations. *Royal Society Open Science*, *5*(2), 171520.

Engel, A., Hoefle, S., Monteiro, M. C., Bramati, J. E., Lima, D. O., Keller, P. E., & Moll, J. (2014). Feeling the groove: perceiving asynchronies in drumming sounds typical for carnival parades of Brazilian Samba Schools. Paper presented at the *The Neurosciences and Music - V,* Dijon, France.

Engel, A., Hoefle, S., Monteiro, M. C., Moll, J., & Keller, P. E. (2014). The influence of loudness and synchrony on pleasure during listening to the drumming section of a Brazilian Samba School. Paper presented at the *Simpósio de Cognição e Artes Musicais* – SIMCAM, UNICAMP, Campinas, Brazil.

Farnebäck G. (2003). Two-frame motion estimation based on polynomial expansion. In: J. Bigun & T. Gustavsson (Eds.) *Image Analysis. SCIA 2003. Lecture Notes in Computer Science*, *vol. 2749*. Berlin/Heidelberg: Springer. doi:10.1007/3-540-45103-X_50.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, *8*(1), 8-11.

Glowinski, D., Mancini, M., Cowie, R., Camurri, A., Chiorri, C., & Doherty, C. (2013). The movements made by performers in a skilled quartet: A distinctive pattern, and the function that it serves. *Frontiers in Psychology*, 4, 841.

Goebl, W., & Palmer, C. (2009). Synchronization of timing and motion among performing musicians. *Music Perception: An Interdisciplinary Journal*, *26*(5), 427-438.

Goebl, W., & Parncutt, R. (2001). Perception of onset asynchronies: Acoustic piano versus synthesized complex versus pure tones. Paper presented at the *Meeting of the Society for Music Perception and Cognition (SMPC2001)*, Kingston, Ontario, Canada.

Hadley, L. V., Sturt, P., Moran, N., & Pickering, M. J. (2018). Determining the end of a musical turn: Effects of tonal cues. *Acta Psychologica, 182,* 189-193.

Hagen, E. H., & Bryant, G. A. (2003). Music and dance as a coalition signaling system. *Human Nature, 14*(1), 21-51.

Halpern, A. R., & Bartlett, J. C. (2010). Memory for melodies. In M. R. Jones, R. R. Fay, & A. N. Popper (Eds.), *Music perception.* (pp. 233-258). New York, NY, US: Springer Science + Business Media.

Hirsh, I. J. (1959). Auditory perception of temporal order. *The Journal of the Acoustical Society of America, 31(6)*, 759-767. doi:10.1121/1.1907782

Holcombe, A. O. (2009). Seeing slow and seeing fast: Two limits on perception. *Trends in Cognitive Sciences, 13(5)*, 216-221.

Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation*. Cambridge, MA: The MIT Press.

Iyer, V. (2002). Embodied mind, situated cognition, and expressive microtiming in African-American music. *Music Perception: An Interdisciplinary Journal*, *19*(3), 387-414.

Jakubowski, K., Eerola, T., Alborno, P., Volpe, G., Camurri, A., & Clayton, M. (2017). Extracting coarse body movements from video in music performance: a comparison of automated computer vision techniques with motion capture data. *Frontiers in Digital Humanities*, *4*, 9.

Jones, M. R., Kidd, G., & Wetzel, R. (1981). Evidence for rhythmic attention. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(5), 1059-1073.

Kawase, S. (2014). Gazing behavior and coordination during piano duo performance. *Attention, Perception & Psychophysics*, 76(2), 527-540. doi:10.3758/s13414-013-0568-0.

Keller, P. E. (2014). Ensemble performance: Interpersonal alignment of musical expression. In D. Fabian, R. Timmers, & E. Schubert: *Expressiveness in music performance: Empirical approaches across styles and cultures* (pp. 260-282). Oxford: Oxford University Press.

Keller, P. E., & Appel, M. (2010). Individual differences, auditory imagery, and the coordination of body movements and sounds in musical ensembles. *Music Perception, 28*, 27-46. doi:10.1525/mp.2010.28.1.27

King, E., & Ginsborg, J. (2011). Gestures and glances: interactions in ensemble rehearsal. In E. King (Ed.), *New Perspectives On Music And Gesture* (pp. 177-201). Aldershot: Ashgate Press.

Labbé, C., & Grandjean, D. (2014). Musical emotions predicted by feelings of entrainment. *Music Perception: An Interdisciplinary Journal*, *32*(2), 170-185.

Large, E. W., & Jones, M. R. (1999). The dynamics of attending: How people track time-varying events. *Psychological review*, *106*(1), 119.

Lartillot, O., Eerola, T., Toiviainen, P., & Fornari, J. (2008). Multi-feature modeling of pulse clarity: Design, validation and optimization. In *ISMIR* (pp. 521-526).

Lartillot, O., Toiviainen, P., & Eerola, T. (2008). A Matlab toolbox for music information retrieval. In *Data analysis, machine learning and applications* (pp. 261-268). Berlin/Heidelberg: Springer.

Latif, N., Barbosa, A. V., Vatiokiotis-Bateson, E., Castelhano, M. S., & Munhall, K. G. (2014). Movement coordination during conversation. *PLoS One*, *9*(8), e105036.

Lehne, M., & Koelsch, S. (2015). Toward a general psychological model of tension and suspense. *Frontiers in Psychology, 6*, 79-79. doi:10.3389/fpsyg.2015.00079

Levinson, S. C. (2016). Turn-taking in human communication–origins and implications for language processing. *Trends in Cognitive Sciences*, *20*(1), 6-14.

Macrae, C. N., Duffy, O. K., Miles, L. K., & Lawrence, J. (2008). A case of hand waving: Action synchrony and person perception. *Cognition*, *109*(1), 152-156.

MacRitchie, J., Varlet, M., & Keller, P. E. (2017). Embodied expression through entrainment and co-representation in musical ensemble performance. In M. Lesaffre, P. J. Maes, & M. Leman (Eds.), *The Routledge Companion to Embodied Music Interaction* (pp. 150-159). New York: Routledge.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314-324.

Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H., & Gross, J. J. (2005). The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion, 5(2)*, 175-190.

Metallinou, A., Katsamanis, A., & Narayanan, S. (2013). Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image and Vision Computing, 31(2)*, 137-152.

Meyer, L. B. (1956). *Emotion and meaning in music*. Chicago: University of Chicago Press.

Moran, N., Hadley, L. V., Bader, M., & Keller, P. E. (2015). Perception of 'Back-Channeling' Nonverbal Feedback in Musical Duo Improvisation. *PLoS ONE*, 10(6), e0130070. doi:10.1371/journal.pone.0130070

Moran, N., Jakubowski, K., Keller, P.E. (2017). *Improvising Duos - visual interaction collection*, 2011 [moving image]. University of Edinburgh. http://dx.doi.org/10.7488/ds/2153.

Palmer, C. (1989). Mapping musical thought to musical performance. *Journal of Experimental Psychology: Human Perception and Performance, 15*(2), 331-346.

Petrini, K., Dahl, S., Rocchesso, D., Waadeland, C. H., Avanzini, F., Puce, A., & Pollick, F. E. (2009a). Multisensory integration of drumming actions: musical expertise affects perceived audiovisual asynchrony. *Experimental brain research*, *198*(2-3), 339.

Petrini, K., Russell, M., & Pollick, F. (2009b). When knowing can replace seeing in audiovisual integration of actions. *Cognition*, *110*(3), 432-439.

Phillips-Silver, J., & Keller, P. E. (2012). Searching for roots of entrainment and joint action in early musical interactions. *Frontiers in Human Neuroscience, 6*, 26. doi:10.3389/fnhum.2012.00026

Platz, F., & Kopiez, R. (2012). When the eye listens: A meta-analysis of how audio-visual presentation enhances the appreciation of music performance. *Music Perception: An Interdisciplinary Journal*, *30*(1), 71-83.

Ragert, M., Fairhurst, M. T., & Keller, P. E. (2014). Segregation and integration of auditory streams when listening to multi-part music. *PLoS ONE, 9*, e84085. doi:10.1371/journal.pone.0084085

Rasch, R. A. (1988). Timing and synchronization in ensemble performance. In J. A. Sloboda (Ed.), *Generative Processes in Music: The Psychology of Performance, Improvisation, and Composition* (pp. 70-90). Oxford, UK: Clarendon Press.

Repp, B. H. (1992). Probing the cognitive representation of musical time: Structural constraints on the perception of timing perturbations. *Cognition, 44*, 241-281.

Repp, B. H. (1995). Detectability of duration and intensity increments in melody tones: A possible connection between music perception and performance. *Perception & Psychophysics, 57*, 1217-1232.

Repp, B. H. (1999). Detecting deviations from metronomic timing in music: Effects of perceptual structure on the mental timekeeper. *Perception & Psychophysics*, *61*(3), 529-548.

Repp, B. H. (2006a). Musical synchronization. In E. Altenmüller, M. Wiesendanger, & J. Kesselring (Eds.), *Music, motor control, and the brain* (pp. 55-76). Oxford, UK: Oxford University Press.

Repp, B. H. (2006b). Rate limits of sensorimotor synchronization. *Advances in Cognitive Psychology, 2,* 163-181. doi:10.2478/v10053-008-0053-9

Repp, B. H., & Keller, P. E. (2008). Sensorimotor synchronization with adaptively timed sequences. *Human movement science*, *27*(3), 423-456.

Repp, B. H., & Su, Y. H. (2013). Sensorimotor synchronization: a review of recent research (2006–2012). *Psychonomic Bulletin & Review, 20*(3), 403-452.

Richardson, D. C., Dale, R., & Kirkham, N. Z. (2007). The art of conversation is coordination. *Psychological Science*, *18*(5), 407-413.

Rosen, S., & Howell, P. (1987). Is there a natural sensitivity at 20 ms in relative tone-onset-time continua? A reanalysis of Hirsh’s (1959) data. In *The psychophysics of speech perception* (pp. 199-209). Dordrecht: Springer.

Schubert, E. (2004). Modeling perceived emotion with continuous musical features. *Music Perception: An Interdisciplinary Journal, 21*(4), 561-585. doi:10.1525/mp.2004.21.4.561

Shaffer, L. H. (1984). Timing in solo and duet piano performances. *Quarterly Journal of Experimental Psychology, 36A*, 577-595.

Shockley, K., Richardson, D. & Dale, R. (2009). Conversation and coordinative structures. *Topics in Cognitive Science*, *1*, 305 - 319. 10.1111/j.1756-8765.2009.01021.x.

Shockley, K., Santana, M.-V. & Fowler, C. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance, 29,* 326-32. 10.1037/0096-1523.29.2.326.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., ... & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587–10592.

Stupacher, J., Hove, M. J., Novembre, G., Schütz-Bosbach, S., & Keller, P. E. (2013). Musical groove modulates motor cortex excitability: A TMS investigation. *Brain and cognition*, *82*(2), 127-136.

Thompson, S., Williamon, A., & Valentine, E. (2007). Time-dependent characteristics of performance evaluation. *Music Perception: An Interdisciplinary Journal, 25*(1), 13-29.

Trost, W. J., Labbé, C., & Grandjean, D. (2017). Rhythmic entrainment as a musical affect induction mechanism. *Neuropsychologia*, *96*, 96-110.

Tsay, C. J. (2013). Sight over sound in the judgment of music performance. *Proceedings of the National Academy of Sciences*, *110*(36), 14580-14585.

Upham, F., & McAdams, S. (2018). Activity analysis and coordination in continuous responses to music. *Music Perception: An Interdisciplinary Journal, 35*(3), 253-294.

van Der Steen, M. C. & Keller, P. E. (2013). The ADaptation and Anticipation Model (ADAM) of sensorimotor synchronization. *Frontiers in Human Neuroscience, 7*, 253.

Vesper, C., Butterfill, S., Knoblich, G., & Sebanz, N. (2010). A minimal architecture for joint action. *Neural Networks, 23*, 998-1003. doi:10.1016/j.neunet.2010.06.002

Vicary, S., Sperling, M., Von Zimmermann, J., Richardson, D. C., & Orgs, G. (2017). Joint action aesthetics. *PloS ONE, 12(7)*, e0180101.

Wallach, H., Newman, E. B., & Rosenzweig, M. R. (1949). A precedence effect in sound localization. *The Journal of the Acoustical Society of America*, *21*(4), 468-468.

Wanderley, M.M., Vines, B.W., Middleton, N., McKay, C., & Hatch, W. (2005). The musical significance of clarinetists' ancillary gestures: An exploration of the field. *Journal of New Music Research, 34,* 97–113. doi:10.1080/09298210500124208

Wapnick, J., Ryan, C., Campbell, L., Deek, P., Lemire, R., & Darrow, A. A. (2005). Effects of excerpt tempo and duration on musicians' ratings of high-level piano performances. *Journal of Research in Music Education*, *53*(2), 162-176.

Williamon, A., & Davidson, J. W. (2002). Exploring co-performer communication. *Musicae Scientiae*, *6*(1), 53-72.

Wing, A. M. (1993). The uncertain motor system: Perspectives on the variability of movement. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance 14: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 709-744). Cambridge, MA: MIT Press.

Wing, A. M., Endo, S., Bradbury, A., & Vorberg, D. (2014). Optimal feedback correction in string quartet synchronization. *Journal of The Royal Society Interface, 11(93)*, 20131125.

Wöllner, C. (2018). Call and response: Musical and bodily interactions in jazz improvisation duos. *Musicae Scientiae*, https://doi.org/10.1177/1029864918772004.

Wöllner, C., & Canal-Bruland, R. (2010). Keeping an eye on the violinist: motor experts show superior timing consistency in a visual perception task. *Psychological Research*, 74, 579-585. doi:10.1007/s00426-010-0280-9

Wren, C. R., Azarbayejani, A., Darrell, T., & Pentland, A. P. (1997). Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *7*, 780-785.

Zera, J., & Green, D. M. (1993). Detecting temporal onset and offset asynchrony in multicomponent complexes. *The Journal of the Acoustical Society of America*, *93*(2), 1038-1052.