Received 21 July 2022, accepted 20 August 2022, date of publication 29 August 2022, date of current version 8 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3202959

RESEARCH ARTICLE

huSync - A Model and System for the Measure of Synchronization in Small Groups: A Case Study on Musical Joint Action

SANKET RAJEEV SABHARWAL[1], MANUEL VARLET[2], MATTHEW BREADEN[2], GUALTIERO VOLPE[1], ANTONIO CAMURRI[1], AND PETER E. KELLER[2,3]

[1]DIBRIS, University of Genoa, 16145 Genova, Italy
[2]MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Penrith, NSW 2751, Australia
[3]Center for Music in the Brain, Department of Clinical Medicine, Aarhus University & The Royal Academy of Music Aarhus, 8000 Aalborg, Denmark

Corresponding author: Sanket Rajeev Sabharwal (sabharwalsanket@gmail.com)

This work was supported by the European Union's Horizon 2020 (H2020) Research and Innovation Program through the H2020 Future and Emerging Technologies Proactive (FETPROACTIVE) ENtrainment and Synchronization at Multiple TIME Scales in the MENTal Foundations of Expressive Gesture (EnTimeMent) Project, under Grant 824160.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Human Research Ethics Committee at Western Sydney University under Protocol No. H10487, and performed in line with the Declaration of Helsinki.

ABSTRACT Human communication entails subtle non-verbal modes of expression, which can be analyzed quantitatively using computational approaches and thus support human sciences. In this paper we present huSync, a computational framework and system that utilizes trajectory information extracted using pose estimation algorithms from video sequences to quantify synchronization between individuals in small groups. The system is exploited to study interpersonal coordination in musical ensembles. Musicians communicate with each other through sounds and gestures, providing nonverbal cues that regulate interpersonal coordination. huSync was applied to recordings of concert performances by a professional instrumental ensemble playing two musical pieces. We examined effects of different aspects of musical structure (texture and phrase position) on interpersonal synchronization, which was quantified by computing phase locking values of head motion for all possible within-group pairs. Results indicate that interpersonal coupling was stronger for polyphonic textures (ambiguous leadership) than homophonic textures (clear melodic leader), and this difference was greater in early portions of phrases than endings (where coordination demands are highest). Results were cross-validated against an analysis of audio features, showing links between phase locking values and event density. This research produced a system, huSync, that can quantify synchronization in small groups and is sensitive to dynamic modulations of interpersonal coupling related to ambiguity in leadership and coordination demands, in standard video recordings of naturalistic human group interaction. huSync enabled a better understanding of the relationship between interpersonal coupling and musical structure, thus enhancing collaborations between human and computer scientists.

INDEX TERMS Entrainment, interpersonal synchronization, joint actions, pose estimation, musical ensemble performance, social interaction, social signal processing, nonverbal communication.

I. INTRODUCTION

Machines have undergone major advances in their capability to interact with users. These advances are being further propelled with applications in human motion analysis and understanding coordination of human behaviors [1], [2]. With a wide range of methods to track human motion today, there is great potential in utilizing them to understand various behavioral aspects and responses of the human body. Humans exhibit phenomenal capabilities in synchronizing joint actions and coordinating at the interpersonal level in a non-verbal manner. This is observed particularly in musical

The associate editor coordinating the review of this manuscript and approving it for publication was Stavros Souravlas.

ensembles where co-performers coordinate their movements precisely yet flexibly, and this coordination often seems effortless [3]. A natural response to music is to move and synchronize to the rhythmic elements, and such spontaneous entrainment can be observed when individuals move to music being played around them, often without the intention to do so [4]. In group settings where multiple individuals interact, the mix of musical sounds and corresponding body movements can trigger social bonding effects reflected in feelings of affiliation, trust, and cooperativity [5], [6].

In musical ensembles, the interaction between performers is visual in addition to the audio cues associated with musical notes. Co-performers thus communicate with each other non-verbally by the motion of their heads and other upper- body movements. Such visual communication can convey information about initiating a musical piece at a certain time, as well as conveying how musical notes should be played to produce specific musical effects. This phenomenon can be observed clearly in musical conductors, who traditionally serve as messengers for the composer of a musical piece by using gestures that guide the performers in recreating the intended emotions and sentiments, allowing the group as a whole to reproduce an immersive experience [7]. Similarly, among musical performers, conveying these messages is often considered to be crucial for the co-creation of a meaningful musical performance. Therefore, musicians continually move during a performance to augment the creation of sound, express their artistic intentions, communicate with their fellow group members, and achieve states of synchronization [8], [9].

Analysis on interpersonal coordination in musical ensembles has implications that go beyond the specific area of music. Tal-Shmotkin & Gilboa, for example, show how a string quartet resembles working groups in organizational units (self- managed teams), i.e., groups of interdependent individuals, acting within an organizational setting, self-regulating their behavior to perform a joint task [10]. Computational approaches have already been used to study this communicative phenomenon. Researchers typically make use of motion capture (MoCap) technologies that can record and extract features from body movements exhibited by performers in musical ensembles. Conventional setups consist of linked optical cameras to track multiple markers that researchers attach to the performers' bodies prior to the recording session. While MoCap has facilitated research on joint actions and group behaviors [11], [12], [13], this technology bears limitations that preclude widespread use, in particular to the extent that it is an intrusive method for capturing trajectories of joints and limbs.

In this paper we introduce huSync (Human Sync), a computational framework and system that is intended to assist with the automated analysis of synchronization in small groups from conventional video recordings by making use of a multi-person pose estimation algorithm to extract body joint coordinates [14]. huSync is designed in recognition of the need to study interpersonal coordination within groups in ecological settings in order to ensure that findings are representative of everyday joint action. With this goal in mind, we apply huSync to video recordings of a professional musical ensemble, which enables the investigation of musicians' movements and interaction in naturalistic contexts. Musical performances serve as an ideal test bed to examine non-verbal communication because they are readily controlled micro-environments where, in many cultural traditions, interactions are scripted in musical scores. Capitalizing on this convention, we analyze and assess how the movements of ensemble performers evolve over the course of structures specified in musical scores, and huSync is used to address research questions about the effects of musical structural features on objective measures of ensemble coordination.

This paper is organized as follows: in Section II, we highlight the hypothesis and research questions that are raised, in Section III we present existing computational approaches for the analysis of synchronization and relevant studies that have examined interpersonal synchronization and entrainment in small groups, particularly musical ensembles; Section IV describes the huSync computational framework and system as well as an instance of the framework, with a detailed methodology and calculation routine, explained using a simulated example, to compute dyadic synchronization; Section V presents the dataset, with a sub-section dedicated to the implementation of huSync on this dataset and parameters utilized for our use case to perform the analysis; We then present statistical results in Section VI followed by Section VII where we discuss them; We conclude the paper by highlighting limitations and possible future research in Section VIII.

II. THE PROBLEM

Our first objective is to develop a computational framework and a system, for the automated analysis of interpersonal coordination in small groups, considering cases of clear leadership by an individual member as well as cases of egalitarian leadership distributed throughout the group. In our computational approach, we get motor, postural, and acoustic data in a non-intrusive manner, to compute synchronization of motor and postural features by applying consolidated techniques, and to provide outputs which are robust with respect to the different conditions addressed (e.g., either clear or egalitarian leadership). Our second goal is to exploit such computational approach to investigate the effects of musical texture and position within musical phrases, and how it affects interpersonal coordination in a professional music group performing in two constellations that are common in Western chamber music: a string quartet (consisting of two violins, viola, and cello) and a clarinet quintet (i.e., a string quartet with an added clarinet soloist). This is intended to at the same time provide evidence of the robustness of the proposed framework and system and increase knowledge of the mechanisms that underly interpersonal coordination in small groups.

Musical phrases are analogous to phrases or sentences in speech to the extent that they are meaningful organizational units that would be perceived as coherent or complete if presented in isolation. We consider phrases to be sections of musical pieces that consist of unified thematic material presented in uniform texture. On this view, phrase boundaries are marked by changes both in thematic material and texture. While this definition can result in longer structural units than what are usually designated as phrases in musicological analyses, this definition serves our research questions related to effects of structural change on interpersonal coordination. The strength of interpersonal coupling in body motion was expected to be influenced by musical texture and phrase position. Based on previous research [15], [16], [17], coupling strength is hypothesized to be stronger for ambiguous textures where no leader is implied (polyphonic) versus textures with an unambiguous melodic leader (homophonic) due to heightened mutual adaptation, anticipation, and joint attention in the former. However, whether coupling would be stronger at the beginning and end of phrases than in the middle [18], [19] was an open question due to potentially counteracting effects of coordination demands and compensatory strategies. These demands and strategies could, furthermore, vary as a function of texture. Specifically, the presence of a leader may be more influential at the beginning and ending of phrases than in the middle, in which case we would expect a statistical interaction of the two factors. To pave the way for naturalistic research conducted in ecological settings, we examined these research questions by analyzing public concert recordings taken with conventional video using our proposed approach which utilizes computer vision techniques involving pose estimation and quantifying synchronization. In addition, video-based coupling measures were correlated with audio features of the performances to assess cross-modal relations [20], [21]. Some degree of dissociation was anticipated, due to auditory coordination being most pronounced at timescales related to the musical beat, whereas visual-based coordination of body motion is most pronounced at longer timescales [22].

III. BACKGROUND AND RELATED WORK

In this Section, we summarize background and related work with respect to the two major objectives of this work, i.e., (i) existing computational approaches to analysis of interpersonal synchronization in small groups as well as (ii) theoretical and experimental background on interpersonal coordination in musical joint action.

A. COMPUTATIONAL APPROACHES TO ANALYSIS OF SYNCHRONIZATION IN SMALL GROUPS

Interpersonal synchronization in small groups is of key interest since it serves as a useful indicator of dyadic, and group-level behavior and coordination. The analysis of synchronization is complex and requires integrating multimodal communicative signals. Many studies in this area are based on manual annotations, and the analysis is done by directly inspecting and coding the data by trained observers. To avoid this tedious process, automated methods can be used to process relevant social signals in small groups and thereby measure interpersonal synchrony.

Analyzing social dynamics and interpersonal synchronization have been studied in many fields. For example, in psychotherapy settings, studies analyzed temporal changes in global body movement using video-based quantification techniques such as motion energy analysis (MEA), a frame differentiating method, to measure synchrony between the patient and counselor during psychotherapeutic sessions [23], [24], [25], [26]. While MEA is a simple approach, a critical issue noted is that since it quantifies frame-differences based on the region of interest (ROI), it is not sensitive to the direction of movement within a ROI. Thus, someone who touches their face often, will exhibit higher head-movement as compared to someone who does not [23]. During unidirectional face-to-face communications, Yokozuka and colleagues [27] made use of wireless accelerometers attached to the forehead of the speaker and listener to analyze head motion synchronization and empathy, using phase and frequency differences. The use of instruments attached to the body makes participants uncomfortable which impedes naturalistic movements.

Among small group ensembles, MoCap systems have been extensively used to study interpersonal coordination with the use of non-linear methods particularly between performers playing music together [28], [29], [30], conductors' gestures inducing entrainment in a musical ensemble [31] or participants moving to the beat of the music [32], [33], [34]. In Burger *et al.*, MoCap data was processed to represent whole-body swaying and bouncing motions among participants. Period and phase-locking behavior was observed in full-body music-induced movements by calculating the circular mean of movement phases and beat locations for each participant, with results informing our understanding of how humans entrain to music. While data can be captured with MoCap systems at high frequencies, good accuracy, and low noise, such specialized systems can be expensive, pose methodological issues [35], and restrict movement due to the use of tight-fitting motion-tracking suits. Marker-less methods are emerging as good alternatives to MoCap systems for synchronization studies in small groups, as seen, for example, in a study in Hadjakos *et al.* [36], who used a Kinect camera to analyze head movements and study synchronization in a violin duet performance.

With huSync, we present a system that instead utilizes a pose estimation algorithm on video sequences and computes Phase-Locking Values (PLV) to study the interaction of social signals in small group setups. As compared to the computational approaches discussed above, huSync is a non-intrusive method to study interaction in small groups in naturalistic contexts and eliminates the dependency on any hardware for tracking body movements. PLVs have been used to quantify interpersonal coordination at the level of body motion and brain activity in a wide range of social interaction tasks [37], [38], [39], [40], [41], [42], suggesting that it is a reliable measure for studying cognitively mediated contributions to the synchronization process. Indeed, phase locking is generally a pervasive concept in computing interactions in non-linear and complex systems, and PLV in particular, is a commonly used interaction measure in diverse domains [43], [44]. When applied to evaluating body movement synchronization, PLV has proven to be sufficiently sensitive to detect subtle and unintended coordination under a range of manipulations, including different leadership conditions [17]. A large body of previous work therefore suggests that PLV is appropriate for assessing functional level connectivity between performers during naturalistic musical ensemble performance to test our hypotheses about relations between musical structure and ensemble coordination.

Following Mormann *et al.* [45], we compute PLVs using phase values that are extracted from the spectrum of the analyzed motion trajectory signals. These phase values are obtained for each frequency by applying a Fast Fourier Transform (FFT). For the purpose of our study, we utilize relative phase values, as in previous research [46] that assessed the dyadic synchronization between individuals within pairs of co-performers. To this end, we calculated PLVs by computing the phase difference between the head motion trajectories of two co-performers in each possible pair using the formula in 1. In this procedure, the phase difference is represented as a complex unit-length vector [43] and the absolute value of the mean is then a measure of the magnitude of the vector, which indicates the degree of synchrony.

$$PLV = \left| \frac{\sum_{t=1}^{n} e^{i(\Theta_1 - \Theta_2)}}{n} \right| \tag{1}$$

Here n is the total number of data points, t represents equally distributed discrete time steps, and Θ_1 and Θ_2 are the phase angles of the two signals for a specific frequency being analyzed. The degree of synchrony as computed here is in the range of [0,1], where the highest state of synchrony is 1.

B. INTERPERSONAL COORDINATION AND ENTRAINMENT IN MUSICAL JOINT ACTION

Small groups of musicians provide a valuable domain to investigate interpersonal coordination and entrainment from multiple perspectives, ranging from bio-mechanical and computational to psychological and neuroscientific [47], [48], [49]. As a microcosm of social interaction, ensemble co-performers coordinate their body movements and sounds with high degrees of precision and flexibility to communicate musical structure and expressive information among themselves and also with their audience [19], [50], [51], [52]. Although auditory information is generally primary in music, visual information can influence musical communication in live and recorded performances. In performance research, a distinction is drawn between instrumental movements, which are directly related to the production of musical sounds (e.g., the bowing of a violinist), and ancillary movements, which are not technically required for sound production but nevertheless take place during performance (e.g., head nods and swaying of the torso) [53].

Ancillary motion may be the key to understanding social communicative effects of group music making. Results outside the music domain indicate that greater head motion synchronization occurs during moments of high empathy in face-to-face communication [27]. This finding suggests that the degree of empathy can be assessed by the correlation between phase and frequency of head motion synchronization in setups where co-actors are in visual contact. Empathy can be considered to be an innate capacity for understanding others thoughts and feelings, and among the core components that enable musicians to engage socially with one another during performances [54], [55]. Empathy contributes to feelings of social bonding and behavioral contagion among individuals in groups, leading to higher states of synchronization in upper-body/head movements [27]. Musical ensembles can therefore be considered to be more than groups of synchronized individuals, but instead as systems for social connection in which empathy facilitates the information transfer between performers by enhancing synchronization states. Rhythmic synchronization of upper body movements and particularly the head is pertinent and sometimes inevitable in a musical ensemble - presumably emerging from high degrees of empathy, agreement, and shared joint goals.

Ancillary body motion also plays a role in regulating an individual's performance, conveying musical structure, expressive intentions, and underlying musical meaning to others in a group or even the audience [56], [57], [58], [59], [60], [61]. In musical ensembles, ancillary motion provides visual cues that assist co-performers to coordinate their actions, and interpersonal coupling can be therefore observed at the level of body movements as well as sounds [61], [62]. Previous research in small group interactions has demonstrated that the coordination of head motion and body sway is positively correlated with coordination of sound onsets, although the relation is not perfect [4], [28], [63], suggesting that visual and auditory information provide parallel channels for musical communication [57]. Additionally, the synchronization of non-verbal elements of expression takes place across multiple temporal scales, with head motion in particular being associated with higher states of connectedness [64], [65]. Correspondingly, the head movement synchronization of performers in a group can serve as a good metric to identify whether or not they are performing cohesively. The rate of synchronization can be logged to note the eventual increase or decrease in performance synchronization [66], [67], [68], [69].

Indeed, the coordination of co-performers' sounds take place at short timescales (millisecond range), while movements such as body sway are aligned at longer timescales associated with higher-order units of musical structure (e.g., phrases) [22]. Furthermore, interpersonal coupling in both body motion and sounds is dynamic in the sense that it varies overtime, and this variation is systematic, that is, not entirely random [20], [70], [71]. The present study addresses how such variation in co-performer coupling relates to two aspects of musical structure - texture and phrasing. Musical texture refers to the hierarchical arrangement of instrumental parts in the ensemble; specifically, how separate parts relate to one another in terms of salience (i.e., tendency to capture attention) and complexity (e.g., degree to which separate parts contain redundant vs unique information) [22], [72]. In some musical textures, there is a clear distinction between a melody and accompaniment parts, where the melody is relatively high in salience (i.e., homophonic textures). It is often assumed that in such cases the melody player serves a leadership role in the ensemble (even if only temporarily) [11], [12], [13]. In other textures, separate parts can each have independent melodic content that proceeds simultaneously (i.e., polyphonic textures). In these cases, the situation is more egalitarian in the sense that leadership is distributed or free to roam around the ensemble [22], [73], [74], [75].

It is currently unclear how such textural variations affect coordination. Laboratory studies of interpersonal coordination suggest that coupling can be stronger without a designated leader – when a form of co-leadership characterized mutual adaptation, anticipation, and joint attention emerges [52] – but this work mainly entails improvised (as opposed to scripted) performances and participants without formal musical training [15], [16], [17]. Naturalistic studies of experienced musicians have yielded mixed results with regard to whether interpersonal synchronization is influenced by leadership instructions and the degree of independence between parts in terms of melody and accompaniment roles [12], [50], [76]. Overall, these findings suggest that the degree to which ensemble coordination is resilient to different conditions might vary with task demands and levels of expertise [77], and that the method of quantifying coupling and the timescale(s) at which it is applied might influence results.

The second structural aspect of interest relates to the segmentation of musical pieces into phrases and higher-order sections. Previous research suggests that phrase entries and endings present challenges for interpersonal coordination due to heightened uncertainty associated with increased timing variability at these points [74]. As a compensatory strategy, ensemble co-performers hence increase the use of visual cues, including gestures and eye contact, to assist coordination at phrase boundaries [3], [9], [12], [51], [78]. Furthermore, improvising jazz musicians have been found to become more synchronous prior to structural boundaries (i.e., transitions where the musical content or style changes), suggesting an increase in the intensity of joint attention and communication at these points [18].

IV. COMPUTATIONAL FRAMEWORK, SYSTEM ARCHITECTURE AND METHODOLOGY

To investigate the research questions raised in section II, we propose huSync as a computational framework and system for evaluating interpersonal synchronization between dyads in small groups. Fig. 1 presents an overview of the

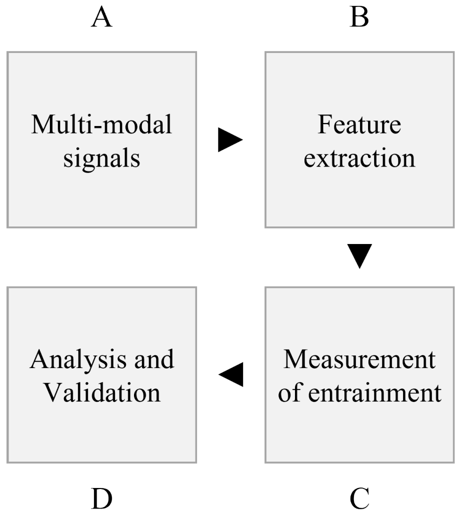


FIGURE 1. The huSync computational framework for the analysis of interpersonal synchronization in small-group setups.

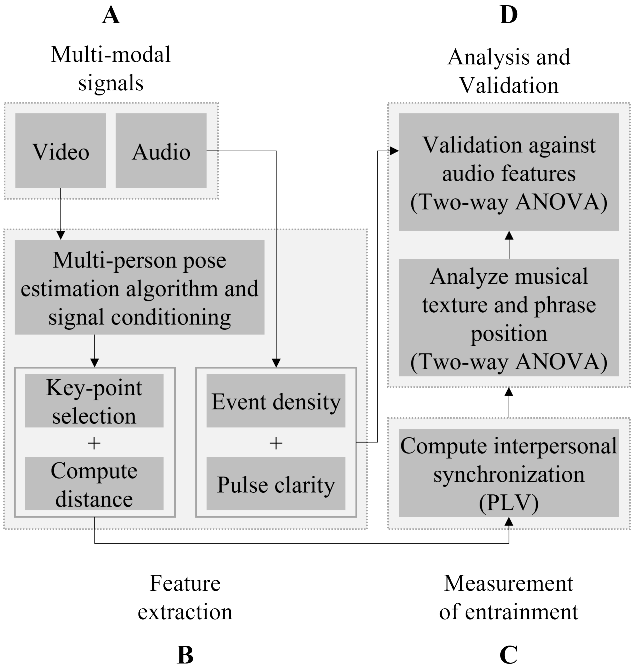

A
Multi-modal signals
D
Analysis and Validation
Video
Audio
Validation against audio features (Two-way ANOVA)
Multi-person pose estimation algorithm and signal conditioning
Analyze musical texture and phrase position (Two-way ANOVA)
Key-point selection
+
Compute distance
Event density
+
Pulse clarity
Compute interpersonal synchronization (PLV)
Feature extraction
B
Measurement of entrainment
C

FIGURE 2. An instance of the computational framework and the huSync system architecture.

computational framework adopted for huSync. It includes four blocks and is grounded on a well-established conceptual framework for the analysis of expressiveness conveyed using body movements and gestures alike [56], [79]. The first block, multi-modal signals (Fig. 1 (A)), consists of information and data that can be sourced from different modalities (e.g., audio, video, heart-rate, respiration rate, and so on). The second block, feature extraction (Fig. 1 (B)), entails extracting raw data from these multi-modal signals and could include pre-processing steps (e.g., up or down-sampling, interpolation, realignment, and normalization) to make sure that all signals are compliant with one another when performing a detailed analysis. It also involves extraction of essential features from the signals that can better describe movements exhibited during small-group interactions (e.g., acceleration peaks, kinetic energy, and distance computation). The third block, measurement of entrainment (Fig. 1 (C)), involves examining the overall behavior of participants in a small group and how they may adjust and adapt their behavior. The fourth block, analysis and validation (Fig. 1 (D)), involves performing a statistical procedure on the results obtained to ascertain the influence of variables present in our data (e.g., one-way ANOVA, two-way ANOVA, and multivariate ANOVA) and validating the results to test the sensitivity, reliability and practical usefulness of our framework and corresponding system.

The structure and architecture of the huSync system is illustrated in Fig. 2 and is represented as an instance of the computational framework in Fig. 1. It follows a structured funnel of steps beginning with selecting videos of interest that satisfy our analysis criteria. The first block (Fig. 2 (A)) is responsible for reading video and audio signals from standard video recordings. The video data are pushed into the second block (Fig. 2 (B)), where they are processed with multi-person pose estimation algorithms to detect key-points on participants' bodies in each frame of the video being analyzed, and it generates a json file in sequential order of the people in each frame, and each person is represented with an array of key-points. Using this trajectory information, and depending on the specific use case, a relevant key-point is selected to obtain kinematic information and processing the data to extract relevant features. For our system instance, we decided to compute the distance between each time step traversed by the key-point. As illustrated, we utilize the audio signals from the video recordings to extract acoustic features such as pulse clarity and event density, and has been explained in more detail in section VI-B. The distance data then moves into the third block which involves computing interpersonal synchronization between participants in the group (Fig. 2 (C)). This is followed by performing a statistical analysis and validation (Fig. 2 (D)) on the phase-locking value results obtained, which additionally helps answer questions raised in our hypothesis. To help interpret and validate the results which are primarily heterogeneous in nature, a cross-modal validation is performed with the acoustic features obtained from the feature extraction block (Fig. 2 (B)).

A. huSync PROCESS PIPELINE AND METHODOLOGY TO COMPUTE DYADIC SYNCHRONIZATION

The process pipeline of huSync entails an 8-step computational methodology, as illustrated in Fig. 3, and is a subset of blocks B and C in Fig. 2. This pipeline covers the entire range of operations performed on the data extracted, from the json file available after pre-processing with pose estimation algorithms to computing the final dyadic synchronization in small-group setups. Thus, before applying the huSync computational model it is necessary to have the data extracted from json files, after selecting the key-point of interest based on the use case and experimental setup.

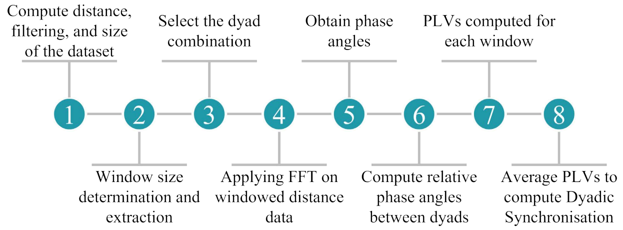

Compute distance, filtering, and size of the dataset
Select the dyad combination
Obtain phase angles
PLVs computed for each window
1
2
3
4
5
6
7
8
Window size determination and extraction
Applying FFT on windowed distance data
Compute relative phase angles between dyads
Average PLVs to compute Dyadic Synchronisation

FIGURE 3. An illustration of the process pipeline for computing dyadic synchronization.

huSync is flexible and allows dyadic synchronization to be computed for the entire duration of a video phrase or over sections, as in the three parts (start, middle, and end) required to address the research questions raised in section II. To provide an intuitive understanding of the process, we explain the 8 steps next, along with an illustration of a simulated example for a dataset with 15 data-points to compute dyadic synchronization between a pair of performers. Fig. 4 covers steps 1 to 6 and Fig. 5 covers steps 7 and 8.

1) STEP 1 – COMPUTE DISTANCE, FILTERING, AND SIZE OF THE DATASET

The key-point of interest can be a single key-point or a computed feature between multiple key-points. As part of our feature extraction step (Fig. 1 (C)), using the data extracted from the json file, we compute the Euclidean distance with the raw coordinate data available in (x,y) format. When processing videos with pose estimation algorithms, the data can be quite noisy and it is important to check if filtering is required. huSync implements the Savitzky-Golay filter, if needed, since it tends to preserve the phase and essential features of a signal [80], [81]. We then ascertain the size of the dataset to be consumed by the huSync model to analyze changes in synchronization level over the time period of interest. In our specific use case, answering the research questions raised in section II requires analyzing the start, middle, and end of musical phrases, and hence the total number of datapoints should be divisible by 3 and also adaptive to the step-size chosen in the next step to fit all data points that fall within the window width. When this condition is not met, extra rows in the data file can be dealt with by truncating the dataset at the end of the phrase segment. Additionally, if there are fewer line items, they can be dealt with by augmenting the existing data at the extremities using polynomial or linear extrapolation. While it did not happen in our case, if loss of information is observed in between a phrase segment, it can be dealt with by making use of a cubic spline interpolation to fill gaps [82].

2) STEP 2 – WINDOW SIZE DETERMINATION AND EXTRACTION

We use a sliding window approach that steps through each portion of the video data so to capture both local and global

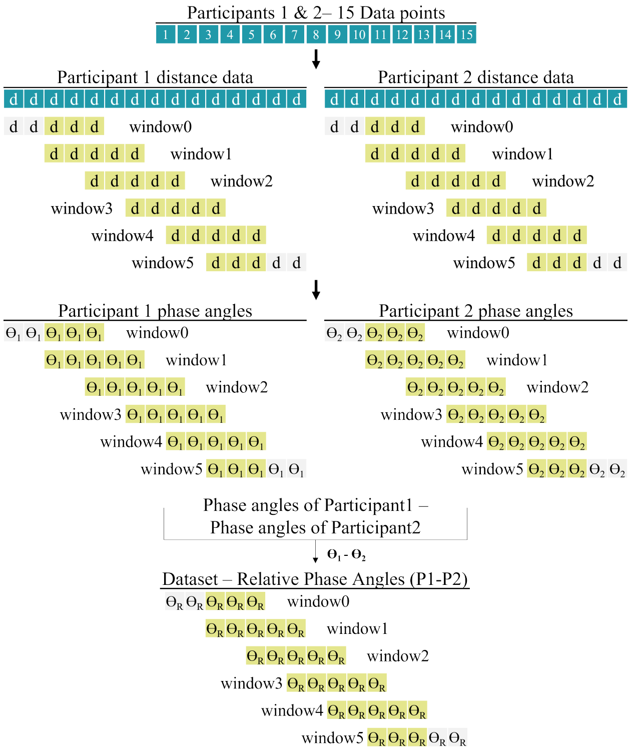

Participants 1 & 2– 15 Data points
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Participant 1 distance data
Participant 2 distance data
d d d d d d d d d d d d d d d
window0
window1
window2
window3
window4
window5
Participant 1 phase angles
Participant 2 phase angles
Θ_1
Θ_2
Phase angles of Participant1 –
Phase angles of Participant2
$\Theta_1 - \Theta_2$
Dataset – Relative Phase Angles (P1-P2)
Θ_R

FIGURE 4. Simulated Example: This figure covers steps 1 to 6.

trends. In our simulated example we use a window size = 5 and step-size =2, and thus we have 6 windows.

3) STEP 3 – SELECT A DYAD COMBINATION

In our simulated example, we consider a situation with two participants, but with huSync we can either select a specific dyad for whom we would like to compute synchronization or do so in an automated manner for all possible pairs. Since the order is not important, the total number of possible pairs can be computed using (2).

$$C\,(n,r) = \binom{n}{r} = \frac{n!}{(r!(n-r)!)} \tag{2}$$

4) STEP 4 – APPLY FFT ON WINDOWED DISTANCE DATA

We apply the FFT algorithm iteratively on all possible dyadic pairs available using the scipy library [81]. On applying FFT to the distance data at each time step, we obtain the spectrum, and proceed with extracting the real and imaginary components.

5) STEP 5 – OBTAIN PHASE ANGLES

By applying FFT over the distance data, we obtain complex-values, from which the magnitude (modulus) and phase values are extracted. For this purpose we utilize the numpy library [83]. Data for all participants undergo FFT individually to obtain phase angles at each frequency bin and time step of the windowed information for all participants. As illustrated in our simulated example, once FFT is applied over the data of Participants 1 and 2, from the complex values we extract the phase angles.

6) STEP 6 – COMPUTE RELATIVE PHASE ANGLE BETWEEN DYADS

After obtaining the phase angles for each participant, we then proceed with computing relative phase angles (difference between the phase angles) for all possible pairs, and for our simulated example it will be between the two participants – by computing for each time step and frequency bin the difference between phase angles of the participants.

7) STEP 7 – PLVs COMPUTED FOR EACH WINDOW

The relative phase values are used by our function to compute PLVs. In a window, each element, or phase angle value, is put together with values present in other windows, but with those having the same position. Thus, we receive a set of PLVs equal to length of each window. In our simulated example, each window element is aligned with those having the same position to obtain PLVs. The colored squares, as seen in Fig. 5, indicate values present at the same position, which are used as inputs in our function to compute a PLV. PLV is then computed for each time step and each frequency bin using all relative phase values of the corresponding window.

8) STEP 8 – AVERAGED PLVs AND DYADIC SYNCHRONIZATION

After we have obtained the PLVs, which as seen in the previous step will result in an array having a length equivalent to the length of a single window, since we have one PLV for each frequency bin. Here, a cut-off frequency can be utilized to discard frequencies beyond a threshold, while excluding the DC component for the computation. As seen in Fig. 5, once the PLVs are calculated, we average them to obtain a single value (avgPLV or averaged PLV), across different frequency bins of interest, and is our final value for dyadic synchronization between a pair. Here, PLV and averagedPLV are computed using (3):

$$PLV_j = \left| \frac{\sum_{i=0}^{n} e^{i\,\theta_{R(ij)}}}{n} \right|$$

$$avgPLV = \frac{\sum_{j=0}^{k} PLV_j}{k} \tag{3}$$

where $i \in \{0..n\}$, $j \in \{0..k\}$ and n are the number of windows, k is the number of relative phase angles in each window, and $\Theta_{R(ij)}$ represents the relative phase angle present in each window i at position j. The value ranges from 0 to 1 where 1 indicates perfect synchrony and 0 no synchrony.

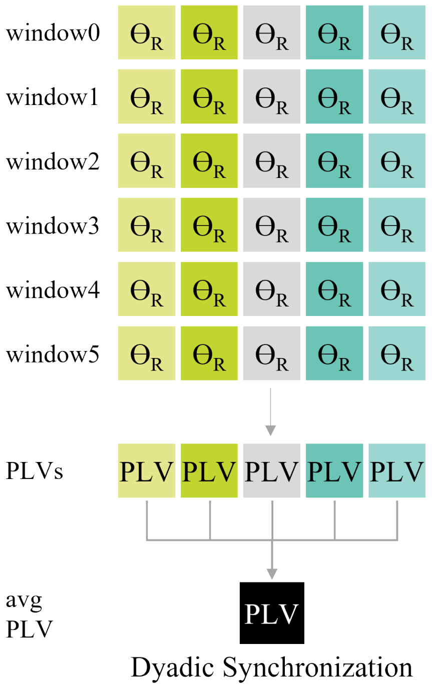

window0 Θ_R Θ_R Θ_R Θ_R Θ_R
window1 Θ_R Θ_R Θ_R Θ_R Θ_R
window2 Θ_R Θ_R Θ_R Θ_R Θ_R
window3 Θ_R Θ_R Θ_R Θ_R Θ_R
window4 Θ_R Θ_R Θ_R Θ_R Θ_R
window5 Θ_R Θ_R Θ_R Θ_R Θ_R
PLVs PLV PLV PLV PLV PLV
avg PLV
PLV
Dyadic Synchronization

FIGURE 5. Simulated Example: This figure covers steps 7 and 8, which are a final synthesis of the computation process.

V. THE TEST DATASET

A. VIDEO RECORDINGS

huSync, as a system, is tested on a dataset consisting of videos from concert performances by the Omega Ensemble, a professional chamber music group from Australia, and consent was given for further use of the data. The concerts took place at City Recital Hall in Sydney in 2017. The videos included musical pieces composed by Alexander Borodin and Johannes Brahms. Videos were recorded by a Canon 1DX camera body and a Canon EF 70-200 1:2.8 L zoom lens as QuickTime movies (.MOV) with dimensions 1920 × 1080 pixels at 25 frames per second. Audio was recorded as 16-bit stereo at 48 kHz. Compression was done with the H.264 video codec and the Linear PCM audio codec synced via Timecode.

For the present study, we chose to perform our analyzes on videos from a concert featuring the Clarinet Quintet in B minor (Op. 115) written in 1891 by Johannes Brahms (1833-1897) (“Brahms Clarinet Quintet”) and String Quartet No. 1 in A major written in 1874-79 by Alexander Borodin (1833-1887) (“Borodin String Quartet”). Fig. 6 includes screenshots from the video sequences. The Borodin String Quartet is scored for violin 1, violin 2, viola, and cello. The Brahms Clarinet Quintet uses the same string instruments plus a clarinet. Both pieces contain four movements with contrasting musical characters. The total duration of the Borodin String Quartet performance lasted 39 minutes and 13 seconds while the duration of the Brahms Clarinet Quintet was 40 minutes and 38 seconds. For our study, specific phrases from each concert recording were selected based on pre-defined parameters, covered in section V-B. In Table 1, we summarise the full dataset and specific phrases selected in terms of phrase duration (min, max, median and average) and count.

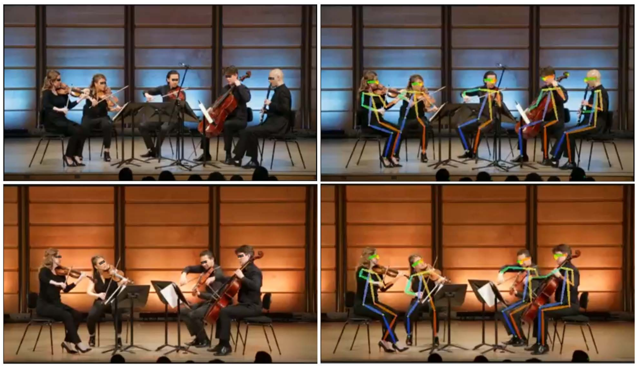


FIGURE 6. Images from the performance of the Brahms Clarinet Quintet (Top Left) and the Borodin String Quartet (Bottom Left) along with the outputs available with tracked key-points using pose estimation algorithm (Top Right and Bottom Right).

B. ANNOTATION PROCESS FOR IDENTIFIED SECTIONS OF INTEREST

Videos were annotated to segment them for analysis addressing the hypothesized effects of musical texture and phrase position on interpersonal synchronization among ensemble co-performers, that is, four individuals for the Borodin performance and five individuals for the Brahms performance. These annotations were done in accordance with the specific aim of testing how the strength of interpersonal coupling is influenced by two factors:

1) Position within the musical phrase (start, middle, and end); and
2) Musical texture (polyphonic, where leadership is ambiguous due to the lack of a clear distinction between melody and accompaniment, versus homophonic, where there is an unambiguous melodic leader).

Each of the videos was annotated using ELAN (an annotation tool for multimedia files) [84] in accordance with a musicological analysis based on the published score. In order to mitigate noise that can be introduced by personal behavioral aspects of performers before or after a phrase has been played such as shaking the legs, rotating the arms, or readjusting their seating position, the annotations should be made carefully and be aligned as accurately as possible with the start and end of musical phrases. Annotated features included phrasing, textural classification, number of instruments currently playing, and instrument roles (e.g., melody, counter melody, or harmonic accompaniment), which were indicated in separate tiers in the ELAN interface. Information from each tier within the annotated ELAN file for each piece was exported to extract video timecodes for each phrase and its textural classification.

Phrases were selected based on the following criteria:

TABLE 1. Summary of the complete dataset and selected phrases for our experiments.

		Dataset					Selected Phrases				
		Duration (s)					Duration (s)				
Concert	Texture	Minimum	Maximum	Median	Average	Count	Minimum	Maximum	Median	Average	Count
Brahms	Homophonic	15.032	38.199	19.742	21.573	27	16.161	30.433	20.615	21.856	12
	Polyphonic	15.488	33.08	23.100	23.528	20	15.488	27.553	20.161	21.105	12
Borodin	Homophonic	15.295	24.973	18.317	19.117	10	15.295	24.973	18.317	19.117	10
	Polyphonic	15.142	29.628	21.271	21.013	11	15.142	29.628	20.924	20.800	10

1) Duration: Phrases were chosen keeping in mind that each one could be split into three equal segments with the same duration, thus giving us a start, middle, and end for each selected phrase, while ensuring that each segment was >5 sec. The segment sizes varied with phrase length. These time intervals are long enough to compute interpersonal synchronization of body motion, but not too long so that musical qualities could be considered stationary.
2) Number of instruments: Phrases that met the duration criterion set above, were filtered based on the number of instruments that were being played. The criterion for selection at this stage was that all instruments in the ensemble were playing throughout most of the phrase.
3) Texture: Phrases longer than the duration specified above, and with all instruments playing, were then filtered based on the musical texture. For our purposes, the focus was on the distinction between textures with a clear musical leader-follower roles versus textures that were less clear in this respect. This resulted in a two-level textural classification. We refer to 'homophonic' textures as having rhythmically differentiated melody and harmonic accompaniment, with the melody instrument assumed to serve a leader role and the accompanying instruments serving as followers. By contrast, 'polyphonic' textures have more than one melody part, and thus more than one potential leader. By our definition, polyphonic textures range from those where there are two melodic parts (e.g., a melody and countermelody) or interdependent melodic material distributed across multiple instrumental parts.
4) Instrument roles: For each phrase selected based on the above criterions, the instrument that was playing a melody line was noted (in the case of homophonic textures). Only phrases where the instrument roles were consistent throughout passed this criterion.

Table 1 reports the number of selected phrases and their min, max, median, and average duration.

C. APPLYING huSync TO THE DATASET

Videos from the dataset are used as a testbed for huSync and to investigate the social signals that lead to states of heightened interpersonal coordination. We pre-processed selected video phrases using AlphaPose (v0.4) and received json files with full-body key-points. From this, as motivated in section III-B, we are interested in the trajectory of the head, and thus extract the nose key-point (key '0'). Data are arranged as a table with x and y coordinates for each participant in separate columns. We analyze dyadic synchronization for all pairs of performers and the total possible dyad combinations is 6 for Borodin (n=4, r=2) and 10 for Brahms (n=5, r=2). We did evaluate the use of a Savitzky-Golay filter for our data, but did not observe any major differences with its use and decided to exclude it during the data processing phase. Using the coordinate information, the Euclidean distance between each time step of the trajectory is computed for every participant and arranged in separate columns. We then proceed with using a sliding window to segregate our data for each participant. Based on previous studies, tests were performed by varying the duration or size of the window to inspect our data across multiple levels of temporal resolution and statistical significance, and decided to proceed with a window size of 30 and step-size of 5 [85]. Based on our window step-size the dataset had to also be divisible by 5 to fit all data points by the window width. We truncate the data in case of extra data-points and extrapolate to fill missing values. For example, if our dataset contains 453 data points, we will truncate it to 450 to arrive to the nearest multiple of 5 and 3, and in case we have 447 data points, we extrapolate 3 data points to arrive to 450. On applying FFT on the windowed distance data, we extract the phase angle and begin to compute relative phase angles for all possible pairs. By analyzing the frequency distribution, and using a window size of 30, a 10Hz cut-off indicates excluding all values above the 11th value and excluding the 1st since it is the DC component. PLV is computed for each frequency bin and then averaged across all frequency bins of interest.

VI. RESULTS

We performed our analyzes on a total of 44 phrases and in Table 1 we share a group summary of the dataset chosen. These phrases met our criteria of a good balance between polyphonic and homophonic textures while also taking into account the duration of each phrase and quality of the data received on pre-processing videos with a pose estimation algorithm.

The PLV results are first presented descriptively and then results of analyzes of Variance (ANOVA) are reported. Performances of the Brahms and Borodin pieces were analyzed separately due to the differing number of performers in each piece. PLVs for all pairs for each piece were entered into an ANOVA that included Phrase Position (Start, Middle, End) as a within subjects factor and Texture (Homophonic, Polyphonic) and Pair (i.e., each separate

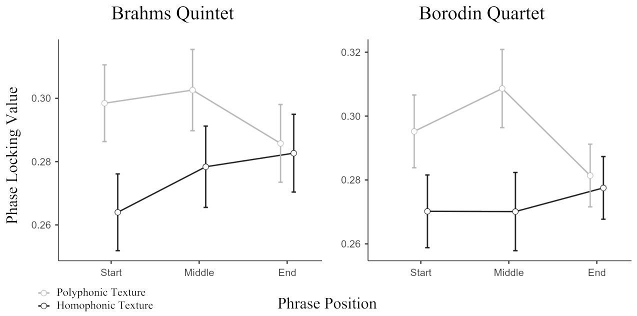


FIGURE 7. Phase locking values, indicating synchronization of co-performers' head motion, across phrase positions for polyphonic and homophonic texture for Position × Texture in the Brahms and Borodin pieces).

pairing of individuals from the ensemble) as between subjects factors. The factor 'Pair' was included in the analyzes since no two instrumentalists were playing the same part and the specific pairing of parts could have systematic effects on PLV, although the detailed analysis of such potential effects is beyond the scope of our study. In addition to the main factors, phrase duration was included as a covariate in the analyzes to control for its potential effects on PLV. The ANOVAs were run in jamovi [86].

A. EFFECTS OF TEXTURE AND PHRASE POSITION ON INTERPERSONAL COUPLING

PLV results are graphically represented in Fig. 7 for the three phrase positions in the two textures, with data averaged across pairs, for Brahms and Borodin performances, respectively. From these graphs it is seen that polyphonic textures have higher PLVs as compared to homophonic textures, as was hypothesized in the research questions raised in section II. We also observe how the PLVs begin at a lower value in both textures. For the polyphonic texture, values start out relatively high, then tend to rise further at the middle of the phrase, and finally drop towards the end of a musical phrase. For the homophonic texture, values start out lower and remain constant until a slight increase at phrase endings.

In Fig. 8, results are shared as network plots of averaged PLVs, across all phrases, for individual instrument pairings observed for the Brahms and Borodin performances separately. Here it can be seen that most pairs show a higher level of synchronization in polyphonic textures in the start, middle and end of each phrase, suggesting that the effect is a general and not tied to specific instrument pairings.

The ANOVA results are illustrated for the Brahms performance in Table 2, and for Borodin in Table 3. Values highlighted in bold indicate statistical significance ($p<0.05$). For Brahms, the ANOVA revealed a statistically significant main effect of Texture, $F(1, 219) = 16.08$, $p < 0.001$, and a significant two-way interaction between Position and Texture, $F(2, 438) = 6.098$, $p = 0.002$. For Borodin, there was also a statistically significant main effect of Texture, $F(1, 107) = 14.051$, $p < 0.001$, and a significant two-way interaction between Position and Texture, $F(2, 214) = 3.399$, $p = 0.035$. For both Brahms and Borodin, the main effect of position was not statistically significant.

TABLE 2. ANOVA results for between and within subjects effects for the brahms concert.

Between Subjects Effects					
	Sum of Squares	**df**	**Mean Square**	**F**	**p**
Pair	0.0126	9	0.0014	0.25	0.986
Texture	0.0902	1	0.09024	16.081	**<.001**
Pair ∗ Texture	0.0583	9	0.00648	1.155	0.326
Duration	0.2595	1	0.25948	46.244	<.001
Residual	1.2289	219	0.00561		

Within Subjects Effects					
	Sum of Squares	**df**	**Mean Square**	**F**	**p**
Position	0.00598	2	0.00299	1.003	0.368
Position ∗ Pair	0.08754	18	0.00486	1.632	0.049
Position ∗ Texture	0.03634	2	0.01817	6.098	**0.002**
Position ∗ Duration	0.00424	2	0.00212	0.711	0.492
Position ∗ Pair ∗ Texture	0.03341	18	0.00186	0.623	0.882
Residual	1.30509	438	0.00298		

TABLE 3. ANOVA results for between and within subjects effects for the Borodin concert.

Between Subjects Effects					
	Sum of Squares	**df**	**Mean Square**	**F**	**p**
Pair	0.02003	5	0.00401	1.542	0.183
Texture	0.0365	1	0.0365	14.051	**<.001**
Pair ∗ Texture	0.00412	5	0.000825	0.318	0.901
Duration	0.24612	1	0.24612	94.746	<.001
Residual	0.27796	107	0.0026		

Within Subjects Effects					
	Sum of Squares	**df**	**Mean Square**	**F**	**p**
Position	0.01229	2	0.00614	2.849	0.06
Position ∗ Pair	0.01569	10	0.00157	0.727	0.698
Position ∗ Texture	0.01468	2	0.00734	3.403	**0.035**
Position ∗ Duration	0.00972	2	0.00486	2.253	0.108
Position ∗ Pair ∗ Texture	0.0302	10	0.00302	1.4	0.182
Residual	0.46151	214	0.00216		

Overall, these results indicate that for both pieces, PLVs were reliably higher—hence interpersonal coupling between performers was stronger—for polyphonic than homophonic textures, though this effect of texture varied over the course of musical phrases. Specifically, the effect of texture was reduced at the end of phrases due to decreases in coupling strength in polyphonic textures and increases in coupling strength in homophonic textures.

B. ANALYSIS OF AUDIO FEATURES

While our main analysis focuses on ensemble coordination of co-performer body motion, we conducted an additional analysis to examine the relationship between the synchronization of body movements, which provides visual cues, with ensemble sounds.

Because we do not have multitrack audio recordings for each instrument on a separate track, we computed indirect measures of global ensemble synchronization from stereo auditory recordings of the full ensemble sound. Based on previous research [20], [21], [47], we included estimates of 'pulse clarity' and 'event density', which were calculated using the 'mirpulseclarity' and 'mireventdensity' functions from the MIRtoolbox in MATLAB [87]. Pulse clarity is a feature that reflects the strength of rhythmic beats, while event density indicates the average frequency of events (i.e., the

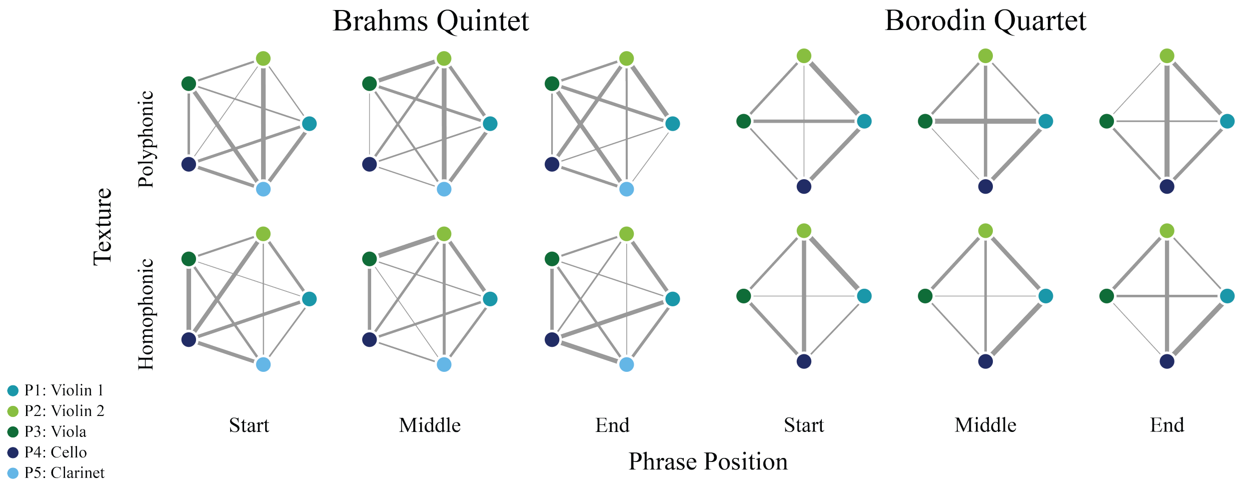


FIGURE 8. Network plots for ensemble PLV data by instrument for each condition (texture and phrase position) in Brahms and Borodin pieces. Edge thickness indicates the coupling strength based on phase locking values averaged across all phrases. Each colored node indicates an instrument played by the performers.

TABLE 4. Mean and standard deviation (SD) of estimates of pulse clarity and event density as a function of texture (homophonic and polyphonic) and phrase position (start, middle, and end) for performances of pieces by Brahms and Borodin.

		Texture					
		Homophonic			Polyphonic		
		Phrase Position					
Piece	Measure	Start	Middle	End	Start	Middle	End
Brahms	**Pulse Clarity**						
	Mean	0.111	0.122	0.167	0.148	0.126	0.135
	SD	0.046	0.057	0.058	0.075	0.047	0.052
	Event Density						
	Mean	1.416	2.098	1.767	1.733	2.105	2.205
	SD	0.844	1.484	0.931	0.846	0.819	0.845
Borodin	**Pulse Clarity**						
	Mean	0.153	0.157	0.152	0.16	0.156	0.144
	SD	0.063	0.044	0.083	0.081	0.054	0.058
	Event Density						
	Mean	2.102	2.202	1.631	2.197	2.025	2.033
	SD	0.946	0.934	0.799	0.798	0.792	1.25

number of events detected per second). Descriptive statistics for these measures are shown in Table 4.

To assess potential effects related to these audio features, we ran a linear mixed effects model analysis using the lmer package [88] in R [89] with PLV as the dependent variable, pulse clarity, event density, texture, and phrase position as predictor fixed effects, and piece as a random effect (with intercepts allowed to vary). Pulse clarity values were arcsine-transformed and event density values were log-transformed prior to analysis. The results revealed a link between PLV and event density. Specifically, a likelihood-ratio test indicated that a model including event density provided a better fit for the data than a model without it (χ^2 (1) = 7.44, p = 0.006), whereas pulse clarity did not contribute significantly to the model (χ^2 (1) = 0.03, p = 0.884). Examination of the output for the full model indicated that PLV values increased with increasing event density (β = 0.031, SE = 0.011, t = 2.767, p = 0.006). These results are consistent with a growing body of evidence that visual and audio cues are both relevant in assessing interpersonal synchronization in musical ensembles [20], [21], [28], [47], [63]. Future work with multitrack audio would allow the relationship between auditory and visual information to be investigated in greater detail, including the assessment of correspondence between leader-follower relations across modalities.

VII. DISCUSSION

The current study had two prime objectives. The first was to develop and present a computational framework and a system to study small-group interactions involving non-verbal social communicative behaviour. huSync can be implemented on video sequences which permits studies to be performed in a naturalistic context without interference associated with motion capture setups. Second, we wanted to put huSync through a test case scenario addressing research questions concerning the relationship between interpersonal coordination of body movements and musical structure. For this specific use case, huSync appears to be a practical alternative technique for quantifying dyadic synchronization between co-performers in musical ensembles based on the automated analysis of human body movements. The outcomes of this investigation are thus methodological and empirical in nature, informing technical aspects and conceptual issues relevant to examining real-time human interaction and non-verbal communication in naturalistic settings.

On the methodological side, our approach progresses through a structured funnel of steps, where kinematic information is gathered from standard video recordings in a marker-less and non-intrusive manner. This kinematic information is then used for quantifying dyadic synchronization between musical performers from within a group ensemble, indexed as phase-locking values, and this routine is done exhaustively for all possible pairs in the group. An advantage of this approach is that it is possible to obtain information about coupling between specific individuals whereas if we take a global measure, we do not necessarily have that level of specificity. The alternative is complicated and rather difficult to interpret when data pertain to natural behavior (in contrast to data from controlled experiments where independent variables are systematically manipulated).

As an empirical case study, we applied the above techniques for body motion analysis to investigate the effects of two aspects of musical structure—texture and phrase position—on the strength of interpersonal coupling in instrumental ensembles. With regard to texture, coupling strength between co-performers was found to be stronger for polyphonic textures than homophonic textures. These textures differ in terms of the presence of a clear leader implied by the relationship between melody and accompaniment parts [32], [33], [34], [35], [36], [37]. Our finding that coupling was stronger for polyphonic textures than homophonic textures could be a consequence of coupling being more evenly distributed across all performers when leadership is ambiguous in polyphonic textures, whereas accompanying performers are more strongly coupled to a single performer serving as a melodic leader in homophonic textures. This interpretation is generally consistent with work on interpersonal coordination in controlled laboratory tasks [15], [16], [17]. Based on this previous work, distributed leadership in polyphonic textures might be associated with greater ensemble synchronization due to heightened mutual adaptation, anticipation, and joint attention [19], [52], [74].

Although position in musical phrases did not have a general effect on the strength of interpersonal coupling, phrase position modulated the effects of musical texture in a manner indicating that the presence of a clear melodic leader was more influential in early than late portions of phrases. Specifically, relatively strong interpersonal coupling for polyphonic textures was evident at the start and middle phrase positions but not at phrase endings. It might be the case that increased coordination demands at phrase endings [74]—that is, just prior to the transition to the next phrase and new musical material—had differential effects in the case of polyphonic and homophonic textures. Specifically, without a clear leader in polyphonic textures, interpersonal synchronization decreased at these challenging coordination points, whereas in homophonic textures, synchronization improved at these points, possibly due to increased attention from the melodic leader. Future work could test this conjecture using eye-tracking technology to monitor eye gaze to quantify eye contact across phrase positions [12], [51], [90].

We evaluated huSync as a system to quantify group coordination by focusing on effects of musical texture and phrase position on interpersonal coupling based on visual information related to body motion. However, we also found evidence for a relationship between ensemble coordination at the level of body motion and sounds in a supplementary analysis of audio tracks from the videos. This correspondence is generally consistent with the results of previous studies of ensemble coordination [4], [28], [63], and more broadly contributes to a growing body of work highlighting the multimodal nature of musical communication [47], [57], [61]. Additionally, it highlights the relevance of both visual and audio cues when assessing interpersonal synchronization in musical groups. Overall findings suggest that huSync is sensitive to modulations of interpersonal coupling related to ambiguity in leadership and coordination demands in standard video recordings of naturalistic human group interaction.

VIII. CONCLUSION

The proposed 'huSync' framework and system provides a reliable and non-intrusive alternative to current methods for the automated analysis of human body movements and associated qualities such as degrees of interpersonal synchronization. It can help in the study of such niche but ecologically valid aspects of human movement sciences, opening an avenue where marker-less technologies can be utilized extensively. This is evident in the use case of musical ensemble performances, where we evaluated the method, and also has potential to be extended to capturing non-verbal social signals in other domains of group behaviour and human interaction more generally. As a concrete outcome, we provide a well-structured jupyter notebook (link) that includes functions designed and implemented to process the data extracted from pose estimation algorithms by converting them into structured csv files, followed by the calculation routine for computing phase locking values, thus quantifying the dyadic synchronization. An especially promising benefit of the huSync model is that it can be applied to standard videos recorded across a wide range of contexts, opening the door to analyzing vast troves of historical material available in archives and on the Internet. The outcomes of the research will thus potentially have broad impact across diverse disciplines including computer science, psychology and cognitive neuroscience, and music psychology. The methodological applications of huSync can be leveraged for further empirical discoveries related to human joint action, group behavior and social cognition [10].

A. OBSERVED LIMITATIONS

There are several areas to improve upon and overcome in future research. At present, there exists a higher amount of noise in tracking conventional video as compared to marker-based systems. This issue becomes particularly acute when examining at higher-order kinematic variables, such as velocity and acceleration (because computing derivatives via differentiation amplifies noise), which is one reason why we focused on distance data. Pose estimation algorithms provide better results with regard to recognizing, isolating, and predicting the pose of participants in videos where the foreground and background are well-differentiated. This suggests that figure-ground differentiation is an important aspect of quality control.

Additionally, the seating position and direction of motion trajectories exhibited by participants is an important aspect to take note of, and plays an influential role in quantifying dyadic synchronization. For our use case, the head moves predominantly in a back-and-forth manner during moments

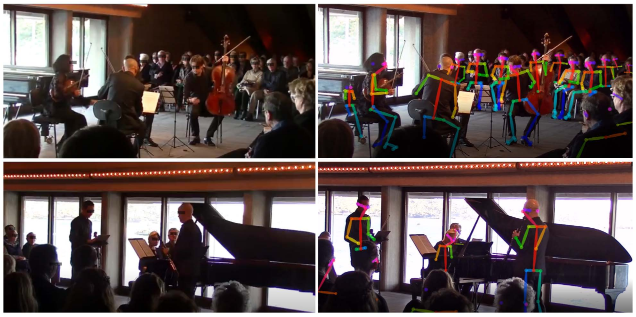


FIGURE 9. Images from a performances of the String Quartet No. 2 composed by Alexander Borodin (Top Left) and a trio for clarinet, viola, and piano composed by Robert Schumann (Bottom Left) along with overlayed keypoints on the right.

of interpersonal coordination and heightened synchronized states. Thus, orientation of the participants relative to the camera view influences the amount of motion information that can be detected and extracted. Fig. 9 shows still images, along with the implementation of a pose estimation algorithm, from videos of performances of different musical pieces, including a trio and quartet, from a concert in a smaller venue than the performances analyzed in the current paper. These video recordings are highly challenging for pose estimation algorithms to be implemented upon. The seating position of the performers makes it unfeasible to track all individuals clearly within a single camera view. Also, each frame is crowded with many individuals in the audience that do not need to be tracked along with multiple occluding objects, such as the piano, chairs, an individual serving as page tuner, and heads of the audience to name a few. These add to the overall visual clutter and complexity that makes it difficult to localize the performers' bodies and to extract reliable data for analysis.

While our focus in this paper remains on small-group setups, problems concerning occlusions and overlaps are evident when analyzing videos of larger groups, in particular those with multi-row ensembles of musicians where occlusions can be caused by both instruments and co-performers. This can invite multiple challenges especially with loss of movement related information. A possible technique that could be investigated in future works might involve the use of multiple cameras at different positions, where data can be reconstructed from multiple perspectives using synchronized multi-view video recordings [91], [92]. This is essentially similar in principle to optical MoCap [93] but is still less invasive and more portable.

B. FUTURE RESEARCH

This study is part of the European Horizon 2020 FETPROACTIVE EnTimeMent Project, on novel time-adaptive technologies operating at multiple time scales in a multi-layered approach. In the future, this work will be extended in line with the overarching goals of the EnTimeMent Project by exploring various techniques to examine how interpersonal coordination unfolds at multiple timescales, which could involve applying these techniques in different experimental setups. The candidate techniques include:

1) Multi-Event Class Synchronization if we have discrete information (landmarks such as points at which co-performer makes eye contact) to help us measure synchronization between two relevant events that belong to different event classes and detected in multiple time series [94]; and
2) Granger Causality to quantify mutual influence / leadership by studying the directionality of coupling (which should be more evident when there is a clear leadership hierarchy, as in homophonic textures), helping us look at effects of musical structure on group coordination and communication simultaneously, at short timescales related to musical beats and longer timescales related to expressive body sway [50], [95], [96].

ACKNOWLEDGMENT

The authors would like to thank the members of the Omega Ensemble, especially Artistic Director and Founder David Rowden. They also thank videographer Bruce Terry for making available the unprocessed video files and Stephanie Laun for assistance with video annotation.

REFERENCES

[1] N. Jaouedi, N. Boujnah, O. Htiwich, and M. S. Bouhlel, "Human action recognition to human behavior analysis," in *Proc. 7th Int. Conf. Sci. Electron., Technol. Inf. Telecommun. (SETIT)*, Dec. 2016, pp. 263–266.

[2] B. Xiao, P. Georgiou, B. Baucom, and S. S. Narayanan, "Head motion modeling for human behavior analysis in dyadic interaction," *IEEE Trans. Multimedia*, vol. 17, no. 7, pp. 1107–1119, Jul. 2015.

[3] L. Bishop, "Collaborative musical creativity: How ensembles coordinate spontaneity," *Frontiers Psychol.*, vol. 9, p. 1285, Jul. 2018.

[4] L. Bishop and W. Goebl, "Communication for coordination: Gesture kinematics and conventionality affect synchronization success in piano duos," *Psychol. Res.*, vol. 82, no. 6, pp. 1177–1194, Nov. 2018.

[5] G. Luck, S. Saarikallio, B. Burger, M. R. Thompson, and P. Toiviainen, "Effects of the big five and musical genre on music-induced movement," *J. Res. Personality*, vol. 44, no. 6, pp. 714–720, Dec. 2010.

[6] J. K. Vuoskoski, W. F. Thompson, D. McIlwain, and T. Eerola, "Who enjoys listening to sad music and why? *Music Perception*, vol. 29, no. 3, pp. 311–317, 2011.

[7] C. Wöllner and P. E. Keller, "Music with others: Ensembles, conductors, and interpersonal coordination," in *The Routledge Companion to Music Cognition*. Evanston, IL, USA: Routledge, 2017, pp. 313–324.

[8] M. M. Wanderley, "Quantitative analysis of non-obvious performer gestures," in *Proc. Int. Gesture Workshop*. Berlin, Germany: Springer, 2001, pp. 241–253.

[9] L. Bishop and W. Goebl, "Beating time: How ensemble musicians' cueing gestures communicate beat position and tempo," *Psychol. Music*, vol. 46, no. 1, pp. 84–106, Jan. 2018.

[10] M. Tal-Shmotkin and A. Gilboa, "Do behaviors of string quartet ensembles represent self-managed teams?" *Team Perform. Manage., Int. J.*, vol. 19, nos. 1–2, pp. 57–71, Mar. 2013.

[11] L. Badino, A. D'Ausilio, D. Glowinski, A. Camurri, and L. Fadiga, "Sensorimotor communication in professional quartets," *Neuropsychologia*, vol. 55, pp. 98–104, Mar. 2014.

[12] L. Bishop, C. Cancino-Chacón, and W. Goebl, "Moving to communicate, moving to interact: Patterns of body motion in musical duo performance," *Music Perception, Interdiscipl. J.*, vol. 37, no. 1, pp. 1–25, 2019.

[13] D. Glowinski, M. Mancini, R. Cowie, A. Camurri, C. Chiorri, and C. Doherty, "The movements made by performers in a skilled quartet: A distinctive pattern, and the function that it serves," *Frontiers Psychol.*, vol. 4, p. 841, Nov. 2013.

[14] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2334–2343.

[15] G. Novembre, M. Varlet, S. Muawiyath, C. J. Stevens, and P. E. Keller, "The E-music box: An empirical method for exploring the universal capacity for musical production and for social interaction through music," *Roy. Soc. Open Sci.*, vol. 2, no. 11, Nov. 2015, Art. no. 150286.

[16] L. Noy, E. Dekel, and U. Alon, "The mirror game as a paradigm for studying the dynamics of two people improvising motion together," *Proc. Nat. Acad. Sci. USA*, vol. 108, no. 52, pp. 20947–20952, Dec. 2011.

[17] M. Varlet, S. Nozaradan, P. Nijhuis, and P. E. Keller, "Neural tracking and integration of 'self' and 'other' in improvised interpersonal coordination," *NeuroImage*, vol. 206, Feb. 2020, Art. no. 116303.

[18] B. Schögler, "Studying temporal co-ordination in jazz duets," *Musicae Scientiae*, vol. 3, no. 1, pp. 75–91, Sep. 1999.

[19] P. E. Keller, G. Novembre, and M. J. Hove, "Rhythm in joint action: Psychological and neurophysiological mechanisms for real-time interpersonal coordination," *Phil. Trans. Roy. Soc. B, Biol. Sci.*, vol. 369, no. 1658, Dec. 2014, Art. no. 20130394.

[20] T. Eerola, K. Jakubowski, N. Moran, P. E. Keller, and M. Clayton, "Shared periodic performer movements coordinate interactions in duo improvisations," *Roy. Soc. Open Sci.*, vol. 5, no. 2, Feb. 2018, Art. no. 171520.

[21] K. Jakubowski, T. Eerola, A. Blackwood Ximenes, W. K. Ma, M. Clayton, and P. E. Keller, "Multimodal perception of interpersonal synchrony: Evidence from global and continuous ratings of improvised musical duo performances," *Psychomusicol., Music, Mind, Brain*, vol. 30, no. 4, p. 159, 2020.

[22] P. E. Keller, G. Novembre, and J. Loehr, "Musical ensemble performance: Representing self, other and joint action outcomes," in *Shared Representations: Sensorimotor Foundations of Social Life*. Cambridge, U.K.: Cambridge Univ. Press, 2016, p. 280.

[23] A. Paxton and R. Dale, "Frame-differencing methods for measuring bodily synchrony in conversation," *Behav. Res. Methods*, vol. 45, no. 2, pp. 329–343, Jun. 2013.

[24] F. Ramseyer and W. Tschacher, "Synchrony: A core concept for a constructivist approach to psychotherapy," *Constructivism Hum. Sci.*, vol. 11, nos. 1–2, pp. 150–171, 2006.

[25] C. Nagaoka and M. Komori, "Body movement synchrony in psychotherapeutic counseling: A study using the video-based quantification method," *IEICE Trans. Inf. Syst.*, vols. 91, no. 6, pp. 1634–1640, Jun. 2008.

[26] M. Komori, "A video-based quantification method of body movement synchrony: An application for dialogue in counseling," *Jpn. J. Interpersonal Social Psychol.*, vol. 7, pp. 41–48, 2007.

[27] T. Yokozuka, E. Ono, Y. Inoue, K.-I. Ogawa, and Y. Miyake, "The relationship between head motion synchronization and empathy in unidirectional face-to-face communication," *Frontiers Psychol.*, vol. 9, p. 1622, Sep. 2018.

[28] M. Ragert, T. Schroeder, and P. E. Keller, "Knowing too little or too much: The effects of familiarity with a co-performer's part on interpersonal coordination in musical ensembles," *Frontiers Psychol.*, vol. 4, p. 368, 2013.

[29] D. Glowinski, F. Dardard, G. Gnecco, S. Piana, and A. Camurri, "Expressive non-verbal interaction in a string quartet: An analysis through head movements," *J. Multimodal User Interfaces*, vol. 9, no. 1, pp. 55–68, Mar. 2015.

[30] K. Jakubowski, T. Eerola, P. Alborno, G. Volpe, A. Camurri, and M. Clayton, "Extracting coarse body movements from video in music performance: A comparison of automated computer vision techniques with motion capture data," *Frontiers Digit. Hum.*, vol. 4, p. 9, Apr. 2017.

[31] G. Luck and P. Toiviainen, "Ensemble musicians' synchronization with conductors' gestures: An automated feature-extraction analysis," *Music Perception*, vol. 24, no. 2, pp. 189–200, Dec. 2006.

[32] B. Burger, M. R. Thompson, G. Luck, S. H. Saarikallio, and P. Toiviainen, "Hunting for the beat in the body: On period and phase locking in music-induced movement," *Frontiers Human Neurosci.*, vol. 8, p. 903, Nov. 2014.

[33] P. Toiviainen, V. Alluri, E. Brattico, M. Wallentin, and P. Vuust, "Capturing the musical brain with lasso: Dynamic decoding of musical features from fMRI data," *NeuroImage*, vol. 88, pp. 170–180, Mar. 2014.

[34] C. Cornejo, Z. Cuadros, R. Morales, and J. Paredes, "Interpersonal coordination: Methods, achievements, and challenges," *Frontiers Psychol.*, vol. 8, p. 1685, Sep. 2017.

[35] G. Welch and E. Foxlin, "Motion tracking: No silver bullet, but a respectable arsenal," *IEEE Comput. Graph. Appl.*, vol. 22, no. 6, pp. 24–38, Dec. 2002.

[36] A. Hadjakos, T. Großhauser, and W. Goebl, "Motion analysis of music ensembles with the Kinect," in *Proc. Conf. Interfaces Musical Expression*, 2013, pp. 106–110.

[37] K. Yun, K. Watanabe, and S. Shimojo, "Interpersonal body and neural synchronization as a marker of implicit social interaction," *Sci. Rep.*, vol. 2, no. 1, pp. 1–8, Dec. 2012.

[38] A. C. E. Onslow, R. Bogacz, and M. W. Jones, "Quantifying phase–amplitude coupling in neuronal network oscillations," *Prog. Biophys. Mol. Biol.*, vol. 105, nos. 1–2, pp. 49–57, Mar. 2011.

[39] O. Jensen and J. E. Lisman, "An oscillatory short-term memory buffer model can account for data on the sternberg task," *J. Neurosci.*, vol. 18, no. 24, pp. 10688–10699, 1998.

[40] O. Jensen, "Maintenance of multiple working memory items by temporal segmentation," *Neuroscience*, vol. 139, no. 1, pp. 237–249, Apr. 2006.

[41] J. E. Lisman and O. Jensen, "The theta-gamma neural code," *Neuron*, vol. 77, no. 6, pp. 1002–1016, Mar. 2013.

[42] J. Vosskuhl, R. J. Huster, and C. S. Herrmann, "Increase in short-term memory capacity induced by down-regulating individual theta frequency via transcranial alternating current stimulation," *Frontiers Hum. Neurosci.*, vol. 9, p. 257, May 2015.

[43] S. Aydore, D. Pantazis, and R. M. Leahy, "A note on the phase locking value and its properties," *NeuroImage*, vol. 74, pp. 231–244, Jul. 2013.

[44] S. Cole and B. Voytek, "Cycle-by-cycle analysis of neural oscillations," *J. Neurophysiol.*, vol. 122, no. 2, pp. 849–861, Aug. 2019.

[45] F. Mormann, J. Fell, N. Axmacher, B. Weber, K. Lehnertz, C. E. Elger, and G. Fernández, "Phase/amplitude reset and theta–gamma interaction in the human medial temporal lobe during a continuous word recognition memory task," *Hippocampus*, vol. 15, no. 7, pp. 890–900, 2005.

[46] M. Rosenblum, P. Tass, J. Kurths, J. Volkmann, A. Schnitzler, and H.-J. Freund, "Detection of phase locking from noisy data: Application to magnetoencephalography," in *Chaos In Brain?*. Singapore: World Scientific, 2000, pp. 34–51.

[47] M. Clayton, K. Jakubowski, T. Eerola, P. E. Keller, A. Camurri, G. Volpe, and P. Alborno, "Interpersonal entrainment in music performance: Theory, method, and model," *Music Perception, Interdiscipl. J.*, vol. 38, no. 2, pp. 136–194, 2020.

[48] G. Volpe, A. D'Ausilio, L. Badino, A. Camurri, and L. Fadiga, "Measuring social interaction in music ensembles," *Phil. Trans. Roy. Soc. B, Biol. Sci.*, vol. 371, no. 1693, May 2016, Art. no. 20150377.

[49] A. D'Ausilio, G. Novembre, L. Fadiga, and P. E. Keller, "What can music tell us about social interaction?" *Trends Cognit. Sci.*, vol. 19, no. 3, pp. 111–114, Mar. 2015.

[50] A. Chang, H. E. Kragness, S. R. Livingstone, D. J. Bosnyak, and L. J. Trainor, "Body sway reflects joint emotional expression in music ensemble performance," *Sci. Rep.*, vol. 9, no. 1, pp. 1–11, Dec. 2019.

[51] S. Kawase, "Gazing behavior and coordination during piano duo performance," *Attention, Perception, Psychophys.*, vol. 76, no. 2, pp. 527–540, Feb. 2014.

[52] E. P. Keller, "Joint action in music performance," in *Enacting Intersubjectivity: A Cognitive and Social Perspective to the Study of Interactions*. Amsterdam, The Netherlands: IOS Press, 2008.

[53] M. Nusseck and M. M. Wanderley, "Music and motion—How music-related ancillary body movements contribute to the experience of music," *Music Perception*, vol. 26, no. 4, pp. 335–353, Apr. 2009.

[54] E. King and C. Waddington, *Music and Empathy*. New York, NY, USA: Routledge, 2017.

[55] J. Decety and W. Ickes, *The Social Neuroscience of Empathy*. Cambridge, MA, USA: MIT Press, 2011.

[56] A. Camurri, B. Mazzarino, M. Ricchetti, R. Timmers, and A. Volpe, "Multimodal analysis of expressive gesture in music and dance performances," in *Proc. Int. Gesture Workshop*. Berlin, Germany: Springer, 2003, pp. 20–39.

[57] J. W. Davidson, "Movement and collaboration in musical performance," in *The Oxford Handbook of Music Psychology*. Oxford, U.K.: Oxford Univ. Press, 2009, pp. 364–376.

[58] J. Davidson and M. C. Broughton, "Bodily mediated coordination, collaboration, and communication in music performance," in *The Oxford Handbook of Music Psychology*. Oxford, U.K.: Oxford Univ. Press, 2016.

[59] A. Jensenius, M. Wanderley, R. Godøy, and M. Leman, "Concepts and methods in research on music-related gestures," in *Musical Gestures: Sound, Movement, and Meaning*. New York, NY, USA: Routledge, 2010, pp. 12–35.

[60] J. MacRitchie, B. Buck, and N. J. Bailey, ''Inferring musical structure through bodily gestures,'' *Musicae Scientiae*, vol. 17, no. 1, pp. 86–108, Mar. 2013.

[61] B. Vines, C. Krumhansl, M. Wanderley, and D. Levitin, ''Cross-modal interactions in the perception of musical performance,'' *Cognition*, vol. 101, no. 1, pp. 80–113, Aug. 2006.

[62] W. Goebl and C. Palmer, ''Synchronization of timing and motion among performing musicians,'' *Music Perception*, vol. 26, no. 5, pp. 427–438, Jun. 2009.

[63] P. E. Keller and M. Appel, ''Individual differences, auditory imagery, and the coordination of body movements and sounds in musical ensembles,'' *Music Perception, Interdisciplinary J.*, vol. 28, no. 1, pp. 27–46, 2010.

[64] N. Latif, A. V. Barbosa, E. Vatiokiotis-Bateson, M. S. Castelhano, and K. G. Munhall, ''Movement coordination during conversation,'' *PLoS ONE*, vol. 9, no. 8, Aug. 2014, Art. no. e105036.

[65] K. L. Marsh, M. J. Richardson, and R. C. Schmidt, ''Social connection through joint action and interpersonal coordination,'' *Topics Cognit. Sci.*, vol. 1, no. 2, pp. 320–339, Apr. 2009.

[66] Z. Néda, E. Ravasz, Y. Brechet, T. Vicsek, and A.-L. Barabási, ''The sound of many hands clapping,'' *Nature*, vol. 403, no. 6772, pp. 849–850, 2000.

[67] S. Kirschner and M. Tomasello, ''Joint drumming: Social context facilitates synchronization in preschool children,'' *J. Experim. Child Psychol.*, vol. 102, no. 3, pp. 299–314, Mar. 2009.

[68] G. Varni, M. Mancini, L. Fadiga, A. Camurri, and G. Volpe, ''The change matters! Measuring the effect of changing the leader in joint music performances,'' *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 700–712, Apr. 2022.

[69] M. J. Hove and J. L. Risen, ''It's all in the timing: Interpersonal synchrony increases affiliation,'' *Social Cognition*, vol. 27, no. 6, pp. 949–960, Dec. 2009.

[70] A. Chang, S. R. Livingstone, D. J. Bosnyak, and L. J. Trainor, ''Body sway reflects leadership in joint music performance,'' *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 21, pp. E4134–E4141, May 2017.

[71] A. E. Walton, M. J. Richardson, P. Langland-Hassan, and A. Chemero, ''Improvisation and the self-organization of multiple musical bodies,'' *Frontiers Psychol.*, vol. 6, p. 313, Apr. 2015.

[72] P. E. Keller, ''Attentional resource allocation in musical ensemble performance,'' *Psychol. Music*, vol. 29, no. 1, pp. 20–38, Apr. 2001.

[73] E. Goodman, ''Ensemble performance,'' in *Musical Performance*. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[74] P. E. Keller, ''Ensemble performance: Interpersonal alignment of musical expression,'' in *Expressiveness in Music Performance: Empirical Approaches Across Styles and Cultures*. Oxford, U.K.: Oxford Univ. Press, 2014, pp. 260–282.

[75] G. Varni, G. Volpe, and A. Camurri, ''A system for real-time multimodal analysis of nonverbal affective social interaction in user-centric media,'' *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 576–590, Oct. 2010.

[76] S. D'Amario, H. Daffern, and F. Bailes, ''Synchronization in singing duo performances: The roles of visual contact and leadership instruction,'' *Frontiers Psychol.*, vol. 9, p. 1208, Jul. 2018.

[77] D. Glowinski, F. Bracco, C. Chiorri, and D. Grandjean, ''Music ensemble as a resilient system. Managing the unexpected through group interaction,'' *Frontiers Psychol.*, vol. 7, p. 1548, Oct. 2016.

[78] T. Yoshida, S. Takeda, and S. Yamamoto, ''The application of entrainment to musical ensembles,'' in *Proc. 2nd Int. Conf. Music Artif. Intell. (ICMAI)*, Edinburgh, Scotland, 2002, pp. 1–11.

[79] A. Camurri, G. Volpe, S. Piana, M. Mancini, R. Niewiadomski, N. Ferrari, and C. Canepa, ''The dancer in the eye: Towards a multi-layered computational framework of qualities in movement,'' in *Proc. 3rd Int. Symp. Movement Comput.*, Jul. 2016, pp. 1–7.

[80] A. Savitzky and M. J. E. Golay, ''Smoothing and differentiation of data by simplified least squares procedures,'' *Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, 1964.

[81] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, and J. Bright, ''SciPy 1.0: Fundamental algorithms for scientific computing in Python,'' *Nature Methods*, vol. 17, no. 3, pp. 261–272, 2020.

[82] A. Lillywhite, D. Glowinski, A. Camurri, and F. Pollick, ''Using fMRI and intersubject correlation to explore brain activity during audiovisual observation of a string quartet,'' *Frontiers Hum. Neurosci.*, 2015.

[83] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, and R. Kern, ''Array programming with numpy,'' *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020.

[84] *Nijmegen: Max Planck Institute for Psycholinguistics, the Language Archive*, Elan (Version 6.2)[Comput. Softw.], Max Planck Inst. Psycholinguistics, Lang. Arch., Nijmegen, The Netherlands, 2019.

[85] J. M. Hurtado, L. L. Rubchinsky, and K. A. Sigvardt, ''Statistical method for detection of phase-locking episodes in neural oscillations,'' *J. Neurophysiol.*, vol. 91, no. 4, pp. 1883–1898, Apr. 2004.

[86] (2021). *The Jamovi Project*. Jamovi (Version 1.6) [Computer Software]. [Online]. Available: https://www.jamovi.org

[87] O. Lartillot, P. Toiviainen, and T. Eerola, ''A MATLAB toolbox for music information retrieval,'' in *Data Analysis, Machine Learning and Applications*. Berlin, Germany: Springer, 2008, pp. 261–268.

[88] D. Bates, M. Maechler, B. Bolker, S. Walker, R. H. B. Christensen, H. Singmann, B. Dai, F. Scheipl, and G. Grothendieck, ''Package 'LME4,''' *Linear Mixed-Effects Models Using S4 Classes. R Package Version*, vol. 1, no. 6, pp. 1–48, 2011.

[89] *A Language and Environment for Statistical Computing*, R Found. Stat. Comput., Vienna, Austria, 2021.

[90] E. King and J. Ginsborg, ''Gestures and glances: Interactions in ensemble rehearsal,'' in *New Perspectives on Music and Gesture*. Evanston, IL, USA: Routledge, 2016, pp. 203–228.

[91] Z. Tang, R. Gu, and J.-N. Hwang, ''Joint multi-view people tracking and pose estimation for 3D scene reconstruction,'' in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

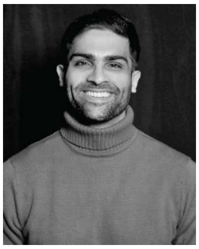
[92] M.-H. Nguyen, C.-C. Hsiao, W.-H. Cheng, and C.-C. Huang, ''Practical 3D human skeleton tracking based on multi-view and multi-Kinect fusion,'' *Multimedia Syst.*, vol. 28, no. 2, pp. 529–552, Apr. 2022.

[93] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel, ''Free-viewpoint video of human actors,'' *ACM Trans. Graph.*, vol. 22, no. 3, pp. 569–577, 2003.

[94] P. Alborno, G. Volpe, M. Mancini, R. Niewiadomski, S. Piana, and A. Camurri, ''The multi-event-class synchronization (MECS) algorithm,'' 2019, *arXiv:1903.09530*.

[95] P. M. Hilt, L. Badino, A. D'Ausilio, G. Volpe, S. Tokay, L. Fadiga, and A. Camurri, ''Multi-layer adaptation of group coordination in musical ensembles,'' *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, Dec. 2019.

[96] A. D'Ausilio, L. Badino, Y. Li, S. Tokay, L. Craighero, R. Canto, Y. Aloimonos, and L. Fadiga, ''Leadership in orchestra emerges from the causal relationships of movement kinematics,'' *PLoS ONE*, vol. 7, no. 5, May 2012, Art. no. e35757.



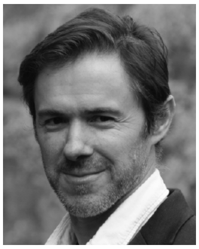
SANKET RAJEEV SABHARWAL received the Bachelor of Engineering degree (B.E.) in information technology from the University of Mumbai, the Bachelor of Law (LLB) degree specializing in intellectual property, and the Master of Computer Science from the University of Genoa, where he is currently pursuing the Ph.D. degree in computer science. He is a part of the Research Group, Casa Paganini–InfoMus. He worked in the tech industry for close to five years. His main research interests include multimodal interfaces, and computational models for extracting non-verbal expressive features and social signals.

MANUEL VARLET is currently an Associate Professor at the MARCS Institute for Brain, Behaviour and Development. His research investigates the perceptual-motor processes underlying human performances and their changes throughout life, with expertise and pathologies, using behavioral, neuroimaging and brain stimulation methods. He is particularly interested in identifying the neural, informational and biomechanical mechanisms that support and enhance agent-environment and multi-agent coordination. He employs a range of motion capture, EEG, TMS, tDCS/tACS, and virtual reality technologies to investigate these perceptual-motor processes, as well as a wide variety of contemporary linear and nonlinear time-series analysis and dynamical modeling techniques.

MATTHEW BREADEN received the Bachelor of Music degree from the University of Melbourne, in 1994, the Secondary Diploma of Education degree from Australian Catholic University, in 1995, and the Graduate Diploma and Ph.D. degrees in creative music therapy from Western Sydney University, in 2004 and 2020, respectively. He is currently a Research Officer at Western Sydney University, and is also a Registered Music Therapist in Australia. His research interests include investigating how music can help people with autism develop social interaction skills, and how interactive music-making can improve quality of life for people with dementia. He is a member of the Australian Music Therapy Association and the Australian Music Psychology Society, and is the Chair-Elect of the Special Music Education and Music Therapy Commission of the International Society for Music Education.

GUALTIERO VOLPE received the M.Sc. degree in computer engineering and the Ph.D. degree in electronic and computer engineering from the University of Genoa, Italy, in 1999 and 2003, respectively. Since 2014, he has been an Associate Professor at the DIBRIS, University of Genoa. His research interests include intelligent and affective human–machine interaction, social signal processing, sound and music computing, modeling and real-time analysis of expressive content, and multimodal systems.

ANTONIO CAMURRI received the Ph.D. degree in computer engineering. He is currently a Full Professor at the University of Genoa. He is also the Scientific Director of Casa Paganini-InfoMus, DIBRIS, University of Genoa. He is also the Co-Director of the Joint Research Laboratory ARIEL (Augmented Rehabilitation Laboratory, Giannina Gaslini Children Hospital, and the University of Genoa). He is also a Co-ordinator of six European funded projects, such as FP5, FP7, and Horizon 2020 (http://dance.dibris.unige.it; http://entimement.dibris.unige.it); a principal investigator in about 20 EU-funded projects and in contracts with industry and cultural institutions; and a co-owner of software patents. He is the author of over 150 scientific publications in international scientific journals and conferences. His main research interests include human–computer interaction inspired by artistic and humanistic research, and multimodal interactive systems and computational models of non-verbal full-body expressive gesture, emotion, and social signals; interactive multimodal systems for performing arts, active experience of cultural content, cultural wellbeing, and cognitive-motor rehabilitation. He is a member of the Editorial Boards of the *Journal of New Music Research* and *PLOS One*, and the ESF College of Expert Reviewers. He is a member of the Board of Directors of Museo Palazzo Reale of Genoa.



PETER E. KELLER received the Bachelor of Music, B.A., and Ph.D. degrees in psychology from the University of New South Wales, Sydney, NSW, Australia, in 1994, 1995, and 2001, respectively. He is currently a Professor of neuroscience with the Center for Music in the Brain and the Department of Clinical Medicine, Aarhus University, Denmark, and a Professor of neuroscience of music with the MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Australia. Previously, he held research positions at Haskins Laboratories (New Haven, CT, USA), the Max Planck Institute for Psychological Research (Munich, Germany), and the Max Planck Institute for Human Cognitive and Brain Sciences (Leipzig, Germany), where he led the Max Planck Research Group for Music Cognition and Action. His research is aimed at understanding the behavioral and brain bases of human interaction in musical contexts, with specific focus on the cognitive and motor processes that enable ensemble musicians to coordinate with one another. He is a member of the Australasian Cognitive Neuroscience Society and the Society for Music Perception and Cognition, and a Founding Member of the Australian Music Psychology Society. His past academic honors include an Australian Research Council Future Fellowship, a Leverhulme Trust Visiting Professorship at Durham University (U.K.), a Visiting Professorship at Central European University in Budapest (Hungary), and a European Institutes for Advanced Study (EURIAS) Fellowship at the Wissenschaftskolleg zu Berlin (Germany). He served as an Editor of the interdisciplinary journal *Empirical Musicology Review*. He was an Editorial Board Member at *Advances in Cognitive Psychology*, *Royal Society Open Science*, and *Psychological Research*. He is also on editorial boards at *Music Perception* and *Psychomusicology: Music, Mind, and Brain*.