Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-021-00751-8 OPEN

The Cyborg Philharmonic: Synchronizing interactive musical performances between humans and machines

Sutirtha Chakraborty [1]✉, Sourav Dutta [2] & Joseph Timoney[1]

Music offers a uniquely abstract way for the expression of human emotions and moods, wherein melodic harmony is achieved through a succinct blend of pitch, rhythm, tempo, texture, and other sonic qualities. The emerging field of "*Robotic Musicianship*" focuses on developing machine intelligence, in terms of algorithms and cognitive models, to capture the underlying principles of musical perception, composition, and performance. The capability of new-generation robots to manifest music in a human-like artistically expressive manner lies at the intersection of engineering, computers, music, and psychology; promising to offer new forms of creativity, sharing, and interpreting musical impulses. This manuscript explores how real-time collaborations between humans and machines might be achieved by the integration of technological and mathematical models from *Synchronization* and *Learning*, with precise configuration for the seamless generation of melody in tandem, towards the vision of *human–robot symphonic orchestra*. To explicitly capture the key ingredients of a good symphony—synchronization and anticipation—this work discusses a possible approach based on the joint strategy of: (i) *Mapping*— wherein mathematical models for oscillator coupling like Kuramoto could be used for establishing and maintaining synchronization, and (ii) *Modelling*—employing modern deep learning predictive models like Neural Network architectures to anticipate (or predict) future state changes in the sequence of music generation and pre-empt transitions in the coupled oscillator sequence. It is hoped that this discussion will foster new insights and research for better "real-time synchronized human-computer collaborative interfaces and interactions".

[1] Department of Computer Science, Maynooth University, Maynooth, Ireland. [2] Huawei Research Center, Dublin, Ireland. ✉email: sutirtha.chakraborty@mu.ie

Introduction

Music has enraptured humans for ages and offers an unbridled avenue for the expression of emotions, intellect, passions and moods, not only for us but also for other creatures of nature. A heart touching melody is reached through the perfect meld of tempo, rhythm, texture, pitch, and other sonic features. Being one of the oldest creative art forms[1], music has inspired research among scholars across different disciplines to study the underlying aesthetics and characteristics for achieving "the harmony of spheres" (Gaizauskas, 1974).

Musicology, the traditional field for the study of music, explores various dimension of music in terms of theory and evolution (Boretz, 1995), along with its societal associations (Soley and Hannon, 2010), therapeutic benefits (Davis et al., 2008), and connections between physics and artistic expressions (Perc, 2020). Recent developments in the field of artificial intelligence (AI) has led to research in *Cognitive Musicology*, which involves the modelling of musical knowledge with computer models, enabling the combination of cognitive science and musicology (Laske, 1999) for developing physical–mechanical systems to produce music.

Musical Performances, on the other hand, represent a different manifestation of physical expression of music, wherein several musicians perform in a precise synchronized fashion involving a temporal change of roles and possibly improvised chord progression—posing a varied set of challenges for seamless interaction between humans and computational models. It is here that we seek to investigate how real-time cooperation between machines and humans could be achieved through technologies and models from *synchronization* and *learning*, with their exact configuration for the generation of melody alongside each other, to achieve the vision of *human–robot symphonic orchestra*.

Technical challenges. One of the major characteristics of *musical performances*, or such *multi-ensemble topologies*, is that musicians do not engage with rigidity—rather, they play, move, and act as per the 'feel' of the music, and in tandem. Further, the role of different musicians in an orchestra changes following a rhythm, communicated in real-time either through mutual non-verbal gestures or by the principal conductor. Thus, there is a smooth transition of the musicians between the roles of 'leader' and 'follower'. This inherent inter-dependency can be imagined as a fully connected dynamic network topology among the musicians as shown in Fig. 1. This multitude of challenges, in terms of (a) synchronized generation of the desired musical chord, (b) dynamic detection of role fluidity, along with (c) understability of gestures and non-verbal communications in real-time, necessitate an advanced cognitive operating level for musical robots in such interactive, synchronized and collaborative environments (Chakraborty and Timoney, 2020). We discuss how recent developments in Machine Learning approaches can be integrated with traditional oscillator coupling techniques to capture the underlying model and dynamics and predict fine-grained temporal variations in group musical performances for achieving enhanced human–robot synchronization.

Proposal outline. To explicitly capture the key features of musical performances and alleviate the above challenges, this manuscript proposes a joint strategy of: (i) *Mapping*—responsible for ensuring control and sensing of the components of musical instruments along with the parameters in sound synthesis, and (ii) *Modelling*—helps in anticipating state changes in the music generation sequence and captures the overall representation of

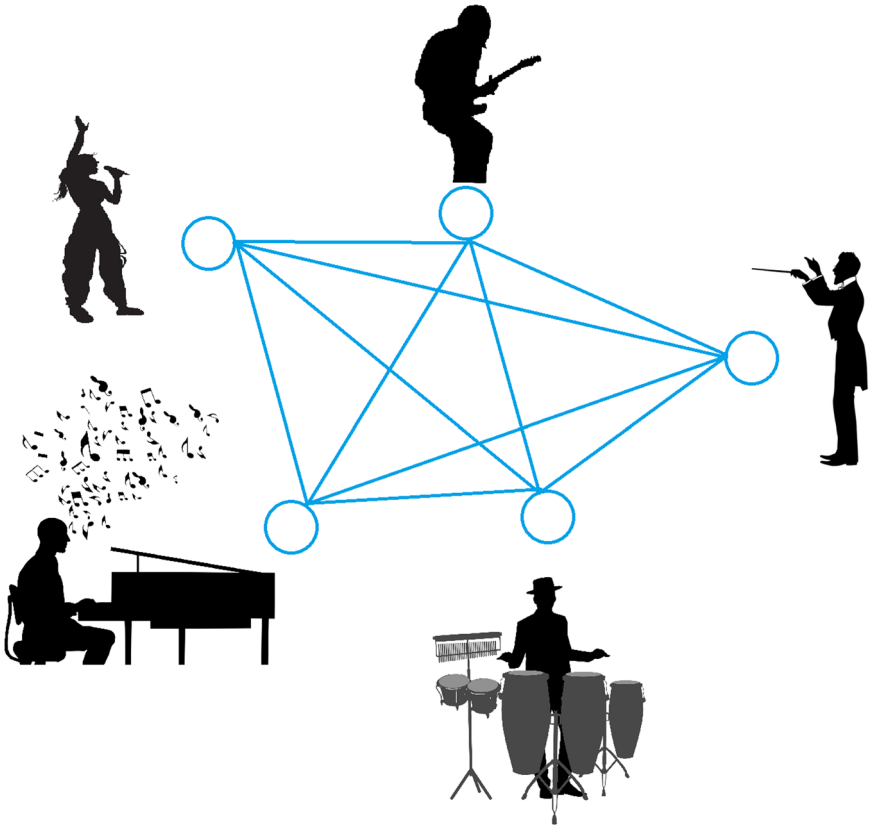

Fig. 1 Modelling the interdependencies of an ensemble musical performance as a dynamic network. Interconnected network among musicians during a musical ensemble depicts individual musician to be connected to others to synchronize their performances with the leader.

the musical process. We consider each musician (human or machine) to act as an independent oscillator, wherein mathematical models for oscillator coupling like *Kuramoto* could be used for establishing and maintaining synchronization. This takes into account the generation of in-sync musical parameters like beat and tempo for creating a harmonious melody, and forms the mapping phase. The modelling stage employs deep learning predictive models like Neural Network architectures to capture long-distance patterns and associations in the music sequence to learn the overall musical feature space comprising beat, onset, sustain, decay and pitch. This would enable the system to anticipate or predict future state changes in the sequence of music generation, enabling faster synchronization of the coupled oscillator sequences, and seamless prediction of role transitions.

Finally, as food for thought, we put forth the possibility of future use of *Swarm Robotic Intelligence* for musical ensembles with only a troupe of robots, providing scalability, adaptability to the environment, and cooperative robustness—a *fully automated orchestra*. We hope that our discussion will foster new insights and research in the direction of human-robot ensemble-based music creation and its underlying dynamics, and in future inspire better "real-time synchronized human–computer collaborative interfaces".

Background and preliminaries

The vision of "creative" intelligence instilled in machines has fanned the imagination of humans for decades, and has inspired inter-disciplinary work across diverse domains such as arts, music, engineering, and psychology (Boden, 1998). The recent advancements in AI, Machine Learning and Quantum Computing have re-invigoration this research domain (Kurt, 2018) for capturing and modelling higher cognitive capabilities for machines. The perceived "humanness" in art forms can be manifested in myriad ways, ranging from art generation, poem or music composition to creative writing.

We initially present to the readers a brief literature overview of the evolution from classical study of music to the modern development of robotic musicianship.

Evolution of robotic musicianship. *Musicology*, the study of music, explores various traditional dimensions in terms of musical beauty (Kant in 18th century), melody theory and classifications (Boretz, 1995), compositional evolution (Wallin et al., 2001), historical associations, religion and societal cultures (Soley and Hannon, 2010). Recent research in this domain explores the benefits of music for medicinal diagnostics and therapy (Davis et al., 2008; Silverman, 2008), as well as in cognitive psychology (Tan et al., 2010). *Cognitive Musicology* involves the modelling of musical knowledge coupled with the use of computer models, and provides the roots for combining AI, cognitive science and musicology (Laske, 1999). Owing to the advancement in the fields of AI and Robotics, physical–mechanical systems for producing music via mechatronic systems have been used to develop musical robots (Kemper and Barton, 2018); while Deep Learning approaches have shown promise in music composition (Payne, 2019).

"Robotic Musicianship" has always been a challenging domain and has garnered significant interest among researchers over recent years for the generation of musical pieces. Along with sonic features like beats and chords, *time* plays a crucial role in terms of the rhythmic aspect of musical performance. Robotic musicians have been developed over the years and have been taught to play different musical instruments such as piano (Kato et al. 1987), strings (Chadefaux et al., 2012), percussion (Kapur et al., 2007), and wind instruments (Solis et al., 2005). However, musical performances involving multiple musicians further complicates the issue of synchronization. Hoffman and Weinberg (2010) introduced the concept of *musical improvisation* and developed 'Shimon', a robotic marimba player that could effectively interacted with other musicians.

Musical Performances, on the other hand, represent a different manifestation of physical expression of music, wherein several musicians perform in a precise synchronized fashion involving a temporal change of roles and possibly improvised chord progression. To this end, the arising research field of "Robotic Musicianship" is centered on creating machine intelligence, using algorithms and cognitive models to obtain novel musical perception and composition, for achieving a harmonious and interactive human–machine performance. The ability of new generation robots to express music artistically in the same way as a human being lies at the intersection of engineering, psychology, music, and computers; and promises to bring about better sharing, creativity, and interpretation of musical impulses.

To alleviate the technical challenges towards human–robot symphonic orchestra, in this manuscript, we propose the joint strategy of Mapping and Modelling. This involves the coupling of oscillator-based synchronization models from Cognitive Musicology with predictive algorithms from AI. We next provide a summary introduction to the concepts and background for both techniques.

Synchronization models. *Synchronization techniques* provide an effective mathematical model for harmonizing several oscillators, each with a possibly different intrinsic natural frequency. The problem of precise interaction between the players in musical performances or for ensemble settings can be aptly formulated as optimizing the synchronization between coupling oscillators, by considering each musician as an independent oscillator. The audio features of the different sources could then be used to achieve global sync, via detection of temporal frequency and phase changes. *Kuramoto's model* is a mathematical model for synchronization problems (Kawamura et al., 2010), and can be used for efficient and scalable synchronization in such musical ensemble scenarios. Mizumoto et al. (2010) developed a robotic thereminist, which uses a Kuramoto's oscillator to synchronize with a human drummer. For establishing synchrony, the human drummer's onset should be correctly adjudged in practice, and the oscillator concept was found to reduce the onset time error by 10–38%, even with sounds with varying tempo. We propose to augment the Kuramoto model with AI techniques for predicting such onsets and other features that could lead to futher reduction in timing and sync error.

Ren et al. (2015) explored the agent-based *'firefly'* model to study the synchronization dynamics between agents and their deviation from the original written music. It was found that the model attained synchronization with a lesser total deviation, even in the presence of ambient noise or multiple players, due to the incorporation of conditional interactions between the musician group. Such models could potentially also be used to find synchronization among the musicians' networks.

Incorporating additional cues from gestures or movements among the musicians has been shown to further reduce the deviations in the above oscillator models. Currently, researchers combine gesture recognition with auditory features for enabling machines to "interpret" the visual cues and interactions between musicians during a performance. In such multi-ensemble topological settings, humans do not play in a rigid fashion, and the roles of 'leader' and 'follower' within the musicians keep changing in a fluid manner (Kawase, 2014). Mizumoto et al. (2012) reported that choosing the correct leader at any given time

is a challenging problem for two main reasons—firstly, each participant is mutually affected by the rhythm of others, and secondly, they compete with as well as complement each other. Although, they showed that Kalman filters can be used to predict musical onset and potential leader states, the estimates were significantly less accurate as compared to that of humans.

The work of Bishop and Goebl (2018) demonstrated the importance of gestural communication, such as head nods or foot taps, and the leader–follower model as useful features in an ensemble environment. They found that gestural communications helped in establishing early synchronization, sometimes even from the very first chord, and that the identification of the leader had a very high impact on maintaining the beat and timing in-sync with the others. A three-phase trial experiment to study the effect of such interactions between musicians was conducted by Maezawa and Yamamoto (2017) by using cameras, motion sensors, and microphones.

Predictive learning models. Artificial neural networks (ANN) have been long studied (Hopfield, 1982) in the literature of AI, for mimicking the functionality of the human brain to create "intelligent machines." ANN comprises a network or circuit of nodes (termed artificial neurons), and the importance of their interactions are depicted by weights between the node connections. A positive weight refers to an excitation, while a negative weight denotes inhibition—wherein a combination coupled with a final activation function controls the output of the network. Such ANNs have been shown to be useful for predictive modelling tasks after an initial training process, where the weights are computed based on a dataset containing annotations and ground-truth information.

In this context, with the current advancement in AI, recurrent neural networks (RNN) were proposed for capturing the node connections in terms of a directed graph along a temporal sequence, enabling the network to model temporally dynamic behaviours on variable input lengths (Tealab, 2018). In fact, RNNs have been shown to be effective in handwriting recognition (Graves et al., 2009). With the advent of the field of Deep Learning, an enhanced ANN was proposed to solved predictive tasks of a higher order complexity. Long short-term memory (LSTM), a variant of RNN with memory and feedback, were shown to be able to capture long-term interactions among a sequence of data points (Sepp and Schmidhuber, 1997), thereby highly suited for classification and predictive applications on temporal data series. In fact, the work of Shlizerman et al. (2018) provides an interesting insight as to how body movements can be accurately predicted from audio inputs by using LSTM models.

Interestingly, it can be observed that our above discussion automatically outlines a possible approach to tackle the dual problems of feature synchronization and predicting transitions in leader–follower roles in a temporal setting, for robotic musical performances. Thus, in this manuscript, we outline how traditional oscillator coupling techniques and modern deep learning approaches could be integrated to capture the underlying dynamics of music analysis and to predict fine-grained changes with musical rhythm for achieving enhanced human–robot synchronization.

The Cyborg Philharmonic architecture

In this section, we describe the components and the working of the proposed system to enable a synchronized human–robot collaborative musical performance.

The creation of melody hinges on two aspects as discussed previously—synchronization (a steady interaction among players) and anticipation (predict what is going to happen)—both of which are moulded by learning and experience in humans. To capture the intrinsic interplays between the components of harmony, the proposed *Cyborg Philharmonic* leverages a joint strategy based on the following modules (as pictorially depicted in Fig. 2, described later in this section):

(i) *Mapping*—responsible for ensuring the control and sensing of the components of musical instruments along with the parameters in sound analysis for synchronization.
(ii) *Modelling*—focuses on predicting the temporal transitions of leader–follower roles via the understanding of gesture

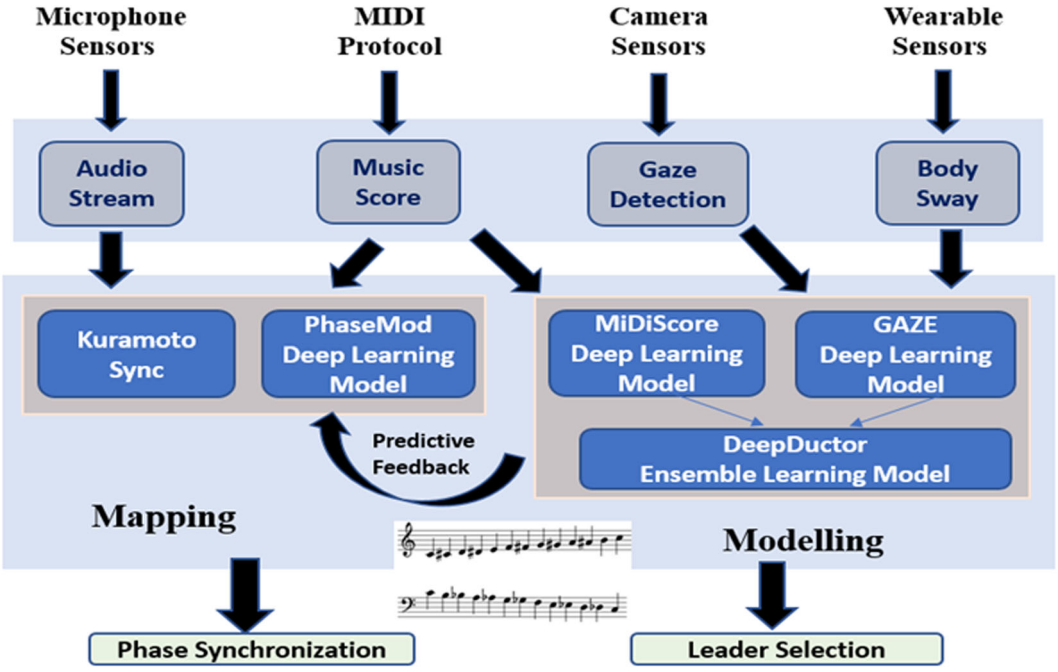

Fig. 2 System architecture and interaction flow between the modules of Cyborg Philharmonic. The proposed architecture showing the modules of Cyborg Philharmonic and their inter-play to achieve a synchrony for human–robot musical performances.

and non-verbal interaction among the other musicians, thereby capturing the overall representation and pre-empting the evolution of the musical process.

The deep learning models employed in *Cyborg Philharmonic*, the proposed musical robot, are initially trained by emulating a real-time orchestra ensemble performance under controlled and supervised conditions. A set of musical performances are selected, preferably across a diverse combinations in terms of beats, tempo, chord progression and other sonic features for robustness. Evolutions in leader–follower role or key temporal changes are appropriately marked for each of the pieces. The machine is now provided with the relevant musical chord notations and the annotated temporal transitions as primary input, while the auditory and visual features will be captured as secondary inputs in real-time during the simulated performance (Keller and Appel, 2010). A group of musicians would then be entrusted to perform the selected musical piece in as realistic manner as possible, simulating an actual performance for training the model. The associated training and performance of the individual modules are described next.

Mapping module. The mapping module ensures the control and sensing of different musical instruments along with the parameters for sound synthesis. Ensemble performance is always a well-rehearsed activity, and due to its complex nature musicians are trained on the musical piece as well as on coordination with other players and conductors to understand and maintain rhythm. To capture the above interactions, the *mapping* module relies on *three* vital channels:

I. the *audio stream* for 'listening' to detect auditory information like major chord, frequency, etc.;
II. *MIDI notations* to understand the different instrumental sheet for capturing a global 'score-follow' or interaction information between the auditory inputs from different instruments input and the actual musical rendition (Bellini et al., 1999).
III. *wearable sensors* capturing body sway for beat or tempo information from movements like foot taps or head jerks.

The audio stream receives raw audio data using an array of microphone sensors. For a small ensemble, a single microphone would be sufficient, but for a large group, an array of several microphones placed at different positions would be necessary to capture all the audio signals. Observe that positioning of the microphone is a vital issue, as it would record nearby instruments louder than the others, resulting in loss of information. Ambient noises are also captured, thus minimizing the number of the microphone is vital for optimizing the time complexity in such a real-time system. Howie et al. (2016) proposed a novel way of recording an orchestra of using the dedicated microphone to create a three-dimensional image of an orchestra. The captured recording is subsequently denoised by the independent component analysis (ICA) filtering technique (Otsuka et al., 2010; Takeda et al., 2008).

Researchers have found (as mentioned in the previous section) that instrumentalists moves their body to produce an expressive artwork (Bishop and Goebl, 2018; Davidson, 2012; Sakata et al., 2009), and furthermore, the sway information would add confidence to the audio-based phase synchrony due to additional cues from wearable sensors. Demos et al. (2018) studied the relationship between body movements and expressive music performance using a weighing sensor. Similarly, we propose the use of Wii weighing plate to measure the overall body movement, and the fluctuation in sensor values during movements would be used to decode movement patterns and link them to a musical beat and other features necessary for synchronization. A Kalman filter or a complementary filter (for faster performance) can be used to remove noise from the obtained readings and understand the movement dependencies (Valade et al., 2017; Wolpert and Ghahramani, 2000).

In such real-time music systems even a small latency could initiate a butterfly effect (Lorenz, 2000), potentially throwing the system out-of-sync. In this regard, MIDI information is beneficial for reducing the time delay over a data transmission (Moog, 1986; Yuan et al., 2010), with the use of 'MidiScore', a score-aligning model based on a bootleg score synthesis, for alignment (Tanprasert et al., 2020). This would help in tracking the portion of the song being played and detect the tempo, as well as provide a global overview of the sonic properties.

The major objective of the mapping module is to enable our Cyborg framework to dynamically synchronize itself with the tempo and beat of the leader musician (detected by the modelling module) during the performance of the orchestra, as described next.

Phase synchronization. The primary goal of Cyborg Philharmonic is to play music harmoniously via a balanced synergy between the human and robot musicians. To this end, information from all the above channels is combined to achieve phase synchronization by using an online beat-tracking algorithm on the audio input filtered by ICA, to calculate the tempo of the song at each time period.

However, data inconsistency and latency in real-time beat tracking algorithms is observed due to the listening window size. To alleviate the above problem, we propose the use of *Kuramoto's model* to maintain stability in beat tracking. Here, each musician is considered as individual oscillators, and their frequency and phase of audio and body sway output would represent the tempo of the song and time of the beat, respectively.

Mathematically, consider a system of N oscillators (in our case musicians), where ω is the frequency (or tempo) of the Kuramoto oscillator, K is the coupling constant and θ is the phase value (or beat timing). The oscillator would then follow the beat tracker to obtain the phase and frequency of the music being played and compute the synchronized system phase value using Eq. (1) below.

$$\dot{\theta}_i = \omega_i + \sum_{j=1}^{N} K_{ij} \sin\left(\theta_j - \theta_i\right), \qquad i = 1, \ldots, N. \tag{1}$$

The solution to the above equation provides the basis for tuning our framework for partial or full synchronization between the musicians during the performance. This, in combination with information from the MIDI standard, enables Cyborg to generate the desired musical notes in-sync with the other musicians for a seamless interactive performance.

It is interesting to note, additional information from the *modelling phase*, employing deep learning techniques, are used as informative cues for enhanced synchronization. Specifically, to help reduce latency, a predicted synchronization state is provided as initialization to the Kuramoto model for faster convergence. Also, information about the identified leader role is provided to focus more on the audio stream and body sensor input from the leader for a better understanding of the current primary chord and beat, to reduce the phase difference from the identified leader. In this context, the body swaying information of the musicians (or solely the identified leader) can be captured using *PhaseMod*, a deep learning model, to understand the relationship between movements and the audio phase (Shlizerman et al., 2018). The trained LSTM network used in PhaseMod provides additional confidence for real-time music phase detection.

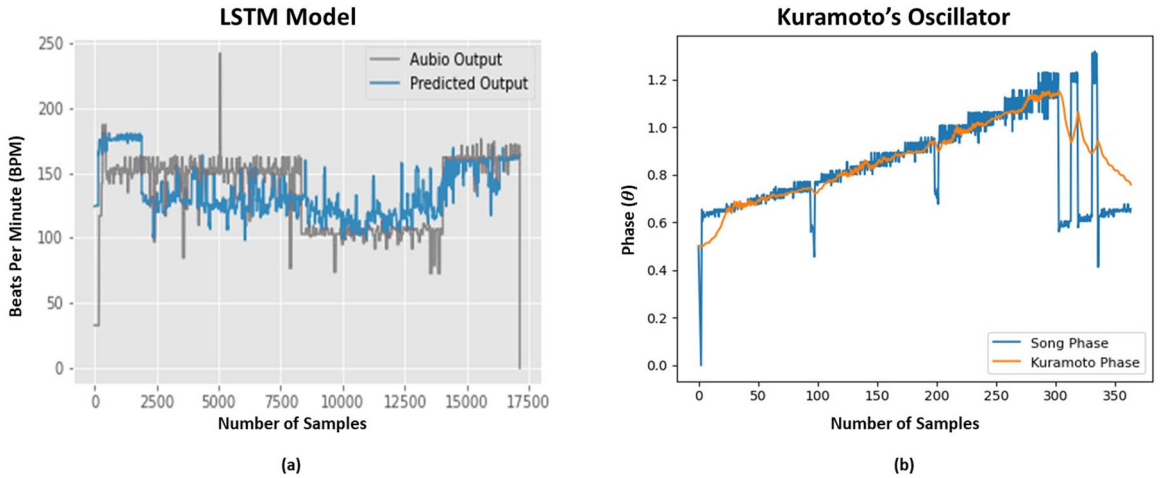

Fig. 3 Beat detection and phase synchronization for single instrumental piece. a Accurate Beat Detection by predictive LSTM model on an input audio stream. The model is seen to closely follow the Aubio's (Brossier, 2006) output obtained. **b** Phase Synchronization achieved between an audio stream and Kuramoto Oscillator (Blue line shows the output phase from Aubio and orange line denotes the Kuramoto's phase). The LSTM model was trained and tested to follow the tempo of the dominant instrument in terms of beat per minute (BPM) on the MUSDB18 STEM dataset (obtained from sigsep.github.io/datasets/musdb.html#musdb18-compressed-stems). The dataset consists of 150 full songs encoded at 44.1 kHz, and ~10 h duration of different genres with their isolated drum, vocal, bass and other stem. In this experiment, the pitch, tempo and volume features were extracted from each stem in real-time. A LSTM model was created to predicit and follow the tempo of the dominant instrument which was guiding the rhythm of the full mix. The Kuramoto's Oscillator was explored to see how multiple instruments can achieve synchronization. As input to the model, we used a a drum beat with varying tempo between the ranges of 85 BPM to 170 BPM with a increase by 5 units at every 64 beats (obtained from www.youtube.com/watch?v=LVJ32rQXDbw&t=8s&ab_channel=MessiahofFire), and as detected by the Aubio Python library. The Kuramoto's Oscillator was allowed to synchronize with the drum beat.

To this end, we show in Fig. 3a an instantiation of the effectiveness of beat detection from audio streams using LSTM models as proposed by Chakraborty et al. (2020). Further, the use of such beat information helps the Kuramoto oscillators to achieve phase synchronization with better convergence, as depicted in Fig. 3b with one instrument (considered as the leader).

Modelling module. The modelling module is responsible for providing the cognitive capabilities to the Cyborg Philharmonic framework, by predictive analysis of the music generation sequence. In practice, an orchestral conductor plays a crucial role in managing and synchronizing the whole system. Further, the gestures of the conductor establish the leader-follower relationship among the troupe. Specific instruments lead the performance for a certain period of time, and then the leadership changes seamlessly. In the absence of a conductor, the leader or dominant musician is established using directional gazes and gestures among musicians towards the lead (Capozzi et al., 2019). This fluidity of role or state transition of musicians enables the creation of the orchestra. However, detecting and following such leader transitions in real-time by integrating diverse information from different sources is a major challenge, the detection of which comprises one of the major focus of the modelling module, as presented next.

Leader detection. Attention to gaze tracking and conductor gestures for establishing the leader–follower can be obtained from cameras in our proposed Cyborg Philharmonic framework. Specifically, we utilize the *GAZE* system (Kawase, 2014), which uses a machine learning model based on support vector machines to distinguish between leaders and followers based on multi-party gaze features using the *PredPsych* tool (Koul et al., 2018). Further, multi-variate cross-classification (MVCC) can be used to achieve greater accuracy in recognizing the lead performer and followers based on leadership styles and situational conditions.

The second challenge is to track and follow the correct leader transitions in real-time during a performance. Traditionally in a large orchestra, there is always a conductor who directs the whole production. GAZE would predict the leader musician based on a statistical approximation of the individual musician's visual gazes. However, the visual stream information would be less efficient for a large orchestra with more musicians, resulting in more computation time. On the other hand, the conductor is not necessary for a smaller group. Relying solely on the detection of subtle gestures between musicians, in such cases, might be prone to errors and involve large latencies—both detrimental for synchronization.

Thus, to overcome the above issues, the modelling module can internally simulate the role of a conductor, by using an ensemble learning model based on the visual features of GAZE and MidiScore among the musicians, as in Qiu et al. (2014). The MidiScore of the different instruments can be used to calculate the 'leadership index' of each instrument at any given point of time, based on its dominance. In fact, such deep learning ensemble models can benefit from the introduction of additional features like the leadership index and phase identification from mapping module, to learn time series-based relationships between chord progression and leader transition. This would enable the early prediction of 'candidate leaders' for transition, that can later be refined in real-time, reducing the overall latency.

Beat prediction. Another aspect of the modelling module, apart from identifying the temporal leader transitions, is its predictive capabilities in terms of overall representation of the musical process. Specfically, by providing a healthy prediction of the synchronized state characteristics (like beats per minutes), the modelling module can reduce the latency in Kuramoto oscillator convergence in the mapping phase. To this end, a deep learning-based regression model (like LSTM) can be trained to predict sonic feature values using the identified leader and chord

recognition the musical notation sheet as features. This information will enable the robot to interpret the musical score in a human-like manner and generate appropriate MIDI data, including control information, that will drive a naturalistic synthesis engine, so as to behave and be perceived like any other musician.

To study the performance of leader detection and synchronization in a natural orchestral setting, we consider the instrumental composition titled "The Art of Fugue" from Bach (Li et al., 2018). A fugue is a contrapuntal composition for a number of separate parts or voices. A short melody or phrase (the subject) is introduced by one part and successively taken up by the other parts and developed by interweaving between them. Thus, the piece appears to be 'led' by different parts at different times throughout, while the other parts 'follow'. The musical piece consists of four different woodwind instruments (flute, oboe, clarinet and bassoon) with changing leadership roles at different timeslots and time-varying tempo, as shown in Fig. 4a. Further, our Kuramoto model was observed to successfully achieve synchronization throughout the musical piece, by accurately identifying leader transitions and beat detection, as demonstrated in Fig. 4b—providing a practical application of the proposed idea.

We thus observe an intricate interaction and feedback of information between the two modules, wherein the detected leader and the predicted beat from the modelling module enables the mapping module to provide better synchronization and interaction between the human–robot musicians. On the other hand, MIDI information and gesture sensory data from the mapping phase provides more efficient leader detection in the modelling module.

Thus the functioning of Cyborg Philharmonic can be outlined as—(i) Leader Detection, (ii) Beat Prediction, (iii) Music Synchronization—showcasing a unique fusion between traditional mathematical models and recent AI predictive techniques for achieving a synchronous human–machine musical orchestra performance.

Discussion

It is interesting to observe that such AI-based automated approaches for human–robot synchronized musical performance have several possible positive societal ramifications. The integration of such frameworks would enable even individuals to compose and perform musical pieces as a troupe (formed by robots). This would provide even people with medical issues to seek solace in creativity and performance—as music has been shown to demonstrate medical benefits (Davis et al., 2008).

Since, human performances are perfected through rigourous training and practice, it might be difficult to find musicians for varied instruments (e.g., for niche traditional instruments, and dependence on musician's skill level or their time schedule) that is associated with ensemble settings. In this scenario, the use of robotic musicianship might reduce such dependences, enabling "musical performance on demand".

Further, considering times of natural calamities, like even in the present context of a viral pandemic, musical performances are practically infeasible, due to travel and other societal restrictions. In such brooding situations, a troupe of robotic musicians might be a viable option—where music has been shown to help improve mental health and instill motivation and hope in communities.

One important aspect is the genre of reproductive art forms, defining the nature and format of the art piece. For example, jazz significantly differs from classical music which differs from rap. In fact, the ease offered by the use of robotic musicians in ensembles would encourage research in terms of compositional aspects as well as inter-mixing of different genres to provide new musical art forms.

Another important consideration in this domain of research is development of an evaluation framework to quantify the objective as well as subjective performance of such robotic performances. Musical note generation, rythym, tempo and other sonic qualities along with synchronization (by the robots) can be detected and objectively evaluated against the notation compositional sheet. However, musical performances have far more abstract dimensions in terms of evoking feelings of joy, nostalgia, calm, harmony and excitement. To this end, development of human evaluation metrics would benefit the further development of this area. This would provide measures to explore and judge as to which aspects (in terms of genres, instruments, etc.) are more amenable to interactive and synchronized human–machine music composition and performances.

Conclusion and future work

In a nutshell, this manuscript proposes a potential architecture for creating human–robot ensemble-based musical performances. We discuss the possibility of integrating mathematical modelling and machine-learning models to efficiently tackle the pivotal challenge of *human–robot synchronization* in real-time for a musical ensemble. This manuscript puts forth how existing technologies like time series predictive models, gaze and gesture detection and synchronization techniques can enable the unique experience of human–robot live orchestra performance, a leap forward towards true "Robotic Musicianship". We hope that our discussions and insights on *Cyborg Philharmonic* would fuel novel inter-disciplinary research avenues

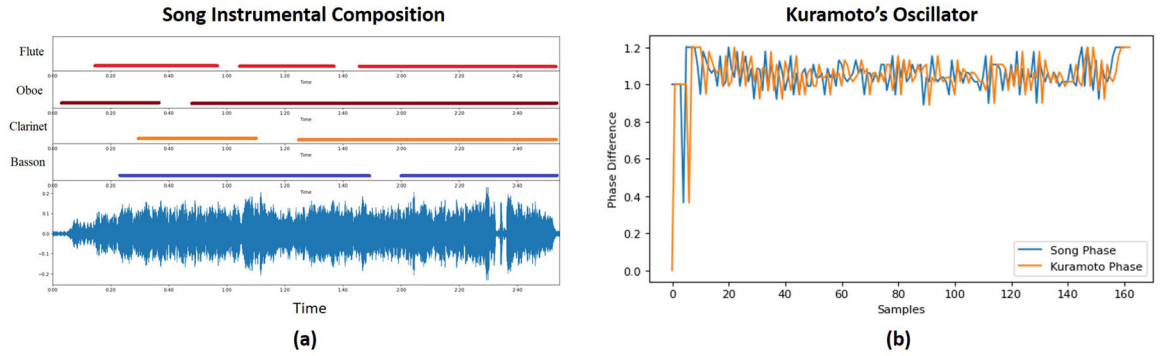

Fig. 4 Phase synchronization for multi-instrumental composition with dynamic leader changes. a Multi-instrumental orchestra composition with leader transition across four woodwind instruments along with time-varying tempo. **b** Phase Synchronization achieved between musical piece with different leader and Kuramoto Oscillator for generating tap sound on every beats is seen to observe close similarity (Blue line shows the musical composition and orange line denotes the Kuramoto's phase). Bach's composition titled "The Art of Fugue", which is 2:55 min long, was obtained from www2.ece.rochester.edu/projects/air/projects/URMP.html.

for better "real-time synchronized human–computer collaborative interfaces and interactions".

A natural direction of future work would involve studying the effects of more recent synchronization techniques like *Janus oscillators* (Nicolaou et al., 2019) shown to concurrently handle phenomena like network synchronization and asymmetry-induced synchronization for scenarios with (a) explosive synchronizations and (b) extreme multi-stability of chimaera states. Further, use of *Quantum Neural Networks* (Narayanan and Menneer, 2000) might provide enhanced predictive power for better leader–follower estimation in our architecture. A futuristic *purely robotic musical performance* can be envisioned by the use of *Swarm Intelligence* algorithms to model a collective behaviour of decentralized and self-organized system (Beni and Wang, 1993; Schranz et al., 2020). Such agglomerative robotic ensembles would employ communication protocols to broadcast leader–follower information, and allow for predictive tasks based on a global information overview—providing scalability, adaptability, and cooperative robustness.

Data availability

The datasets analysed during the current study are available from the following public domain resources: https://sigsep.github.io/datasets/musdb.html#musdb18-compressed-stems; https://www.youtube.com/watch?v=LVJ32rQXDbw&t=8s&ab_channel=MessiahofFire; http://www2.ece.rochester.edu/projects/air/projects/URMP.html.

Received: 14 October 2020; Accepted: 17 February 2021;
Published online: 17 March 2021

Note

1 Musical instruments have been found even from Paleolithic or Old Stone Age archaeology sites.

References

Bellini P, Fioravanti F, Nesi P (1999) Managing music in orchestras. Computer 32 (9):26–34
Beni G, Wang J (1993) Swarm intelligence in cellular robotic systems. In: Proceedings of the NATO advanced workshop on robots and biological systems. Springer, Berlin, Heidelberg, pp 703–712
Bishop L, Goebl W (2018) Communication for coordination: gesture kinematics and conventionality affect synchronization success in piano duos. Psychol Res 82(6):1177–1194
Boden MA (1998) Creativity and artificial intelligence. Artif Intell 103(1–2):347–356
Boretz B (1995) Meta-variations: studies in the foundations of musical thought. Open Space
Brossier PM (2006) The aubio library at mirex 2006. Synthesis
Capozzi F, Beyan C, Pierro A, Koul A, Murino V, Livi S, Bayliss AP, Ristic J, Becchio C (2019) Tracking the leader: gaze behavior in group interactions. iScience 16:242–249
Chadefaux D, Le Carrou JL, Vitrani MA, Billout S, Quartier L (2012) Harp plucking robotic finger. In: IEEE/RSJ international conference on intelligent robots and systems. IEEE, pp 4886–4891
Chakraborty S, Timoney J (2020) Robot human synchronization for musical ensemble: progress and challenges. In: Proceedings of the International Conference On Robotics and Automation Engineering (ICRAE). IEEE, pp 93–99
Chakraborty S, Kishor S, Patil SN, Timoney J (2020) LeaderSTeM-A LSTM model for dynamic leader identification within musical streams. In: Proceedings of the joint conference on AI music creativity. AIMC, Zenodo, Stockholm, Sweden, p 6. https://doi.org/10.5281/zenodo.4285378
Davidson JW (2012) Bodily movement and facial actions in expressive musical performance by solo and duo instrumentalists: two distinctive case studies. Psychol Music 40(5):595–633
Davis WB, Gfeller KE, Thaut MH (2008) An introduction to music therapy theory and practice, 3rd edn. The Music Therapy Treatment Process, Silver Spring
Demos AP, Chaffin R, Logan T (2018) Musicians body sway embodies musical structure and expression: a recurrence-based approach. Music Sci 22(2):244–263
Gaizauskas BR (1974) The harmony of the spheres. J R Astronom Soc Canada 68:146
Graves A, Liwicki M, Fernandez S, Bertolami R, Bunke H, Schmidhuber J (2009) A novel connectionist system for improved unconstrained handwriting recognition. IEEE Trans Pattern Anal Mach Intell 31(5):855–868
Hoffman G, Weinberg G (2010) Shimon: an interactive improvisational robotic marimba player. In: CHI extended abstracts on human factors in computing systems. ACM, pp 3097–3102
Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci USA 79(8):2554–2558
Howie W, King R, Martin, D (2016) A three-dimensional orchestral music recording technique, optimized for 22.2 multichannel sound. In: Audio Engineering Society convention 141. Audio Engineering Society
Kapur A, Trimpin E, Singer A, Suleman G, Tzanetakis, G (2007) A comparison of solenoid-based strategies for robotic drumming. In: ICMC. ICMA
Kato I, Ohteru S, Shirai K, Matsushima T, Narita S, Sugano S, Kobayashi T, Fujisawa E (1987) The robot musician 'wabot-2' (waseda robot-2). Robotics 3(2):143–155
Kawamura Y, Nakao H, Arai K, Kori H, Kuramoto Y (2010) Phase synchronization between collective rhythms of globally coupled oscillator groups: noiseless nonidentical Case. Chaos 20(4):43–110
Kawase S (2014) Assignment of leadership role changes performers' gaze during piano duo performances. Ecol Psychol 26(3):198–215
Keller PE, Appel M (2010) Individual differences, auditory imagery, and the coordination of body movements and sounds in musical ensembles. Music Percept 28(1):27–46
Kemper S, Barton S (2018) Mechatronic expression: reconsidering expressivity in music for robotic instruments. In: New Interfaces for Musical Expression (NIME). Virginia Tech, pp 84–87. https://www.nime.org/archives/
Koul A, Becchio C, Cavallo A (2018) PredPsych: a toolbox for predictive machine learning-based approach in experimental psychology research. Behav Res Methods 50(4):1657–1672
Kurt DE (2018) Artistic creativity in artificial intelligence. Master's thesis. Radboud University, Netherlands
Laske O (1999) A.I. and music: a cornerstone of cognitive musicology. In: Balaban M, Ebcioglu K, Laske O (Eds.) Understanding music with A.I.: perspectives on music cognition. MIT Press, Cambridge
Li B, Liu X, Dinesh K, Duan Z (2018) Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. IEEE Trans Multimedia. 21(2):522–535
Lorenz E (2000) The butterfly effect. World Sci Ser Nonl Sci Ser A 39:91–94
Maezawa A, Yamamoto K (2017) MuEns: a multimodal human-machine music ensemble for live concert performance. In: CHI conference on human factors in computing systems. ACM, pp 4290–4301
Mizumoto T, Ogata T, Okuno HG (2012) Who is the leader in a multiperson ensemble? Multiperson human–robot ensemble model with leaderness. In: IEEE/RSJ international conference on intelligent robots and systems. IEEE, pp 1413–1419
Mizumoto T, Otsuka T, Nakadai K, Takahashi T, Komatani K, Ogata T, Okuno HG (2010) Human–robot ensemble between robot thereminist and human percussionist using coupled oscillator model. In: IEEE/RSJ international conference on intelligent robots and systems. IEEE, pp 1957–1963
Moog RA (1986) Midi: Musical Instrument Digital Interface. J Audio Eng Soc 34 (5):394–404
Narayanan A, Menneer T (2000) Quantum artificial neural network architectures and components. Inf Sci 128:231–255
Nicolaou ZG, Eroglu D, Motter AE (2019) Multifaceted dynamics of Janus oscillator networks. Phys Rev X 9(1):011–017
Otsuka T, Mizumoto T, Nakadai K, Takahashi T, Komatani K, Ogata T, Okuno HG (2010) Music-ensemble robot that is capable of playing the theremin while listening to the accompanied music. In: International conference on industrial, engineering and other applications of applied intelligent systems. Springer, Berlin, Heidelberg, pp 102–112
Payne C (2019) MuseNet. OpenAI, openai.com/blog/musenet
Perc M (2020) Beauty in artistic expressions through the eyes of networks and physics. J R Soc Interface 17:20190686
Qiu X, Zhang L, Ren Y, Suganthan PN, Amaratunga G (2014) Ensemble deep learning for regression and time series forecasting. In: IEEE symposium on computational intelligence in ensemble learning. IEEE, pp 1–6
Ren IY, Doursat R, Giavitto JL (2015) Synchronization in music group playing. In: International symposium on Computer Music Multidisciplinary Research (CMMR). Springer, pp 510–517
Sakata M, Wakamiya S, Odaka N, Hachimura K (2009) Effect of body movement on music expressivity in jazz performances. In: International conference on human–computer interaction. Springer, Berlin, Heidelberg, pp 159–168
Schranz M, Umlauft M, Sende M, Elmenreich W (2020) Swarm robotic behaviors and current applications. Front Robot AI 7:36
Sepp H, Schmidhuber J (1997) Long short-term memory. Neural Comput 9 (8):1735–1780
Shlizerman E, Dery L, Schoen H, Kemelmacher-Shlizerman I (2018) Audio to body dynamics. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, pp 7574–7583

Silverman MJ (2008) Quantitative comparison of cognitive behavioral therapy and music therapy research: a methodological best-practices analysis to guide future investigation for adult psychiatric patients J Music Ther 45:457–506
Soley G, Hannon EE (2010) Infants prefer the musical meter of their own culture: a cross-cultural comparison Dev Psychol 46:286–292
Solis J, Chida K, Isoda S, Suefuji K, Arino C, Takanishi A (2005) The anthropomorphic flutist robot WF-4R: from mechanical to perceptual improvements. In: IEEE/RSJ international conference on intelligent robots and systems. IEEE, pp 64–69
Takeda R, Nakadai K, Komatani K, Ogata T, Okuno HG (2008) Barge-in-able robot audition based on I.C.A. and missing feature theory under semi-blind situation. In: IEEE/RSJ international conference on intelligent robots and systems. IEEE, pp 1718–1723
Tan S, Pfordresher P, Harré R (2010) Psychology of music: from sound to significance. Psychology Press
Tanprasert T, Jenrungrot T, Müller M, Tsai TJ (2020) Midi-sheet music alignment using bootleg score synthesis. arXiv preprint arXiv:2004.10345
Tealab A (2018) Time series forecasting using artificial neural networks methodologies: a systematic review Future Comput Inf J 3(2):334–340
Valade A, Acco P, Grabolosa P, Fourniols JY (2017) A study about Kalman filters applied to embedded sensors Sensors 17:2810
Wallin NL, Björn M, Brown S (2001) An introduction to evolutionary musicology. In: Wallin NL, Björn M, Brown S (eds) The origins of music. MIT press, pp 5–6
Wolpert D, Ghahramani Z (2000) Computational principles of movement neuroscience Nat Neurosci 3:1212–1217
Yuan S, Lu Y, He H (2010) Midi-based software for real-time network performances. In: International symposium on cryptography, and network security, data mining and knowledge discovery, e-commerce and its applications, and embedded systems. IEEE, pp 226–230

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to S.C.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.