.

D1.6 Models and Algorithms

Project No	GA824160
Project Acronym	EnTimeMent
Project full title	ENtrainment & synchronization at multiple TIME scales in the MENTal foundations of expressive gesture
Instrument	FET Proactive
Type of action	RIA
Start Date of project	1 January 2019
Duration	48 months

TABLE OF CONTENTS

1. INTRODUCTION

This document presents the first results on multi-temporal predictive models elaborated in the context of dyadic and group synchronisation in human ensembles, integrating fast-time signatures of participants (e.g., eigen frequency, amplitude), multisensory coupling functions (of visual, auditory, haptic and tactile types), and their consequences on entrainment and synchronization (phase and period), as well as low-time psychological and social modulators (e.g., mood and attitudes, likeability, rapport, social competences and emotion), and expressive qualities of gesture.

Specific predictions will be tested in WP2, based on modelling and experimental effort to uncover the emergence of dyadic and group synchronisation (e.g., Alderisio et al., 2017; Zhong et al., 2018) at different levels in the EnTimeMent multiscale approach.

The models and algorithms that are being developed by the consortium can be divided into two families: approaches based on machine learning, and approaches based on computational models and algorithms based on different techniques, such as the integration of graph theory and game theory to measure the joint origin of human movement (Kholikalova et al 2020), and computational models of the individual motor signature (Slowinski et al 2016).

2. MACHINE LEARNING FOR MOVEMENT ANALYSIS

To allow for efficient movement analysis over varying time scales, machine learning-based methods play an essential role in EnTimeMent. Motivations for and promises of machine learning in movement analysis have previously been given in D1.3. Here we summarize these in light of what has so far been studied as part of the project. Depending on the purpose of the study, both traditional methods that rely on feature engineering and more recent deep learning-based methods are considered, since they both bring different sets of benefits. Traditional methods that are typically based on feature engineering can more easily be made explainable, if features that are used provide some semantic meaning, and are less reliant on large sets of data for training.

Deep learning-based methods that involve feature learning do not require features to be engineered but learn features in a data-driven manner less influenced by the designer. Features that emerge from training might resemble those that were earlier engineered but could also be novel and point to qualities in movement that require further investigation. With recent methods trained end-to-end, i.e., methods that are trained directly from raw data without a layer of engineered features, it is possible to discover relevant movement patterns that are indicative of a particular prediction. Methods known as attention-based methods do this explicitly by highlighting indicative events in space and time, while other methods focus more on the general nature of the movement over longer time horizons, without necessarily pinpointing any particular localized events.

Another benefit is the potential for representation learning using e.g., encoder-decoder networks with which large amounts of movement data can be compressed into a more compact form by removing redundancies in data over both space and time. This is of particular importance when learning predictors from small data sets for which predictors easily become overtrained and biased towards the particular data, without the ability to generalize. Representations can further be combined with generative methods for predictions of multiple possible futures. This can be exploited in multi-agent settings where an agent is responding to the movement of others in a more fluent manner. By comparing predicted and true future movements a generative model may also be used to assess the quality of the representation and detect changes in the qualities of movement over time.

Deep learning may also provide means of transferring models from one domain to another to overcome the limitations of existing data sets, that are often too small for the full range of bodily expressions to be captured, even in constrained domains. Much research has so far focused on the activities that agents are engaged in, but relatively few collected datasets include affective qualities of movement, which is the focus of EnTimeMent. The lack of data is particularly critical in clinical studies, an area for which reliable data are hard to collect, and the quality of predictions may have a considerable impact. For that reason, data of people with clinical conditions will be collected as part of the project. Another potential remedy, that is explored in the project, is to transfer models from source domains for which large amounts of data exist to target domains for which a sufficient amount of data is hard to collect, such as going from motion capture data with affective qualities transferred to predictions of such qualities from video data.

2.1 Summary of work using machine learning reported in other deliverables

Here we summarize work on movement analysis using machine learning methods done by partners and point to relevant deliverables where such work is reported in greater detail.

2.1.1 Prediction of group behaviour in conversational groups

Using motion capture data collected for the purpose, KTH has explored machine learning methods for representation learning of agents moving in groups to predict whether a newcomer will be accommodated or ignored when approaching a group of agents already in conversation. A more extensive description of the developed and evaluated methods can be found in D3.5 with experiments on conversational groups detailed in D2.3. Three methods were explored, two attention-based networks, AGNet and AGTransformer, and a method based on graph convolution networks (GCN). All three methods were extended to groups of agents, while the GCN method was also extended to multiple temporal scales. While GCNs keep the graph structure all through processing, the attention-based networks collapse the structure by focusing on indicative events in space and time. If the overall movement is more important than any localized event, which could be the case in our particular task, this could explain why the GCN method was shown to perform the best. The multi-temporal version turned out to be even better at predicting the behaviour of the group with an accuracy of 91.5%. To test whether the same kind of structure can be used also for generation of behaviours, the GCN method was applied to a framework with reinforcement learning to learn policies for a collaborative robot, with the goal of making the robot more proactive by predicting the intentions of its home partner. The policy was derived from a Q-function using the GCN representation as input state, after being processed through an LSTM to make it less dependent on the choice of time windows. It is likely that a very similar solution can be applied to promote curiosity-based social interaction between a human subject and an artificial agent.

2.1.2 Three novel individual action prediction architectures based on encoding of multiple timescales

We proposed three novel neural network architectures: *Body Attention Network* (BANet), *Movement in Multiple Time* (MiMT) neural network, and *Global Workspace Network* (GWN). Detailed descriptions of these networks and results of their validation have been reported in deliverable D3.1 (architectures, for the BANet and MiMT) and D2.1 (experiments and results, for all three).

The BANet implements a hierarchically weighted fusion of time encoding for the individual joint angles in body movement data. The MiMT further explores the use of separate but shared time encoding for different groups of anatomical segments in the data together with multiple timescales of the label being modelled. The GWN, on the other hand, addresses the challenge of fusing multimodal movement data and uses the approach of recurrent self-weighting (self-attention), including a mapping component to deal with multiple temporal resolution across modalities. As the GWN was not introduced in D3.1, we provide an overview below.

The GWN architecture (now published in ICMI 2020 (Bao et al. 2020)) was developed in collaboration with a company and addresses one of the inherent challenges of multimodal prediction, unknown dynamic noise in the different modalities, i.e. multiple timescales of noise. The architecture additionally accounts for different temporal and dimension resolutions for the different modalities. The GWN is a neural network that derives from the Global Workspace Theory of (Baars 1997, 2002) by which mental processes are managed. According to the theory, at each time *t*, multiple mental processes compete, the winner then broadcasts its signal to a global workspace that is shared by all the processes, and this is archived in an external storage for later use.

Our GWN, as shown in Figure 1, similarly comprises of a *compete and broadcast* network (shared across time) by which multiple modalities compete at each time *t* but instead of a hard competition where only one modality wins the right to broadcast at *t*, the competition is a soft one. This module is implemented as a transformer (Vaswani et al. 2017) with softmax ‘weights’ applied as additions. The other component of the GWN is an external storage that enables broadcast data at time *t* to influence the data at time *t*+1. The storage is implemented using a long short-term memory neural network (LSTMNN) (Hochreiter and Schmidhuber 1997; Gers et al. 1999). The two other elements of the GWN are a mapper and a classifier. The mapper is a pretrained autoencoder that transforms multimodal data of different number of dimensions and sampling rates to uniform dimension and temporal sizes. The classifier, on the other hand, takes in the time encodings from the external storage to predict a given label.

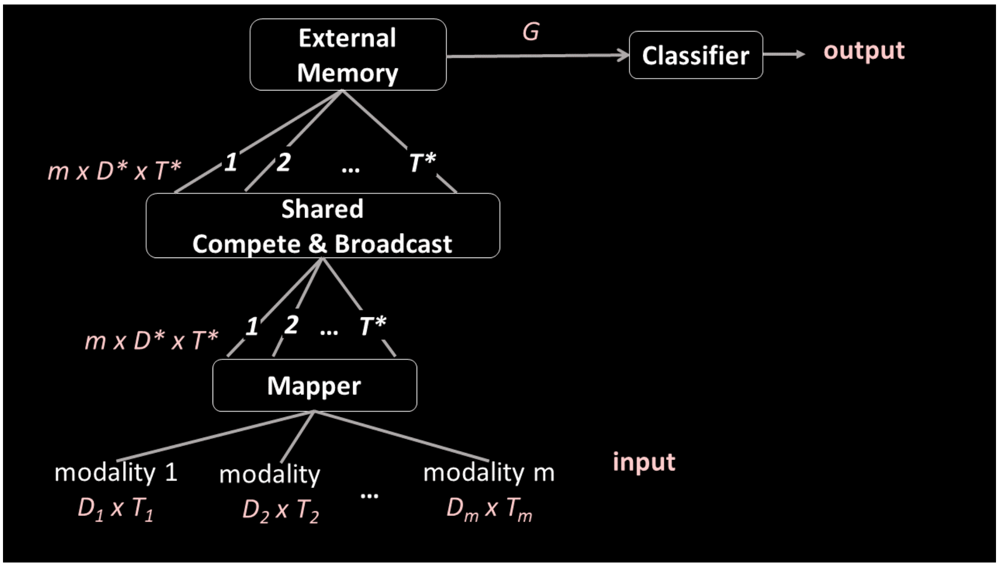
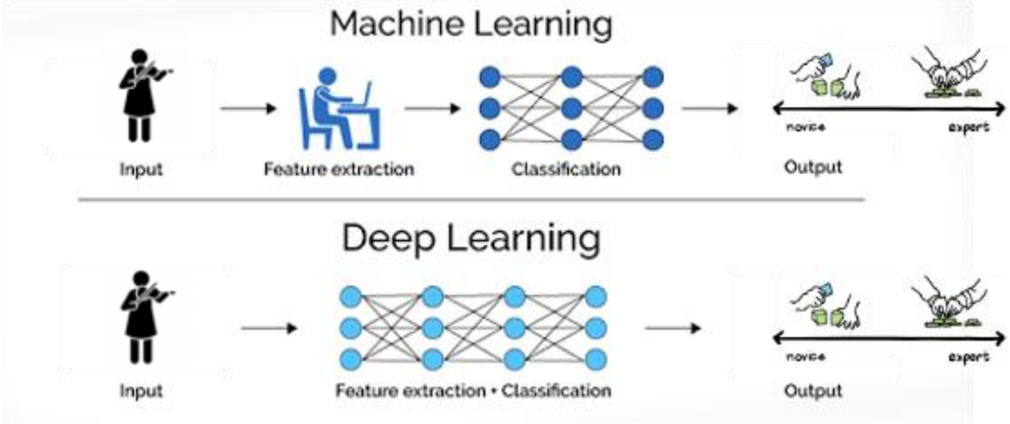


Figure 1. An Overview of the Global Workspace Network (GWN)

- REMOVE THE FOLLOWING TEXT THAT WAS EARLIER USED IN D1.3



2.2 Discovering relevant movement patterns

Traditional methods for human movement analysis have relied heavily on feature engineering, using predefined features that have shown to be indicative of the quality of movement. However, with engineered features, information deemed to be irrelevant by the designer is discarded from further consideration, leading to a potential bias in the analysis. The analysis cannot say more than what is contained in the features, which are in turn based on what the designer believes to be relevant. It is for example increasingly clearer that psychology frameworks (e.g. FACs) used to inform the design of features for automatic facial expressions recognition are limited in capturing everyday facial expressions (Barret et al., 2017). More relevant to EnTimeMent is the work by Avizier et al. (2012) showing that people are convinced to make most of their evaluations of other people’s emotions on the basis of their facial expressions while often they are using features of the body movement to reach their conclusions. With the introduction of modern representation learning methods, typically based on deep neural networks, such bias can be avoided (Bengio et al., 2013). Instead of engineering features, features are learned from the raw sensory data with the influence of the designer kept to a minimum.

Features that emerge during the learning process may or may not resemble those of earlier designed features. Regardless of which, the learned features may be of interest for further investigation. Some features may be used to confirm the relevance of patterns that are believed to be indicative of particular movement qualities. It is also possible for new relevant patterns to be discovered, in particular patterns that are hard for a designer to conceptualize and define as a feature. If the learning framework allows for signals to be represented over multiple scales, the temporal scale of the pattern can potentially also be discovered.

Unfortunately, most learning frameworks do not explicitly indicate the movement patterns that are most indicative of a particular prediction. Notable exceptions are attention-based methods that apply weights to place more or less emphasis on the data given how important they are assumed to be for the prediction (Liu et al, 2017). For feed-forward neural networks, a common approach to understand the inner workings of the predictor is to propagate back gradients from a predicted concept to the most salient part of the input data, resulting in a heatmap that can easily be visualized ((Selvaraju et al. 2017), (Stergiou et al., 2019)). It should be noted though that much is still unknown what deep networks actually learn and what leads to a particular prediction. This can be illustrated by their lack of resistance towards so-called adversarial attacks (Akhtar and Mian, 2018), where small unnoticeable changes to the input can significantly alter the prediction.

2.3 Predicting multiple futures

Another application of modern machine learning methods is to predict future movements. Given the recent history of movements and qualities derived from it, it can be predicted how an agent will move in the near future. If movements are modeled in a probabilistic fashion, future predictions can be described as distributions that capture multiple possible futures. Even in cases where only the most probable future will eventually be used, the variance of the distribution can be viewed as a measure of the uncertainty of the prediction.

There could be many reasons why a system would benefit from future predictions. You might have multiple interacting agents, where predictions are used for planning for an agent to move in accordance with the movements of others. A prediction can also be used as a means to verify whether the system's understanding of the movement is likely to be correct by comparing it to the true future movement. Another possible reason is to detect events that suggest a change in state when the agent changes from one type of movement to another.

Most research of movement prediction has so far been done on relatively short time scales, typically not more than a few seconds. This is true at least if predictions are given in the space of the original input data ((Fragkiadaki et al., 2015), (Bütepage et al., 2017)). Most such frameworks rely on some kind of encoder-decoder structure, where the encoder compresses recent movement into a low-dimensional representation, which is then used by the decoder to predict multiple futures. For human-robot interaction, the decoder can often be used as a generator to generate appropriate movements in response to the movement of a partner. In the context of physical rehabilitation (one of the cases of EnTimeMent), such decoders could be used to drive real-time generation of movement sonification to enhance movement awareness or even help correct movement trajectory before they take place by using embodied mechanisms of sound (Newbold et al., 2016).

For longer time scales, predictions are typically given in terms of some higher-level movement qualities, such as the activity the agent is currently engaged in. To reduce the complexity of the learning framework that would otherwise grow, as longer time scales are used and more data is processed for prediction, learning is often done in stages. Data are first represented as streams of learned spatio-temporal movement features that are later aggregated for prediction ((Ryoo, 2011), (Li and Fu, 2014)). Since this division into two stages might lead to bias by design, there is a trend towards end-to-end training, where features and predictors are learned at the same time (Kong et al., 2020).

2.4 Abstracting representations of movements

Regardless of which machine learning method you use, over-training easily becomes a critical concern as the temporal scales from which predictions are made increase. Learning frameworks increase in complexity and with more unknown free parameters to train, the need for more training data increases far beyond what is feasible to capture within a reasonable time. Without a sufficient amount of training data, the excess learning capacity of frameworks will instead be used to learn particular patterns in the training data that are of no real significance, making the frameworks unable to generalize beyond the training data. A common solution in computer vision and machine learning is to pre-train networks for an auxiliary task for which large amounts of data are already available. This is not always the case though. Annotated training data with human participants is always expensive to acquire in particular if it involves partners that are interacting or data from clinical populations.

Fortunately, the repetitive and cycling patterns of movement data tend to make the data highly redundant. If the training data include annotations of movement qualities to predict for a particular task, information redundant for this task can be eliminated by training the framework in a supervised setting, which in turn reduces the complexity of the predictor and the need for large amounts of data. This does, however, prevent representations from being used for other tasks, at least as long as the two tasks are different enough with respect to the nature of the information required by their respective predictors. An alternative is to use unsupervised training and reduce the dimensionality of data while representing it in a form in which more information is preserved.

Given the nature of movement data, it comes as no surprise that traditional methods for dimensionality reduction have been applied to reduce its size, e.g. using wavelet transforms (Beaudoin et al., 2007). The size is reduced by discarding dimensions that have the least impact on a reconstruction from further analysis. However, whether data is redundant or not depends on the application. Dimensions might be removed despite being essential for the prediction of particular movement qualities if these dimensions have little effect on the overall movement. It thus becomes important to reduce the complexity of data, while at the same time ensuring that information necessary for prediction remains. Another way to compress the data is to learn some kind of sparse representation of the data using primitives known to be relevant for the target application. These primitive can be localized in time and space modeling individual joints or markers ((Schaal, 2006), (Li et al., 2010)) or could be a more global temporal segmentation of the movement (Lu and Ferrier, 2004). However, there is a risk for introducing a bias by design given by the limitations of the approach chosen, in particular for the temporal scale during which significant events are expected to occur.

More modern methods in machine learning try to limit the influence the human designer has on what is eventually learned. Many such methods are based on variational autoencoders, feed-forward neural networks that are trained to predict what it has on its input but does so through a narrow waist represented by a layer consisting of a small set of neurons. Once the network has been trained, the activations of these neurons can be regarded as a low- dimensional latent representation of the original data (Kingma and Welling, 2014). These autoencoders are flexible in the sense that they can be easily extended with additional objectives related to the particular task, allowing e.g. ground truth data of known movement qualities to affect the way the latent representation is structured through semi-supervised learning (Bütepage et al., 2018). By doing so representation can remain compact, while still keeping dimensions of the data known to be important for prediction.

2.5 Coupling and transfer of information

Another benefit of probabilistic modeling, possibly using variational autoencoders or similar networks, is to model the coupling between different units. These units could e.g. be different moving agents, different tasks or sensor modalities. For example, in (Bütepage et al., 2019) it was tested whether observations of human partners engaged in interaction, different types of hand-shakes in this case, could help a robot to learn similar movement patterns. Learned probabilistic models were used to describe the movement of each individual participant, but these models were conditioned on a latent space representation of the joint task that was also learned. In experiments, it was shown that the robot benefits from such a coupling between models when trying to learn a new behavior, which is true even if the coupling was learned from observations only. This suggests that observations can help an agent learn to engage in an activity, even with the observations do not include the agent itself, but two other interacting agents.

In recent years, transfer learning has become an important tool to allow transfer knowledge from one domain to another (Cook et al., 2013), typically from a domain for which you have a sufficient amount of data to a domain in which it is easier to draw conclusions on relevant movement qualities. One such possibility is to transfer knowledge from video data to motion capture data ((Mehta et al., 2017), (Zhou et al., 2017)) which may have direct practical application in e.g. rehabilitation, which is a target area of EnTimeMent. In addition, as rehabilitation moves from exercise-based sessions into functional activities, there is also the need to transfer to a personalized version of wearable (possibly reduced) mocap technology. Other applications for which transfer learning can be beneficial is to transfer observed qualities of movement into a form that is more easily interpreted by human users, using e.g. sonification.

2.6 Providing insight into the data available and the need for more data

The sections above have discussed some of the possibilities the new machine learning techniques offer to the study and modeling of body movement and body movement qualities at different time scales. The application of such machine learning techniques to available human activity and affective body expressions datasets can already help shed more understanding on important movement patterns, beyond the one available from the psychology literature (e.g., for a survey see Kleinsmith et al., 2013; Karg et al., 2013) to better understand how emotion is expressed through the body by comparing and modeling datasets from multiple contexts rather than independent single individual datasets and what movement parts are cues to understand people interaction dynamics.

Unfortunately, motion-capture-based body movement datasets are still very sparse and quite limited to very simple activities. They often include a very small number of sensors and they are generally unimodal (see Deliverable 1.1 for a survey). This is even more the case when we consider affective qualities of movement rather than activities. Most of these datasets consist of acted rather than naturalistic data. This sparsity of data limits the possibility to apply more advanced machine learning techniques and even to develop new algorithms that address the specific characteristics of such data and of the related modeling applications.

A specific need emerges from the fact that affective body expressions do not happen outside a context and such context contributes to shape them (Barret, 2017). A particular context is the activity the person is doing. To fully understand the affective quality of body expressions, there is a need to gather those within the context of a large variety of activities rather than isolation. The trigger of an emotional response and the environmental and cultural context (what is socially acceptable) the person is in are also factors that contribute to modify the expressions. This variability is critical as affect recognition and activity recognition systems are starting to be deployed in specific real-life applications beyond entertainment and research. While voice and face are the main modality industry is currently considering, we can already see how assumptions are made by industry on the generalization capabilities of this technology. However, facial expression recognition can rely on advances made in computer vision and on the large datasets that are now being collected using convenient front-of-camera settings (e.g., laptop). The work on affective body expressions suffers from the fact that such settings do not easily fit full-body activities (general just face and shoulder) and as said above HAR is also under-developed as often these datasets are privy of emotional responses as data are conducted in quite repetitive sessions.

A number of full-body datasets have been created in the contexts of dance, music playing, and computer games. Again these are mainly sparse activities and there has been very little attempt to understand how these resources can be brought together and exploited. This is an activity that EnTimeMent may aim to address given that many of the available datasets have been developed by the consortium partners. In doing so, EnTimeMent may attempt to set guidelines for such recording, sharing and building a benchmark dataset platform.

A domain that is even more critical to cover is clinical datasets ((Lucy et al., 2013), (Riva et al., 2018), (Aung et al., 2016)). These are critical for building effective applications, however, they are very much time consuming and suffer from the fact that, unless built in the hospital (e.g., Joshi et al., 2013), they may be biased toward a clinical population that is able to cope with the physical and psychological demand of such dataset. For example, outside of the clinic data collection, may contain less people that suffer from depression as such a population may be more reticent to engage in a psychologically demanding social activity of a typical data collection session. To address such limitations (among giving also the possibility to gather more real-life data and useful data), in EnTimeMent, we are aiming to collect data of people with clinical conditions (chronic pain) in their own homes. This raises technical challenges to data collection as complete and more reliable sensors architecture cannot be used. In such settings, for example wearable with a limited number of sensors (to address acceptability, comfort and reduce sensor interference) can only be used (Olugbade et al., 2018).

Another important point to consider is the need to go beyond simple movement sensors to capture the skeleton of the body. EnTimeMent sees movement as a more complex process. Muscle activity, respiration or even neural activation patterns provide different information about movement and about its affective drives. Such rich platforms create interesting new questions for machine learning (their different role, off-set, temporal scale, etc.).

Whilst all movement datasets can be said to be built mainly for very short temporal scales in mind, wearable devices could offer new opportunities in terms of long term temporal scales. Such data would allow the study of movement perception and modeling at a very different level (e.g., habits, perception of capabilities, long term prediction). A proxy to such data collection can be seen in a multi-session data collection for the same person. Such datasets are still very limited and mainly considered only in stationary situations or for a very limited number of sensors (e.g., accelerometer in a smartphone).

3. COMPUTATIONAL MODELS AND ALGORITHMS BASED ON THE INTEGRATION OF GRAPH AND GAME THEORIES

This section presents the results of the models and algorithms developed by UNIGE based on the integration of graph and game theory. The approach is a first contribute towards the automated analysis of the perceived origin of full-body human movement and its propagation. The study was published in the IEEE Transactions on Human-Machine Systems (Kolykhalova et al, 2020). While this first study focuses on the origin of movement in an individual movement, the same techniques are currently adopted to model the behavior of small groups: the origin of movement in an individual movement (i.e., the joint where the movement originates and propagates to the other body parts) is a promising metaphor to study the social relations in dyads and in small groups of people, where the origin of movement is a candidate to express the leader of the group in that moment, and could be analysed at different temporal scales to study the dynamics of leadership. In this section, we present the computational model developed on individuals.

The analysis of the origin of movement in the movement of an individual is an important component in the understanding and modeling expressivity. For example, in rehabilitation the detection of the origin of movement can help in enabling a patient to learn how to perform a movement (e.g., how to stand up from a chair) correctly to avoid injuries. For example, the leaning forward of an arm can have very different expressive meanings depending on the origin of movement: a “punch” originates from the foot, a “push away” may originate from the shoulder, and a “caress,” from the hand. All these movements are basically a leaning forward of an arm, the very different dynamics of which are explained also in terms of the origin of movement.

The Approach Pipeline

The approach, which is grounded on the combination of cooperative game theory and graph theory, consists of the following steps.

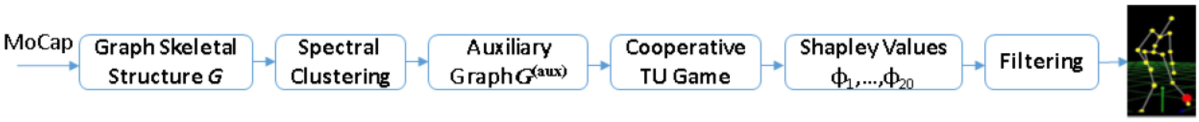

MoCap
Graph Skeletal Structure G
Spectral Clustering
Auxiliary Graph $G^{(aux)}$
Cooperative TU Game
Shapley Values $\phi_1,\dots,\phi_{20}$
Filtering

Fig. 1. Conceptual architecture of the proposed method.

The human body is represented by means of an undirected graph, in which the vertices are the joints and the edges are both physical and non-physical connections between these body joints. Moreover, the edges are associated with weights, the values of which depend on a feature extracted from motion capture data. On the one hand, physical links represent connections between consecutive physical body joints, such as the forearm. On the other hand, non-physical links model the dependencies between joints that are not physically connected, solely derived from correlations observed in the chosen movement feature between these joints: for example, a hand moving towards the head, followed by a movement of the head in response in the same direction, reveals a non-physical link between the hand and the head. Non-physical links therefore play the role of potential bridges joining body parts that are not directly connected within the skeletal structure but exhibit correlated dynamics during the movement performed

Starting from the graph representing the skeletal structure augmented by non-physical links, we define a suitable mathematical game (Maschler et al., 2013) in which the vertices (i.e., the body joints) are the players and the edges model the communication channels (through which movement can propagate) between these players. Body movement is therefore represented by a game constructed on the graph. A cooperative game model is proposed, since both the vertices and the edges contribute to the overall movement.

Then, the Shapley value (Maschler et al., 2013) – which is a classical solution concept from cooperative game theory able to provide a ranking of the players that represents their relevance in the game - is computed for all the players of the game and adopted as a measure of vertex relevance in the graph to estimate how much each vertex contributes to a shared goal (i.e., to the way in which a specific movement-related feature is transferred among the joints). The possibility to know, moment by moment, which joint(s) is the most representative in the ongoing full-body movement (i.e., those with the highest Shapley value) constitutes precious information for the automated analysis of expressiveness in movement. The joints with the highest Shapley value are candidates to be the perceived origin of movement propagating in the body and they can provide useful cues to detect which parts of the body are most relevant for the analysis of expressive movement and worth a detailed observation by means of further analysis techniques (possibly at a finer scale), as well as to inform automated techniques of movement prediction.

Implementation

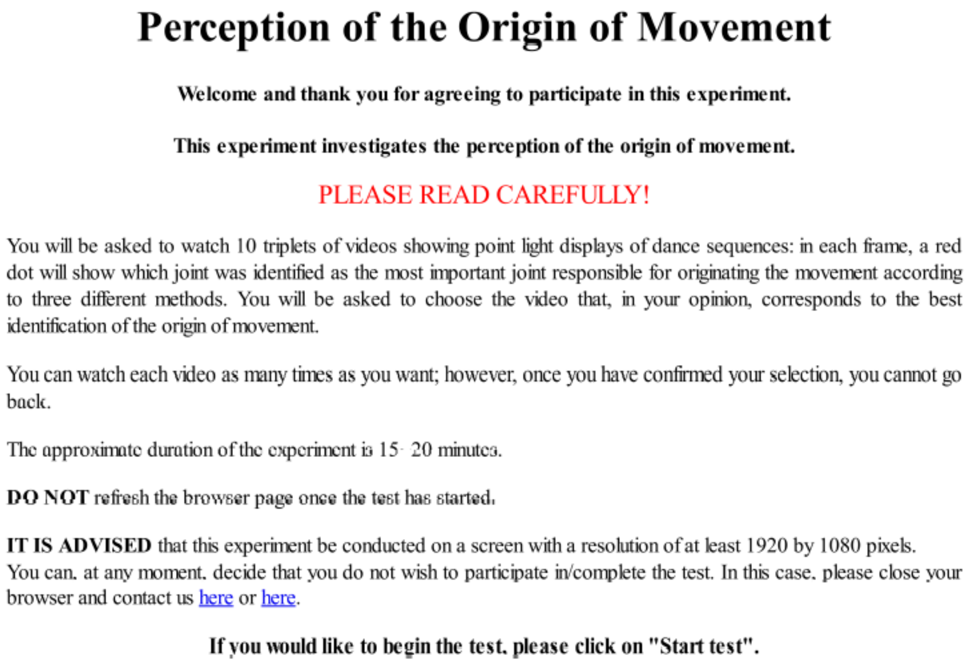
We used a recorded multimodal data set composed of 127 trials (or recordings), acquired with the goal of analysing movement, determining the features associated with it, and designing computational techniques for their evaluation. The recordings were acquired using a Qualisys motion capture system with 13 infra-red cameras synchronized with 2 video cameras in the frontal and lateral views. The two professional dancers were equipped with 1 microphone, 5 accelerometers, and 64 infra-red reflective markers.

After their acquisition, the data were cleaned and post-processed via the Qualisys Track Manager native software using a cubic polynomial interpolation for trajectories with gaps in the data.

Finally, annotations of the origin, path, and destination of each movement were produced by experts. The expressive movements performed were not related to a specific dance style, being normal full-body movements, e.g., leaning an arm towards a target or turning towards a direction, characterized by a clear origin of movement, enabling the detection of the origin even by a non-expert observer (though its automatic detection is still not a trivial task). The choice of dancers as movement executors was motivated by their full awareness and control of movement details and their higher motor skills with respect to non-trained people, which allowed reducing the amount of noise with respect to alternative performances by non-experts.

Validation

Validation of the approach included an on-line survey (based on the data repository) in which participants with different levels of expertise in dance took part. A survey website was developed to collect user ratings (see Fig. 2).



Perception of the Origin of Movement

Welcome and thank you for agreeing to participate in this experiment.

This experiment investigates the perception of the origin of movement.

PLEASE READ CAREFULLY!

You will be asked to watch 10 triplets of videos showing point light displays of dance sequences: in each frame, a red dot will show which joint was identified as the most important joint responsible for originating the movement according to three different methods. You will be asked to choose the video that, in your opinion, corresponds to the best identification of the origin of movement.

You can watch each video as many times as you want; however, once you have confirmed your selection, you cannot go back.

The approximate duration of the experiment is 15- 20 minutes.

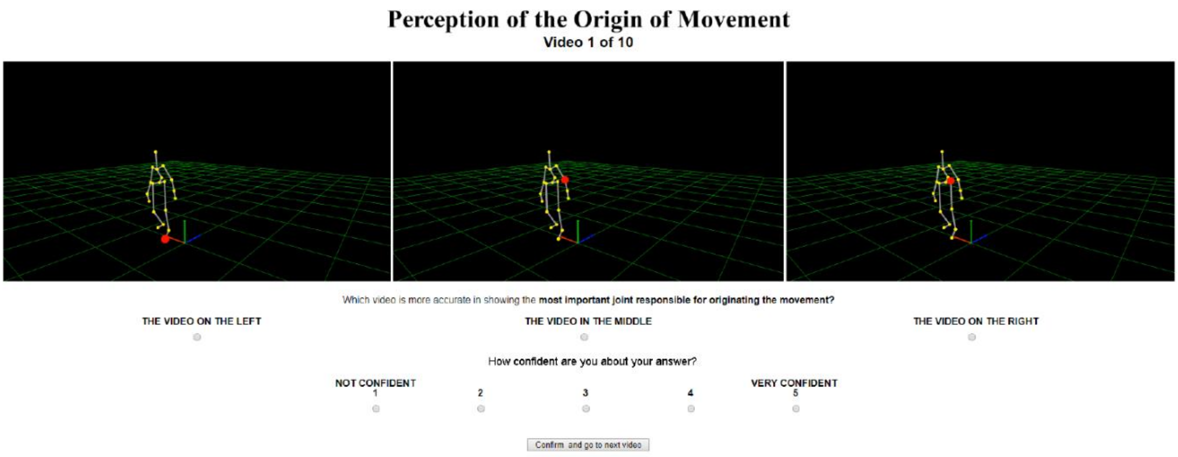
DO NOT refresh the browser page once the test has started.

IT IS ADVISED that this experiment be conducted on a screen with a resolution of at least 1920 by 1080 pixels.
You can, at any moment, decide that you do not wish to participate in/complete the test. In this case, please close your browser and contact us here or here.

If you would like to begin the test, please click on "Start test".

Figure 2. Introductory page of the web tool for the evaluation of the method.

Once the user agrees to participate and perform the task, a series of triplets of videos is presented.



Perception of the Origin of Movement
Video 1 of 10

Which video is more accurate in showing the **most important joint responsible for originating the movement?**

THE VIDEO ON THE LEFT	THE VIDEO IN THE MIDDLE	THE VIDEO ON THE RIGHT
○	○	○

How confident are you about your answer?

NOT CONFIDENT 1	2	3	4	VERY CONFIDENT 5
○	○	○	○	○

Figure 3. Website for the evaluation of the proposed method: best method selection.

Each of the three videos in a triplet displays a skeletal representation of a dancer performing the same full-body expressive movement. Each video has one highlighted joint (in red). This joint corresponds to the most relevant joint according to one of the following criteria: (i) joint with the maximum Shapley value; (ii) joint with the maximum speed; and (iii) random choice. The identity of the joint highlighted in red is possibly updated by each criterion every second (making it difficult for the user to guess when the criterion applied in a specific video is, e.g., a random choice). The order of the three criteria is randomized among the three videos so that the specific criterion applied to each video is not predictable by the user. For a fair evaluation, the criteria themselves are also completely unknown to the user (i.e., the user has no idea how they are named and how they work).

During the survey, the participant is asked to choose the video that better represents the evolution of the most relevant joint responsible for originating the dancer's movement. Once a user has selected one video, he/she is asked to declare how confident he/she is in his/her choice by selecting a value from 1 to 5 on a 5-point Likert scale (levels: not confident, not so confident, neutral, confident, very confident). The participant can see all the videos as many times as desired and has to answer both questions (video choice and confidence level) before proceeding to the next triplet of videos. Each participant has to rate ten triplets of videos proposed from a selection of one hundred triplets using a Latin square selection method.

The website was submitted to people with three different levels of expertise in dance: professionals, semi-professionals, and novices/non-dancers. A total of 22 people took part in the evaluation. Each participant self-evaluated his/her own level of expertise. The general information about the participants is as follows:

Professionals: 8 participants (3 male, 5 female), with a mean age of 42.75 years (std 9.56 years).

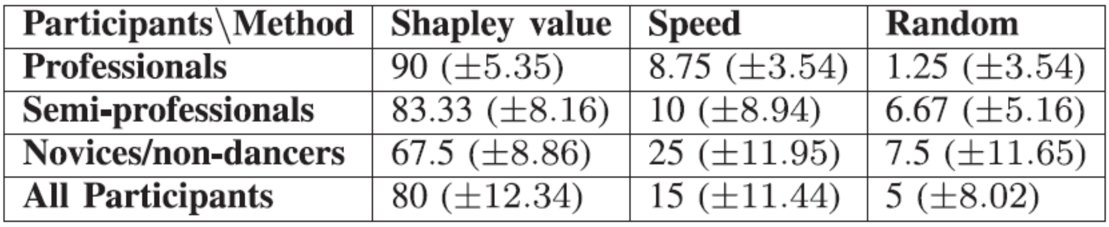
Semi-professionals: 6 participants (3 male, 3 female), with a mean age of 30 years (std 4.47 years).

Novices/non-dancers: 8 participants (6 male, 2 female), with a mean age of 35.5 years (std 7.4 years).

In the first two cases, the dancers were, respectively, experts and amateurs in contemporary dance.

Examples of the obtained results

Some results of the chosen type from among the three types of different stimuli are presented in the Table 1 and in the diagram shown in Figure 4. Both demonstrate that the results of the validation of the proposed method are promising. Indeed, the Shapley value method was selected in the large majority of cases.



Participants\Method	**Shapley value**	**Speed**	**Random**
Professionals	90 (±5.35)	8.75 (±3.54)	1.25 (±3.54)
Semi-professionals	83.33 (±8.16)	10 (±8.94)	6.67 (±5.16)
Novices/non-dancers	67.5 (±8.86)	25 (±11.95)	7.5 (±11.65)
All Participants	80 (±12.34)	15 (±11.44)	5 (±8.02)

Table 1.

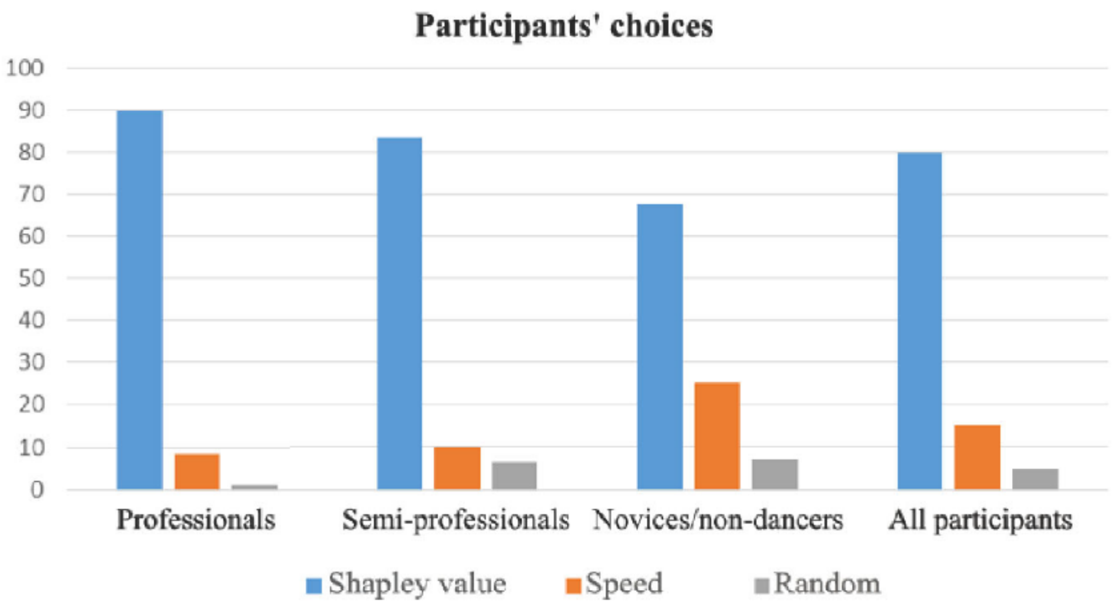


Figure 4. Participants' choices.

Table 2 shows, for two selected frames, the first 5 joints, ordered non-increasingly with respect to their Shapley values, normalized with respect to the maximum Shapley value in each frame. The associated normalized Shapley values are also reported in the table. The movements associated with the two frames are, respectively:

a) a sudden leaning down to the left with the trunk and the head, followed by a rotation to the right, with a final rising of the trunk and head, where the shoulder centre is clearly the origin of movement;

b) a rotation and extension toward the right of the performer, where the right elbow and right shoulder are clearly leading the movement of the whole body.

In both cases, the origin of movement is correctly identified by the proposed approach.

The two frames illustrate, respectively, the following situations, which were quite often observed during the analysis of the specific motion capture data set:

a) a case in which the first and second largest Shapley values are very well separated and (at least) the first largest one is uniquely achieved;

b) a case in which there are two joints with the largest Shapley value, but these joints are connected by a physical edge in the body graph, and the second largest Shapley value (which, in the second column, corresponds to the third joint) is very well separated from the first one.

1st frame	2nd frame
shoulder centre (1.00)	Right elbow (1.00)
head (0.46)	Right shoulder (1.00)
right ankle (0.40)	Right hip (0.53)
right knee (0.40)	Right knee (0.53)
left elbow (0.38)	left ankle (0.27)

Table 2.

Developments

In March 2020, UNIGE and EuroMov started a collaboration to extend and further develop this approach with novel ideas. Several directions are under exploration by the two research teams, including the following ones:

Exploiting a more complex skeletal structure (for which each cluster of joints is associated to a specific joint in the simpler 20-joint skeletal structure), making it possible to analyze movement in parallel at a finer interacting spatio-temporal scale in a multiple-scale approach.

Using movement-related features different from speed (or of a higher dimensional feature vector) to compute the Shapley value for a comparison with the results obtained using speed as a feature.

Incorporating multiple temporal scales. For example, one can look at a fast temporal scale at the very first moment of the origin of movement and at a slower temporal scale where one can analyse the origin of movement at a higher level.

Applying the developed methodology to analyse the emergence of the origin of movement when two persons or small groups are involved in the movement itself. In this case the graph nodes are not anymore the joint of an individual person, but each person of the group is a node of a more complex organism formed by the social group.
This extension of the theoretical framework is in the direction of observing movement at different spatio-temporal scales.

A recent paper including preliminary results on some of these extension of the theory will be presented in a joint UNIGE-EuroMov paper at the ACM ICMI 2020 EnTimeMent Workshop.

References

S. Cohen, G. Dror, and E. Ruppin, “Feature selection via coalitional game theory,” Neural Computation, vol. 19, no. 7, pp. 1939-1961, 2007.

K. Kolykhalova, G. Gnecco, M. Sanguineti, G. Volpe, and A. Camurri, “Automated analysis of the origin of movement: An approach based on cooperative games on graphs,” IEEE Transactions on Human-Machine Systems, 2020. DOI: 10.1109/THMS.2020.3016085

M. Maschler, E. Solan, and S. Zamir, Game Theory. Cambridge, UK: Cambridge University Press, 2013.

T. P. Michalak, K. W. Aadithya, P. L. Szczepánski, B. Ravindran, and N. R. Jennings, “Efficient computation of the Shapley value for game-theoretic network centrality,” Journal of Artificial Intelligence Research, vol. 46, pp. 607-650, 2013.

4. MECS - THE MULTI-EVENT-CLASS SYNCHRONIZATION ALGORITHM (UNIGE)

Synchronization is a fundamental component of computational models of human behavior, at both intrapersonal and inter-personal level. Event synchronization analysis was originally conceived with the aim of providing a simple and robust method to measure synchronization between two time series. In this paper, we propose a novel method extending the state-of-the-art of event synchronization techniques: Multi-Event-Class Synchronization (MECS). MECS measures synchronization between relevant events – belonging to different event classes – that are detected in multiple time series. Using MECS, synchronization can be computed between events belonging to the same class (intra-class synchronization) or between events belonging to different classes (inter-class synchronization). In our paper (Volpe et al, submitted paper) , we also show how our technique can deal with macro-events (i.e., agglomerations of events satisfying specific temporal constraints) and macro-classes (i.e., agglomerations of classes). Finally, our submitted paper presents a case study in which we exploit MECS to compute synchronization between multimodal channels “on-the-fly”. In particular, the proposed example shows how synchronization between respiration and full-body movement of a person explains different movement qualities such as fluidity and impulsivity. Next steps in EnTimeMent will include the extension of our MECS algorithm to compute event synchronization at multiple time-scales. An approach to this problem was proposed recently (Eero et al 2017).We believe MECS can support research on synchronization processes both at an intra-personal and at an inter-personal level. Implementing MECS as a module in the EyesWeb platform makes it freely available to the scientific community, contributing to shed light on phenomena that range from neuronal activity, to human behavior analysis, up to social interaction and cooperation in teams.

References

Gualtiero Volpe, Paolo Alborno, Maurizio Mancini, Radoslaw Niewiadomski, Stefano Piana, Antonio Camurri (2019) MECS – The multi-event-class synchronization algorithm. JAM8 – Intl Joint Actions Meeting, Genoa.

Gualtiero Volpe (submitted) MECS – The multi-event-class synchronization algorithm.

Eero Satuvuori, Mario Mulansky, Nebojsa Bozanic, Irene Malvestio, Fleur Zeldenrust, Kerstin Lenk, and Thomas Kreuz. 2017. Measures of spike train synchrony for data with multiple time scales. Journal of Neuroscience Methods 287 (2017), 25–38. https://doi.org/10.1016/j.jneumeth.2017.05.028

5. CAPTURING HUMAN MOVEMENT AND SHAPE INFORMATION FROM SMALL GROUPS TO EXTRACT EXPRESSIVE AND SOCIAL FEATURES – USING MARKER-LESS TECHNIQUES

Human beings exhibit phenomenal capabilities in synchronizing joint actions and coordinating at the inter-personal level in a non-verbal manner. This is observed specifically in musical ensembles where co-performers are seen to coordinate their movements effortlessly (Bishop, 2018). Perhaps the most natural response to music is to move and synchronize to the rhythmic elements in and inter-twined in music. When listening to music, we tend to raise our hands, tap our feet, dance, and shake our heads. It has been observed in a musical ensemble, that when a musical piece is being played, there are parts or phrases, basis which the members tend to coordinate their movements with the rhythmic behavior of other group members.

We, as individuals, usually have a general feeling of entrainment, but almost unknowingly, yet spontaneously, we move to music being played around us (Bishop et al. 2018). We may react genuinely to music, and sometimes showcase unique movements in response to what is being heard (Luck et al. 2010; Vuoskoski et al. 2011). The mix of music and the corresponding movement seems to trigger a social bonding effect. These movements or bodily gestures tend to convey certain subtle messages, and this conveyance of messages is crucial for the co-creation of a musical piece, and musicians are seen to perpetually move during a performance to augment the creation of sound, express themselves, communicate with their fellow group members, and transition into states of synchronization (Wanderley, 2001; Bishop et al. 2018). This coordination often takes place in the context of different musical texture that vary in terms of whether or not there is a clear hierarchy with a leader playing the melody while others play the accompaniment.

Music permits the study of these subtle facets of the human body, using current state-of-the-art methods, while producing minimal noise during experiments. This noise is the random variability one may find in a signal. One way to study such an area is extracting the signals that emerge out of human movements. But when one must perform experiments in-the-wild, noise is a major concern – which adds to the numerical complexities. Another benefit of experimenting with music is that it allows research to be carried out with more care, attention to detail, and control. Music ensemble performances are in turn special examples of joint actions with key advantages. In our experiments, we analyze each of the videos in phrases. These phrases are units of information at relatively long musical timescales. Interpersonal coordination measures using a windowing approach captures shorter timescales. During our analysis we examine different positions of the phrase (start, middle, end). With this, we explore the opportunity to answer questions related to multiple timescales – especially how coordination at shorter timescales changes over the course of longer timescales.

We explore the relationship between music and movement by proposing a computational model to compute the synchronization of dyadic pairs in a musical ensemble using marker-less computer vision techniques. Our methods involve the use of human pose estimation algorithms. The human body in pose estimation algorithms is looked at as a system with many elements. Each of these elements, called a key-point, is tracked in a consistent manner through the visual sequences. Eventually, on discovering the position of a group of anatomical joints such as elbows, nose, shoulders, knees, etc., these key-points help identify a blue- print where a skeleton-like structure of the body can be super-imposed. The algorithm then proceeds with completing a pair-wise connection between these key-points to provide a human skeletal structure, super-imposed on the input, known as a Pose. On implementing these algorithms on our dataset, we receive an output as a json file which contains the coordinates of essential body joints.

Interestingly, in a musical ensemble, due to being seated, musicians communicate with each other not using speech, but head and other upper-body movements. They can interact with each other with visual or audio cues. This interaction usually conveys a message to initiate a musical piece at a certain time and could also convey how musical notes need to be played. Thus, we focus on investigating techniques for the automatic analysis of synchronization by tracking the movement of the human head

(Xiao et al. 2001; Yokozuka et al. 2018). The head is tracked by making use of the coordinates of the human nose as provided by the pose estimation algorithm. The kinematic information extracted from the head movements of the performers are then subject to certain signal processing techniques which also find use in neuroscientific research activities to assess connectivity and synchronization of different regions in the brains. We use the data to compute dyadic synchronization between the performers using a metric called the Phase-Locking Value.

5.1 Dataset and Hypothesis of the experiments

The dataset used for these experiments consist of videos from concert performances by the Omega Ensemble, a professional chamber music group from Australia. Each of the videos have been annotated as determined by a musicological analysis based on the published score. For our research the concert video on which experiments were performed was “Brahms Clarinet Quintet” in B minor (Op.115) written in 1891. This work has three movements, which are each designated as a “piece” in the analysis below. Each of these videos being experimented upon have 5 participants.

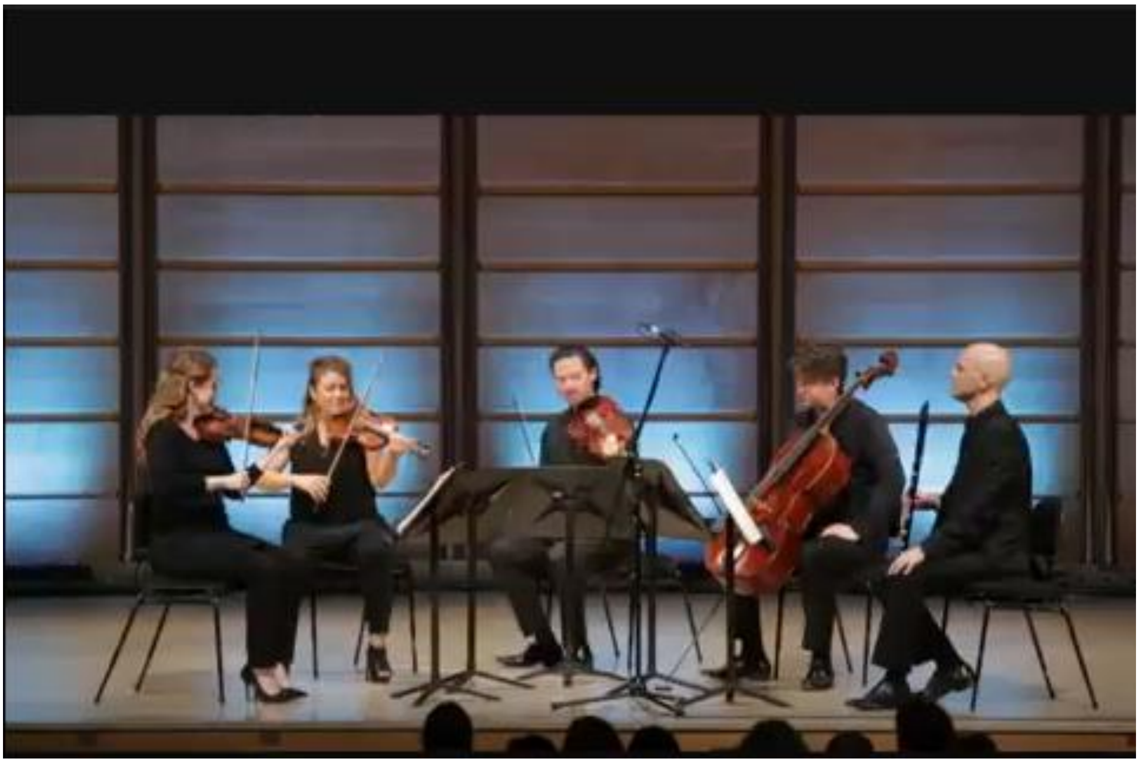


Figure 1: Image from a musical piece composed by Johannes Brahms

The goal of our study is to investigate factors that influence the quality of interpersonal coordination between group members. The planned analyses address:
1. Strength; and
2. Directionality of interpersonal coupling in musical ensemble performance.

The specific aim of this analysis is to test how strength and directionality of coupling are influenced by two factors:
1. Position within the musical phrase (Start, Middle, and End); and
2. Musical textures (Homophonic and Polyphonic)

Previous research (Scho¨gler, 1999; Keller et al. 2014) suggests that coupling may be stronger at the beginning and end of phrases than in the middle. The expectation is that when leadership is not assigned there is better interpersonal coordination. This would suggest that coupling strength will be stronger when leadership is distributed – stronger in polyphonic than homophonic texture conditions (Novembre et al. 2015; Noy et al. 2011; Varlet et al. 2020). Also, 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.

Our study is strongly based on agreements that performers in a musical ensemble produce more coordinated gestures, and by finding the dyadic synchronizations using performer head – movements, we get closer to answering some of these open-ended questions (King et al. 2011). Using non-intrusive techniques, we also present a new computational method to test and present results on the above hypothesis.
The inter-personal coordination is observed by obtaining the dyadic synchronization on selecting a dyad combination of two co-performers in a musical group. For a set of 5 participants, we can have a total of 10 combinations.

They can be:
1. Participants 1 and 2
2. Participants 1 and 3
3. Participants 1 and 4
4. Participants 1 and 5
5. Participants 2 and 3
6. Participants 2 and 4
7. Participants 2 and 5
8. Participants 3 and 4
9. Participants 3 and 5
10. Participants 4 and 5

Thus, we end up with 10 eventual dyadic synchronization values for each video sample, and a total of 26 video samples have been studied. We as experimenters have decided which parameters to base our study on. This includes deciding within the pairs who shall be the melody and accompaniment for between subject effects. This was established on the basis of the musicological analysis, which classified whether each part was playing the melody or accompaniment in the musical score. We then decided that texture will be a between subjects' factor in the Analysis of Variance because we did not have control over the textures since they were specified in the musical score. Also, no two people are playing the same part – thus also permitting the inclusion of factors such as instruments and understanding their individual effects on a musical phrase.

5.2 Results

Our results investigate and demonstrate the variation in body movements during different kinds of musical textures, and delicate transmission of messaging at multiple temporal scales. The Phase Locking Values for all pairs were entered into an Analysis of Variance that tested for effects of position in the phrase, texture, and pair.

Position * Texture

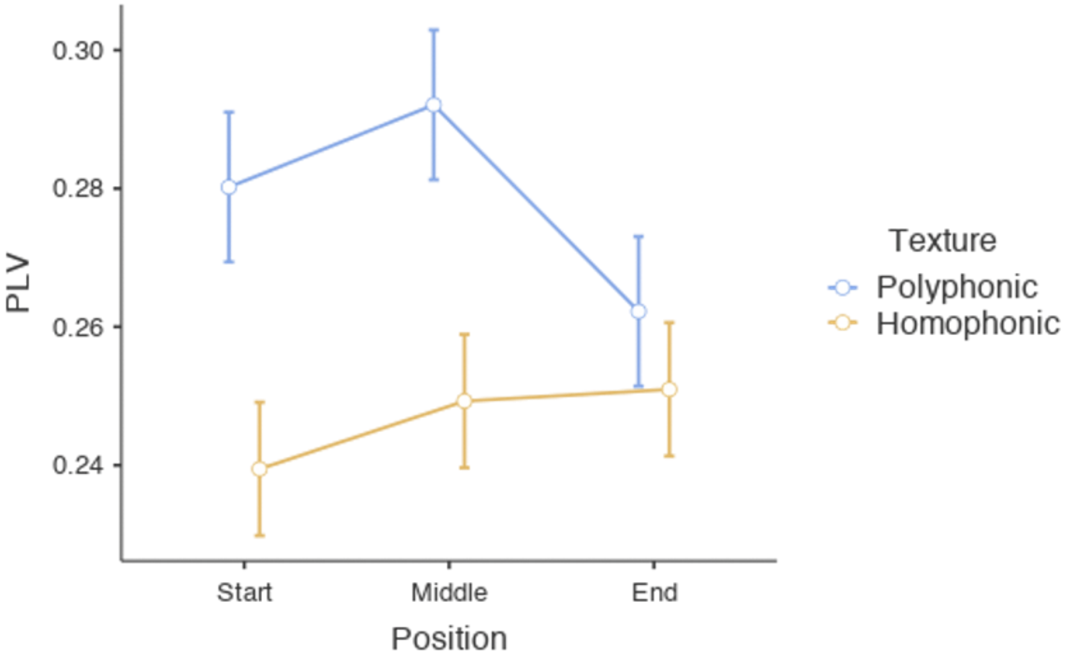


Figure 2: Measure output for Position x Texture

From Figure 2, we observe how the PLV begins at a lower value in both textures, Polyphonic and Homophonic. These PLVs while they start out higher, it tends to rise until the middle of the phrase, and then begins to drop in value towards the end of a musical phrase. It is typically seen from the data plotted that polyphonic textures have stronger coupling than homophonic textures.

5.2.1 Texture * Pair

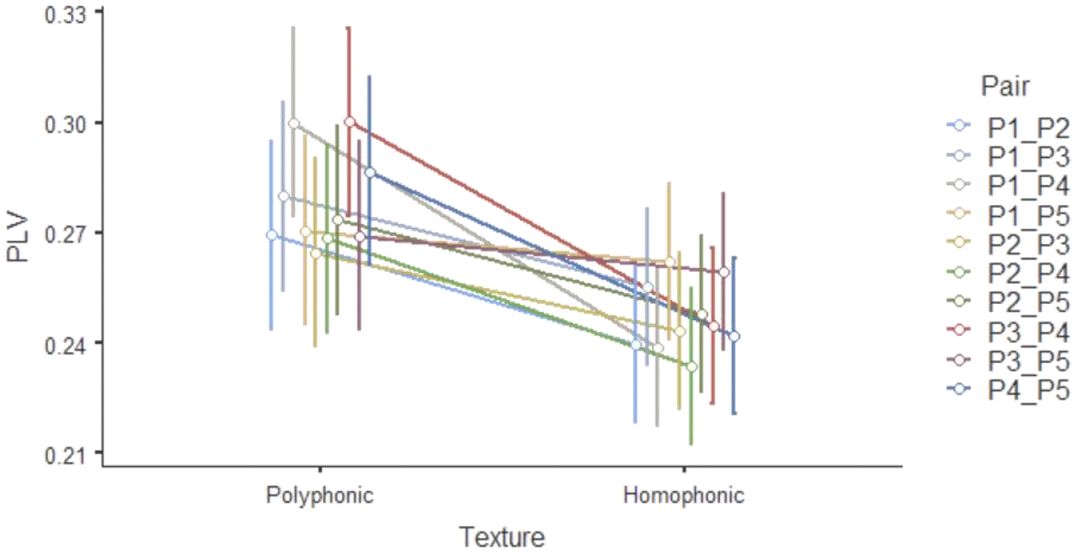
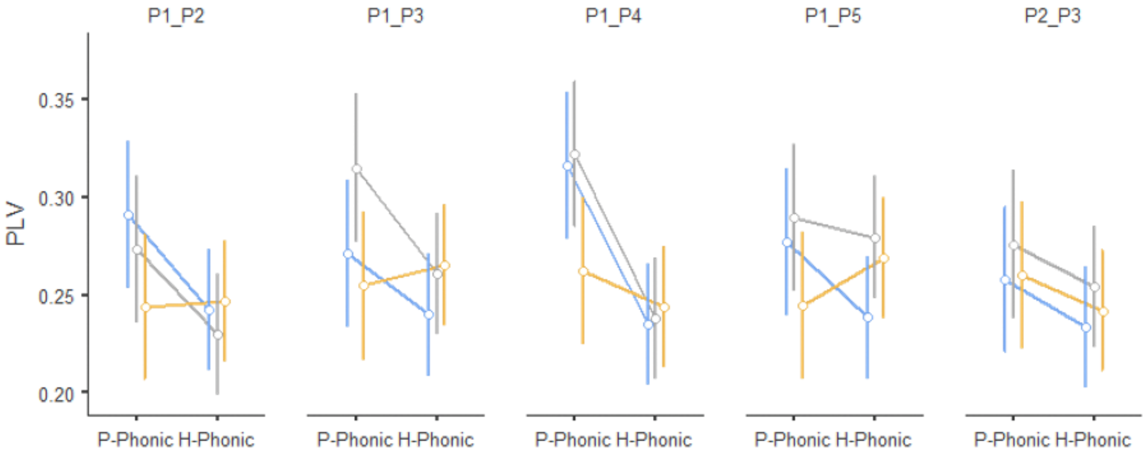


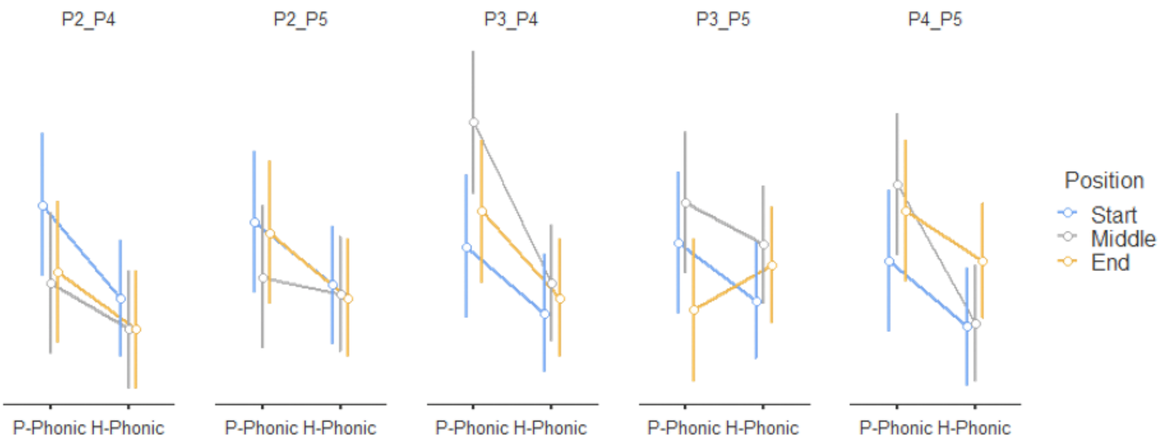
Figure 3: Measures output for Texture x Pair

There are 5 players synchronizing their notes in both the homophonic and polyphonic textures. As seen in Figure 3, pertaining to a pair – wise analysis, we notice that all pairs show a higher level of synchronization in polyphonic textures. We see a sharp or marginal drop in all pairs of performers, suggesting that there may exist sufficient conditions to initiate a leader and follower relationship. This could be because the four performers in a homophonic texture are coupled to the melody player, whereas in a polyphonic texture the coupling is evenly distributed across all performers.

5.2.2 Texture * Position * Pair



(a)

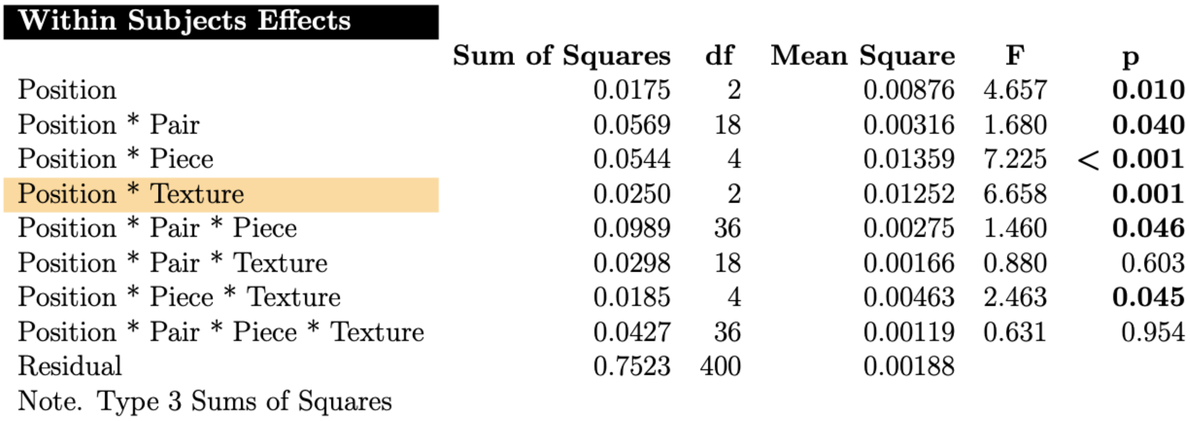


(b)

Figure 4: Measure output for Texture x Position x Pair

In Figure 4, P-Phonic stands for polyphonic while H-Phonic stands for homophonic, and the results have been arranged in a pairwise order. They depict the results of marginal means of the position as well. We observe that nearly all pairs are seen to have higher PLV values in the middle section of the phrases. In some examples such as P2_5, P2_4, and P1_2, this may differ because while there may be head movements involved, our video is being captured from the front. Thus, the ones seated on the extreme left and extreme right will exhibit more numerical movement due to being seated in a different orientation which can capture better trajectories. Additionally, the drop in Phase Locking Values at the end of the phrase for polyphonic textures could indicate that one of the instruments takes over as leader at that point, thus making it like a homophonic texture and reducing the symmetry of coupling across the ensemble.

5.2.3 Repeated Measures ANOVA

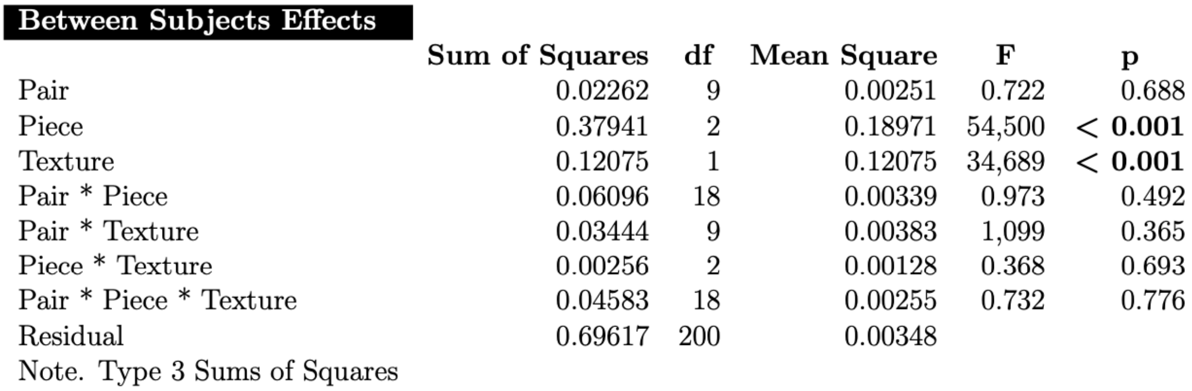


Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Position	0.0175	2	0.00876	4.657	**0.010**
Position * Pair	0.0569	18	0.00316	1.680	**0.040**
Position * Piece	0.0544	4	0.01359	7.225	< **0.001**
Position * Texture	0.0250	2	0.01252	6.658	**0.001**
Position * Pair * Piece	0.0989	36	0.00275	1.460	**0.046**
Position * Pair * Texture	0.0298	18	0.00166	0.880	0.603
Position * Piece * Texture	0.0185	4	0.00463	2.463	**0.045**
Position * Pair * Piece * Texture	0.0427	36	0.00119	0.631	0.954
Residual	0.7523	400	0.00188		

Note. Type 3 Sums of Squares

Table 1: ANOVA results for Within Subjects Effects



Between Subjects Effects	Sum of Squares	df	Mean Square	F	p
Pair	0.02262	9	0.00251	0.722	0.688
Piece	0.37941	2	0.18971	54,500	**< 0.001**
Texture	0.12075	1	0.12075	34,689	**< 0.001**
Pair * Piece	0.06096	18	0.00339	0.973	0.492
Pair * Texture	0.03444	9	0.00383	1,099	0.365
Piece * Texture	0.00256	2	0.00128	0.368	0.693
Pair * Piece * Texture	0.04583	18	0.00255	0.732	0.776
Residual	0.69617	200	0.00348		

Note. Type 3 Sums of Squares

Table 2: ANOVA results for Between Subjects Effects

The ANOVA measures found in Tables 1 and 2 reveal statistically significant main effects of Position, F(2, 400) = 4.657, p = 0.010, and Texture, F(1, 200) = 34.689, p ¡ 0.001. A significant effect of Piece was also observed, F(2, 200) = 54.500, p ¡ 0.001, but this is currently beyond our theoretical interest. Additionally, the two-way interaction between Position and Texture is statistically significant, F(2, 400) = 6.658, p = 0.001. This indicates that there is a matching effect between the Position and Texture in musical phrases that have been analyzed – which is in line with the specific aim of our hypothesis and confirms the reliability of the overall pattern of results described in the sections above. Note: Values in bold indicate statistical significance (p<0.05).

Thus, based on our results, we can say that the effect of the texture changes with respect to the position of the phrase, further helping us study the relationship between musical textures and overall coupling strengths of performers over the course of a musical phrase.

References

L. Bishop, "Collaborative musical creativity: How ensembles coordinate spontaneity," Frontiers in psychology, vol. 9, p. 1285, 2018

L. Bishop and W. Goebl, "Communication for coordination: gesture kinematics and conventionality affect synchronization success in piano duos," Psychological Research, vol. 82, no. 6, pp. 1177–1194, 2018

G. Luck, S. Saarikallio, B. Burger, M. R. Thompson, and P. Toiviainen, "Effects of the big five and musical genre on music-induced movement," Journal of Research in Personality, vol. 44, no. 6, pp. 714–720, 2010

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

M. M. Wanderley, "Quantitative analysis of non-obvious performer gestures," in International Gesture Workshop, pp. 241–253, Springer, 2001

L. Bishop and W. Goebl, "Beating time: How ensemble musicians' cueing gestures communicate beat position and tempo," Psychology of music, vol. 46, no. 1, pp. 84– 106, 2018

T. Yokozuka, E. Ono, Y. Inoue, K.-I. Ogawa, and Y. Miyake, "The relationship be- tween head motion synchronization and empathy in unidirectional face-to-face communication," Frontiers in psychology, vol. 9, p. 1622, 2018

B. Scho ̈gler, "Studying temporal co-ordination in jazz duets," Musicae scientiae, vol. 3, no. 1 suppl, pp. 75–91, 1999

P. E. Keller, G. Novembre, and M. J. Hove, "Rhythm in joint action: psychological and neurophysiological mechanisms for real-time interpersonal coordination," Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 369, no. 1658, p. 20130394, 2014

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," Royal Society open science, vol. 2, no. 11, p. 150286, 2015

L. Noy, E. Dekel, and U. Alon, "The mirror game as a paradigm for studying the dynamics of two people improvising motion together," Proceedings of the National Academy of Sciences, vol. 108, no. 52, pp. 20947–20952, 2011

M. Varlet, S. Nozaradan, P. Nijhuis, and P. E. Keller, "Neural tracking and integration of 'self' and 'other' in improvised interpersonal coordination," Neuro Image, vol. 206, p. 116303, 2020

E. King and J. Ginsborg, "Gestures and glances: Interactions in ensemble rehearsal," New perspectives on music and gesture, pp. 177–201, 2011