EXPLORING THE PERCEPTION OF EXPRESSIVITY AND INTERACTION WITHIN MUSICAL DYADS

Georgios Diapoulis

Master's thesis

Music, Mind & Technology

Department of Music

June 15, 2016

University of Jyväskylä

JYVÄSKYLÄN YLIOPISTO

Tiedekunta – Faculty Humanities	Laitos – Department Music
Tekijä – Author Georgios Diapoulis	
Työn nimi – Title Exploring the perception of expressivity and interaction within musical dyads	
Oppiaine – Subject Music, Mind & Technology	Työn laji – Level Master's Thesis
Aika – Month and year June 2016	Sivumäärä – Number of pages 33

Tiivistelmä – Abstract

Bodily gestures in music performance play an important role in the perception and appreciation of music by the audience. Performance variations can be identified by observers even when no auditory cues are available; visual kinematic information has been demonstrated to be crucial in identifying expressive intentions. The current thesis explores violin dyads performing an unfamiliar song. High quality optical motion capture was employed to record full body movement. We applied the standard paradigm to study the perception of expressive performance. Our aim was to predict perceptual ratings of expressivity from movement computations. That is, to deduce secondary aspects of musical gestures, namely intentions, from primary aspects, namely physical movement. Our hypotheses were that perceived expressivity is dominated by vision and that ancillary gestures carry significant cues for observers. For this purpose, a web-based perceptual experiment was conducted using short duration stimuli that were rendered as point light representations. Sound and vision, vision only, and sound only stimuli were employed to explore interactions across different modalities. Preliminary movement analysis showed that the musicians performed with greater amounts of kinetic energy in the more exaggerated expressive conditions. We extracted low level kinematic features based on instantaneous velocities of the markers of interest. We applied principal component analysis on the motion capture timeseries data, and we performed multiple linear regression and linear discriminant analysis to assess our hypotheses. We extend previous findings about visual perception in dyadic context, and we provide an account of idiosyncratic movements in violin performance.

Asiasanat – Keywords
perception, music performance, expressivity, gesture, dyad, movement features

Säilytyspaikka – Depository

Muita tietoja – Additional information

CONTENTS

1 INTRODUCTION

Music performance is the epitome of music making, and the later is the most intriguing of the musical activities. Musicians have to perform highly demanding cognitive and motor functions, while structural elements unfold within musical experience (Palmer, 1997). Music perception involves multimodal percepts and crossmodal interactions (Vines, Krumhansl, Wanderley, & Levitin, 2006; Vuoskoski, Thompson, Spence, & Clarke, 2016), and vision appear to have major importance in the perception of expressive music performance (Davidson, 1993; Dahl & Friberg, 2007; Davidson, 1993; Vuoskoski, Thompson, Clarke, & Spence, 2014). Bodily gestures have been proposed that play a major role in perception of expressive music performance. The notion of *musical gestures* under the prism of embodied music cognition unifies bodily movement and meaning (Cadoz, Wanderley, et al., 2000; Leman, 2008). Therefore, it provides a bridge to study subjective experience from bodily movement (Jensenius, Wanderley, Godøy, & Leman, 2009; M. Thompson, 2012). The study of expressive intentions in music heavily depends on the *standard paradigm* (Juslin & Timmers, 2010). This paradigm assigns expressive manners to musicians which are being validated via ground truth knowledge of perceptual judgements.

The focal point of this thesis was to explore the role of bodily gestures in the perception of expressivity, interaction, and synchronization within dyadic violin performance. The main experiment is based on infrared optical motion capture technology. Three pairs of violinists performed a short duration arrangement of a folk song. We conducted a perceptual experiment based on stimuli that we created using stick figures from the motion capture data and we performed regression models to predict perceptual ratings. Our approach stems from experimental psychology research, and the main point was to predict behavioural ratings from physiological data that we collected from full body motion capture.

Our aim was to sketch out performance variations of violin dyads while performing a folk composition. We followed a cognitive paradigm and we approached the problem from both a top-down and a bottom-up perspective. Our view stems from the embodied

cognition research program (Maturana & Varela, 1987), which differs in many aspects from the classical cognitive paradigm. In that respect we consider the Cartesian division between mind and body as an unfortunate event in scientific evolution. We see that physiological bodily movement has a joint role with neuronal processes in order to make our understanding. In that respect, we focus on bodily movement with a view to describe intersubjective experience such as expressive intentions in music performance.

The dyadic context of our study points to the social interaction. That is of particular importance as cognitive sciences have mainly focused on the study on individuals. Gestures are considered as prelinguist assets of human communication (Sebanz, Bekkering, & Knoblich, 2006). Gestures in music performance are primary responsible for sound production, but they may carry emotional content, or they may serve as means of communication with coperformers or with the audience (Jensenius et al., 2009).

The advent of music computing technologies opened news possibilities for the study of music using objective measures. Music research is being transformed from its artistic roots to an interdisciplinary, and in some cases to transdisciplinary research field. Besides that, the main focus have been placed on auditory information, though the last decades there is a growing interest in studying bodily gestures in the context of music making. As a consequence there is a growing interest in body movement and dance studies in music listening conditions.

The levels of cognitive processing in music can be studied in three different domains, the behavioural level of subjective experience, physiological measures and neural correlates (Luck, 2016). The present study aimed to interwind the first two levels. The neural correlates of music performance are fairly unexplored due to apparatus constraints. Subjective behavioural experience such as the perception of expressive performance is a real challenge when we are interested to present quantifiable results. The most ecologically valid perspective to study subjective feeling is the method of self-responses (Zentner & Eerola, 2010). The design of this experiment thus aimed to quantify subjective ratings about expressive music performance using physiological data of bodily movement.

2 THEORETICAL BACKGROUND

Musical expression was traditionally studied from the perspective of the musical composition. That has changed since 80s where the focus was placed on the performer (Dogantan-Dack, 2014). Musical expression studies both the musical content of a composition, that is the *composer's layer*, and the emotional content of the musician, that is the *performer's layer* (Schubert & Fabian, 2014). In the current research we focused on the later. That might be a trade-off, but our focus was to examine novel movement measures for quantifying expressivity and interactivity in music performance. We investigated attributes of performers' gestures in the context of music information retrieval for body movement, which are not so well developed as the study of expressivity based on acoustical features. We took advantage of the advent of high resolution optical motion capture systems, which offered new opportunities to quantify human movement with precision of millimeter, consequently we were able to explore fine grained gestural control in violin performance. On the other hand, qualitative analysis of human movement has been developed for the purposes of choreographers, and one of the most important qualitative approach is that of Laban movement analysis[1]. In our study we ignored any qualitative descriptions of body movement and we have fully relied on quantifiable movement measures.

2.1 Music Performance

Music performance has been studied for over a century; it is a higly demanding and sophisticated task which involves complex motor control and cognitive skills (Palmer, 1997). The research field of music performance is multilayered and it involves aspects of performance planning, that is mental representations of music, and performance practice, that is performance plans and strategies (Gabrielsson, 2003). Early research in music performance had mainly focused on issues of synchronization (Repp, 2006), though recently several models have been presented either for computational applications (Widmer & Goebl, 2004), or psychological accounts with implications in music education and research (Juslin, 2003).

[1] `https://en.wikipedia.org/wiki/Laban_Movement_Analysis`

2.1.1 Musical Expression

Etymologically "expression" derives from the French "exprimere" which means to 'press out' (Dictionary, 2004), and it has Latin root. In Greek language expression is "ek-phrasis", the prefix *ek* has identical meaning with the *ex*, that is 'out of', whereas *phrasis* is the root word for *phrase*. In Greek Wiktionary[2] *phrasis* (*φράσις*) has as root the verb *phrazo* (*φράζω*), though a possible affinity might be that of *phrin* (*φρήν*) meaning 'mind' or 'soul'. Translational loans between Latin and Greek have a long history, and it might be possible that the Latin root word is a translational loan from the Greek language. In that respect the minor affinity to *phrin* relates *expression* with high-level cognitive processing.

Expressive music performance is among the few musical universals (Davies, 2011). With the advent of computer music technology we were able to focus on sonic features, and in that respect the research community made an effort to explain expressivity in performance based on acoustical features (Dogantan-Dack, 2014). Whereas acoustical features indeed have a crucial role, they are not enough to fully describe the phenomenon of expressiveness. This fact is empowered by the view that movement should be taken into account in music information retrieval (Godøy & Jensenius, 2009). A methodological problem in the study of expressivity is that different studies use different units to quantify their results, which makes the comparison of different studies a difficult task (Palmer, 1997).

2.1.2 Motor Control

The architecture of motor control can be divided into *low-level* motor control and *high-level* motor control of gesture. Low-level control relies on the *theory of perception* (action-perception), with which we can state the hypothesis that our perception is linked to the gestural expression of sound-producing gestures. High-level control deals with the question of how the nervous system deals with the body movement. Three main concepts are involved: *motor equivalence*, *flexibility*, and *prediction*. Motor equivalence suggests that there are levels of controls for the mapping between the central nervous system commands and the muscle commands. Flexibility refers to the idea that the same planning strategies for the accomplishment of a movement could involve different synergies of muscles (eg. air-guitar, drumming with hands, or drumming with sticks). Prediction refers to that the capability to make predictions is possible if there is an internal representation of the gestures sequence (Gibet, 2009).

[2] `https://el.wiktionary.org/wiki/%E1%BC%94%CE%BA%CF%86%CF%81%CE%B1%CF%83%CE%B9%CF%82`

Gibet (2009) presented three different approaches in motor control; the motor program, the biomechanical approach, and the non-linear dynamics approach. The former approach in musical terms is referred in the bibliography as the timekeeper model (Repp, 2005). This approach considers an internal clocking mechanism in each individual that is responsible for sensorimotor synchronization. Questions arouse to whether or not this approach lays on an open loop or a closed loop (Gibet, 2009). The second approach stems from the *equilibruim point hypothesis* and considers that movement arise from perturbations around the equilibruim. The later approach considers that movement results from dynamic interaction with the environment, and it is in the focus of the present study.

2.1.3 Embodiment in Music Performance

The current study is designed from the first-, second-, and third-person perspective (Leman, 2008); that is the subjective level of experience, the social interaction, and the objective level of movement measures respectively. Leman (2008) argues that music does not always have a referential point. In that respect corporeal understanding accounts for recognition of expressiveness with no need for reference, whereas cerebral understanding, or cognition, often accounts for something that is imitated. This view suggests that corporeal imitation along with motor resonance may account for the perception of expression.

2.2 Music Perception

A pioneer of music perception was Aristoxenos, as he was likely the first person who realized that musical pitch is a perceptual phenomenon. Perception, just like knowledge, is an extrinsic property. In plain language that means that each individual is a unique perceiver. If we make an analogy from physics, mass is a intrinsic property of a body, but weight is an extrinsic as it depends on gravitational acceleration.

Music perception is neither purely mental, neither purely physiological phenomenon (Palmer, 1997). Perceptual studies are used as *ground truth* knowledge in order to evaluate intuitions and insights about phenomena. The perception of music performance suggests that the interpretation of performers and listeners is affected by structural elements of the musical composition. That might be a trade off that we have to overcome, as we did not performed any structural analysis over the musical composition, but our intuition is that this will have a minor effect as we are not investigating any metrical levels and we have segmented our melodic parts into musical closures (ie. complete musical phrases).

2.2.1 The Role of Vision

Abstract visual representations of biological motion using the point-light technique has been demonstrated that provide adequate information in order to recognize different activities, such as running, walking, or playing a musical instrument (Johansson, 1973). Visual cues have primary importance for the perception of expressive intentions (Davidson, 1993). Davidson studied solo pianists and showed that the visual channel dominates our music perception. Visual information seems adequate to recognize nominal emotional qualities, such as happy, sad, anger (Dahl & Friberg, 2007), and it has been used to recognize expressive conductors' gestures (Luck, Toiviainen, & Thompson, 2010). Furthermore, the importance of appreciation of music is enhanced using audio-visual stimuli (Platz & Kopiez, 2012), and several reports have shown crossmodal interactions between vision and sound (Chapados & Levitin, 2008; Vuoskoski et al., 2016).

2.3 Music Interaction

Representation is how we conceptualize phenomena in order to facilitate our understanding, whereas interaction is the way that our experience is being shaped by them. Interaction in musical terms is multifaceted and it may refer to different aspects of the musical realm, from instrument-performer and human-computer interaction, to performer-audience, performer-performer and teacher-learner interaction. As we have noted in the section 2.1.3 the current study explored the aspect of social interaction within dyadic performance. In that respect we investigated music interaction in the context of performer-performer interaction.

It is prerequisite that both individuals are aware that they are sharing a common experience in order to achieve a collaborative action (Tomasello & Carpenter, 2007). Gestures are considered as having a primitive role compared to that of concepts (Gill, 2015), and it has been noted that they may served as a prelinguistic form of communication (Sebanz et al., 2006). The evolution of language is of particular importance in the human evolution, as it enabled the flourish of art and science (Daniels & Bright, 1996, p. 265).

2.4 Musical Gestures

Music and movement are tightly bind with each other. Nonverbal cues have a catalytic role in music performance. We move our bodies in response to music, and we are being moved from musicians expressive intentions. In that respect musical gestures play an important role in the perception and appreciation of music by the audience. Musical

gestures have been conceptualized as having two interacting layers, the first is the focus on *extension* or physical movement, and the secondary focus is on *intention*, which is related to expressivity and meaning in music performance (Leman & Godøy, 2010).

2.4.1 Gestures in Music Performance

The physical anatomy of the musician is important, as smaller limbs like fingers are capable for performing fast movements with high dexterity, whereas the torso is related to slower and more rigid movement patterns. Expert performers use more repetition in the gestures than novices, due to the concept of motor equivalence and to the principle of least energy. Ancillary gestures carry significant information for identifying performers' expressive intentions (M. R. Thompson & Luck, 2012), and they may account for sound production as the level of expressivity increases. They may also referred to as *non-obvious* performer gestures (Wanderley, 2001).

Bowing gestures are more expressive because they involve more precise control of timbral features and duration of the notes. Anatomical differences in bowing movements are reflected in variation in the wristl and elbow of performers, and variations decrease as the tempo increases (Dahl et al., 2009). An interesting metaphor between speech and bowing is that vowels are the sustained part and consonants the strokes (Galamian & Thomas, 2013). This description constucts "diphones" out of bow strokes, which can be seen as a starting point for segmentation of bowing gestures. Gestures in electronic instruments differ because often there is no direct physical control of the produced sound. In that way electronic instruments are more closely to conductor's gestures (Dahl et al., 2009).

3 METHOD

3.1 Dyadic Violin Performance and Motion Capture

The current experiment is part of a study which focused on the effect of tempo and vision on interpersonal coordination (M. Thompson, Diapoulis, Johnson, Kwan, & Himberg, 2015). The expressive conditions that are explored in the current thesis were added as an extra part at the end of the aforementioned experiment.

3.1.1 Participants and Procedure

Three violin dyads participated in this study (6 musicians total; 4 females; age: M = 24.1, SD = 1.7). The violinists were recruited from student populations at the University of Jyväskylä and the Jyväskylä University of Applied Science. Musicians had received on average 15.8 (SD = 2.3) years of instrumental training on the violin.

The dyads performed a short piece arranged for two violins: "De Kleinste", composed by J. Beltjens (16 bars, 6/8 time signature). The score is available in the Figure 12. The selection of this song was done with a view to be unfamiliar to the performers. After a short rehearsal period, each dyad performed the piece nine times in a 3×3 task design: three expressive intentions (deadpan, normal, exaggerated) performed using three timing conditions (60-BPM, 90-BPM, free tempo). In the current study we ignored the effect of tempo, as a factor that might affect the overall perception of expressivity and interaction of the music performance.

The score of the song is in the Appendix (see section 7.2). We would like to thank Susan Johnson and Pui Yin Kwan for the selection and arrangement of the song.

3.1.2 Experimental Apparatus

Audio of the experimental trials was recorded using two AKG C417 L wireless microphones. The microphones were positioned around each violinist's right ear lobe and secured with adhesive tape. The recording sampling rate was 48 kHz, and recording

performed using ProTools digital audio workstation.

Optical motion capture data was produced using 8 Qualisys Oqus infrared cameras at 120 Hz sampling rate. Twenty-six markers were placed on the joints of each musician, and five markers were placed on the violin (2 on the bow, and 3 on the violin itself; see section 7.1). The data was labeled within Qualisys' Track Manager software and analyzed in MATLAB using functions within the MoCap Toolbox (Toiviainen & Burger, 2010).

3.2 Perceptual Experiment

A perceptual experiment is a deductive approach. From laws and theories we make predictions to explain certain phenomena. Following to the performance conditions that we assigned to musicians, we conducted the perceptual experiment to evaluate our hypothesis about the perception of expressive intentions. No explicit meaning was given to the participants for the concepts of *expressivity*, *interaction*, and *synchronization*.

For the development of the perceptual experiment we had two main directions. The first was to use short duration stimuli (5-10 seconds) in order to avoid continuous response ratings (Luck et al., 2010), and the second to include three types of stimuli: audio-visual (AV), visual-only (VO), and audio-only (AO). The study of mutlimodal perception is a "standard paradigm" in music research (Davidson, 1993; Juslin & Timmers, 2010; Vines et al., 2006; Vuoskoski et al., 2014). We performed two experiments, a pilot and the proper perceptual test.

The stimuli were based on stick figures, in order to avoid any biases from performers' appearance. We exported the video segments from the raw `tsv` files using MoCap Toolbox. We rendered the stimuli using REAPER digital audio workstation, and we segmented the stimuli and annotated the start and end position of each segment. We segmented the score on four parts (the score is available in section 7.2). The first segment was from the beginning of the score up to the end of the fourth meter, in the corona. The second segment started from the beginning of the fifth meter and ended on the next corona in the 8th meter. The third segment started from the following notes next to the aforementioned corona and ended on the next corona in the 12th meter, and the remaining part up to the end of the score was the fourth segment.

We ignored the effect of *mental rotation*[1] in our stimuli. We observed that there was no

[1] `https://en.wikipedia.org/wiki/Mental_rotation`

depth perception in the video segments, consequently our visual perception constructs a symmetrical representation of the two violinists. In fact both performers are right handed, and in that manner we see the posterior view of the performer on the right side (see Figure 1).

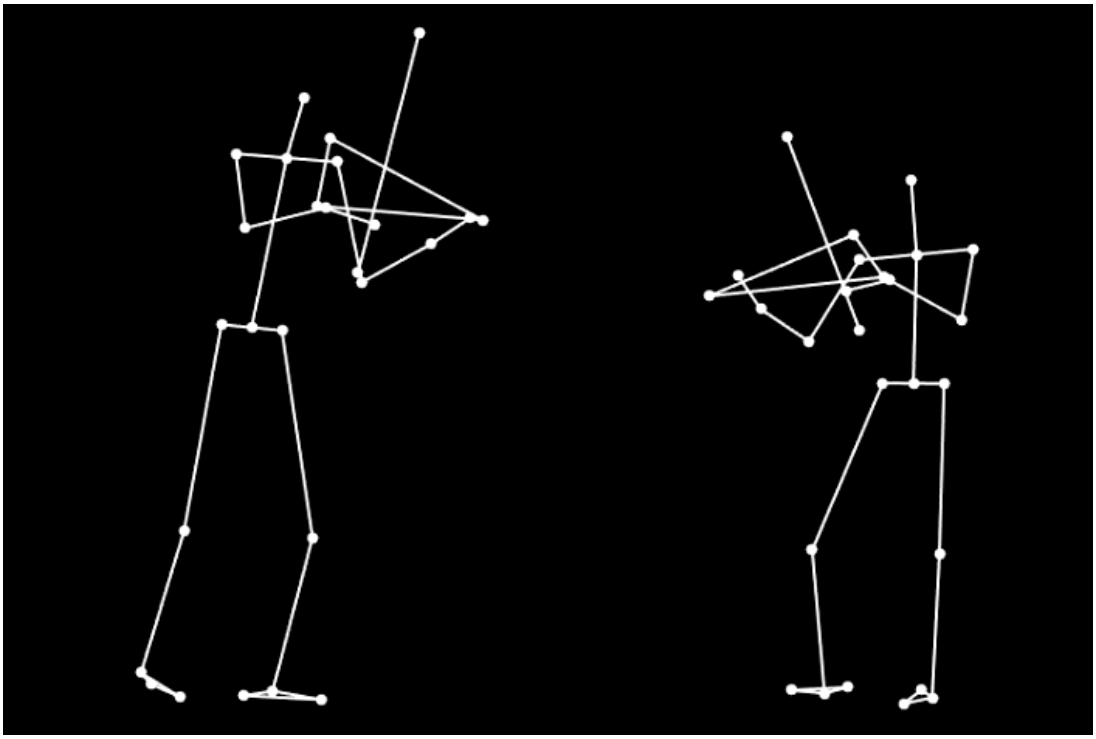

FIGURE 1. Snapshot from the stimuli of the preceptual test.

3.2.1 Pilot Perceptual Experiment Design

We performed a pilot perceptual experiment to assess our main hypothesis, which is that different levels of expressivity can be perceived in a dyadic context. Our aim was the experiment to be as sort as possible. In that respect we included 27 stimuli following a $3 \times 3 \times 3$ design for three expressive intentions, three types of modalities, and three dyads. We announced the experiment during the "Music, mind and technology" colloquium and we made announcements in mailing lists of the University of Jyväskylä. Because the web platform did not provide stimuli randomization we made a pseudo-randomization theme based on the season that the participants born[2].

In the pilot test the participants were asked to rate in an elevel-point Likert scale on two questions:

- In this segment, how expressive is the performance?

[2] `http://users.jyu.fi/~gediapou/pilot/test.html`

- In this segment, how interactive are the musicians to each other?

3.2.2 Perceptual Experiment Proper Design

For the proper experiment we segmented the musical composition in four phrases based on the score, and we followed a 2×3×3×4 design, in order to take into account exhaustively all possible combinations of two expressive conditions, three types of modalities, three dyads, and four melodic segments. We included only the deadpan and exaggerated expressive conditions with a view to reduce the average completion time of the experiment. That was done because we did not provided any incentives to the participants. The experiment was online for a period of two months.

The complete test had 72 stimuli, and for each stimulus we asked three questions:

- In this segment, how expressive is the performance overall?
- In this segment, how interactive are the musicians to each other?
- In this segment, how synchronized are the musicians to each other?

We used a low cost web-based platform to conduct the survey. Our view was to use very few instructions, in order to make to experiment user friendly for the participants. After the introduction page the participants watched one stimuli from each modality, that is an AV stimuli, a VO stimuli, and an AO stimuli. Next to that, the experiment started. The stimuli were presented in randomized order. At the end of the experiment we asked the participants to provide their age, sex and if they are musicians.

The videos were uploaded in an online streaming service and we didn't make use of the autoplay option. Each stimuli had three seconds of black screen (silence) in the beginning of the video. The participants had the opportunity to save their responses and continue later. A nine-point Likert scale was used for the ratings of expressivity, interaction, and synchronization, which is also known as *stanine*, for standard nine (Likert, 1932).

Design of the Stimuli

Table 1 shows the number of stimuli for each different category. We designed the experiment by making all the combinations of intentions, modalities, dyads, and melodic segments in an exhaustive manner. That is $2 \times 3 \times 3 \times 4$ equals 72 unique stimuli. Each

column represents the total number of the stimuli, that is 72 stimuli.

TABLE 1. Design of the stimuli for the perceptual proper.

Number of stimuli	Dyads	Intentions	Modalities	Segments
Dyad 1	24			
Dyad 2	24			
Dyad 3	24			
Deadpan		36		
Exaggerated		36		
Audio-visual			24	
Visual-only			24	
Audio-only			24	
Segment 1				18
Segment 2				18
Segment 3				18
Segment 4				18
Total number of stimuli	72	72	72	72

3.3 Movement Computations

Our aim was to extract low-level kinematic features with a view describe high-level kinematic features that account for expressive music performance. In that respect we computed the velocities of the four markers that are not related to obvious sound-producing gestures, particularly the head, the root, and the left & right shoulder (M. R. Thompson & Luck, 2012; Wanderley, 1999), and we performed dimension reduction techniques in order to identify the eigenmovements of violin performance (Toiviainen, Luck, & Thompson, 2010). More specifically, we applied principal component analysis (PCA) on the motion capture timeseries and we extracted global descriptors (standard deviation and kurtosis) that we used as predictors in regression models and as feature vectors in linear discriminant analysis (LDA).

3.3.1 Preprocessing of Movement Data

The total number of markers were 31 markers per performer, 26 on the human body and three markers on the violin body and two markers on the bow. The markers' labels as they were labeled in Qualisys software, their position over the body, and their relationship with the Dempster model are shown in Table 8. For the Dempster model we were using the markers' enumeration as documented in MoCap Toolbox Manual (Toiviainen & Burger, 2010).

Dempster Model

Dempster studied the properties of body segments using empirical measurements on the human body, and presented a model for applications in ergonomics and music research among others (Dempster & Gaughran, 1967). From this model we are able to find the center of mass for example, and to estimate kinetic features of human movement. In Figure 2 we see a snapshot from the experiment based on the Dempster model. We reduced the total number of the 26 markers for each performer to 20 *joints*, by averaging the position as shown in Table 8.

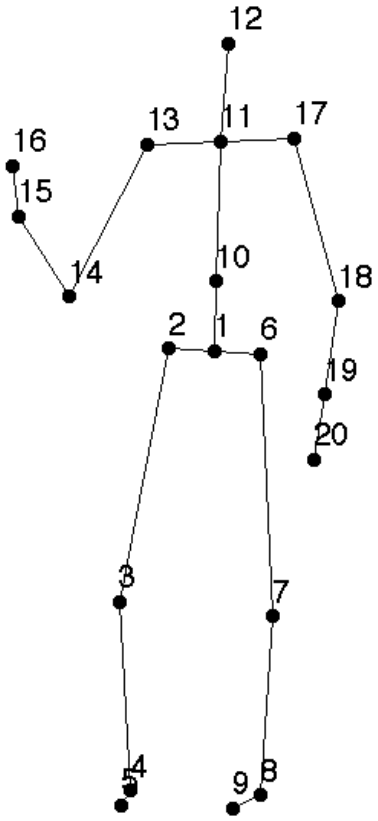

FIGURE 2. The Dempster model. This is posterior view of the performer, on the left hand side is the left side of the musician.

3.3.2 Movement Dynamics

Movement dynamics refer to kinematic and kinetic features of body movement. For the kinematic feature extraction we relied on the motion capture timeseries data and we extracted higher order derivatives of position data, and following to that we applied dimension reduction techniques. The main point was to reduce the complexity of the high dimensionality of human movement.

We standardized the motion capture timeseries data using local coordinate system for each performer. For that purpose we used the MoCap Toolbox function `mc2frontal` to assign anterior view for each performer using as a reference the left and right hips and by applying the function `mccenter` which calculates the centroid of all markers over time. We performed concatenation across the different segments that we used as stimuli in the perceptual test and we computed the first derivative of the position data, in order to estimate instantaneous velocities. For that purpose we focused on markers that are not related to obvious sound producing gestures (Wanderley, 1999, 2001). Specifically we focused on four joints of the Dempster model: head, root, left shoulder, right shoulder (M. R. Thompson & Luck, 2012). Afterwards we concatenated the timeseries data of the perceptual stimuli and we applied joint principal component analysis (PCA) on the velocity data (Toiviainen et al., 2010; Burger, Saarikallio, Luck, Thompson, & Toiviainen, 2012). The first five principal components (PCs) explained more than 95% of the variance. The final step was to calculate global descriptors for each PC in order to use them as predictors in the regression model. We calculated standard deviation and kurtosis for each PC, which are the second and fourth order of the statistical moment of the probability distribution (Glowinski, Camurri, Volpe, Dael, & Scherer, 2008). We ignored the mean value of the velocities because deviates around zero. The standard deviation shows the deviation of the velocities from the mean. Kurtosis is a measure which describes if the distribution of the velocities is narrow or widespread. Figure 3 shows step-by-step the procedure of the kinematic feature extraction. On the last step we applied LDA (Pedregosa et al., 2011) to predict classes of expressive intentions and performing pairs of violinists.

3.4 Statistical Learning

The statistical reasoning that we have relied is that of Bayesian statistics. In that respect we didn't follow any sophisticated falsificationism accounts (Chalmers, 2013). The main purpose of this study was to predict perceptual ratings from movement computations. We were looking for the most likely outcome, and we were not trying to reject any null hypothesis in order to empower our research hypothesis (Duda, Hart, & Stork, 2001).

3.4.1 Regression

Linear regression is a statistical technique to make predictions, that is a deductive approach. Multiple linear regression has more than one inputs as independent variables, the predictors, and predicts a single value, the response or dependent variable. Our predictors were the standard deviation and kurtosis of the first five PCs. Furthermore, we were us-

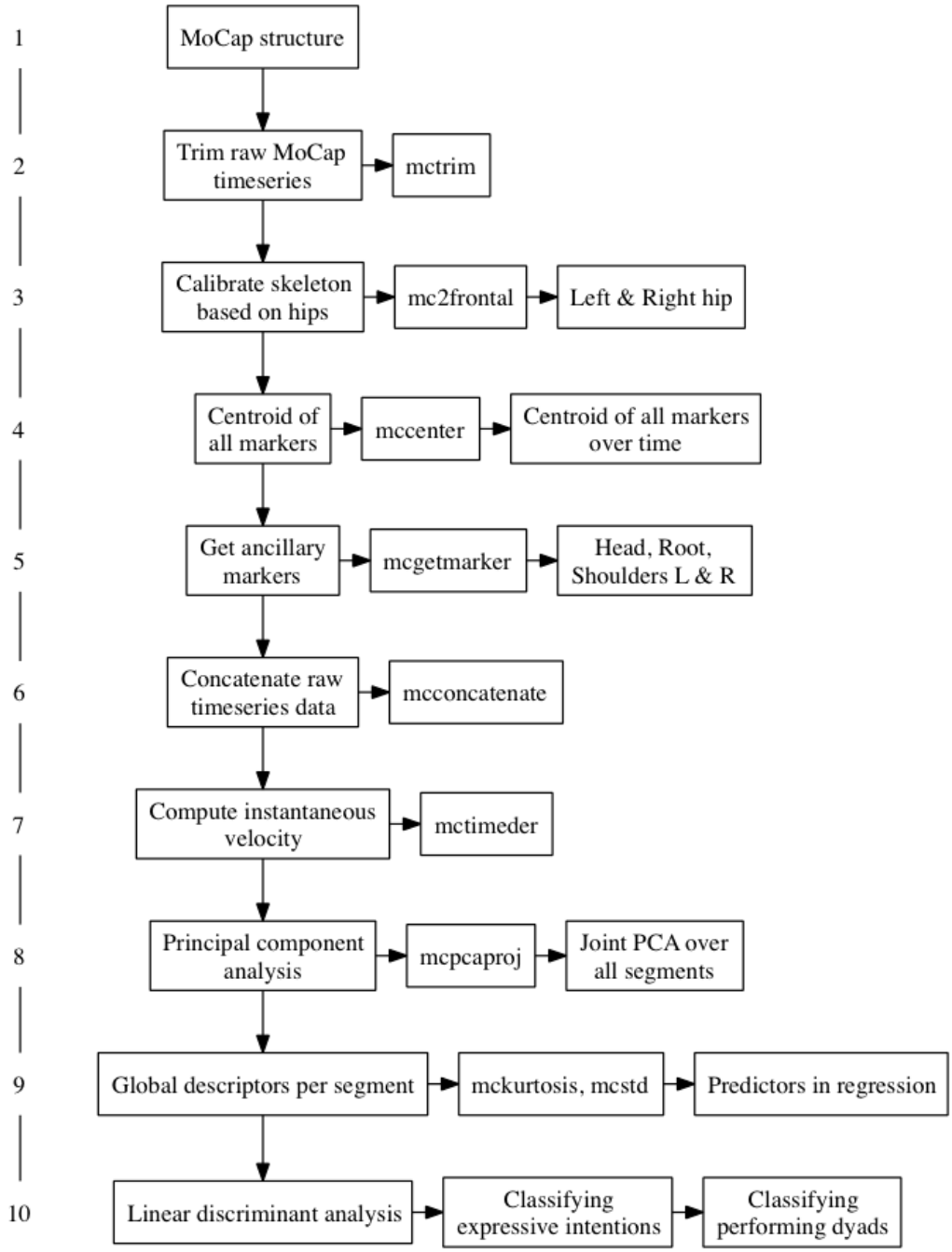

FIGURE 3. Analytical procedure of the kinematic feature extraction using MoCap Toolbox in MATLAB, and scikit-learn library in python (only in step 10).

ing cross-validation which is a method for estimating precision error (Hastie, Tibshirani, Friedman, & Franklin, 2005). Multiple linear regression can be described in mathematical language as the following equation 1, where y_i are the responses, β_0 the intercept, β_1, β_2 the regression coefficients, w_i and x_i the predictors, and ϵ_i the residuals. This is the simplest case of multiple linear regression using only two predictors.

$$y_i = \beta_0 + \beta_1 w_i + \beta_2 x_i + \epsilon_i \tag{1}$$

3.4.2 Dimension Reduction Techniques and Classification

Dimension reduction is used in feature extraction techniques in order to reduce randomness of variability within the data set. The main approach is to make a new dataset in which the new synthetic dimensions minimize the correlation of the variables. For example PCA is a linear tranformation which make a new coordinate system of components that are ordered based on the percent of explained variance. Dimensionality reduction is useful in regression models and classification.

Whereas regression is used to predict quantitative responses, classification is used to predict qualitative responses (Hastie et al., 2005). For example, in the Iris dataset, R. Fisher (1936) classified the three different species of Iris, Virginica, Setosa and Versicolor, using four features, petal length and petal width, and sepal length and sepal width. In this discrimination example Fischer introduced the method of *Linear Discriminant*, which was generalized later to *Linear Discriminant Analysis* (LDA).

PCA & LDA

Similar to PCA, LDA is another technique for dimension reduction. The aforementioned techniques are both linear transformations, which are fully reversible. PCA can be described as an unsupervised technique which maximizes the variance in the dataset, whereas LDA is a supevised technique which maximizes the distance between the classes in order to perform the discrimination. We applied PCA on the motion capture timeseries data and we extracted global measures, or global descriptors, for each perceptual segment. We used the global measures as predictors in the regression model. Then we applied LDA on the global predictors in order to predict the classes.

4 RESULTS

4.1 Perceptual Experiment Analysis

We conducted two perceptual experiments, a pilot study and the perceptual proper. We briefly report the main results from the pilot experiment and we continue to the proper experiment.

4.1.1 Pilot Perceptual Experiment

The pilot experiment was a web-based test which had in total 27 stimuli. We recruited 26 participants (age: M = 32.03, SD = 7.32, 50% female). The average completion time was 13 minutes, and the average duration of the stimuli was 7.71 seconds. For more information see section 3.2.1.

The stimuli design was 3×3×3 for each dyad, each modality, and each expressive condition. We standardized the perceptual ratings for each participant in order to remove perceptual biases. Mean ratings for the expressive ratings were M = 5.11, and the standard deviation of the means SD = .91. Mean ratings for the interactive ratings were M = 5.45 and SD = .77 respectively. Table 2 shows the mean ratings for each pair of musicians.

TABLE 2. Mean perceptual ratings per dyad

Mean ratings	Dyad1	Dyad2	Dyad3
Expressive	4.89	5.19	5.28
Interactive	5.31	5.55	5.54

4.1.2 Perceptual Experiment Proper

We remind the reader that the perceptual experiment was a web-based survey. The experiment was available online for a period of two months and we recruited 51 participants (N = 51, Female = 61.1%, musicians = 47.2%, age: M = 32.68, SD = 5.97), 36 provided complete responses and 15 were partial responses, which correspond to 28% of the length of the experiment. Our threshold for taking into account partial responses was 20 stimuli, which corresponds to duration of approximately 10 minutes of perceptual effort.

Mean Ratings

Figure 4 shows the mean expressive ratings for all the participants. The average of the means for the expressive, interactive and synchronization ratings are shown in Table 3. It is interesting to notice that the Pearson correlation of the mean perceptual ratings was greater than 99% between the unstandardized responses and the responses after we applied standard score transformation (zscore). The table below shows the non standardized responses, but for the regression model (see section 4.3.1) and the ANOVAs (see section 4.1.2) we reported standardized perceptual ratings using zscore.

TABLE 3. The average of the mean ratings per participant for each question.

	Expressive	Interactive	Synchronization
Average of means	5.10	5.26	5.78

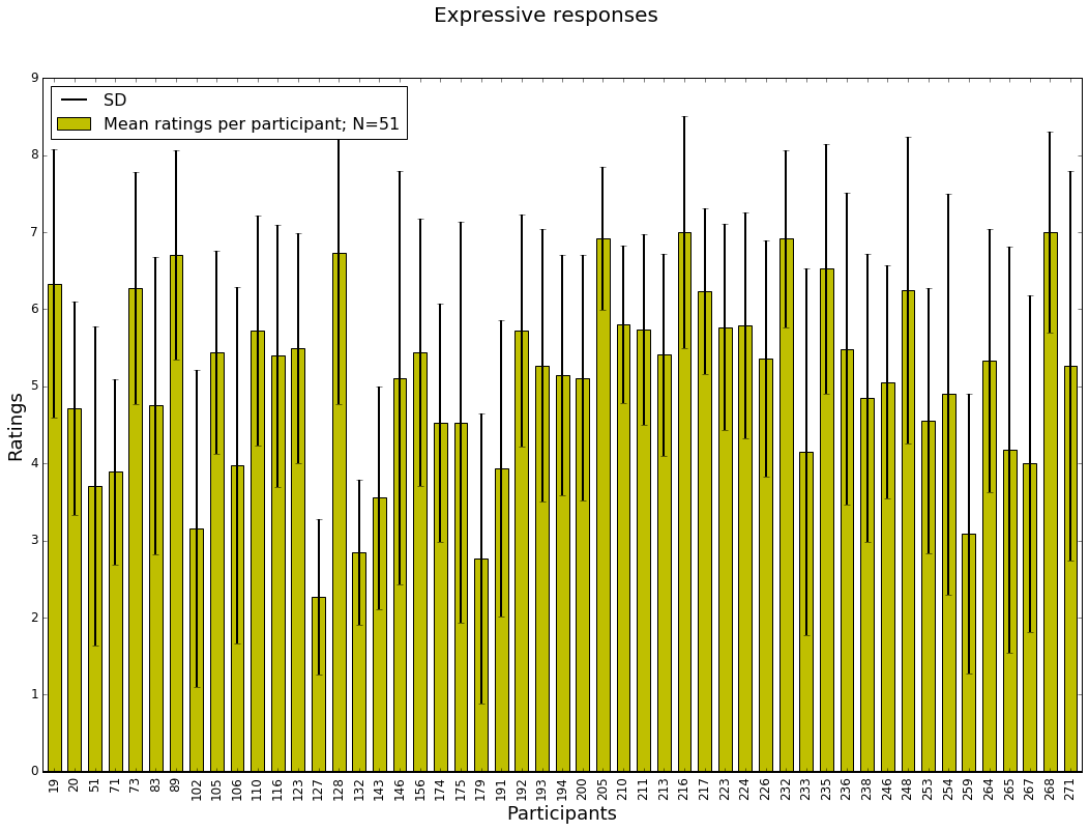

FIGURE 4. The mean ratings per participant for the expressive question.

Skewness and Kurtosis

We calculated skewness and kurtosis for every participant. The average mean skewness and kurtosis are shown in Table 4.

TABLE 4. Mean values for skewness and kurtosis.

	Expressive	Interactive	Synchronization
Skewness	-.13	-.26	-.43
Kurtosis	-.36	.19	.26

Cronbach's Alpha of Perceptual Ratings

The Cronbach's Alpha (α) is a measure to validate the reliability of psychometric tests (Cronbach, 1951). The estimates for each category of questions are shown in Table 5. α estimates if random samples are correlated, and it depends on the length of the test.

TABLE 5. Cronbach's Alpha for each category of ratings.

	Expressive	Interactive	Synchronization
Cronbach's Alpha	.91	.88	.85

Mean Ratings per Category

Figure 5 shows the mean ratings for each different category of perceptual responses (expressivity, interaction, synchronization), that were grouped using a different subset of stimuli (see Table 1 for the number of stimuli per cluster). Whereas the figure shows

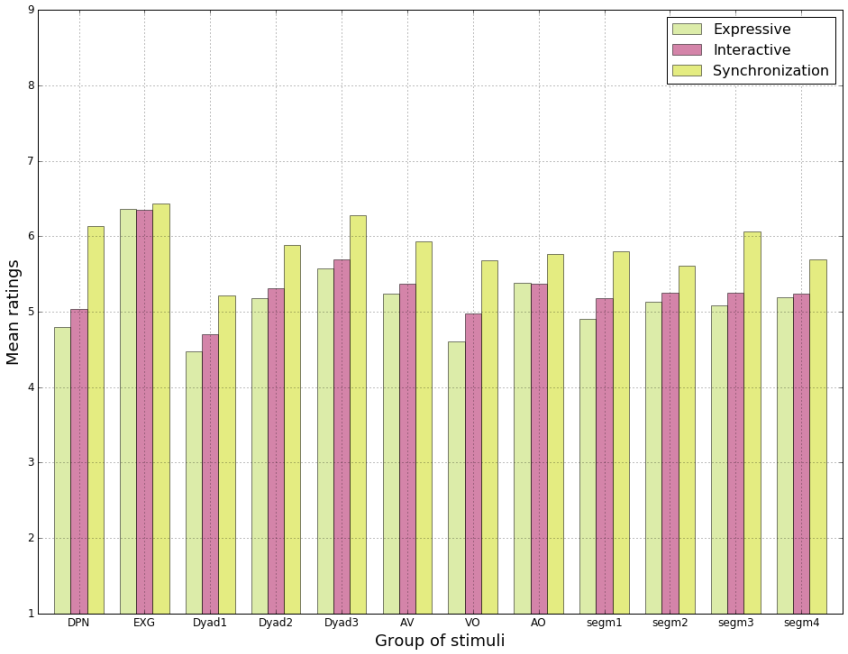

FIGURE 5. Mean ratings for all categories and clusters.

the average of the means for each participant, intuitively it is clear that the three different questions are strongly correlated to each other. The Pearson correlation coefficient between expressivity and interaction questions were $r = .96$, for expressivity and synchronization $r = .69$, and for interaction and synchronization $r = .77$. This might be due to demand characteristics, as the first question in the perceptual experiment was about expressivity, or it might be related to the fact that we did not provide explicit description for the concepts of expressivity, interactivity and synchronization.

Analysis of Variance between Groups

We performed one-way multivariate analysis of variance (MANOVA) to examine possible interactions between different modalities and intentions for ratings of expressivity, interaction and synchronization responses. For dependent variables (DV) we used the performance conditions and the modalities. For independent variables (IV) we used the ratings for expression, interaction and synchronization.

We estimated significant effects between interaction and synchronization ratings for the expressive intentions $F(1, 71) = 17.33, p < .001$, also between expression and synchronization $F(1, 71) = 13.95, p < .001$, and between expression and interaction $F(1, 71) = 10.22, p < .01$. Furthermore, using as DV the modalities and IV the ratings of expressivity, interaction and synchronization we estimated effects for the intersection of AV and AO modalities $F(1, 47) = 10.96, p < .001$, and between VO and AO $F(1, 47) = 22.95, p < .001$. No significant effect was found between AV and VO modalities. Using as DV all the modalities we estimated significant effects between expressivity and interaction $F(2, 71) = 13.07, p < .001$, and between expressivity and synchronization $F(2, 71) = 8.71, p < .001$, but no effect between interaction and synchronization. Also, significant effects were estimated between responses about expressivity and interaction. For the intersection of VO and AO we estimated $F(1, 47) = 21.36, p < .001$, for VO and AV $F(1, 47) = 5.75, p < .05$, and for AV and AO $F(1, 47) = 10.83, p < .001$. Also, significant effects were estimated between responses about expressivity and synchronization for the intersection between VO and AO $F(2, 47) = 8.71, p < .001$. No significant effect found between VO and AV, and AO and AV. No significant effect was found between interaction and synchronization for the different modalities.

4.2 Movement Analysis

Preliminary movement analysis showed that the musicians performed with greater amounts of kinetic energy in the more exaggerated expressive conditions. The mean instantaneous kinetic energy for all dyads per expressive condition was .27, .78, 1.15 Joules for deadpan, normal and exaggerated respectively. This fact was an evidence that the musicians embodied the different levels of expressivity. The mean kinetic instantaneous energy per dyad was .34, .98, and 1.03 Joules for Dyad 1, 2 & 3 respectively.

4.2.1 PCA on the Movement Timeseries

For the kinematic feature extraction, the study focused on markers that are not related to obvious sound producing gestures. Instantaneous velocity was estimated and principal component analysis was applied on the motion capture timeseries data to reduce the number of predictors that were used in multiple linear regression. Higher order derivatives over the position data, such as acceleration, failed to explained any variance in the PC loadings. As a result we focused on the velocity space, and particularly on the markers of head, root, left and right shoulder. Following to the procedure in Figure 3 we projected back to three-dimensional space the velocity data of the PCs (see Figure 6). The projections showed the amount of variation of the velocities for each marker for the first five PCs (see online video[1]). For example the greater velocities appeared on swaying as showed from the projection of PC1. While it might be a bit arbitrary to project the velocities back to the position data, we found that useful as velocities are the rate of change of the position, thus they maintain orientation across the different axis.

Table 6 shows the percent of explained variance for each PC. We put as threshold the 95% of the explained variance and we took into account the first five PCs that explained 95.7% of the variance.

TABLE 6. Percent of explained variance of the first five PCs, that explained more than 95% of variance.

	PC1	PC2	PC3	PC4	PC5
Explained variance (%)	55.6	20.8	11.3	4.6	3.3

[1] `https://youtu.be/QVaGYQasVhU`

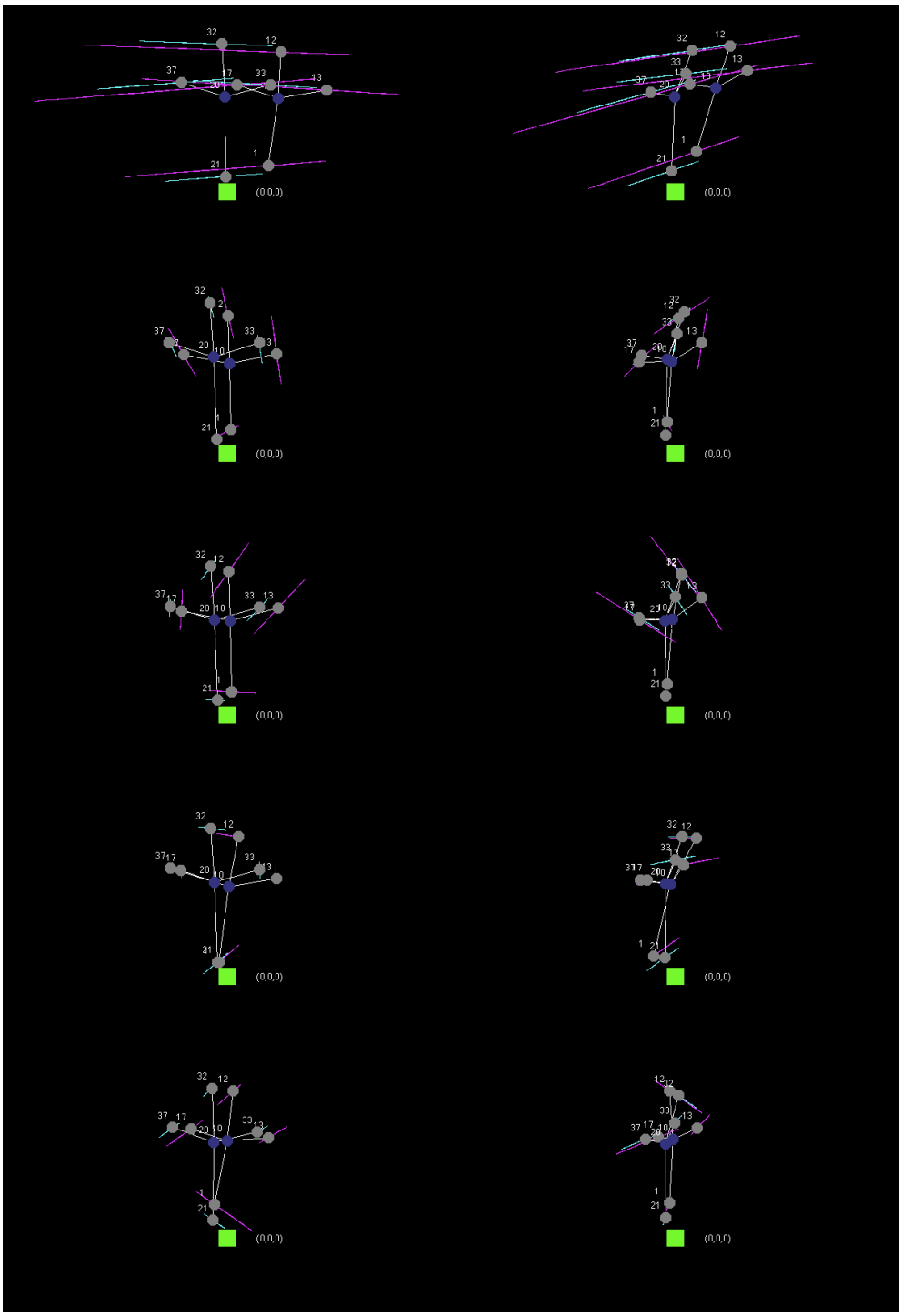

FIGURE 6. Projections of the first five PCs on the position data for both performers. The stick-figures have anterior view. The green spot is the axis origin (0,0,0) and represents the local coordinate system based on `mccenter` function. The marker names and numbers are: root (1,21), head (12,32), left shoulder (13,33), right shoulder (17,37). The first row represents the first PC (PC1) and so forth. The left column corresponds to view settings az = 0, el = 0 in MoCap Toolbox, and the right column to az = 60, el = 20 (where az: azimuth, el: elevation). Video available: `https://youtu.be/QVaGYQasVhU`

Principal Component Loadings Matrix

The PC loadings matrix is a visual representation of the contribution of the features on each PC (Alluri et al., 2012). Figure 7 shows the PC loadings matrix, based on varimax rotation. The first PC (PC1) showed that the major contribution was from mediolateral

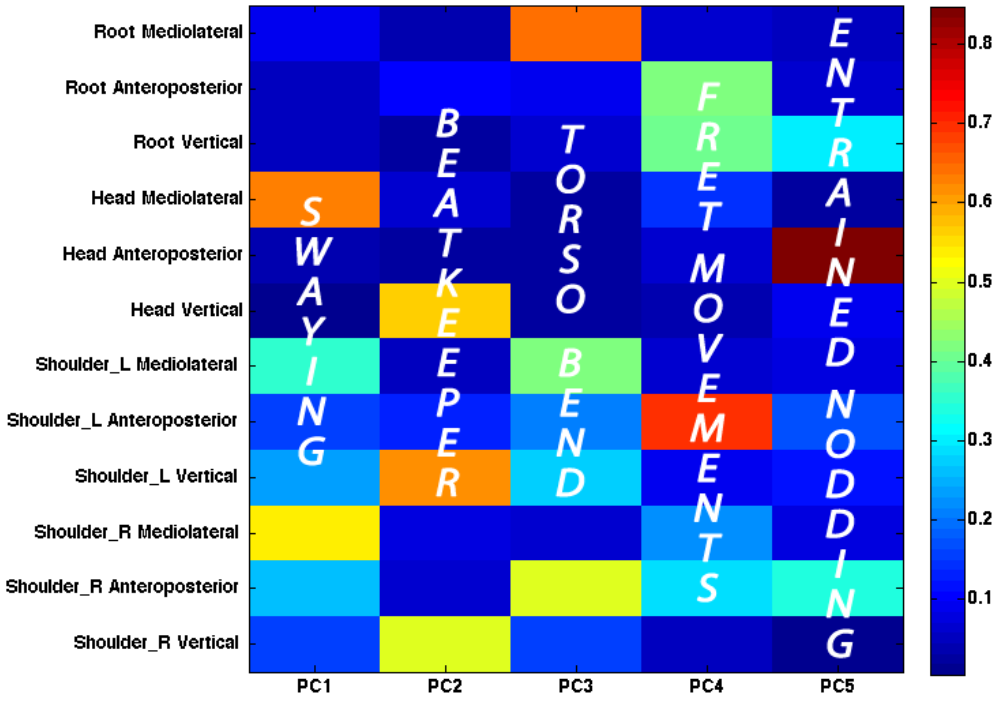

FIGURE 7. Principal component loadings matrix of the movement analysis based on varimax rotation.

movement, especially of the head and right shoulder. Our interpretation is that PC1 is the "swaying" component, and it is associated with durations on the meter level (Toiviainen et al., 2010). PC2 is related with performers' movement on the vertical axis, thus it is associated with the durations on the beat level and serve as a "beatkeeper". PC3 was an interesting component as it involves movements on different axes, that is mediolateral sway of the root marker and anterior-posterior movement of the right shoulder. The right shoulder had the biggest contribution in the production of sound, because all musicians were performing bowing gestures with their right hand. Our interpretation is that it is a *sound-facilitating bending* gesture (Jensenius et al., 2009) that support the musician to perform pousse and tire bowing movements (upbow, downbow). PC4 had major contribution from the non-bowing shoulder. Respectively it is associated with performers' movements on the neck of the violin (Dahl et al., 2009). Finally, we interpreted PC5 as

the "entrained" component, as it involved anteroposterior movement of the head. Our interpretation for the PCs 1-5 is below:

- PC1: *swaying* (mediolateral movement)
- PC2: *beatkeeper* (vertical movement)
- PC3: *torso bend* (sound-facilitating torso movements)
- PC4: *fret movements* (sound-facilitating fret movements)
- PC5: *entrained nodding* (sound-facilitating entrained gestures)

PCA based on Distance Between Markers

We examined as well the potential of the distance between markers. This approach is really effective for reducing the high dimensionality of the data. The drawback of this approach is that it looses any phase information about the gestural control. In that respect it is more meaningful feature for dance studies, where the performers are constantly changing their position. Furthermore, distance cannot be a measure of synchronization as long as there is no phase information, in that respect it remains unclear whether or not it can be used as a measure of synchrony (Glowinski, Mancini, Cowie, & Camurri, 2013).

4.3 Predictions of Perceptual Ratings

We applied multiple linear regression using the builtin function in MATLAB and k-fold crossvalidation. Our predictors were the global descriptors of kurtosis and standard deviation of the PCs. We ordered the predictors based on the PCs (ie. the two first predictors were the standard deviation and kurtosis of the PC1 and so forth).

4.3.1 Multiple Linear Regression

We applied multiple linear regression for the 72 stimuli for the expressive intentions. The regressor's scatterplot is shown in Figure 8. In order to find the optimal prediction for our regression model we added one predictor at a time. The R^2 and RMSE values are shown on the right and left panel of Figure 9 respectively. The R^2 values represent percentage from 0 to 1, and RMSE values were calculated with respect to the standardized perceptual mean ratings of the nine-point Likert scale. The maximum value of R^2, and the minimum for RMSE respectively, appear for the third predictor.

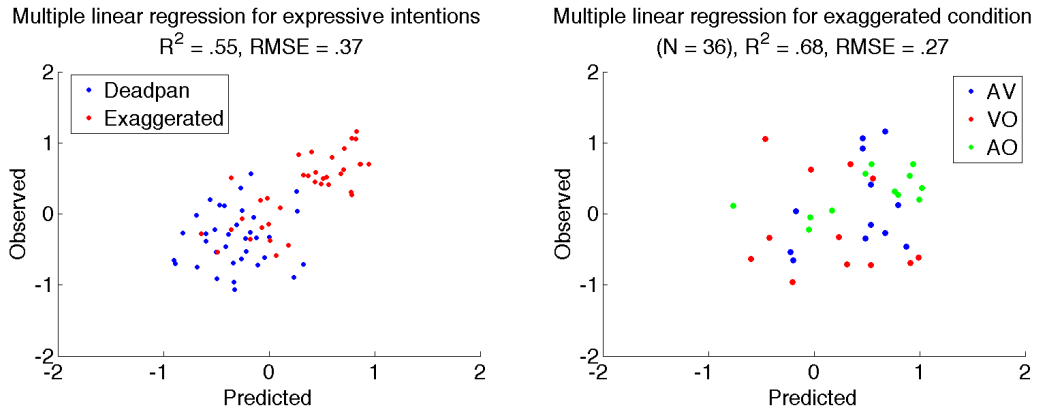

FIGURE 8. Multiple linear regression for expressive intentions. Left panel shows all perceptual segments (N=72). Right panel shows the exaggerated condition (N=36).

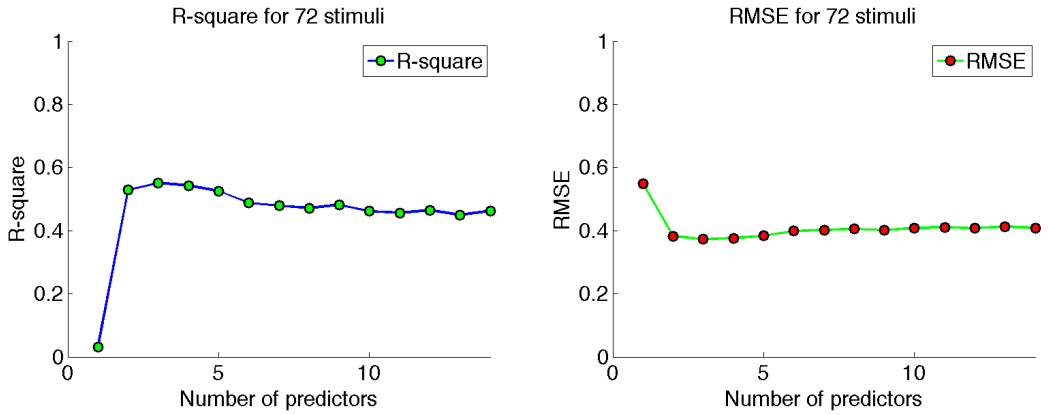

FIGURE 9. R-square and RMSE for the number of predictors.

From each motion capture timeseries, that corresponds to a perceptual stimuli, we extracted global descriptors of kurtosis and standard deviation for each PC. We did that for the first five principal components. The global descriptors were the low-level kinematic features that we used as predictors for the regression model, and as feature vectors in the LDA (see section 4.4).

Table 7 shows the R^2 and RMSE values for all different categories of stimuli. The last column of the table `cvpartition` refers to the MATLAB's function for cross-validation. Thus these columns show the number of stimuli that were used in the cross-validation model. For all computations we used 6-fold cross-validation. The Table 7 presents the results for the expressive ratings only. Each `cvpartition` with equal number of observations refers to same partition of data.

TABLE 7. R-square and RMSE values for all the groups of stimuli for the expressive ratings.

Expressive ratings	R^2	RMSE	`cvpartition`	Number of predictors
All stimuli	.55	.37	72	3
Deadpan	0	-	36	-
Exaggerated	.68	.28	36	9
Audio-visual	.62	.37	24	2
Visual-only	.83	.24	24	10
Audio-only	.38	.28	24	6
Dyad 1	0	-	24	-
Dyad 2	.71	.28	24	2
Dyad 3	.43	.43	24	2
Segment 1	.39	.39	18	2
Segment 2	.72	.30	18	2
Segment 3	.65	.31	18	2
Segment 4	.44	.47	18	4

4.4 Linear Discriminant Analysis

The feature vectors that we used for the LDA were the predictors of the regression (ie. global descriptors of standard deviation and kurtosis of the PCs for each perceptual segment). We applied linear discriminant analysis in python, using scikit-learn library and singular value decomposition as solver. The results for the expressive intentions are shown in Figure 10. On the left panel is shown the LDA for the expressive conditions for all

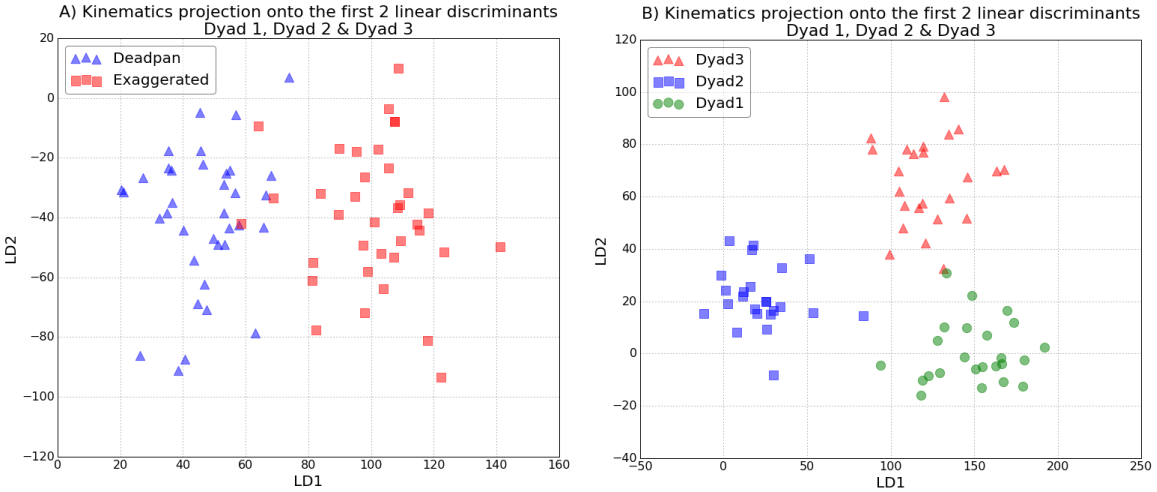

FIGURE 10. Linear discriminant analysis using the global descriptors (kurtosis, std) of the first 5 PCs.

dyads, with classification accuracy 94.44% and explained variance LD1 = 100%, LD2 = 0%, and on the right panel is shown the LDA for the three pairs of violinists, with classification accuracy 100%, and explained variance LD1 = 61.33%, LD2 = 38.67%.

5 DISCUSSION

5.1 Perceptual Experiment

Perceptual experiments are expensive. This fact applies both for financial matters and for perceivers' cognitive load. We conducted two web-based perceptual experiments, a pilot and the proper. The results for the perceptual proper showed valuable insights about the perception of expressivity, interaction and synchronization. Furthermore, the experiment was examined by Cronbach's α and the expressive responses were highly reliable.

We hypothesized that expressivity can be quantified using one dimensional scale. The fact that we did not provide any explicit interpretation for the concepts of expressivity, interaction and synchronization might be reflected in the ratings for interaction between the performers. In that respect our interpretation is that the perceivers were unclear how to rate intersubjective aspects without explicit interpretation. Furthermore, more than half of the perceivers were not musicians, and in that respect it is reasonable to assume that interaction has different meaning for a physicist and a musician. Analysis of variance identified interesting effects between the perceptual ratings, which support that a closer investigation will reveal valuable insights about the ratings for the perception of interactivity across different modalities.

5.2 Interpretation of Movement Analysis

The PC loadings matrix in Table 6 showed that swaying accounts for 56% of explained variance in violin performance. In dance studies swaying showed to be not an easy task to synchronize to music (Burger, Thompson, Luck, Saarikallio, & Toiviainen, 2014). Swaying showed to differ considerably from vertical movement, as the later responds to beat level whereas the former to the meter level (Toiviainen et al., 2010).

From Figure 9 we can see that the first three predictors of the dataset performed the highest accuracy, which is $R^2 = 55\%$. Whereas 55% accounts for approximately the 30% of the standard deviation explained (Nau, 2016), we believe that with a bigger sample of

movement data we could have performed much better predictions. The regression analysis showed that we can predict quite accurately the exaggerated expressive intention (R^2 = 68%), but we cannot predict at all the deadpan performing condition. That was probably the consequence of the fact that Dyad 1 was an outlier. The kinetic analysis showed that Dyad 1 performed with less than the one third of the kinetic energy of Dyad 3. We can acknowledged that from the performance experiment. That is reflected in every part of the analysis, though we decided to take into account all the dyads as the sample was already small. An advice for future research is that the musicians should have the same musical skill level. From the regression analysis our view was that the musicians should not be virtuosos, but skilled student musicians, as Dyad 2 was predicted with R^2 = 71%. For the VO set of stimuli we achieved the highest prediction (R^2 = 83%), which is in agreement with research in solo performance (Davidson, 1993). Furthermore, the first and last segment achieved the lowest scores, which suggests that intro and outro segments should not be used in perceptual experiments.

The movement timeseries data for the expressive conditions showed that the standard deviation was smaller for the deadpan condition across all the PCs. This measure showed that the deviation of the velocities from the mean was larger in the exaggerated condition, which sounds reasonable as the performers were performing with greater amounts of kinetic energy in the exaggerated expressive condition. On the other hand, kurtosis followed the reverse pattern for the PC2, PC3, PC4, and PC5, but there was no clear trend for PC1. This suggest that the musicians were performing using a broad spectrum of different velocities in the deadpan condition, which show that the musicians were unclear how to embody this expressive manner.

5.3 Levels of Motion Processing

The crux of the current study was to make inference about high-level kinematics, that is expressive intentions in music performance, from low-level kinematic features, such as velocity of the upper torso. Figure 11 shows the levels of motion processing in music information retrieval (Godøy & Jensenius, 2009) and it is based on the model of the levels of music processing by Toiviainen (2015). The model that we present below is incomplete, but it serves as an approach to conceptualize the different steps in computational processing for movement feature extraction. This hierarchical diagram cannot account for our visual perception, as the model by P. Toiviainen which presents the levels of processing of the auditory system.

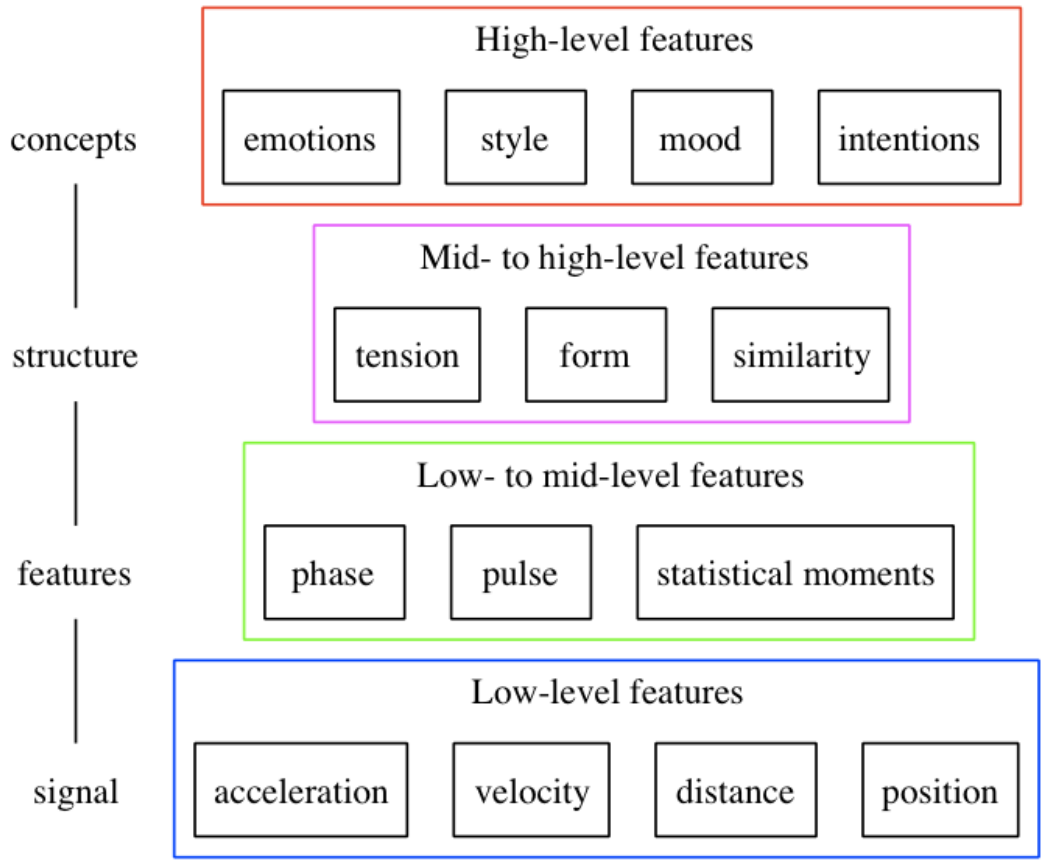

FIGURE 11. Levels of motion processing, based on the model about the "Levels of Music Processing" by P. Toiviainen.

In the present thesis we explored the lower level features of velocity, distance and position. Velocity was our best option to achieve our goal. Distance is an interesting feature, but it cannot describe efficiently fine grained gestural control in music performance. That's because it does not carry any phase information. We examined different local coordinate systems and we concluded that the most valid local coordinate system was to center the performers' position across all frames and over time. The reason for that is that our performers were situated in a certain "performance space", as such the axis origin that we extracted can describe pretty good the position of the performer. In that respect, this local coordinate system is a valid option only for music performance studies. For other types of studies that the subject is moving in space to a greater extend other options should be taken into account to standardize the motion capture timeseries data.

From the aforementioned low level kinematic features, we were able to formulate mid-level kinematic features based on the PC loadings matrix (see Figure 7). Further research is required to identify which components contribute to the first two linear discriminants. In

order to move from mid-level to high level movement features we performed the perceptual experiment to acquire ground truth knowledge, which we validated using multiple linear regression with predictors kinematic features from the upper torso.

5.4 Future Work

Further research involves the projection of the LDs on cartesian coordinates in order to identify which kinematic features had the larger contribution in the classification. Future work should focus on feature selection in order to identify which specific markers are better predictors. It is also interesting to compare feature selection with PCA to examine if they are in agreement or not.

6 CONCLUSION

The movement analysis showed that the musicians embodied the different expressive intentions with greater amounts of kinetic energy in the more exaggerated expressive conditions. Our study contributes quantitative results which describe idiosyncratic movements of expressive gestures in violin performance (Wanderley, 2001). Mediolateral swaying appeared to have major effect along with movement on the vertical axis. The former factor is associated to the meter level, whereas the later with movement on the beat level (Toiviainen et al., 2010). The regression analysis showed that the first two PCs were adequate for the regression analysis, and we predicted the mean perceptual ratings with $R^2 = 55\%$. That might be an indicator that rhythmic structures have primary importance in the perception of expressivity. Furthermore, visual-only stimuli were predicted with the highest accuracy ($R^2 = 83\%$). The kinematic feature of kurtosis showed that the musicians used a broad range of different velocities in the deadpan condition, which suggest that the musicians were unclear how to embody this expressive manner. More close investigation is required to validate interactions across different modalities and between the perception of expressivity, interaction and synchronization.

7 APPENDIX

7.1 Appendix A

TABLE 8. Labels of the markers for Player 1. The leftmost column shows the original markers labels during the recording, the rightmost shows the joints (combinations of markers) for adjusting the skeletons to the Dempster's model.

	Marker Label	Body Position	Dempster Model
1	P1Head_FL	Head front-left	Head (12)
2	P1Head_FR	Head front-right	Head (12)
3	P1Head_BL	Head back-left	Head (12)
4	P1Head_BR	Head back-right	Head (12)
5	P1Shoulder_L	Left shoulder	Left shoulder (10, 11 & 13)
6	P1Shoulder_R	Right shoulder	Right shoulder (10, 11 & 17)
7	P1C7	Chest	Chest (10)
8	P1Shoulder_BR	Asymmetrical marker	-
9	P1Elbow_L	Left elbow	Left elbow (14)
10	P1Elbow_R	Right elbow	Right elbow (18)
11	P1Wrist_L	Left wrist	Left wrist (15)
12	P1Wrist_R	Right wrist	Right wrist (19)
13	P1Finger_L	Left finger	Left finger (16)
14	P1Finger_R	Right finger	Right finger (20)
15	P1Hip_FL	Front-left hip	Left hip (1, 2 & 10)
16	P1Hip_FR	Front-right hip	Right hip (1, 6 & 10)
17	P1Hip_BL	Back-left hip	Left hip (1, 2 & 10)
18	P1Hip_BR	Back-right hip	Right hip (1, 6 & 10)
19	P1Knee_L	Left knee	Left knee (3)
20	P1Knee_R	Right knee	Right knee (7)
21	P1Ankle_L	Left ankle	Left ankle (4)
22	P1Ankle_R	Right ankle	Right ankle (8)
23	P1Heel_L	Left heel	Left ankle (4)
24	P1Heel_R	Right heel	Right ankle (8)
25	P1Toe_L	Left toe	Left sole (5)
26	P1Toe_R	Right toe	Right sole (9)
27	P1Bow_up	Bow upper end	-
28	P1Bow_down	Bow lower end	-
29	P1Vio_up	Violin head	-
30	P1Vio_down	Violin body	-
31	P1Curl	Violin body	-

7.2 Appendix B

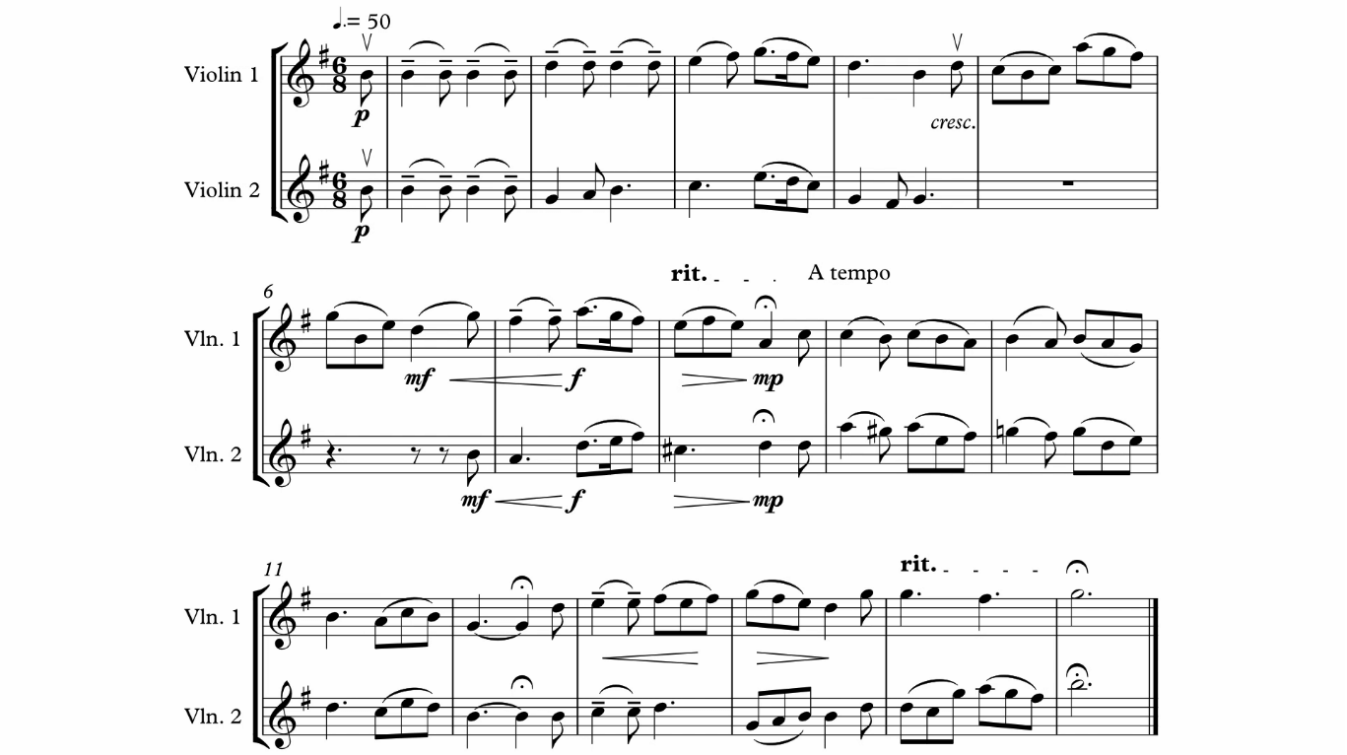

FIGURE 12. The score of the song De Kleinste composed by J. Beltjens; arrangement by Susan Johnson and Pui Yin Kwan.

References

Alluri, V., Toiviainen, P., Jääskeläinen, I. P., Glerean, E., Sams, M., & Brattico, E. (2012). Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *Neuroimage*, *59*(4), 3677–3689.

Burger, B., Saarikallio, S., Luck, G., Thompson, M. R., & Toiviainen, P. (2012). Emotions move us: Basic emotions in music influence people's movement to music. In *Proceedings of the 12th international conference on music perception and cognition and 8th triennial conference of the european society for the cognitive sciences of music* (pp. 177–182).

Burger, B., Thompson, M., Luck, G., Saarikallio, S., & Toiviainen, P. (2014). Hunting for the beat in the body: on period and phase locking in music-induced movement.

Cadoz, C., Wanderley, M. M., et al. (2000). Gesture-music. *Trends in gestural control of music*, *12*, 101.

Chalmers, A. F. (2013). *What is this thing called science?* Hackett Publishing.

Chapados, C., & Levitin, D. J. (2008). Cross-modal interactions in the experience of musical performances: Physiological correlates. *Cognition*, *108*(3), 639–651.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, *16*(3), 297–334.

Dahl, S., Bevilacqua, F., Bresin, R., Clayton, M., Leante, L., Poggi, I., et al. (2009). Gestures in performance. *Musical gestures: Sound, movement, and meaning*, *36*.

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

Daniels, P. T., & Bright, W. (1996). *The world's writing systems*. Oxford University Press on Demand.

Davidson, J. W. (1993). Visual perception of performance manner in the movements of solo musicians. *Psychology of music*, *21*(2), 103–113.

Davies, S. (2011). Emotions expressed and aroused by music: Philosophical perspectives.

Dempster, W. T., & Gaughran, G. R. (1967). Properties of body segments based on size and weight. *American journal of anatomy*, *120*(1), 33–54.

Dictionary, O. E. (2004). Oxford english dictionary online. *Mount Royal College Lib., Calgary*, *14*.

Dogantan-Dack, M. (2014). Philosophical reflections on expressive music performance. *Expressiveness in music performance: Empirical approaches across styles and cultures*, 1–21.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern classification. 2nd. *Edition. New York*.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, *7*(2), 179–188.

Gabrielsson, A. (2003). Music performance research at the millennium. *Psychology of music*, *31*(3), 221–272.

Galamian, I., & Thomas, S. (2013). *Principles of violin playing and teaching*. Courier Corporation.

Gibet, S. (2009). Sensorimotor control of sound-producing gestures. *Musical Gestures: Sound, Movement, and Meaning*, 212–237.

Gill, S. P. (2015). *Tacit engagement: Beyond interaction*. Springer.

Glowinski, D., Camurri, A., Volpe, G., Dael, N., & Scherer, K. (2008). Technique for automatic emotion recognition by body gesture analysis. In *Computer vision and pattern recognition workshops, 2008. cvprw'08. ieee computer society conference on* (pp. 1–6).

Glowinski, D., Mancini, M., Cowie, R., & Camurri, A. (2013). How action adapts to social context: the movements of musicians in solo and ensemble conditions. In *Affective computing and intelligent interaction (acii), 2013 humaine association conference on* (pp. 294–299).

Godøy, R. I., & Jensenius, A. R. (2009). Body movement in music information retrieval.

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, *27*(2), 83–85.

Jensenius, A. R., Wanderley, M. M., Godøy, R. I., & Leman, M. (2009). Musical gestures. *Musical gestures: Sound, movement, and meaning*, *12*.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & psychophysics*, *14*(2), 201–211.

Juslin, P. N. (2003). Five facets of musical expression: A psychologist's perspective on music performance. *Psychology of Music*, *31*(3), 273–302.

Juslin, P. N., & Timmers, R. (2010). Expression and communication of emotion in music performance. *Handbook of music and emotion: Theory, research, applications*, 453–489.

Leman, M. (2008). *Embodied music cognition and mediation technology*. Mit Press.

Leman, M., & Godøy, R. I. (2010). Why study musical gestures. *Musical gestures: Sound,*

movement, and meaning, 3–11.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

Luck, G. (2016, January). *Lecture notes in the x-factor in music.* University of Jyväskylä.

Luck, G., Toiviainen, P., & Thompson, M. R. (2010). Perception of expression in conductors' gestures: A continuous response study. *Music Perception: An Interdisciplinary Journal*, *28*(1), 47–57.

Maturana, H. R., & Varela, F. J. (1987). *The tree of knowledge: The biological roots of human understanding.* New Science Library/Shambhala Publications.

Nau, R. (2016). *Statistical forecasting: notes on regression and time series analysis.* Available from `http://people.duke.edu/~rnau/411home.htm` ([Online; accessed 28-May-2016])

Palmer, C. (1997). Music performance. *Annual review of psychology*, *48*(1), 115–138.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

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

Repp, B. (2005). Sensorimotor synchronization: A review of the tapping literature. *Psychonomic bulletin & review*, *12*(6), 969–992.

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

Schubert, E., & Fabian, D. (2014). A taxonomy of listeners' judgements of expressiveness in music performance. *Expressiveness in music performance: Empirical approaches across styles and cultures*, 283.

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends in cognitive sciences*, *10*(2), 70–76.

Thompson, M. (2012). *The application of motion capture to embodied music cognition research.* University of Jyväskylä.

Thompson, M., Diapoulis, G., Johnson, S., Kwan, P. Y., & Himberg, T. (2015). Effect of tempo and vision on interpersonal coordination of timing in dyadic performance. In *Proceedings of the 11th international symposium on cmmr, plymouth, uk, june 16-19, 2015.*

Thompson, M. R., & Luck, G. (2012). Exploring relationships between pianists' body movements, their expressive intentions, and structural elements of the music. *Musicae Scientiae*, *16*(1), 19–40.

Toiviainen, P. (2015, January). *Lecture notes in music perception 1.* University of Jyväskylä.

Toiviainen, P., & Burger, B. (2010). Mocap toolbox manual. *Online at: http://www.jyu.fi/music/coe/materials/mocaptoolbox/MCTmanual*.

Toiviainen, P., Luck, G., & Thompson, M. R. (2010). Embodied meter: hierarchical eigenmodes in music-induced movement. *Music Perception: An Interdisciplinary Journal*, *28*(1), 59–70.

Tomasello, M., & Carpenter, M. (2007). Shared intentionality. *Developmental science*, *10*(1), 121–125.

Vines, B. W., Krumhansl, C. L., Wanderley, M. M., & Levitin, D. J. (2006). Cross-modal interactions in the perception of musical performance. *Cognition*, *101*(1), 80–113.

Vuoskoski, J. K., Thompson, M. R., Clarke, E. F., & Spence, C. (2014). Crossmodal interactions in the perception of expressivity in musical performance. *Attention, Perception, & Psychophysics*, *76*(2), 591–604.

Vuoskoski, J. K., Thompson, M. R., Spence, C., & Clarke, E. F. (2016). Interaction of sight and sound in the perception and experience of musical performance. *Music Perception: An Interdisciplinary Journal*, *33*(4), 457–471.

Wanderley, M. M. (1999). Non-obvious performer gestures in instrumental music. In *Gesture-based communication in human-computer interaction* (pp. 37–48). Springer.

Wanderley, M. M. (2001). Quantitative analysis of non-obvious performer gestures. In *Gesture and sign language in human-computer interaction* (pp. 241–253). Springer.

Widmer, G., & Goebl, W. (2004). Computational models of expressive music performance: The state of the art. *Journal of New Music Research*, *33*(3), 203–216.

Zentner, M., & Eerola, T. (2010). Self-report measures and models. *Handbook of music and emotion*, 187–221.

List of Figures

List of Tables