# Dynamics in Music Conducting: A Computational Comparative Study Among Subjects

Álvaro Sarasúa
Escola Superior de Musica de Catalunya
Padilla 155, Barcelona, Spain
alvaro.sarasua@esmuc.cat

Enric Guaus
Escola Superior de Musica de Catalunya
Padilla 155, Barcelona, Spain
enric.guaus@esmuc.cat

## ABSTRACT

Many musical interfaces have used the musical conductor metaphor, allowing users to control the expressive aspects of a performance by imitating the gestures of conductors. In most of them, the rules to control these expressive aspects are predefined and users have to adapt to them. Other works have studied conductors' gestures in relation to the performance of the orchestra. The goal of this study is to analyze, using simple motion capture descriptors, how different subjects move when asked to conduct on top of classical music excerpts, focusing on the influence of the dynamics in the performance. Twenty-five subjects were asked to conduct on top of of three classical music fragments while listening to them and recorded with a commercial depth-sense camera. The results of different linear regression models with motion capture descriptors as explanatory variables show that, by studying how descriptors correlate to loudness differently among subjects, different tendencies can be found and exploited to design models that better adjust to their expectations.

## Keywords

expressive performance, classical music, conducting, motion capture

## 1. INTRODUCTION

The conductor-orchestra paradigm has been used in the New Interfaces for Musical Expression field for many years now. The simple idea behind it is to consider the computer as a *virtual orchestra* which the user conducts with gestures (captured by different sensors) that somehow resemble those of a real conductor.

One of the first approaches within the conductor-orchestra paradigm is the one by Max Mathews [12], where he uses radio batons to control the beat with strokes of the right hand and the dynamics with the position of the left hand. Ever since then, there have been many different approaches with the same basic principles of using gestures to control tempo and dynamics with different refinements. Some of these refinements are on the direction of improving input devices to be able to capture more subtle information about the movements. The *Digital Baton* by Martin [11], for example, measures the pressure on different parts of the handle. Similarly, Nakra's *Conductor's Jacket* [13] uses up to sixteen extra sensors to track muscular tension and respiration mapping those to different musical expressions. Other works make use of the conducting grammar (i.e. predefined gestures with a concrete meaning) to convey richer meanings from movements. A good example of this can be found in the work by Satoshi Usa [17], where Hidden Markov Models (HMM) are used to recognize gestures from the baton hand and some of their variability to allow different articulations (*staccato/legato*). Additionally, *ad hoc* refinements for specific scenarios can be found in works such as *You're The Conductor* [7], where children are allowed to control tempo and loudness using a robust (difficult to break and able to respond to erratic gestures of children) baton. The appearance of depth-sense cameras has resulted on new approaches such as Rosa-Pujazon's [15], where the user can control tempo by moving the right hand on the x-axis and the dynamics of each of the sections by first pointing at them and then raising or lowering the position of the left hand. The *Conductor Follower* [1] by Bergen is able to follow the tempo from the motion of a conductor correctly using different gestures of one hand.

On a different yet related scope, some works try to analyze the gestures of conductors and their relationship to the resulting music from a computational point of view. In a work by Luck et al.[9], cross-correlation of descriptors extracted from movement and the pulse of a performance has shown that beats tend to be synchronized with periods of maximal deceleration along the trajectory of the baton hand. The same authors analyzed the relationships between kinematics of conductors' gestures and perceived expression by presenting subjects with point-light representations of different performances and asking them to provide continuous ratings of perceived valence, activity, power and overall expression [10]. They found that gestures with high amplitude, greater variance and higher speed were those that conveyed higher levels of expressiveness. Nakra et al. [14] presented a computer-vision based method to analyze conductors' gestures. It allows to align gestures with musical features and perform correlation analysis of them, but it is not robust to lighting conditions.

In this work, we study how different subjects move when asked to "conduct" without further instructions. More concretely, we look for relationships between movement and loudness in the performance. This kind of analysis can help to design systems where users can more easily be engaged by the experience, without requiring specific instructions and thus being able to focus on the expressive aspects of their performance. This is one of the goals of the PHENICX[1] project, where we look for ways of engaging interesting experiences for new audiences in the context of classical music.

With this goal, we recorded the movement of twenty-

[1] http://phenicx.upf.edu/


*NIME'14,* June 30 – July 03, 2014, Goldsmiths, University of London, UK.

five subjects with different musical backgrounds using Microsoft's depth-sense camera, Kinect. They were asked to conduct on top of three fragments from a performance of Beethoven's *Eroica* 1st Movement by the Royal Concertgebouw Orchestra[2] for which multimodal data (including high quality audio for every section, multi-perspective video and aligned score) is available within the PHENICX project. The fact that subjects are not controlling the music is intentional and necessary to study spontaneous conducting movements without any predefined rules for control.

The rest of the paper is organized as follows: in Section 2, we describe how the recordings were done. The features extracted from them are explained in Section 3. The results of posterior analysis are presented in Section 4 and discussed in Section 5. Directions for future work are pointed in Section 6.

## 2. RECORDING PROCEDURE

### 2.1 Selection of the Excerpts

We selected 35 seconds fragments so we could have enough data while still allowing users to memorize them in a short period of time. The fragments were chosen to have some variability regarding different aspects such as dynamics, timbre or tempo (see [16] for another study dealing with beat information). All files were converted to mono so users did not also have to pay attention to spatialization. Specifically for the scope of this work, loudness values were computed using libxtract [2] and resampled to 30Hz (the rate of the motion capture data) in order to be able to compare them with descriptors extracted from motion capture.

### 2.2 Recording of Subjects

We designed the experiments in a way that subjects could become familiar with the excerpts while not making the recording too long. For each of the 35 seconds fragments we allowed subjects to listen to them twice (so they could focus on learning the music) and then asked them to conduct them three times (so they could keep learning the music while already practicing their conducting). The total time for the recording was approximately 10 minutes for all subjects. To get rid of the effects of initial synchronization, we focused the analysis on the 30 seconds from second 4 to second 34. For each of the fragments, we are using just the last take, where users are supposed to be able to anticipate changes. This means that for the analysis in this work we are considering a total of 90 seconds of conducting for each subject.

In addition, subjects had to fill out a survey about their age (avg=32.08, std. dev.=7,33), sex (5 female), handedness (5 left-handed), musical background and their feelings about the experiment (including familiarity with the piece, ability to recognize the time signature...). We removed those subjects that claimed not to be able to anticipate changes in the last take or not to use dynamics while conducting, using a final subset of 17 subjects. The musical background of the subjects before and after this filtering is summarized in Table 1. In Section 4 we comment on how the descriptors were indeed not capturing any correlation between their movement and loudness.

An application was specifically developed for the recordings. It is built on openFrameworks[3] with the ofxOpenNI[4] module, a wrapper for the OpenNI framework[5] and the proprietary middleware NiTE[6], which provides skeletal tracking. The skeleton provided by NiTE, with a sampling rate of 30Hz, has 15 joints corresponding to head, neck, torso, shoulders, elbows, hands, hips, knees and feet.

**Table 1: Summary of subjects' background. Between parentheses, number of users after removing those who claimed not to use dynamics while conducting. (nft = non formal training, <5 = less than five years of formal training, >5 = more than five years of formal training)**

| Musical background | | | | Conductor training | | |
|---|---|---|---|---|---|---|
| None | nft | <5 | >5 | None | Basic | Expert |
| 3 (2) | 4 (3) | 4 (4) | 13 (9) | 19 (12) | 5 (5) | 1 (1) |

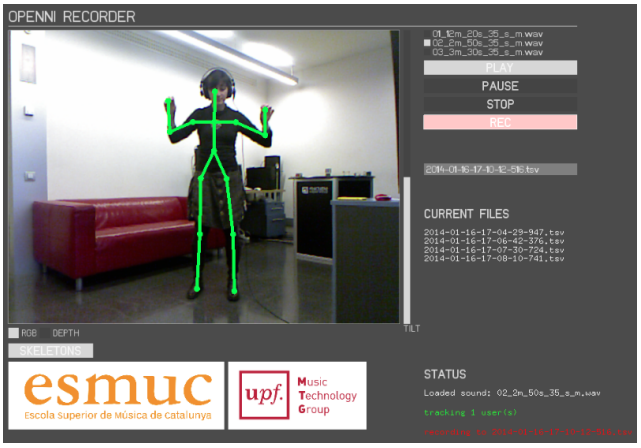


**Figure 1: Motion-capture recorder.**

The program basically consists on a GUI (see Figure 1) that allows to select different audio files that can be played, paused or stopped. When the "REC" button is played, the audio starts to play and, if any user is being tracked, the position of all the joints is stored in a `.tsv` file which in each row contains the index of the skeleton frame, the audio playback time in milliseconds, a timestamp and the position of every joint. The skeleton(s) being tracked can also be visualized on top of the RGB or RGBD images to make sure the recording is correct.

All the recordings, aligned with the audio and the motion capture descriptors explained in Section 3 are available online[7].

## 3. MOTION CAPTURE ANALYSIS

All the descriptors presented in this Section are computed frame by frame. They are classified into *joint descriptors*, computed for every joint, and *general descriptors*, describing general characteristics of the whole body movement.

### 3.1 Joint Descriptors

Similar features to the ones that proved to be relevant in [8] and [10] were extracted for every joint: position, velocity, acceleration and jerk (derivative of acceleration). For the last three, not only the components along each dimension but also the magnitude was computed. The acceleration along the trajectory was also calculated by projecting the acceleration vector on the direction of the velocity vector. The calculation was done, again similarly to how they were

[2] `http://www.concertgebouworkest.nl/`
[3] `http://www.openframeworks.cc/`
[4] `https://github.com/gameoverhack/ofxOpenNI`
[5] `http://www.openni.org/`
[6] `http://www.openni.org/files/nite/`
[7] `http://alvarosarasua.wordpress.com/research/beat-tracking/`

**Table 2: Summary of joint descriptors.**

| Name | Symbol(s) |
|---|---|
| Position | $x, y, z$ |
| Velocity | $v_x, v_y, v_z, v$ |
| Acceleration | $a_x, a_y, a_z, a$ |
| Jerk | $j_x, j_y, j_z, j$ |
| Velocity mean | $v_{mean}$ |
| Velocity standard deviation | $v_{dev}$ |
| Distance to torso | $d\_tor$ |
| Relative position to torso | $x\_tor, y\_tor, z\_tor$ |

**Table 3: Summary of general descriptors.**

| Name | Symbol(s) |
|---|---|
| Quantity of Motion | $QoM$ |
| Contraction Index | $CI$ |
| Maximum hand height | $Y_{max}$ |

computed in the cited works, by fitting a second-order polynomial to 7 subsequent points centered at each point and computing the derivative of the polynomial.

Additionally, we compute the mean and standard deviation of the velocity magnitude 1.03 seconds (31 frames, to be symmetric) around the point. They are expected to account for the "quantity" and the "regularity" of the joint movement, respectively. Also, we extract the distance to the torso ($d\_tor$). Last, we compute normalized relative positions of the joint in respect to the torso ($x\_tor, y\_tor, z\_tor$) that takes into account the height of the user and for which we have defined the equations empirically checking that it works for subjects with different heights. For the $x$ axis, points to the left (from the subject perspective) of the torso are positive and points to the right are negative. In the $y$ axis, points over the torso are positive and points below are negative. In the $z$ axis, points in front of the torso are positive and points behind are negative. For all axes, values are approximately 1 for positions where hands are completely extended in the corresponding axis. The empirically obtained equations are:

$$x\_tor_j = ((x_j - x_{torso})/h)/1.8 \quad (1)$$

$$y\_tor_j = ((y_j - y_{torso})/h)/1.8 \quad (2)$$

$$z\_tor_j = ((z_j - z_{torso})/h)/1.4 \quad (3)$$

Where $h$, the squared distance between the head and the torso, is the term that makes these features proportional to the height of the subject.

$$h = (x_{torso} - x_{head})^2 + (y_{torso} - y_{head})^2 + (z_{torso} - z_{head})^2 \quad (4)$$

## 3.2 General Descriptors

Some other features describing the general characteristics of the body movement were extracted, including Quantity of Motion ($QoM$) and Contraction Index ($CI$) (conceptually equivalent to the ones extracted from video [3, 5]). To compute $QoM$, we averaged the mean velocity values, previously explained, for all joints. For the $CI$, we looked at maximum and minimum values along every axes and empirically derived an equation to make its value approximately 1.0 when arms and legs are completely stretched out:

$$CI = \frac{-4.0 + \frac{abs(x_{max} - x_{min}) + abs(y_{max} - y_{min}) + abs(z_{max} - z_{min})}{h}}{6.0} \quad (5)$$

Also, after observing subjects in the recordings and realizing that some of them tended to raise one or both hands in loud parts, we decided to add another simple descriptor describing highest hand position with respect to the torso.

$$Y_{max} = max(y\_tor_{LHand}, y\_tor_{RHand}) \quad (6)$$

# 4. RESULTS

In order to study the relationship between the movement of the subjects and the loudness of the fragments we performed least squares linear regression, using the movement features as predictors and the loudness as the independent variable.

We created different linear models for different levels of specificity (from all subjects to subject-specific) in order to derive conclusions about how this kind of analysis could be used to create a general model for any user or to exploit user-specific characteristics to better adjust to his/her expectations.

In all cases, we started from maximal models including all descriptors which were correlated by more than 0.5 for some user and kept simplifying by removing non-significant explanatory variables until all remaining variables were significant in order to get the minimal adequate model.

The complete information of the regression models can be found in the already mentioned website[7]. In this paper, we show those results that better explain the conclusions that can be derived from them.

## 4.1 Looking for a general model

As a first step, we tried to find a model for all the subjects to understand how much they had in common. In preliminary observations subject by subject where we looked at descriptors which were correlated to loudness by more than 0.5, we found out that for some subjects there were not such descriptors. In addition, we noticed that most of these subjects where those that in the survey claimed that they did not use dynamics to guide their conducting. For these reasons, we excluded those subjects for this study.

The resulting model, summarized in Table 4 shows that, at least with the descriptors we are using, it is hard to obtain a linear model that generalizes to all subjects. Also, the average predicted values for the first fragment are shown in Figure 3.

**Table 4: Summary of the regression model for all subjects.**

| $F$ Statistic | $df$ | $R^2$ | $(R^2)_{adj}$ |
|---|---|---|---|
| 1,459.110 ($p$<0.01) | 18; 48635 | 0.351 | 0.350 |

## 4.2 Clustering subjects by styles

In the preliminary observation of features correlated to loudness by more than 0.5 for all subjects we observed an effect that is also clearly noticeable when playing back the 3D models of the motion capture recordings: while most subjects show high $cor(loudness, QoM)$, some show high $cor(loudnes, Y_{max})$ instead. This is in accordance to the observations we did during recordings, where we noticed how in loud parts most subjects did movements with higher amplitudes and some just raised their hands higher.

Having observed this difference, which is captured by the descriptors, we created linear regression models for each of

**Table 5: Regression Models for $QoM$ and $Y_{max}$ clusters. (LH = left hand, RH = right hand)**

| Coefficients | QoM cluster | | $Y_{max}$ cluster | |
|---|---|---|---|---|
| | Estimate | Pr(>\|t\|) | Estimate | Pr(>\|t\|) |
| (Intercept) | 3.633 | < 2e-16 | 4.713 | < 2e-16 |
| $a_{LH}$ | 4.499 | 0.01145 | 14.294 | 4.39e-06 |
| $a_{RH}$ | 1.835 | < 2e-16 | 12.297 | 9.01e-05 |
| $d\_tor_{LH}$ | 0.001 | < 2e-16 | 0.002 | < 2e-16 |
| $d\_tor_{RH}$ | 0.0001 | 0.00257 | 0.002 | 5.55e-10 |
| $y\_tor_{LH}$ | - | - | 1.162 | < 2e-16 |
| $y\_tor_{RH}$ | - | - | 3.289 | < 2e-16 |
| $v_{mean_{LH}}$ | 0.910 | 1.79e-06 | 0.190 | 2e-16 |
| $v_{mean_{RH}}$ | 0.771 | < 2e-16 | 0.337 | < 0.000124 |
| $v_{dev_{RH}}$ | 1.090 | < 2e-16 | - | - |
| $Y_{max}$ | - | - | 2.922 | 2e-16 |

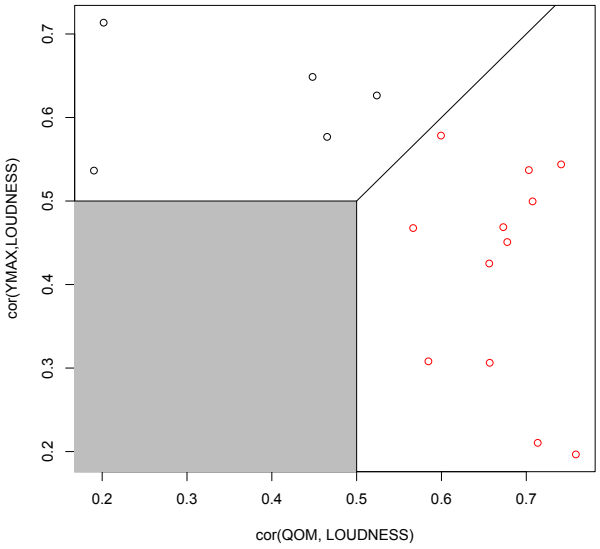


**Figure 2: Subjects clustered by correlation of loudness ($L$) to $QoM$ and $Y_{max}$. The shaded area corresponds to points where neither $cor(L, QoM)$ nor $cor(L, y_{max})$ are $> 0.5$. The line divides areas for both clusters: Red, $cor(L, QoM) > cor(L, y_{max})$; Black, $cor(L, y_{max}) > cor(L, QoM)$**

the clusters (12 subjects for the "$QoM$ cluster" and 5 for the "$Y_{max}$ cluster"). This clustering is illustrated in Figure 2. Loudness values are better explained by these two models according to the results shown in Table 6 if we consider the statistical scores of the model (both $(R^2)_{adj}$ values are over 0.4 while in the case of the global model it was 0.350). Although this $(R^2)_{adj}$ increment could be caused by the fact that the number of subjects in each cluster is reduced with respect to the case were we tried to create a model for all of them, we did not observe the same kind of improvement when different reduced groups of 5 subjects were created with subjects from both groups.

The resulting models (shown in Table 5) are in accordance with the strategy with which the clusters have been created. For the case of the "$QoM$ cluster", none of the variables relating to the position in the y axis of the hands has appeared as significant. Regarding the fact that $QoM$ does not show in any of the two models, this is not contradictory: $v_{mean}$ values for both hands are correlated to $QoM$ by definition (the latter is calculated as the mean $v_{mean}$ for all joints). When creating the model by removing variables that were not significant, what the model was actually telling us is that including $QoM$ in the model was not significantly improving the prediction. The same goes for $CI$, which is correlated to the $d\_tor$ values of both hands.

In addition to this, we also observed how subjects in the "$QoM$ cluster" did not have the same *dynamic range*, meaning that while all of them performed movements with different amplitudes for soft and loud parts, the amplitude of these movements was not the same for all of them. In order to balance this effect, we normalized the values of $QoM$ and $v_{mean}$ compressing or expanding them so all subjects had the same dynamic range (maximum and minimum $QoM$ values of one subject coincide with maximum and minimum $QoM$ values for another subject).

As expected, this supposes a clear improvement in the "$QoM$ cluster" and some improvement in the "$Y_{max}$ cluster". This makes sense given that as shown in Table 5, while descriptors related to the position of the hands did not appear as significant in the model for the "$QoM$ cluster", descriptors related to the $QoM$ did appear as significant for the "$Y_{max}$ cluster".

**Table 6: Summary of the Regression Models for different clusters. Cluster with "N" have normalized $QoM$ and $v_{mean}$.**

| Cluster | $F$ Ratio | $df$ | $R^2$ | $(R^2)_{adj}$ |
|---|---|---|---|---|
| $QoM$ | 2,278.599 ($p$<0.01) | 7; 32428 | 0.413 | 0.413 |
| $QoM$ N | 2,705.160 ($p$<0.01) | 7; 32428 | 0.455 | 0.455 |
| $Y_{max}$ | 571.727 ($p$<0.01) | 9; 13505 | 0.459 | 0.458 |
| $Y_{max}$ N | 630.012 ($p$<0.01) | 9; 13505 | 0.470 | 0.469 |

## 4.3 Subject-specific models

Finally, we created subject-specific models for each of the subjects. As expected, these models are capable of better predicting the loudness from the movement, with an average $(R^2)_{adj}$ of 0.620 (std. dev. = 0.08). Nevertheless, although this suggests that the descriptors are able to learn better for specific subjects, the clear improvement in the statistical score of the models may also be related to the fact that these models are created from fewer observations than those created for all subjects or clusters. In any case, in the same way that the descriptors we are using were capable of identifying different tendencies among subjects, it is not strange that when a model is created from a single subject it is able to predict the loudness from his/her movements more accurately.

Figure 3 illustrates the average predicted values for the

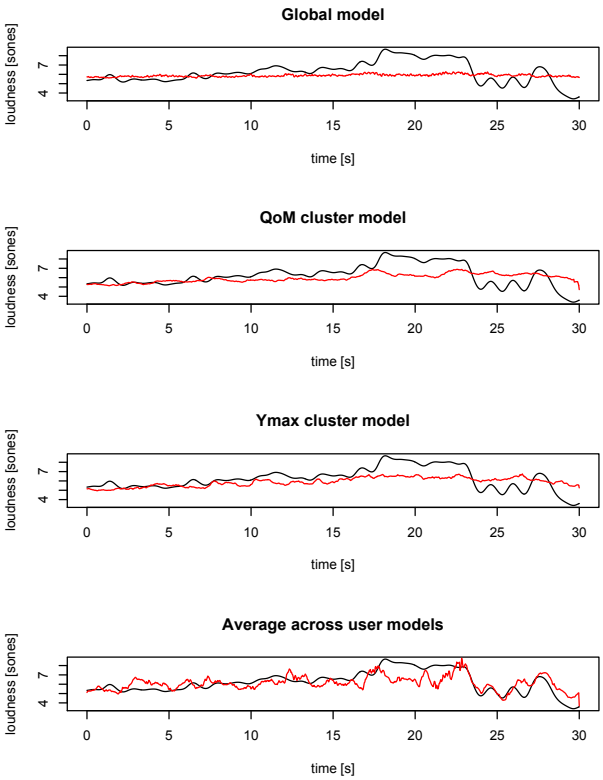


**Figure 3: Different model predictions for the first fragment. Black: loudness extracted from audio, Red: average predicted values.**

different regression models we have explained in this Section. In the first graph, we can see how the values predicted from the global model do not follow the groundtruth curve at all: it is almost a straight line. For the models extracted for different clusters, this tendency is partially corrected and we can observe some variations following the loudness curve, but still not being able to predict values in all the dynamic range. The average behavior of subject-specific models is far better in terms of being able to approximate the original curve.

## 5. DISCUSSION

The goal of this study was to analyze the movement of subjects when asked to conduct without specific instructions. We focused on the influence of the dynamics of the performance. We believe that this kind of analysis is useful to determine how these descriptors can be used in designing a system for *virtual orchestra* conducting where the user can more naturally control the expressiveness of the performance.

In this context, the results of our study bring different conclusions:

- Trying to build **a general model is not the right approach**. The observations that we did during the recordings, where we could see clear differences across subjects, was confirmed in the correlation analysis. Descriptors highly correlated to loudness were not the same for different subjects. Also, only 35% of the variability was explained in the linear regression model built for all subjects.

- A system with some predefined models could be designed and easily adjusted for a specific user considering that:

  – The **descriptors are able to find different tendencies by which users can be classified**. While in loud parts most users were performing movements with higher amplitude, some were just raising their hands higher. Conforming different groups of subjects according to these two tendencies (relying only on the differences that the descriptors were able to capture), we were able to build two models that accounted for more variability than the general model with reasonable values for models that try to explain a human behavior [4]. In terms of applying *a priori* assumptions on which group users would belong to depending on their background, the only clear tendency was that all users who had some basic notions as conductors (5) and the trained conductor belonged to the $QoM$ group.

  – **Accounting for the different dynamic ranges** of subjects can **improve the models**. When the different dynamic ranges of the amplitude of movement across subjects in the $QoM$ group was taken into account by normalizing the values of $QoM$ and $v_{mean}$, the model was significantly improved.

- A learn-by-example system could benefit from the specificities of each subject. The results suggest that simple descriptors as the ones we have used can capture the relationships between the movements of different subjects and loudness when asked to conduct.

There are, however, some issues we want to point out about the presented results.

- Subjects in this study were not controlling the performance with their movements. This was intentionally done in order to study spontaneous movement, but different effects may appear in the learning process where subjects try to adapt their movements to the response of the system.

- The group of subjects (after filtering those who did not use dynamics) in this study was unbalanced in terms of musical training (2 non-musicians, 3 non-formal musicians, 4 trained musicians and 9 expert musicians). Also, the group of users who were removed for this reason do not show a clear pattern in terms of their musical training (2 non-musicians, 1 non-formal musician, 1 trained musician and 3 expert musicians). Creating a bigger and more balanced group of users may help to find clearer influences of the musical background on the conducting style (in terms of loudness) that did not appear in our results.

- The groundtruth we used was the loudness descriptor directly extracted from audio. However, loudness is a perceptual concept and as such it cannot be detached from the user perception. Although we considered manual annotation of loudness from the subjects to actually check if their movements were correlated to *their* perception of loudness, we discarded doing so mainly for three reasons: (a) manual annotation of perceptual phenomena by using for example sliders has some intrinsic problems that would have complicated the study [6], (b) we did not want the recordings

to take more than around 15 minutes and (c) the purpose of this work was not to come up with actual models to be used by the subjects that participated but to observe if the kind of descriptors we are using are able to capture relationships between their movement and the dynamics of the performance.

## 6. FUTURE WORK

Having extracted conclusions and identified open issues from this study, there are different directions for future work in this topic.

First, having learned some possible models, new experiments in which users actually control the performance should be done in order to identify the intrinsic issues of a situation where the subject knows that his movements are controlling the performance. We can think of an experiment where different subjects can conduct by (a) using the global model, (b) using one of the two models for the two identified styles or (c) training their own model by first *impersonating* the conductor of a given performance (i.e. using the method of our study as a training step). The way in which they learn to control the system can derive new conclusions about the validity of global vs. user-specific models. In addition, different pieces should be included to assure the validity of the models for any musical work.

The other most important parameter of expressive performance is **tempo**. In another work [16], we performed beat estimations from motion capture data using the acceleration along the trajectory and the $y$ axis, in accordance with previous work by Luck[9, 8]. The results indicate that different tendencies regarding beat anticipation can also be estimated across users.

## 7. ACKNOWLEDGMENTS

We would like to thank the people who participated in this study for their valuable time and patience. Also, we thank Perfecto Herrera and Julián Urbano for their advice. This work is supported by the European Union Seventh Framework Programme FP7 / 2007-2013 through the PHENICX project (grant agreement no. 601166).

## 8. REFERENCES

[1] S. Bergen. Conductor Follower: Controlling sample-based synthesis with expressive gestural input. Master's thesis, Aalto University School of Science, 2012.

[2] J. Bullock. libxtract: a lightweight library for audio feature extraction. In *Proceedings of the 2007 International Computer Music Conference*, volume 2, pages 25–28, Copenhagen, Denmark, August 2007. ICMA.

[3] A. Camurri, M. Ricchetti, and R. Trocca. EyesWeb-toward gesture and affect recognition in dance/music interactive systems. In *IEEE International Conference on Multimedia Computing and Systems*, pages 643–648, 1999.

[4] J. Cohen and P. Cohen. *Applied multiple regression/correlation analysis for the behavioral sciences*. Lawrence Erlbaum, 1975.

[5] A. Jensenius. *Action-sound: Developing methods and tools to study music-related body movement*. PhD thesis, University of Oslo, 2007.

[6] T. Koulis, J. O. Ramsay, and D. J. Levitin. From zero to sixty: Calibrating real-time responses. *Psychometrika*, 73(2):321–339, 2008.

[7] E. Lee, T. M. Nakra, and J. Borchers. You're the conductor: a realistic interactive conducting system for children. In *Proceedings of the 2004 conference on New Interfaces for Musical Expression*, pages 68–73. National University of Singapore, 2004.

[8] G. Luck. Computational Analysis of Conductors' Temporal Gestures. *New Perspectives on Music and Gesture*, pages 159–175, 2011.

[9] G. Luck and P. Toiviainen. Ensemble musicians' synchronization with conductors' gestures: An automated feature-extraction analysis. *Music Perception*, 24(2):189–200, 2006.

[10] G. Luck, P. Toiviainen, and M. R. Thompson. Perception of Expression in Conductors' Gestures: A Continuous Response Study. *Music Perception*, 28(1):47–57, 2010.

[11] T. Martin. Possibilities for the Digital Baton as a General-Purpose Gestural Interface. In *Extended Abstracts on Human Factors in Computing Systems*, number March, pages 311–312, 1997.

[12] M. V. Mathews. The radio baton and conductor program, or: Pitch, the most important and least expressive part of music. *Computer Music Journal*, 15(4):37–46, 1991.

[13] T. M. Nakra. *Inside the Conductor's Jacket: Analysis, Interpretation and Musical Synthesis of Expressive Gesture*. PhD thesis, Massachusetts Institute of Technology, 1999.

[14] T. M. Nakra, D. Tilden, and A. Salgian. Improving upon Musical Analyses of Conducting Gestures using Computer Vision. In *Proceedings of the International Computer Music Conference, SUNY Stony Brook*, 2010.

[15] A. Rosa-Pujazon and I. Barbancho. Conducting a virtual ensemble with a kinect device. In *Proceedings of the Sound and Music Computing Conference, Stockholm, Sweden*, pages 284–291, 2013.

[16] A. Sarasua and E. Guaus. Beat tracking from conducting gestural data: a multi-subject study. *International Workshop on Movement and Computing - MOCO '14*, Accepted for publishing, 2014.

[17] S. Usa and Y. Mochida. A conducting recognition system on the model of musicians' process. *Journal of the Acoustical Society of Japan*, 4:275–287, 1998.