Beat Tracking from Conducting Gestural Data: a Multi-Subject Study

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

The musical conductor metaphor has been broadly used in the design of musical interfaces where users control the expressive aspects of the performance imitating the movements of conductors. Most of the times, there are predefined rules for the interaction to which users have to adapt. Other works have focused on studying the relation between conductors' gestures and the resulting performance of the orchestra. Here, we study how different subjects move when asked to conduct on top of classical music excerpts, with a focus on the influence of the beat of the performance. Twenty-five subjects were asked to conduct on top of three classical music fragments and recorded with a commercial depth-sense camera. We evaluated predicted beats using ground truth annotations from score-performance alignment by an expert musicologist and a modified F-measure that is able to account for different tendencies on beat anticipation across subjects. The results show that these tendencies can be used for possible improvements in the design of conducting musical interfaces in terms of user adaptation.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Interaction styles, Theory and methods; H.5.5 [**Sound and Music Computing**]: Signal analysis, synthesis, and processing

Keywords

expressive performance, classical music, conducting, motion capture, beat tracking

1. INTRODUCTION

The conductor-orchestra paradigm has commonly been used in the design of new musical interfaces for many years now. Basically, the idea behind it is to consider the computer as a *virtual orchestra*, letting the user *conduct* it by using gestures (captured by different means) that somehow resemble those of a real conductor. Most of the times, the user is allowed to control the expressive parameters (mainly tempo and dynamics) of the performance.

One of the first approaches using this metaphor is the one by Max Mathews [8], 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. Later, the same basic principles of using gestures to control tempo and dynamics have appeared in several approaches with different refinements. Sometimes, these refinements consist on improvements of the input devices that allow to capture more subtle information about the movements. The *Digital Baton* by Martin [7], for example, measures the pressure on different parts of the handle. Similarly, Nakra's *Conductor's Jacket* [10] uses up to sixteen extra sensors to track muscular tension and respiration mapping those to different musical expressions. Some other times, the improvements are achieved by using a conducting grammar (i.e. predefined gestures with a concrete meaning) to convey richer meanings from movements. Satoshi Usa's work [14] is an example of this. 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 appear in works such as *You're The Conductor* [4], 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 (popularized by Microsoft's Kinect) has resulted on new approaches such as Rosa-Pujazon's [12]. In this work, users 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.

Other works try to analyze the gestures of conductors and their relationship to the resulting music from a computational point of view. Luck et al. [5] performed cross-correlation analysis of descriptors extracted from movement and the pulse of a performance, showing that beats tend to be synchronized with periods of maximal deceleration along the trajectory of the baton hand. In another study, the same authors analyzed the relationships between kinematics of conductors' gestures and perceived expression [6]. They presented point-light representations of different conducting performances to some subjects and asked them to provide continuous ratings of perceived valence, activity, power and overall expression, finding that gestures with high am-

MOCO'14, June 16-17 2014, Paris, France.

ACM 978-1-4503-2814-2/14/06 ...$15.00.
http://dx.doi.org/10.1145/2617995.2618016.

plitude, greater variance and higher speed were those that conveyed higher levels of expressiveness. Nakra et al. [11] 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.

We carried a study designed to understand how different subjects move when asked to "conduct" without further instructions. The intention is to use this as a first step in the design of musical interfaces that use the conductor-orchestra paradigm without completely pre-established rules to control (mainly) tempo and dynamics. This is one of the goals of the PHENICX[1] project, which aims at offering ways of engaging interesting experiences for new audiences in the context of classical music. Because of this intention of just observing how different subjects intuitively move when asked to conduct, and although it has some implications that are discussed in Section 6, we recorded the movement of twenty-five subjects *conducting* different musical excerpts being aware that the resulting audio was not being modified. In this sense, subjects are somehow "impersonating" the conductor, as they are moving according to what they think they should do to conduct that performance. The fact that subjects are not controlling the music is hence intentional and necessary to study spontaneous conducting movements without the constraints of some predefined rules for control.

Here, more concretely, we study whether subjects' movements are guided by the beat of the performance. For this, we perform beat tracking from gesture data and we evaluate beat estimations with respect to ground truth annotations based on the score of the piece, manually aligned to the audio recording by an expert musicologist. We used a performance of Beethoven's 3^{rd} symphony (*Eroica*) by the Royal Concertgebouw Orchestra for which multimodal data (including high quality audio for every section, multi-perspective video and aligned score) is available within the PHENICX project. From the score, we have information such as the instruments playing every moment or the position of every note in the performance. Although this information is useful for other analyses, for this concrete work we are only using the beat positions.

The rest of the paper is organized as follows: in Section 2, we describe how we recorded subjects. In Section 3 we explain what kind of features we extract from motion capture data and the method we follow to extract beat information. The proposed evaluation metric is detailed in 4 and its results are shown in Section 5 and discussed in Section 6, where some directions for future work are also pointed out.

2. RECORDING PROCEDURE

2.1 Selection of the Excerpts and Beats Ground-Truth Annotation

We decided to select 35 seconds fragments in order to have enough data while still allowing users to memorize them in a short period of time. We chose three fragments with some variability in terms of dynamics, timbre or tempo (see [13] for a study focusing on dynamics). All audio files were converted to mono so that users did not also have to pay attention to spatialization. From the score-performance alignment, we had the position of every beat in the audio. We repositioned the beats to fit a 30Hz (Kinect's skeleton frame rate) template (e.g. a beat at 17.40s is set to 17.34s).

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

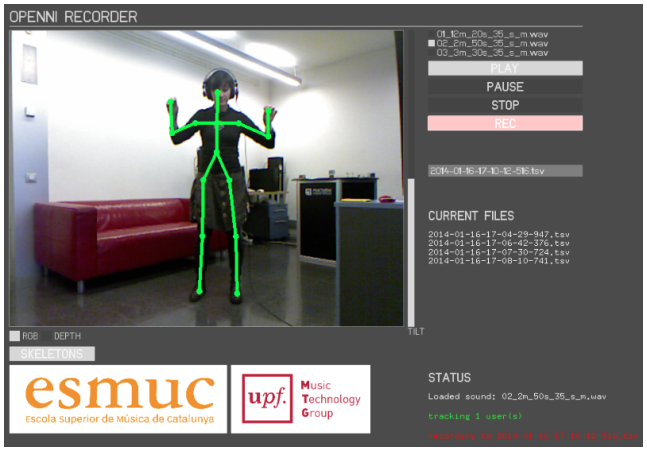

Figure 1: Motion-capture recorder.

2.2 Recording of Subjects

The recordings were designed to allow subjects to become familiar with the excerpts while not taking too long. For each of the fragments, they were asked to listen to the piece twice (so they could memorize it) and then asked to conduct it three times (so they could both practice their conducting and keep learning the fragment). In this study, we are just analyzing the last take, where subjects are supposed to know the excerpt. This makes a total of 105 seconds for each of the subjects. The total time for the recording was approximately 10 minutes for each subject.

In addition, all twenty-five 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...).

2.2.1 *Motion-capture Recorder*

We developed an application to record Kinect motion capture data synchronized with the audio excerpts. It is built on openFrameworks[2] with the ofxOpenNI[3] module, a wrapper for the OpenNI framework[4] and the proprietary middleware NiTE[5]. NiTE provides skeletal tracking with a sampling rate of 30Hz including 15 joints corresponding to head, neck, torso, shoulders, elbows, hands, hips, knees and feet.

The program consists on a GUI (see Figure 1) where different audio files can be played, paused or stopped. The "REC" button activates the motion capture recording and starts the audio. The positions of all joints of users being tracked 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 program also allows to visualize the skeleton that is being tracked on top of the RGB or RGBD images to make sure the recordings are being done correctly.

The software, together with all the recordings aligned with

[2] http://www.openframeworks.cc/
[3] https://github.com/gameoverhack/ofxOpenNI
[4] http://www.openni.org/
[5] http://www.openni.org/files/nite/

audio and motion capture descriptors, the ground-truth and the extracted beats are available for visualization and download online[6].

3. MOTION CAPTURE ANALYSIS

In order to extract meaningful information from motion capture data, we compute descriptors that relate to different aspects of the movement. Here, we focus on explaining those that are relevant for this specific study, although the computed values for all descriptors, aligned with audio and motion capture data, can be visualized and downloaded from the previously mentioned site[6].

Previous work by Luck [5], correlating motion capture data from a real conductor and resulting audio from the orchestra, showed that beats tended to be synchronized with moments of maximal deceleration along the trajectory. Although most of the subjects in our study are not real conductors, preliminary observation of the extracted features confirmed the acceleration along the trajectory as a potentially good candidate for beat extraction.

Acceleration values for all joints were calculated frame by frame by computing the second derivative of the position values. To compute the derivative of position values every frame, we fitted a second-order polynomial to 7 subsequent points centered at the frame and computed the derivative of the polynomial. Additionally, we calculated the acceleration along the trajectory by projecting the acceleration vector on the direction of the velocity vector (first derivative). However, as pointed by some previous work [1] and confirmed by preliminary experiments, the depth (z) value provided by Kinect for the hands is dependent on the hand position and thus is a source of noise. For this reason we are just considering information in the x and y axes. Also, we are just using the information from both hands, assuming those are the joints where information related to the beat can be found.

3.1 Beat Extraction

As discussed in Section 2.2 users were not given any instructions regarding how to communicate the beat or even about the need of communicating it. However, we hypothesize that most subjects will use the beats considering that when asked to "conduct" most people are aware that one of the duties of the conductor is to set the tempo. This was confirmed in the observation of the recordings and in the survey done afterwards, where only two subjects claimed not to have used rhythmic information.

3.1.1 Detecting Beats

Previous experiments by other authors with professional conductors indicated a tendency of moments of maximal deceleration along the trajectory to be synchronized with the beat. However, by looking at the 3D models of the recordings aligned with the descriptors, we observed that for most cases the information on the y axis is the most relevant in terms of beat synchrony. Changes from downward to upward motion (maximum acceleration along the y axes) appeared for subjects that just "stroke" the air at every beat and for those more trained musicians who were actually *drawing* the standard 3/4 time signature figure shown in Figure 2, in which beats indeed correspond to the changes in the y trajectory.

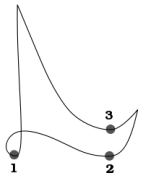

Figure 2: Standard 3/4 time signature figure.

Considering this, depending on whether we are extracting beats from the acceleration along the trajectory or the acceleration along the y axis, we look for local minima (instants of maximum deceleration) or maxima (changes from downward to upward motion), respectively. Additionally, we set a threshold of 0.001 m/s^2 (with opposite sign when looking for minima on the acceleration along the trajectory) to avoid finding points from noise in parts with no actual movement.

3.1.2 Selecting Hand and Descriptor

At this point, we have four different descriptors in which to look for the beats: acceleration along the trajectory and acceleration along the y axis for left and right hands. We denote them as L_t, L_y, R_t and R_y, with the capital letter denoting the hand (Left or Right) and the subindex indicating the descriptor (t for the acceleration along the trajectory and y for the acceleration along the y axis). Although the intention of this study is to see if the beat information is actually present in the movement of the subjects and we could just evaluate the extracted beats for all of them, we used a simple method to decide in each case the hand and descriptor in which to look for the beats. It is based on selecting the descriptor with the highest activity as the best candidate, and we use the energy of the signal as a measure of this activity computed as

$$E = \sum_n x_n^2, \tag{1}$$

where x is the discrete time series of the descriptor.

3.1.3 Accounting for Time Deviations

Another effect we want to take into consideration derives from the fact that different subjects can anticipate the beat in different ways. For example, one subject may move completely in synchrony with the beat while some other may anticipate some time (actually better reflecting what real conductors do [5]). Also, some subjects may move off-beat with respect to the audio. In order to make an estimation of such an effect, we make use of the beat annotations and study the error distribution of the beat predictions. To build the distribution of errors, we check the sizes of the annotations ($\mathbf{a}$) and predicted beats ($\mathbf{p}$) and, depending on which is greatest, for every annotated a_n or predicted p_n beat we store one error value corresponding to the difference to the closest value in the other vector. Summarizing, every e_n value in the errors vector $\mathbf{e}$ is calculated as

[6] `http://alvarosarasua.wordpress.com/research/beat-tracking/`

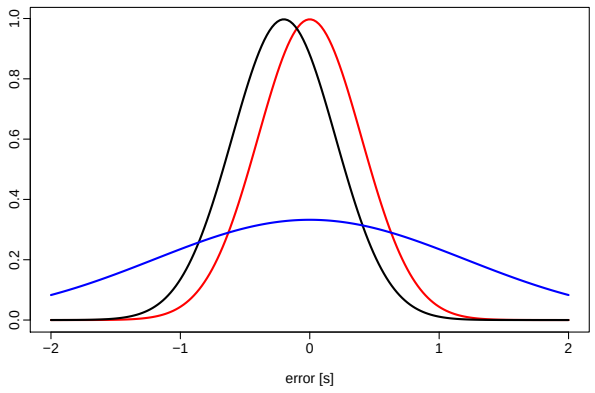

Figure 3: Expected error distributions for subjects moving in synchrony with the beat (red), anticipating the beat (black) and not following the beat (blue).

$$e_n = \begin{cases} (\mathbf{a} - p_n)_{argmin(abs(\mathbf{a}-p_n))} & \text{if } size(\mathbf{p}) \geq size(\mathbf{a}) \\ (\mathbf{p} - a_n)_{argmin(abs(\mathbf{p}-a_n))} & \text{otherwise} \end{cases} \quad (2)$$

In an ideal case where all the estimated beat positions corresponded to the annotated ones, all the values in **e** would be equal to 0. In a case where the estimated beats consistently appeared some time before the beat annotations (subject anticipating the beat), values in **e** would be below 0. In real cases, we can expect narrow error distributions centered at 0 for subjects moving in synchrony with the beat, narrow error distributions centered at a value below 0 for subjects anticipating the beat and wide distributions for subjects not following the beat (either in synchrony or anticipating). This is illustrated by the red, black and blue lines in Figure 3, respectively. We take the **mean of the error distribution** as an estimate of the time deviation (or anticipation) for each subject and fragment. The standard deviation of this distribution is indicative of how consistent this anticipation is.

4. EVALUATION METRIC

We understand this task as similar to that of an audio beat-tracker, in the sense that we are using some descriptors (in our case, from motion capture) to predict the position of the beats in a musical excerpt. For this reason, we checked the different evaluations that have been used in the Music Information Retrieval Evaluation eXchange (MIREX)[7] Audio Beat Tracking task, for which a review can be found in [2]. However, there are some issues that are specific to our problem and which have an influence on the choice of the evaluation method:

- The intention of the work is not to evaluate subjects in terms of their synchronization to the beat but just to check if the beat positions are influencing how they move.
- As explained in Section 3.1.3, different subjects can anticipate the beat in different ways.

With these considerations, the chosen evaluation method needs to (a) be tolerant to a certain deviation from the annotated beat position, taking as "correct" any beat that is close enough to the annotation and to (b) be able to estimate as equally good two subjects with a narrow error distribution regardless of how much they tend to anticipate the beat (i.e. the evaluation should be the same for subjects with distributions as the ones illustrated with the black and red lines in Figure 3).

The first requirement is fulfilled by different evaluation metrics but the second one is not completely fulfilled by any. For this reason, we defined a **time-deviation corrected version of the F-measure** [3] (which considers beats falling within a ±70 ms window around annotated beats as correct detections) that shifts the ground truth by the estimated time anticipation for each subject (see Section 3.1.3). This new F-measure $F*$, unlike the original one, would not assign an F-measure of zero to off-beat movement and is able to assign the same evaluation to subjects that anticipate the beat differently. It is conceptually equivalent to the F-measure but the ground truth is time-shifted to correct the anticipation effect. The reason why we decided to use the F-measure instead of, for example, the PScore [9], is that the *precision* (proportion of predicted beats that are correct) and *recall* (proportion of annotated beats that are correctly predicted) values that are calculated in the process are also informative about the performance. In Section 5, evaluation results for the original and modified versions of the F-measure are provided.

5. RESULTS

Table 1 contains the average results of the evaluation for all subjects. Together with the *precision*, *recall* and F-measure values for the original and time-deviation corrected versions, the mean and standard deviation values of the error distributions used for the corrections are presented. Information from the survey about their ability to recognize the time signature (marked with †) or to anticipate changes in the last recording (marked with ‡) is also displayed for each subject. In addition, the selected descriptor using the activity-detection method is displayed. In our case, the acceleration along the trajectory was never automatically selected. Additional analysis did not show any improvement when choosing it instead of the acceleration along the y axis.

In most cases (those in Table 1 where F^* is highlighted in bold), the value of the corrected version is higher than the original one. This indicates how taking into consideration the way in which users anticipate the beat can help to get a more meaningful estimate on how their movement is related to the beat. The effect is clearly noticeable in subjects such as 4 ($F = 63.67\%$, $F^* = 85.27\%$) or 9 ($F = 39.10\%$, $F^* = 71.25\%$) who are anticipating the beat 58 and 70 milliseconds respectively according to our estimation. There are some other subjects for which this improvement is not meaningful, though. Examples of this are subjects 5 ($F = 37.08\%$, $F^* = 42.66\%$) or 18 ($F = 36.36\%$, $F^* = 37.42\%$). The deviation of their error distributions is high (0.67 and 0.43 respectively), which indicates how the estimated time anticipation is not actually the result of a clear tendency.

We can observe how for all fourteen subjects who recog-

[7] `http://www.music-ir.org/mirex/wiki/MIREX_HOME`

Table 1: Evaluation results for all subjects. S = subject († indicates subjects who claimed not to have recognized the time signature, ‡ indicates users who claimed not to have been able to anticipate changes in the last take, ⋆ indicates users who claimed not to have used rhythmic information), d = descriptor, p = *precision*, r = *recall*, F = F-measure, μ = error distribution mean (in seconds), σ = error distribution std. deviation, $*$ = time-deviation corrected. $F^* > F$ are highlighted in bold.

S	d	p	r	F (%)	μ (s)	σ	p^*	r^*	F^* (%)
1	R_y	0.53	0.47	49.79	-0.016	0.21	0.58	0.51	**54.00**
2	R_y	0.32	0.22	25.93	-0.002	0.27	0.32	0.22	25.93
3	L_y	0.42	0.38	39.89	-0.032	0.16	0.58	0.53	**55.38**
4	R_y	0.64	0.63	63.67	-0.058	0.09	0.86	0.84	**85.27**
$5^{\ddagger\star}$	L_y	0.45	0.33	37.08	-0.057	0.67	0.55	0.38	**42.66**
$6^{\ddagger\star}$	R_y	0.45	0.40	42.11	-0.027	0.18	0.46	0.40	**42.71**
7	R_y	0.51	0.40	43.83	-0.041	0.21	0.68	0.57	**61.03**
8	L_y	0.56	0.50	52.89	-0.051	0.21	0.62	0.56	**58.43**
9	L_y	0.40	0.38	39.10	-0.070	0.14	0.73	0.70	**71.25**
$10^{\dagger}$	R_y	0.52	0.44	47.28	-0.014	0.20	0.49	0.42	45.10
$11^{\ddagger}$	R_y	0.47	0.34	39.77	0.004	0.28	0.47	0.34	39.77
12	R_y	0.55	0.52	53.21	-0.044	0.14	0.74	0.69	**71.28**
13	R_y	0.74	0.73	73.68	0.003	0.14	0.74	0.73	73.68
14	L_y	0.42	0.36	38.87	-0.012	0.19	0.48	0.42	**45.06**
$15^{\dagger}$	R_y	0.45	0.36	39.81	0.015	0.19	0.38	0.31	33.88
16	R_y	0.58	0.58	58.14	-0.016	0.12	0.63	0.62	**62.52**
17	R_y	0.40	0.35	37.35	-0.084	0.20	0.63	0.57	**59.72**
$18^{\ddagger}$	L_y	0.41	0.33	36.36	-0.094	0.43	0.42	0.34	**37.42**
$19^{\dagger}$	R_y	0.38	0.35	36.19	-0.054	0.21	0.47	0.44	**45.49**
20	R_y	0.52	0.50	51.20	-0.041	0.13	0.62	0.59	**60.76**
21	R_y	0.62	0.60	60.83	0.016	0.13	0.68	0.66	**66.78**
$22^{\dagger}$	R_y	0.59	0.55	56.97	-0.020	0.14	0.59	0.56	**57.48**
$23^{\dagger}$	R_y	0.65	0.64	64.43	-0.042	0.09	0.78	0.77	**77.76**
$24^{\dagger}$	R_y	0.49	0.41	44.87	-0.017	0.27	0.54	0.45	**49.17**
$25^{\dagger\ddagger}$	R_y	0.40	0.37	38.42	-0.034	0.19	0.35	0.33	34.04

nized the time signature and were able to anticipate changes in the last take (in Table 1, those without † or ‡) the F^* value is above 50% except for users 2 and 14. This is not the case for any of the subjects who were not able to anticipate changes (marked with ‡). For subjects 22 and 23, who were not able to identify the time signature, the beat tracking performance is still good.

Figure 4 shows the error distribution for subjects 4 (black), 13 (red) and 18 (blue). In these distributions we can observe how the distribution of subject 4 is not centered around 0 (it is centered around -0.058 according to Table 1) while still being narrow. For subject 13, the distribution is similar but centered very close to 0 (0.004). For subject 18, the distribution is very wide and thus is not actually informative about any specific pattern in the location of predicted beat positions. This is in accordance to the expected effects illustrated in Fig 3.

6. DISCUSSION AND FUTURE WORK

The goal of this work was to analyze how different subjects move when asked to conduct on top of classical music fragments. More concretely, we focused on analyzing the relationship of their movement to the beat. The motivation for this kind of analysis is based on the idea that it can be used in designing systems for conducting a *virtual orchestra* letting users move in a more natural way.

Considering this, the presented results allow us to define some conclusions:

- As previous work studying a professional conductor [5] indicated, acceleration extracted from motion capture data is related to the beat of the performance. However, in our experiment, the acceleration along the y axis was more informative for all subjects than the acceleration along the trajectory. For us, this is in accordance to the ways in which the subjects who were marking the beat in our experiment moved, either "hitting" the air or drawing the time signature figure.

- If there is one hand which movement is informative about the beat, a simple method as the activity-estimation based we used can detect which hand it is.

- The different ways in which users anticipate the beat need to be considered. The results show how studying the error distribution gives a good estimation of this anticipation as long as the subject is actually moving hands according to the beat. We presented a time-deviation adjusted F-measure that can be useful for this task or any other in which human subjects are actually asked to follow the beat and anticipation must not be penalized.

- From the results, it seems that some subjects are either not using the beat or not being able to convey

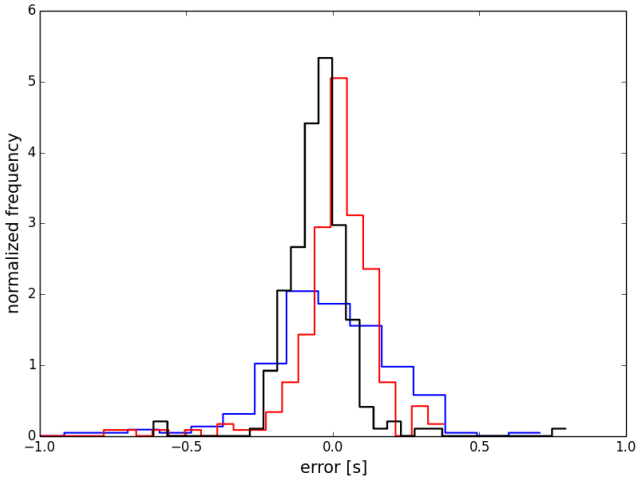

Figure 4: Error distributions of subjects 4 (black), 13 (red) and 18 (blue).

it properly. Depending on the context, this may be used to decide not to let the movement change the beat (and focus on other aspects such as dynamics or articulation) or to inform the user about this issue.

For next steps, experiments where the process of this study is used as a learning step after which users are allowed to actually conduct modifying the tempo of the performance should be done. This would allow to check if this adaptation to the user is indeed improving the way in which the resulting performance adapts to their expectations. Moreover, the learning and adaptation of subjects to a system that is actually responding to their movements may show new effects that were not present in this study. Also, here we assumed that one hand is consistently being the one that carries the beat information, but this may not always be true, especially for cases where the interaction is longer than the recordings of our experiment. Further refinements may be achieved by adapting to these or other changes along time from the user.

In addition, extending the experiments to more pieces and subjects an thus increasing the variability in tempo and musical background can help to derive more meaningful and generally applicable conclusions. Finally, provided that we have performance-score alignment for the excerpts of this work, further studies can be done by including other information such as the instrumentation and the position of notes in the predominant melody, as some users may move in synchrony with this melody instead of doing so with the beat.

7. ACKNOWLEDGMENTS

We would like to thank the people who participated in this study for their valuable time and patience. Also, we thank Agustín Martorell for the effort on the score-alignment and Perfecto Herrera and Julián Urbano for their valuable 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] M. E. Davies, N. Degara, and M. D. Plumbley. Evaluation methods for musical audio beat tracking algorithms. *Queen Mary University of London, Centre for Digital Music, Tech. Rep. C4DM-TR-09-06*, 2009.

[3] S. Dixon. Evaluation of the audio beat tracking system beatroot. *Journal of New Music Research*, 36(1):39–50, 2007.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] M. F. McKinney, D. Moelants, M. E. Davies, and A. Klapuri. Evaluation of audio beat tracking and music tempo extraction algorithms. *Journal of New Music Research*, 36(1):1–16, 2007.

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

[11] 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.

[12] 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.

[13] A. Sarasua and E. Guaus. Dynamics in music conducting: A computational comparative study among subjects. *14th International conference on New interfaces for musical expression - NIME '14*, Accepted for publishing, 2014.

[14] 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.