Machine Learning of Personal Gesture Variation in Music Conducting

Alvaro Sarasua
Escola Superior de Música de Catalunya
Universitat Pompeu Fabra
Barcelona, Spain
alvaro.sarasua@upf.edu

Baptiste Caramiaux
Goldsmiths, University of London
London, UK
b.caramiaux@gold.ac.uk

Atau Tanaka
Goldsmiths, University of London
London, UK
a.tanaka@gold.ac.uk

ABSTRACT
This note presents a system that learns expressive and idiosyncratic gesture variations for gesture-based interaction. The system is used as an interaction technique in a music conducting scenario where gesture variations drive music articulation. A simple model based on Gaussian Mixture Modeling is used to allow the user to configure the system by providing variation examples. The system performance and the influence of user musical expertise is evaluated in a user study, which shows that the model is able to learn idiosyncratic variations that allow users to control articulation, with better performance for users with musical expertise.

Author Keywords
Gesture-based interaction; Machine Learning; Music interfaces; Music Conducting; Expressive interaction

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: User Interfaces—*Interaction styles, User-centered design*; I.5.1 Pattern Recognition: Models—*Statistical*

INTRODUCTION
There are a wide range of technologies for gesture-based interaction. However, understanding non-functional characteristics of gesture, such as expressivity, still remains a challenge for computers. In this note we present a machine learning-based approach to give users expressive control through idiosyncratic variations of gesture execution, and we evaluate the approach through the specific scenario of music conducting.

The orchestra conducting metaphor is a good use case for investigating aspects of expression in gesture-based interaction. In a standard (and simplified) conducting situation, the conductor repetitively performs a gesture. Its speed of execution conveys the information of tempo to the musicians, while the dynamics convey the articulations to be applied between notes in the score. Typical articulations are: *legato*, with smooth and connected notes, or *staccato*, characterized by short and detached notes. In HCI and computer music, prior work has investigated how to recognize conductor gesture for use in interactive pedagogy or performance [11, 12]. Although the practice of conducting is highly codified and entails training, a key challenge arises from the fact that each conductor develops personal style, and communicates expressive intent (including articulation) with different nuances. Their idiosyncratic nature has made exploitation in technology-mediated interactive scenarios difficult.

We believe that this problem, situated in the wider context of expression in gesture-based interaction, would benefit from recent advances at the intersection between Machine Learning (ML) and HCI. Interactive Machine Learning (IML), as introduced by Fails and Olsen [5], puts the human in the training loop. Users can iteratively edit the examples used to train a model until its quality is acceptable [2]. Recently, Fiebrink and Caramiaux [6] brought a human-centered view of ML to HCI, reporting how learning algorithms can actually be used in a creative way by allowing users to convey concepts and intentions to the machine through examples, the approach being of particular interest in music performance or digital musical instrument building.

We draw upon the principles of a human-centered approach to Machine Learning to design an interaction technique that allows users to expressively interact with music in a conducting scenario. We show that users with differing levels of expertise in music can control musical articulation through personal, idiosyncratic variations of gesture execution, these variations being learned by a probabilistic model.

The paper is structured as follows. First, we review related work. Then, we define the proposed model and the study we undertook to test it. Finally, we present the results of the study, discuss their implications and propose directions for future work.

RELATED WORK
Prior work in computational design for expressive, gesture-based interaction has used movement analysis in order to extract suitable features to describe qualities of body movement. Movement theory, such as Laban Movement Analysis (LMA), provides a high-level representation [14] which

CHI 2016, May 7–12, 2016, San Jose, California, USA.

ACM ISBN 978-1-4503-3362-7/16/05 ...$15.00.
http://dx.doi.org/10.1145/2858036.2858328

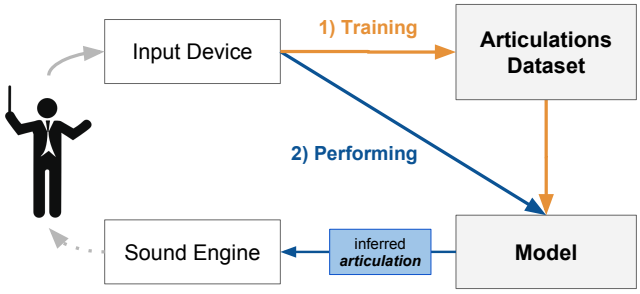

Figure 1. System diagram. During training, the user teaches the system how she embodies different articulations. During following, the model estimates the *inferred articulation* from gesture dynamic variations, driving the sound synthesis.

can be used to train gesture recognition systems for discrete control [15]. In other work, heuristics from experts in the field, such as choreographers, are used to specify movement qualities to train learning algorithms to track gestures in real time [1]. Here, particular emphasis is made on the use of dynamic features to represent movement qualities. In conducting, heuristics and characterization of expressive gestures are made based on empirical research with expert and non-expert conductors [10]. These heuristics are used in conducting systems such as *You're the conductor* [9], providing control over tempo and dynamics. Heuristics-based methods, however, are not suitable for capturing idiosyncratic gesture variations, which are particularly relevant for music articulation.

In this prior work, substantial effort is put on characterizing and describing "classes" (of either gestures, effort, or qualities) to be recognized by the machine learning algorithm. Another approach proposes the design of adaptive systems meant to be flexible enough to deliberate expressive variations of motion input [4, 16]. In both cases, the scope of adaptation is not defined by the user which may prevent considering user specific variation of gesture execution.

Recently, human-centered approaches to machine learning have gained interest in HCI for creative practice. Fiebrink's Wekinator allows creative practitioners to build music systems by iteratively providing examples [7]. Such an approach can be used to teach the system to recognize the personal gesture vocabulary of a performer [13]. Another approach allows gesture-to-sound mapping by demonstration [8] through the use of probabilistic models. These models do not take into account varying input, either for expression or personal idiosyncrasy.

CASE STUDY: ARTICULATION IN MUSIC CONDUCTING

In this section we present the interactive conducting scenario. First we present the system overview that allows users to train a music conducting system with their own gesture articulations. Then we present the modeling of these articulations.

Description of the system

The system is seen in Figure 1. First, the user teaches the system how she embodies music articulations by performing the same gesture with expressive dynamic variations. Each variation in the gesture defines a potential different musical

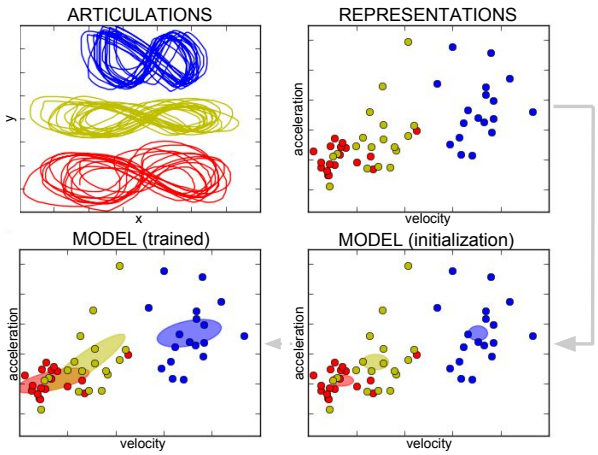

Figure 2. Learning procedure for subject 5 with mouse as input device. Input examples are represented in the velocity-acceleration feature space and associated to an articulation label. The representation feeds a GMM initialized with the means of each class and adapted using Expectation-Maximization.

articulation. Phrase articulation can range from totally *legato* to totally *staccato*. The system is however not constrained to these particular articulations, nor limited to two or three articulations. In this sense, the system is generic.

During performance, the user starts to execute a new gesture with a given dynamic. The trained model takes this gesture as input and infers which articulation the user is doing. The inferred articulation may be one of the learned articulations (from the training dataset), but it may also be a combination of learned articulations. In other words, the articulation space is not discrete, but continuous. The inferred articulation then controls the way the synthesized melody is rendered. For the aforementioned example, the sound engine creates long notes with long attack and release for *legato* and short notes with short attack and release for *staccato*.

Learning a Model of Articulations

The computational design can be formulated as a supervised learning problem: the user provides a set of data input, each one representing an articulation of the same gesture, paired with outputs encoding the articulations. This is *classification*. In the interaction scenario described above, we are specifically interested in having continuous output informing the proportionate level of each articulation within a given gesture. One solution is to interpolate between classes, achieving a form of *soft classification*.

The learning procedure is represented in Figure 2. From the gesture (a shape drawn by the user), we extract dynamic features (velocity and acceleration). These features feed a probabilistic model based on a Gaussian Mixture Model (GMM). We use GMM in a supervised mode, by providing the algorithm with the training dataset and a code for each articulation. The articulation code is an integer index, incremented for each new articulation added to the training dataset. We initialize the model by providing the means of each class (see Figure 2, bottom-right). An Expectation-Maximization algo-

rithm iteratively adapts the covariance matrices, creating a model of each articulation. At test time, incoming gestures are analyzed online. The model will then assign a continuous value to it, representing the relative distance between each articulation[1].

USER STUDY

We carried out a user study to evaluate if users' gesture articulations can be learned with the proposed model and if users can subsequently use the model to control music by varying gesture. We inspect the effect of musical expertise and the input device used to capture gestures.

Twenty participants (7 female, 13 male) aged between 22 and 38 (mean = 29.7, std = 4.2) volunteered to participate. 10 had some musical training; the others had none. We chose two input devices: a computer mouse and a Microsoft Kinect[2].

The study procedure was repeated twice, one for each input device, counterbalancing the order across participants. The participant was briefed that she would control the articulation of a melody (a excerpt from Beethoven's *Ode to Joy* from the 9^{th} Symphony) using figure-eight gestures. The experimenter plays the stimulus melody with articulations *legato*, *normal*, and *staccato* (respectively coded 1, 2 and 3). *Normal* refers to a melody synthesized with parameters in between the ones used for *legato* and *staccato*. In the Training Phase, the participant is asked to draw a figure eight following the beat while the melody is played for each different articulation during 8 bars (16 figure eight gestures). The participant is encouraged to perform these gestures with the variations she feels best match the articulations. She can rehearse until she feels confident and then records the training examples (one gesture variation for each articulation).

In the Task Phase, the participant is presented with one of the melody versions used for training, the articulation of the version being the target articulation. After listening to it, she is asked to start drawing a figure eight in order to control the melody articulation such as to reach the articulation target, *as close to the example as possible* until the melody ends. Two bars with a metronome are played before the melody starts. This process is repeated twice for each of the 3 target articulations appearing in random order. As visual feedback, a screen shows the trace of the gesture. During performance, a slider shows the fixed target articulation value together with the inferred one. The hand position and the estimated articulations by the model are recorded frame by frame for analysis.

At the end of the study, participants were asked to rate different aspects of the task on a scale from 1 (total disagreement) to 7 (total agreement).

RESULTS

Analysis of articulation performance

The questionnaire revealed that participants, in general, felt that they had fulfilled the task according to their answer to "Do you think you managed to fullfill the tasks asked during the study?" (μ=5.3; β=0.9). They also replied positively to "When you were controlling the music, was the audiovisual feedback of your movement variations what you were expecting?" (μ=5.2; β=0.8). There was no significant difference between musicians and non-musicians in response to these questions.

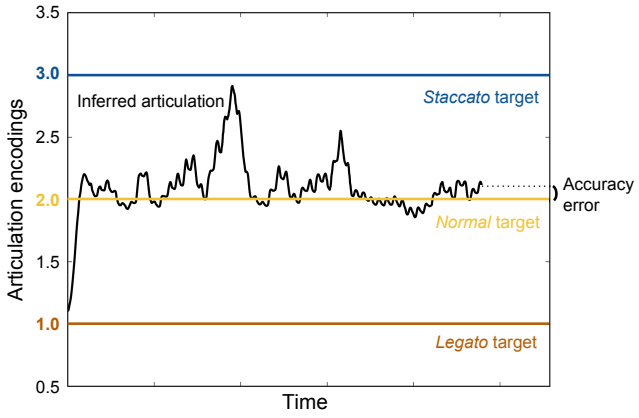

Figure 3. Inferred articulation (black curve) for participant 19, target 2. Colored straight lines represent the three possible target articulations in the study. In this case, the target is articulation 2 or *normal* (yellow line).

The accuracy of the estimated articulation compared to the intended one is assessed computing the mean error between the running articulation estimation (along time) and the given target, during the Task Phase. Figure 3 reports an example of estimated articulation for participant 19, target 2. The mean error is the cumulative difference between the black curve and the yellow straight line. We averaged the mean error across participants, devices and target articulations. The resulting global error is $\epsilon = 0.31$ ($\sigma = 0.21$). To compare with subjective measures, we computed the correlation coefficient between each participant's rating of their perception of task fulfillment and the mean accuracy values over all of that participant's performances. We found that subjective ratings and objective measure are correlated with a coefficient of 0.6.

We then inspect how the accuracy given by the mean error is affected by three factors: the TARGET articulation, the participants musical EXPERTISE and the DEVICE used for the task. A repeated-measure analysis of variance (ANOVA) shows that there is a significant effect of TARGET ($F(2, 108) = 3.992$, $p < 0.05$) and EXPERTISE ($F(1, 108) = 7.264$, $p < 0.01$), while there is no effect of DEVICE. A Tukey's HSD (Honestly Significant Difference) post-hoc analysis shows that the accuracy is higher for TARGET 2 and 3 compared to TARGET 1 ($p < 0.05$), while there is no significant difference between TARGET 2 and 3. For EXPERTISE, the analysis shows that the accuracy is significantly better ($p < 0.01$) for musicians ($\epsilon_m = 0.26$, $\sigma = 0.19$) than for non-musicians ($\epsilon_{nm} = 0.36$, $\sigma = 0.24$).

Analysis of model training

From the Training Phase, we examine the quality of training by computing the separability between articulations in the

[1]Technical details are beyond the scope of this note. The interested reader in the use of GMM for soft classification and regression is encouraged to read [3]

[2]https://dev.windows.com/en-us/kinect

representation space (Figure 2, top-right). The separability measure is defined as the distance ratio between the data belonging to different classes (articulations) to the variance of data within each class[3]. A low separability means that articulations are ambiguous from the model perspective. ANOVA reveals that the EXPERTISE does not affect separability, while the DEVICE does ($F(1, 36) = 5.911$, $p < 0.05$). Also, we found that articulation separability is not correlated to model accuracy (correlation coefficient is 0.12).

We finally examine an important aspect of the gestures considered in the study: the idiosyncrasy of articulations performed by users. For that, we perform cross-validation on the training data: for each participant i, we train the model with the articulations from that participant and test with training data from the other participants $j = 1..20, j \neq i$. From these tests, we compute the average error between the estimated articulation value given by the model and the expected articulation. We found that the global error is $\epsilon_{idiosyn} = 0.80$. We then computed an "individual error" by training the model and testing with the training data from the same participant i, for each participant. The global individual error is $\epsilon_{indiv} = 0.46$. A statistical test (t-test) shows that the two errors $\epsilon_{idiosyn}$ and ϵ_{indiv} are significantly different ($p < 0.001$).

DISCUSSION AND LIMITATIONS

According to the questionnaire results, the model succeeds at providing control over articulation from the user perspective. Also, objective measures provide acceptable errors for both kind of users (although significantly better for musicians). Importantly, the model managed to learn intended articulations even if the way participants performed the articulations to train the system embedded idiosyncratic elements that were not shared across participants. Indeed, while we imposed a particular base gesture (figure-eight) and tempo, users were not told how to vary their gesture to achieve the different articulations. Instead, variations in execution were free, but asked that they be coherent with the sound stimuli. As a result, a model learned on a user's set of data may not be transferable to another user, but embeds a given user's own expressive gesture qualities.

We saw that musicians performed significantly better than non musicians. Although training quality was similar for participants of different musical expertise, a musician's knowledge allowed them to better understand the task from a musical perspective. We think that their musical ability allows them to concentrate on dynamic variations and to better interpret the synthesized sonic differences representing *staccato* and *legato* articulations. This is supported by the fact that most non-musicians reported that they focused exclusively on visual feedback during performance.

The results also showed that participants were able to control the system through the models of articulations previously learned. Interestingly, the individual error of the model, obtained when training and testing offline on a participant's data from the Training Phase (so considering the same data for the training and the testing) is higher than the average accuracy error of the model obtained from the Task Phase (where participants trained the system and performed through it online). We believe that online task execution with audiovisual feedback involves an action-perception loop which helps users adapt their gesture to achieve the task. The results from the questionnaire reveal that the audiovisual feedback was consistent with participants' expectations, indicating that the audiovisual feedback was a reinforcing confirmation to the user on her actions. This, we believe, also shows that participants did not deliberately adapt their performance with unnatural gesture variations. Such aspects of sensorimotor learning that may enter into play constitute an important direction for future research.

Interactive conducting of this kind has potential application in music pedagogy, exhibitions (e.g. public installations) and new music performance. Creative applications beyond music that could leverage this approach to expressive, gesture-based interaction include dance, illustration, and gaming. Our study brings insights on the applicability of such a model to other scenarios. We showed that the input device only plays a minor role in the interaction (at least for position-based input device such as mouse and Kinect) making the use of such an approach viable with other types of input devices.

Finally, we would like to report on limitations of the current work that could be addressed in future research. The proposed scheme considers a single gesture at a fixed tempo. In order to overcome this limitation, we believe that incorporating a temporal model of gesture could extend the current model of articulation. In our other recent work in machine learning for gesture-based interaction design, we have proposed a method for realtime gesture recognition with tracking of variation based on dynamical systems [4]. A combination of both models could afford the user the possibility to train the system to recognize different gestures and a set of potential variations which could then be dynamically explored, in performance, by the user.

CONCLUSION

We have presented a machine learning-based approach for expressive gesture-based interaction, and used the music conducting metaphor as an example to investigate technology-mediated control of music through expressive gesture articulations. We showed that the proposed probabilistic model based on GMM is able to learn participants' intended and personal idiosyncratic articulations and that these participants are able to control an interactive music application through the trained models of articulations.

ACKNOWLEDGMENTS

We would like to thank all the people who participated in the study. This research received funding from the European Union Seventh Framework Programme FP7 / 2007-2013 PHENICX project (grant agreement no. 601166), MusicBricks, Grant Agreement Nr. 644871, and the European Research Council MetaGesture Music project (ERC grant agreement no. FP7-28377).

[3]The separability is a common criterion in machine learning, implemented in well-know classification techniques such as Linear Discriminant Analysis (LDA).

REFERENCES

1. Sarah Fdili Alaoui, Baptiste Caramiaux, Marcos Serrano, and Frédéric Bevilacqua. 2012. Movement qualities as interaction modality. In *ACM DIS*. 761–769.
2. Saleema Amershi, Maya Cakmak, W Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120.
3. Sylvain Calinon, Florent Guenter, and Aude Billard. 2007. On learning, representing, and generalizing a task in a humanoid robot. *IEEE Transactions on Systems, Man, and Cybernetics, Part B* 37, 2 (2007), 286–298.
4. Baptiste Caramiaux, Nicola Montecchio, Atau Tanaka, and Frédéric Bevilacqua. 2014. Adaptive gesture recognition with variation estimation for interactive systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 4, 4 (2014), 18–51.
5. Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *ACM IUI*. 39–45.
6. Rebecca Fiebrink and Baptiste Caramiaux. (In press, 2016). The machine learning algorithm as creative musical tool. In *Oxford Handbook of Algorithmic Music*, Roger Dean and Alex McLean (Eds.). Oxford University Press.
7. Rebecca Fiebrink, Perry R Cook, and Dan Trueman. 2011. Human model evaluation in interactive supervised learning. In *ACM CHI*. 147–156.
8. Jules Françoise. 2015. *Motion-Sound Mapping by Demonstration*. Ph.D. Dissertation. University Pierre and Marie Curie (Paris 6).
9. Eric Lee, Teresa Marrin Nakra, and Jan Borchers. 2004. You're the conductor: a realistic interactive conducting system for children. In *NIME*. 68–73.
10. Eric Lee, Marius Wolf, and Jan Borchers. 2005. Improving orchestral conducting systems in public spaces: examining the temporal characteristics and conceptual models of conducting gestures. In *ACM CHI*. 731–740.
11. Michael Lee, Guy Garnett, and David Wessel. 1992. An adaptive conductor follower. In *ICMC*. 454–454.
12. Declan Murphy, Tue Haste Andersen, and Kristoffer Jensen. 2004. Conducting audio files via computer vision. In *Gesture-based communication in human-computer interaction*. Springer, 529–540.
13. Margaret Schedel and Rebecca Fiebrink. 2011. A demonstration of bow articulation recognition with Wekinator and K-Bow. In *Proceedings of ICMC*.
14. Diego Silang Maranan, Sarah Fdili Alaoui, Thecla Schiphorst, Philippe Pasquier, Pattarawut Subyen, and Lyn Bartram. 2014. Designing for movement: Evaluating computational models using LMA effort qualities. In *ACM CHI*. 991–1000.
15. Dilip Swaminathan, Harvey Thornburg, Jessica Mumford, Stjepan Rajko, Jodi James, Todd Ingalls, Ellen Campana, Gang Qian, Pavithra Sampath, and Bo Peng. 2009. A dynamic bayesian approach to computational laban shape quality analysis. *Advances in Human-Computer Interaction* 2009 (2009), 2.
16. Andrew D Wilson and Aaron F Bobick. 1999. Parametric hidden markov models for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21, 9 (1999), 884–900.