Datasets for the Analysis of Expressive Musical Gestures

Alvaro Sarasúa, Baptiste Caramiaux, Atau Tanaka, Miguel Ortiz

Datasets for the Analysis of Expressive Musical Gestures

Álvaro Sarasúa
Music Technology Group, Universitat Pompeu Fabra
Barcelona, Spain
alvaro.sarasua@upf.edu

Baptiste Caramiaux
McGill University, Montreal, Canada
UMR STMS Ircam-CNRS-UPMC, Paris, France
baptiste.caramiaux@ircam.fr

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

Miguel Ortiz
Queen's University Belfast
Belfast, United Kingdom
m.ortiz@qub.ac.uk

ABSTRACT

In this paper we present two datasets of instrumental gestures performed with expressive variations: five violinists performing standard pedagogical phrases with variation in dynamics and tempo; and two pianists performing a repertoire piece with variations in tempo, dynamics and articulation. We show the utility of these datasets by highlighting the different movement qualities embedded in both datasets. In addition, for the violin dataset, we report on gesture recognition tests using two state-of-the-art realtime gesture recognizers. We believe that these resources create opportunities for further research on the understanding of complex human movements through computational methods.

CCS CONCEPTS

•**Information systems** → **Multimedia databases;** •**Applied computing** → *Sound and music computing;*

KEYWORDS

Database, Motion capture, EMG, myo, Machine learning, Gesture recognition, Hidden Markov Models, Particle Filtering

ACM Reference format:
Álvaro Sarasúa, Baptiste Caramiaux, Atau Tanaka, and Miguel Ortiz. 2017. Datasets for the Analysis of Expressive Musical Gestures . In *Proceedings of 4th International Conference on Movement Computing, London, UK, June 2017 (MOCO'17),* 4 pages.
DOI: http://dx.doi.org/10.1145/3077981.3078032

1 INTRODUCTION

Music performance involves rich body movement that have been studied in music research [9]. However, understanding musical motion remains a challenge for the machine because of complex temporal and spatial variations in their execution. Tackling this challenge requires techniques that are able to capture such variations, as well as datasets upon which these techniques can be evaluated. In this paper we provide two such datasets.

MOCO'17, London, UK

DOI: http://dx.doi.org/10.1145/3077981.3078032

Succesful methods in machine learning have often relied on well-designed datasets that can be used for benchmarking. One famous example is the MNIST dataset[1] which is comprised of binary images of digits. In movement research, there has been less consensus about a unique dataset that can be used for benchmarking. The CMU MoCap database is perhaps the best known online motion library[2] and has been used in a wide variety of research projects. However, none of them deals with expressive movements of the sort involved in musical performance. While systems like RepoVizz [4] allow the sharing of multimodal musical repositories and others like Mova [1] allow the analysis and visualization of movement features, motion capture studies of musical performance have not, to our knowledge, made datasets available beyond the original studies in which they were used.

Gestural expressivity is linked to the notion of variation in body movement execution. For instance, in human-human communication we usually differentiate between the information content (what is communicated) and the expressive information (how it is communicated) [6]. Similarly, in music we can differentiate between what gesture is performed, which is linked to the musical task, and how the gesture is performed, which is linked to the musical expression [5]. The ways in which a gesture recognition system can be robust against, or sensitive to, these variations depends on the task at hand and the classification/adaptation algorithm used.

The contribution of this paper is twofold. First we present two datasets of musical gestures with expressive variations that have been built using a similar experimental procedure. Second we illustrate the potential of these datasets by highlighting intrinsic data variations and by testing state-of-the-art classification techniques. Ultimately, our goal is to advocate for complementary research tackling the problem of motion computing under conditions of expressive variation. We believe that the datasets we provide are useful resources in pursuing such endeavors.

The article is organised as follows. We first describe the datasets: number of participants, material, procedure and equipment. Then we illustrate the data variations embedded in the datasets. In Sections 3 and 4, we test state-of-the-art gesture recognizers against one of these datasets. Finally we discuss the results and propose future research directions relevant to the motion computing community for which we think the provided datasets are useful.

[1] http://yann.lecun.com/exdb/mnist/
[2] http://mocap.cs.cmu.edu/

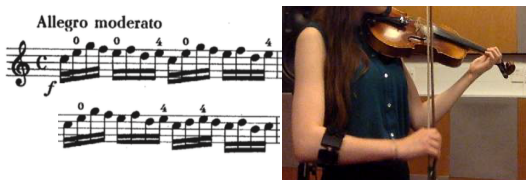

Figure 1: The phrase (L) from the Kreutzer Etude. Placement (R) of the sensor on the bowing arm of the subject

2 DESCRIPTION OF THE DATASETS

We present two datasets that have been built on the same purpose: gathering real-world musical gestures with explicit expressive variations. The first dataset is comprised of violin gestures while the second is comprised of gestures from piano performance. Participants recruited were all experts. Both datasets are available online[3].

2.1 Violin gesture dataset

2.1.1 Participants and materials. We recruited 9 experienced violinists (3 male, 6 female, aged between 17 and 43) and asked them to play an excerpt from standard pedagogical repertoire: Kreutzer Etude No. 2 in C major (see Figure 1, left). All subjects had formal classical music training, from 6 to 36 years of study.

2.1.2 Procedure. Each subject played the excerpt 10 times using each of the following 5 bowing techniques: *détaché*, *legato*, *spiccato*, *staccato*, *martelé*. This set of bowing techniques has been chosen based on previous work [7, 10] and discussion with professional violinists during a pilot study.

After the 50 trials performed, each violinist was asked to play the excerpt with 3 bowing techniques (*détaché*, *legato* and *spiccato*), 10 times each, this time varying the dynamic from *pianissimo* to *fortissimo* (very soft to very loud). Finally, each violinist played the excerpt with the same 3 bowing techniques 10 times each, but now varying the tempo from slow to fast. In the following, the union of the first 50 trials across the 9 participants is called the *original dataset*; the union of the 30 dynamics trials the *piano-forte dataset*; and the union of the last 30 tempo trials the *slow-fast dataset*.

2.1.3 Equipment. We captured the violinists' gestures with the Myo consumer device to acquire 8 channels of electromyogram (EMG), as well as 3-channel accelerometer and 3-channel gyroscope from its inertial measurement unit (IMU). We maintained consistent sensor positioning for all participants (Figure 1 right).

2.2 Piano gesture dataset

2.2.1 Participants and materials. We recruited 2 professional pianists (both female) and asked them to play an excerpt from Schumann's *Träumerei* (Kinderszenen Op.15 No.7) with different variations in speed execution and expressive intention. This piece has been previously used in research to investigate expressive aspects of piano music performance [8].

2.2.2 Procedure. Each subject played the excerpt at 3 different tempi, with and without a metronome: normal (70 beats per minute), slow (40 bpm) and fast (120 bpm). In the no-metronome condition, they also played with *rubato* (continuous expressive tempo alteration). In each of the conditions for metronome and speed, they played with 5 expressive intentions: normal, still (trying to move as little as possible), exaggerated, finger *legato* (melodic consecutive notes smoothly connected) and *staccato* (detached consecutive notes). 3 takes were recorded in each of the conditions, making a total of 105 takes per pianist.

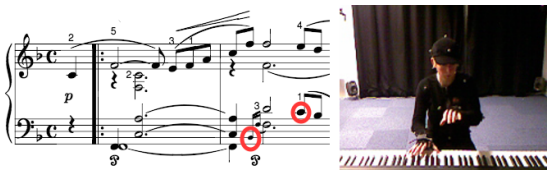

Figure 2: Left hand phrase from *Träumerei* (L), between the circled notes. Still image from video (R) showing normal tempo, exaggerated expression.

2.2.3 Equipment. Recordings were made in a room equipped with an OptiTrack Motion Capture system with which we captured the position and orientation of 22 body limbs[4] at 100 Hz. The pianists played an 88-key electronic piano (with weighted action) from which we recorded audio at 44100 Hz and MIDI data. Video was recorded using a Microsoft Kinect at 30 fps (a still image from this video is shown in Figure 2, right). A MaxMSP patch was developed to record all 4 modalities (motion capture, video, audio and MIDI) aligned into separate files.

3 VARIATION ANALYSIS

We analyzed the variations embedded in the two datasets. For the violin dataset we expect variation in the EMG signal reflecting the muscle inflections required to perform the different articulations. In the piano dataset, we expect spatio-temporal variation in the motion capture position data reflecting the different phrasings.

3.1 Violin dataset

We analyze the variation in the EMG data by computing the average IMU and EMG amplitude across trials and participants for each dataset. Figure 3 reports the statistics. A one-way ANOVA shows that there is no significant difference between IMU amplitudes ($F(1, 108) = 1.10$), $p > 0.05$) but there is a significant difference between EMG amplitudes ($F(1, 108) = 101.53$), $p < 0.001$). More precisely, EMG amplitude for trials played in *original* condition is significantly lower than the EMG amplitude in either the *piano-forte* or the *slow-fast* conditions. Interestingly there is no significant difference in EMG amplitude between the two last conditions, meaning that playing with increasing speed or with increased bow pressure both involves more muscle groups, and consequently more tension.

3.2 Piano dataset

The piano dataset includes position/orientation of multiple joints, as well as aligned audio, video and MIDI, resulting in a complex analysis task that is beyond the scope of this paper. Instead we focus on a specific example that illustrates the potential of the dataset.

[3] http://gitlab.doc.gold.ac.uk/expressive-musical-gestures/dataset

[4] Due to occlusions caused by the piano, the position of body limbs below the abdomen was unstable

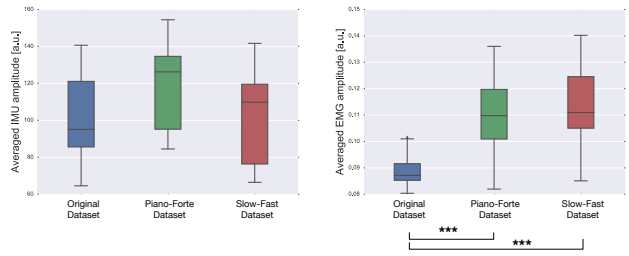

Figure 3: IMU (L), EMG (R) average amplitude across trials

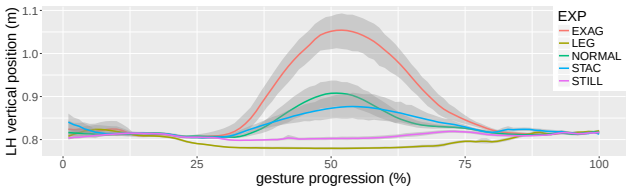

Figure 4: Vertical position of pianist LH across execution of the phrase at different expressive intentions with means across trials (color) and standard deviation (shade).

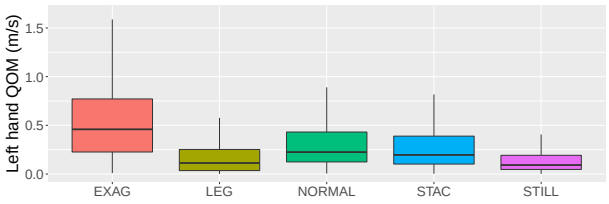

Figure 5: *QoM* of the LH in the analyzed excerpt at each variation task: exaggerated, *legato*, normal, *staccato*, and still.

Here we show, for one of the pianists, how the movement of the left hand varies for different articulation in a single isolated phrase (a leap of a major 10th with grace notes leading to the upbeat of measure 2). Figure 2, left, shows the score of this phrase. We trimmed the mocap data aligning on MIDI note messages. Figure 4 shows the vertical position of the left hand for this phrase for the 5 different expressive variations. It illustrates how the pianist swept the hand away from the keyboard in the *exaggerated* variation (as shown in Figure 2, right) while making a more restrained arch for *normal* and *staccato*. The dip in the curve shows how she dug down for the *still* and *legato* intentions.

We computed the *Quantity of Motion* (*QoM*)[5] of the left hand as the magnitude of its 3-d velocity across trials and execution speeds. Figure 5 reports the statistics. A one-way ANOVA shows that there is a significant effect of the expressive intention on *QoM* ($F(4, 18805) = 642.7$), $p < 0.001$). A Tukey's HSD (Honestly Significant Difference) post-hoc analysis shows that the *Quantity of Motion* computed for each expressive intention is significantly different between all pairs of intentions.

4 GESTURE REALTIME RECOGNITION

In this section we inspect the potential of two state-of-the-art realtime gesture recognizers on one of the datasets as a way to highlight to what extent existing techniques can handle variations as presented above and infer opportunities for further research.

We chose to perform these tests on the violin dataset as it is comprised of several gestures, multiple instances of each of those gestures, performed with explicit variation.

4.1 Procedure

We conducted a within-subject procedure where for each subject we performed 10 tests. In each test we randomly chose a training set from the original dataset and trained two models from the state-of-the-art: a Hidden Markov Model (HMM) adapted for realtime gesture recognition as described in [3] and a dynamical system for realtime gesture recognition and variation tracking [2] based on Particle Filtering (PF).

The HMM was either trained with a single example per gesture class (denoted HMM-1) or 5 examples per class (HMM-5). The PF was trained with a single example per class. We inspected cases of IMU data only or multimodal IMU+EMG. For each test, we stored the likelihood estimations along gesture execution to analyze the classification accuracy at each time step (progression $1 - 100\%$).

4.2 Results

4.2.1 Classification on the original dataset (Fig. 6, Frame 1). IMU data. In classification, HMM-1 (93.1%) outperforms PF (91.3%). HMM-5 is even better with a final classification rate of 98.9%. At the very beginning of the gesture execution (1% of the gesture completed) HMM-1 is more accurate than PF (52.0% vs. 30.8%) and HMM-5 more accurate than HMM-1 with 63.0% accuracy.

IMU+EMG data. PF outperforms both HMM-1 and HMM-5 (94.3% against 80.9% and 90.2% respectively). Multimodal IMU+EMG improves PF classification accuracy compared to results with IMU only. On the contrary HMM-1 and HMM-5 global accuracies decrease when adding EMG features to the dataset. For each model, using the EMG modality significantly improves the early recognition rate. At 1% of the gesture performed, PF, HMM-1 and HMM-5 reach respectively 67.1%, 72.1% and 79.0%.

4.2.2 Classification on the piano-forte dataset (Fig. 6, Frame 2). IMU data. HMM-5 and PF obtain similar classification rates (81.3% and 81.6% respectively), outperforming HMM-1 (75.8%). PF reaches similar performance than HMM-1 and HMM-5 after about 60% of the gesture has been executed.

IMU/EMG data. Results are globally low. PF, HMM-1, and HMM-5 perform similarly with classification accuracies of 41.2%, 39.2%, and 39.5% respectively. Accuracies remain relatively constant at different times along the gesture progression.

4.2.3 Classification on the slow-fast dataset (Fig. 6, Frame 3). IMU data. As in the previous comparison, HMM-5 and PF obtain similar classification rates (81.2% vs. 82.0%), outperforming HMM-1 (73.1%). HMM-5 achieves a better early recognition rate compared to HMM-1 and PF. PF requires about 20% of the gesture to be completed to reach similar performance to HMM-1. The initial classification rate (at 1% of full gesture) is similar for HMM-1 and HMM-5 (49.6% and 51.3% respectively), above PF performance (42.4%).

[5]We do not refer to the magnitude of the 3-d velocity as *speed* to avoid confusion with the speed of execution.

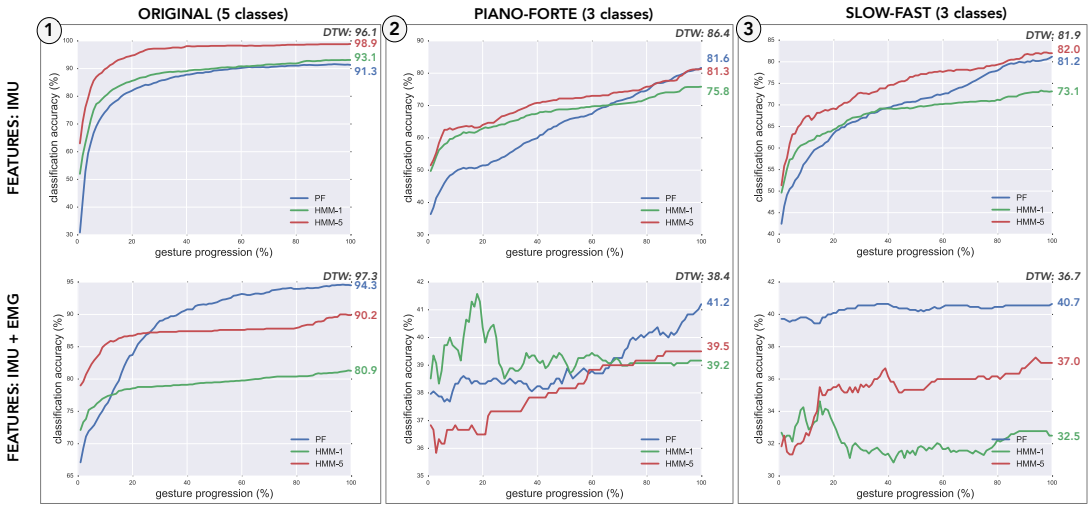

Figure 6: Classification accuracies computed for HMM and PF, trained on the *original* dataset and testing on *original*, *piano-forte*, and *slow-fast*, either using IMU only or IMU and EMG

IMU+EMG data. PF achieves a 40, 7% accuracy while HMM-1 and HMM-5 performances are 32.5% and 37.0%, close to chance (33%). The PF rate is relatively constant along the gesture progression.

5 DISCUSSION

We have presented two datasets of real-world musical gesture performed by expert musicians with explicit expressive variations: 1.) inertial and physiological recording of violin gesture and 2.) motion capture data of piano gesture.

The first dataset has a number of subjects (9), different gesture variations (5), and multiple instances of each (10), it constitutes a good candidate for testing gesture recognition systems. We showed that HMM exhibits best performance when trained and tested on the same dataset. Also, HMM shows a better early recognition rate than PF because PF has to update and propagate a probability distribution at every sample, which leads to slower convergence. Adding the 8-channel EMG modality decreases HMM recognition accuracy while increasing PF accuracy. However, none of the methods managed to adapt when trained on normal condition and tested on (unexpected) variations with complex data (EMG). This limitation offers an opportunity for further research in the design of realtime gesture recognition systems robust to complex variations. This dataset could then be used as a benchmark.

The piano dataset contains the position and orientation values of all body limbs during the performance of a musical excerpt, with variations in tempo and melodic articulation. This dataset is a comprehensive representation of a single gesture and its variations, thus minimizing the classification task and focusing instead on analyzing gesture variation. The dataset is multimodal including aligned video, audio and MIDI thereby offering a wide range of possibilities for analysis. In the illustrative example here, we used MIDI data to automatically segment motion capture data to center the analysis between two notes of interest. We focused on the left hand movement at the end of the first phrase in the excerpt and showed that computed *Quantity of Motion* is affected by expressive intentions. We consider that this dataset offers great potential for more complex analysis. For example, an interesting direction would be to study how expressive intention affects features computed from other modalities (e.g. audio and/or MIDI), and how the variations in features computed from different modalities correlate.

These datasets are made available to the research community. Our procedures are replicable, using readily available interfaces, and repertoire commonly used in the study of musical expression [8]. We hope that they provide a resource for future research in expressive musical gesture.

ACKNOWLEDGEMENTS

This work was supported by the EU project MIM (H2020- MSCA-IF-2014 , GA no. 659232) and MetaGesture Music (ERC FP7-283771).

REFERENCES

[1] Omid Alemi, Philippe Pasquier, and Chris Shaw. 2014. Mova: Interactive Movement Analytics Platform. In *Proc. of the 2014 Int. Workshop on Movement and Computing (MOCO '14)*. ACM, New York, NY, USA, Article 37, 6 pages. DOI: http://dx.doi.org/10.1145/2617995.2618002

[2] Baptiste Caramiaux, Nicola Montecchio, Atau Tanaka, and Frédéric Bevilacqua. 2015. Adaptive gesture recognition with variation estimation for interactive systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 4, 4 (2015), 18.

[3] Jules Françoise, Norbert Schnell, Riccardo Borghesi, and Frédéric Bevilacqua. 2014. Probabilistic models for designing motion and sound relationships. In *Proc. of the 2014 Int conference on New Interfaces for Musical Expression*. 287–292.

[4] Oscar Mayor, Jordi Llop, and Esteban Maestre Gómez. 2011. RepoVizz: a multimodal on-line database and browsing tool for music performance research. In *12th Int. Society for Music Information Retrieval Conference (ISMIR 2011); Miami; 2011 Oct. 24-28.*

[5] Caroline Palmer. 1997. Music performance. *Annual review of psychology* 48, 1 (1997), 115–138.

[6] Catherine Pelachaud. 2009. Studies on gesture expressivity for a virtual agent. *Speech Communication* 51, 7 (2009), 630–639.

[7] Nicolas H Rasamimanana, Emmanuel Fléty, and Frédéric Bevilacqua. 2005. Gesture analysis of violin bow strokes. In *Int. Gesture Workshop*. Springer, 145–155.

[8] Bruno H Repp. 1996. The dynamics of expressive piano performance: Schumann's "Träumerei" revisited. *The Journal of the Acoustical Society of America* 100, 1 (1996), 641–650.

[9] Jonna K. Vuoskoski, Marc R. Thompson, Eric F. Clarke, and Charles Spence. 2014. Crossmodal Interactions in the Perception of Expressivity in Musical Performance. *Atten Percept Psychophys* 76, 2 (Feb. 2014), 591–604. DOI: http://dx.doi.org/10.3758/s13414-013-0582-2

[10] Diana Young. 2008. Classification of Common Violin Bowing Techniques Using Gesture Data from a Playable Measurement System.. In *NIME*. Citeseer, 44–48.