Bow Gesture Classification to Identify Three Different Expertise Levels: A Machine Learning Approach

David Dalmazzo(✉) and Rafael Ramírez(✉)

Universitat Pompeu Fabra, 08018 Barcelona, Spain
{david.cabrera,rafael.ramirez}@upf.edu

Abstract. To acquire new skills in a high-level music context, students need many years of conscious dedication and practice. It is understood that precise motor actions have to be incorporated into the musicians' automatic executions, where a repertoire of technical actions must be learned and mastered. In this study, we develop a computer modelled assistant applying machine learning algorithms, for self-practice musicians with the violin as a test case. We recorded synchronized data from the performer's forearms implementing an IMU device with ambient sound recordings. The musicians perform seven standard bow gesture. We tested the model with three different expertise levels to identify relevant dissimilitudes among students and teachers.

Keywords: Machine learning · Music education · Hidden Markov Model

1 Introduction

1.1 Motivation

To become an expert performer in the context of music education is not only needed natural attitudes, as well, many years of conscious practice. It is understood that specific fine-motor actions must become part of the automatic execution (system 1) [10] in other words, a *"learned technique of the body"* [3], known as musical gesture, has to be developed and incorporated through precise practice and repetition. The standard strategy behind new skills development is based on the coupling of sound qualities, expressiveness and motor executions. However, the standard master-apprentice educative model based in imitation by example has some weaknesses, where the students could develop bad habits in self-practising hours. Therefore, in the context of Telmi (Technology Enhanced Learning of Musical Instrument Performance), we are investigating the implications of applying a computer modelled assistant to novice students, particularly at the moment to acquire new skills practising standard classical gestures

Music Technology Group, and Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502).

P. Cellier and K. Driessens (Eds.): ECML PKDD 2019 Workshops, CCIS 1168, pp. 494–501, 2020.
https://doi.org/10.1007/978-3-030-43887-6_43

with the test case of violin performers. We intend to stretch the gap of *"good-practice"* feedback, providing immediate information about gestural executions in real-time.

1.2 Gesture Recognition in Musical Context

To address the first stage of recognising specific gestures executions, we implemented Machine Learning (ML) techniques broadly found in the literature such a Hidden Markov Models (HMM) [2].

Bevilacqua et al. [1] presented a study in which an HMM system reports gesture time-progressions and its likelihood windowing. The ML model can be adjusted in states; which estimates Gaussian probabilities inside gesture progressions. Authors are not focused on specific gestural analysis; instead, they presented an optimal "low-cost" algorithm without the need for big datasets. Fiebrink and Cook [6] introduced the open-source multi-platform application called Wekinator, which includes a set of ML algorithms for pattern classifications, as well, dynamic time warping algorithms for time-related events. The tool is broadly used in academics and workshops for prototyping, artistic interactive music applications or as an educative reference of ML applicability in research topics. Fiebrink et al. [7] Executed the Wekinator to analyze bow-stroke articulations in a cello player. Authors embedded an IMU device in the bow-frog called K-Bow. The main goal was to allow the performer to interact in real-time through the gestures with a compositional computer-assistant. Françoise et al. [8,9] First exposed a gestural descriptor applying HMM and introduced the concept of mapping-by-demonstration as a principle of teaching with small amount of data the ML algorithms to then be used in the context of music education or real-time music interaction. In the next publication, authors describe probabilistic models such as Gaussian Mixture Models (GMM), Gaussian Mixture Regression (GMR), Hierarchical HMM (HHMM) and Multimodal Hierarchical HMM (MHMM). Dalmazzo and Ramirez [4] Based on IMU device and EMG data recorded from left-hand violinist players, authors estimated fingering disposition in the violin's neck. Two ML approaches (DT and HMM) were compared to determine accuracy. The main goal is to develop a computer-assisted pedagogical tool for self-regulated learners. Tanaka et al. [14] Based on the mapping-by-demonstration principle, authors describe different ML approaches to interact with generative sound and upper limb gestural patterns, applying techniques such as Static Regression, Temporal Modelling (HMM), Neural Network Regression and Windowed Regression, where the ML was feed using an IMU device including electromyogram (EMG) musician muscle-activity of the forearm signals. Dalmazzo and Ramírez [5] presented an ML approach to describe seven standard bow-stroke articulations (Détaché, Martelé, Spiccato, Ricochet, Sautillé, Staccato and Bariolage). A high-level expert violinist recorded the gestures, and then the system was used as a gestural estimator with an accuracy of 94%. ML model is based on HHMM, which is trained using audio descriptors and inertial motion information from the IMU device called Myo. The primary

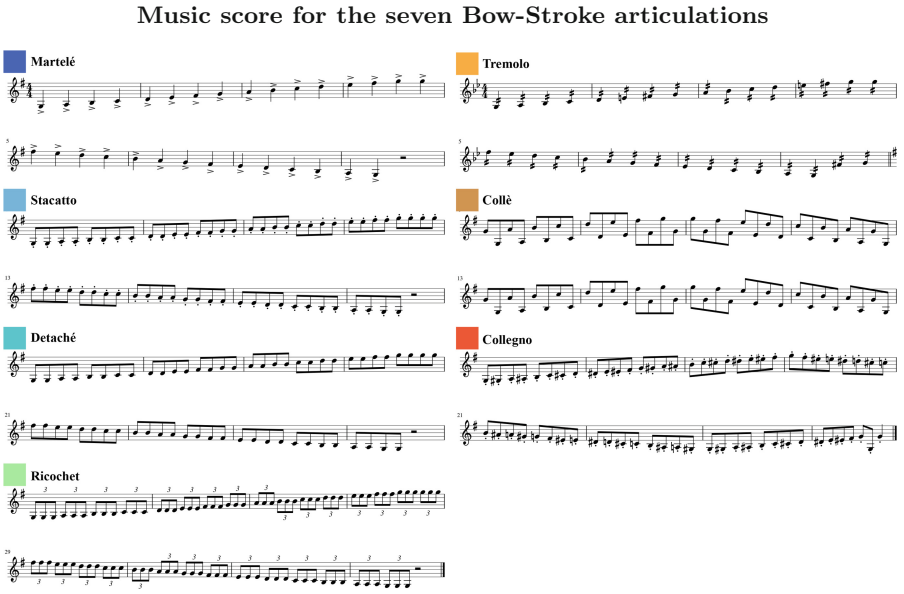

Fig. 1. Music score reference for the seven bow-strokes. Gestures 1, 2, 3, 4, and 6 are in G mayor. Gesture 5 in G melodic-minor and gesture 7 in G chromatic scale. All gestures were recorded with a metronome with a fixed tempo of Square-note 80 BPM.

purpose is to develop a computer-assistant for specific real-time feedback provider for self-regulated music students.

2 Methods and Materials

2.1 Music Score

Seven bow-strokes were recorded following a score with a fixed tempo of quarter-note in 80 bpm. Gestures were recorded in the key of G major, except for Tremolo (G minor) and Collegno (Chromatic G scale). In the violin, two octaves starting from G3 covers the whole neck and also the four strings are needed (Fig. 1).

2.2 Recordings and Synchronization

For the study, nine musicians (4 female) were recorded performing all gestures and a final music piece (Kreutzer 4), which include several bow-strokes examples. The data is composed of two expert performers categorized as **L1**, three high-level students categorized with the **L2** with more than nine years of practice, and four middle-level violin students categorized as **L3** with less than eight years of practice (5–7 years of practice). Data from two IMU devices *Myo* placed on both forearms were recorded using a C++ application which receives Bluetooth

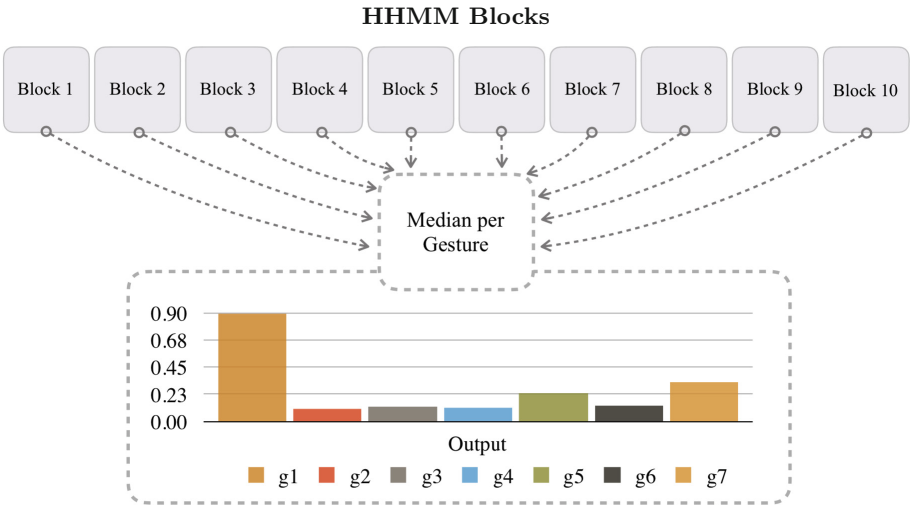

Fig. 2. Each block is an input of an HHMM which then gives as an output seven likelihood progressions and seven classification outputs of the most common number identified by the ten blocks

signals and formats it in a CSV file. Audio samples are synchronized with the *Myo* signals, recording all files with the same length in terms of time-reference. Both files are created and stored in the same time-events triggers. Audio playback has a timing reference in milliseconds, which is directly used to read Myo's data. −5 ms offset is needed to synchronize inertial data with audio sampling. A time reference value is stored with the inertial data which is transmitted at a 200 Hz ratio, that time reference is used from the audio player to sync gestures and sound.

2.3 OpenFrameworks Visualization

An application programmed in C++ using the open-source platform called Openframeworks (OF) [11] is used to visualize the data. From OF the data is send to Max 8 patch (via Open-Sound-Control) which has an HHMM implemented using the MUBU object extension [13] for real-time gesture estimation. For offline analysis, the python library *hmmlearn* is implemented [12].

2.4 Machine Learning Model

In a previous publication, we have implemented an HHMM to recognize gestures based on the *mapping-by-demonstration* principle [5]. In the current model, we intended to design a more generalist probabilistic estimation to be tested by different students. For that we have an architecture based on ten blocks of HHMM sampling ten different dispositions of gestures over the four strings of the violin;

ten sub-blocks are trained with one of the experts L1 and the other ten sub-blocks are trained with the second L1 expert. A median is then extracted as a final output for all likelihood gestures estimations (Fig. 2).

3 Results

Three different performers were selected from the original nine recordings, one for each expertise level, L1, L2, L3, being L1 the expert as a model, L2 high-level students and L3, middle-level student. Confusion Matrix in the Fig. 3 is composed of three different expertise levels: L1 corresponds to a high-level expert. L2 corresponds to an advanced student. L3 corresponds to a beginner-level student. Gestures are distributed as (1) Martelè (2) Staccato (3) Detaché (4) Ricochet (5) Tremolo (6) Collè and (7) Collegno. L1, L2 and L3 identification are at the right part of the matrix.

Weighted probabilities in the Fig. 4 in letters (E), (F), (G) AND (H) plot the output of the average block as a result of the ten HHMM blocks estimations. (E) is Ricochet gesture from L1 and (F) is Ricochet gesture from L2. (G) is the Tremolo gesture from L1, and H) is the Tremolo gesture from L2. Those maps are distributed in a range of 0.0 to 1.0 (normalized), where 1.0 is the highest probability that the current gesture is being recognized.

Confusion Matrix of three different performer's levels

class	1	2	3	4	5	6	7	
1	0.999	0.000	0.000	0.000	0.000	0.001	0.000	
2	0.000	0.650	0.096	0.000	0.008	0.246	0.000	
3	0.000	0.005	0.994	0.000	0.000	0.001	0.000	L1
4	0.024	0.116	0.002	0.695	0.001	0.162	0.000	
5	0.000	0.000	0.002	0.006	0.839	0.001	0.151	
6	0.000	0.029	0.047	0.000	0.005	0.919	0.000	
7	0.000	0.000	0.015	0.000	0.002	0.160	0.823	
1	0.971	0.000	0.000	0.000	0.024	0.004	0.000	
2	0.000	0.261	0.190	0.000	0.220	0.329	0.000	
3	0.002	0.002	0.764	0.005	0.227	0.000	0.000	
4	0.010	0.000	0.000	0.831	0.087	0.072	0.000	L2
5	0.085	0.001	0.009	0.001	0.904	0.001	0.000	
6	0.001	0.011	0.035	0.001	0.264	0.687	0.000	
7	0.000	0.428	0.027	0.011	0.023	0.148	0.363	
1	0.451	0.000	0.000	0.350	0.052	0.147	0.000	
2	0.003	0.204	0.259	0.020	0.337	0.176	0.000	
3	0.017	0.031	0.409	0.309	0.132	0.094	0.008	
4	0.006	0.000	0.011	0.638	0.000	0.344	0.000	L3
5	0.000	0.096	0.073	0.000	0.793	0.032	0.007	
6	0.115	0.077	0.292	0.000	0.439	0.076	0.000	
7	0.000	0.021	0.156	0.281	0.319	0.136	0.089	

Fig. 3. Confusion Matrix figure of the three different levels (L1, L2 and L3) numbers are classes identifications per gesture. The colour code is based on a linear gradient where white is 0.0, and full orange is 1.0 (Color figure online)

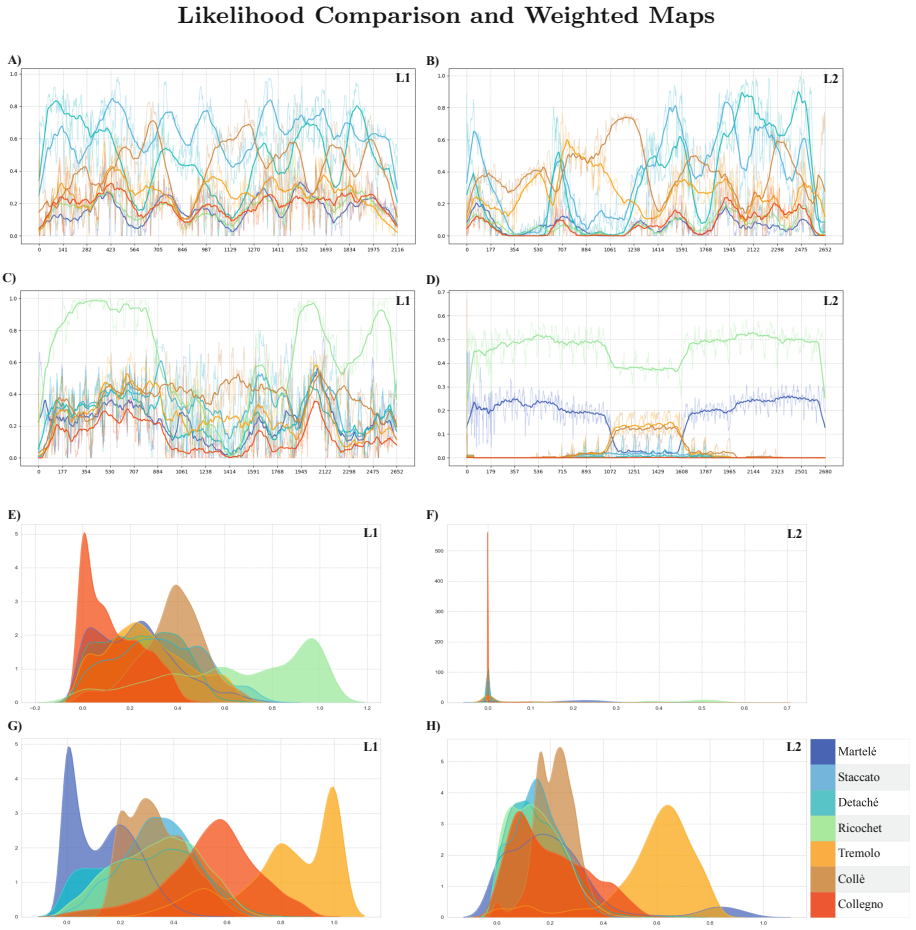

Fig. 4. (A) and (B) corresponds to the second gesture (Staccato) from the L1 and L2 performers; (C) and (D) corresponds to Ricochet from the levels L1 and L2 respectively. (E) and (F) Are weighted-maps (WM) in a range from 0.0 to 1.0 in the X-axis, where 1.0 corresponds to 100% accuracy in gesture estimation. (E) is the WM from gesture 4 (Ricochet) from L1, and (F) is the same WM for gesture 4 in the case of L2. (G) and (H) are WM of the gesture 5 (Tremolo) comparing the levels L1 and L2. Dotted lines in X-axis are markers for each note in the scale where the gesture was performed

4 Discussion and Conclusions

In the case where a small amount of training data is available, HHMM is a robust algorithm for pattern recognition of temporal events. The *mapping-by-demonstration* principles is sufficient for modelling an ML human gestures classifier; as in the case of generative music and gesture interaction [14]. However, for a more generalist model, similar to an MNIST [15], another approach would be needed, perhaps the implementation of Recurrent Neural Networks (RNN),

and bigger datasets. The HHMM approach based on blocks reported accurate results in recognizing the seven gestures explained above. Nevertheless, some curious differences among L1 and L2 were observed for the gestures Ricochet (4) and Tremolo (5). The Confusion Matrix in Fig. 3 in the case of L1 reported 69.5% and 83.9% of accuracy in gestures 4 and 5 consecutively, and for the L2 case it was higher 83.1% and 90%, however, in the Fig. 4 different probabilistic weighted-maps (graph (C) and (D), as well, (E) and (F)), are visible, in (C) L1 gesture estimation oscillates between 100% to bellow 20% and L2 in (D) keeps more stable around 50% of certainty. As the HHMM blocks are build using two experts, we consider that both have some dissimilitudes, particularly when the first string of the violin is played. It opens the discussion that strings two, three and four might have a more constrained range of movement as the bow needs to avoid contact with the neighbour's strings, therefor performers permit some execution-freedom in the first string.

In the Fig. 3, the Confusion Matrix give an insight of the variability among the three levels, where L1 is above 82% in gestures Martelé, Detaché, Tremolo, Collè and Collegno, L2 has some variations especially in the gestures Tremolo, Collè and Collegno; and the L3 has a broader variability. Staccato is a gesture commonly confused with Martelé; it is characterized as an isolated distinct sound; it does not have a strong attack; however, it has some similitude with Detaché. In Fig. 4 this similitude can be seen in the (A) and (B) examples, where L1 model mixes Staccato and Detaché; and (B) L2 case Staccato appears at the beginning of some gestures, but the model also detects Detaché, Collè and even Tremolo.

4.1 Future Work

From the perspective of building a general model for bow-stroke gestural detection, it is needed a broader dataset, also to apply data augmentation, as the motion information is based on an imaginary direction in terms of quaternions, it is possible to expand by extrapolating to many other horizontal angles. A new algorithm based on Long-Short Term Memory (RNN) would be tested in a mixture architecture with Hidden Markov Models.

References

1. Bevilacqua, F., Zamborlin, B., Sypniewski, A., Schnell, N., Guédy, F., Rasamimanana, N.: Continuous realtime gesture following and recognition. In: Kopp, S., Wachsmuth, I. (eds.) GW 2009. LNCS (LNAI), vol. 5934, pp. 73–84. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12553-9_7. http://articles.ircam.fr/textes/Bevilacqua09b/index.pdf
2. Caramiaux, B., Bevilacqua, F., Schnell, N.: Towards a gesture-sound cross-modal analysis. In: Kopp, S., Wachsmuth, I. (eds.) GW 2009. LNCS (LNAI), vol. 5934, pp. 158–170. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12553-9_14

3. Carrie, N.: Agency and Embodiment: Performing Gestures/Producing Culture. Harvard University Press, Cambridge (2009)
4. Dalmazzo, D., Ramirez, R.: Air violin: a machine learning approach to fingering gesture recognition, November 2017, pp. 63–66 (2017). https://doi.org/10.1145/3139513.3139526
5. Dalmazzo, D., Ramírez, R.: Bowing gestures classification in violin performance: a machine learning approach. Front. Psychol. **10**(MAR), 1–20 (2019). https://doi.org/10.3389/fpsyg.2019.00344
6. Fiebrink, R., Cook, P.R.: The Wekinator: a system for real-time, interactive machine learning in music. In: Proceedings of The Eleventh International Society for Music Information Retrieval Conference (ISMIR 2010), vol. 4, no. 3, p. 2005 (2010). http://ismir2010.ismir.net/proceedings/late-breaking-demo-13.pdf?origin=publicationDetail
7. Fiebrink, R., Cook, P.R., Trueman, D.: Human model evaluation in interactive supervised learning, p. 147 (2011). https://doi.org/10.1145/1978942.1978965
8. Françoise, J., Caramiaux, B., Bevilacqua, F.: A hierarchical approach for the design of gesture-to-sound mappings. In: 9th Sound and Music Computing Conference, pp. 233–240 (2012)
9. Françoise, J., Schnell, N., Borghesi, R., Bevilacqua, F., Stravinsky, P.I.: Probabilistic models for designing motion and sound relationships. In: Proceedings of the International Conference on New Interfaces for Musical Expression, pp. 287–292 (2014)
10. Kahneman, D.: Thinking, Fast and Slow (Kindle Edition) (2011)
11. Openframeworks (2017). https://openframeworks.cc/
12. Rabiner, L.R.: tutorial on hmm and applications.pdf
13. Schnell, N., Röbel, A., Schwarz, D., Peeters, G., Borghesi, R.: MuBu & friends - assembling tools for content based real-time interactive audio processing in MAX/MSP. In: Proceedings of the International Computer Music Conference (ICMC), pp. 423–426 (2009)
14. Tanaka, A., Di Donato, B., Zbyszynski, M.: Designing gestures for continuous sonic interaction, June 2019
15. Zhu, W.: Classification of MNIST handwritten digit database using neural network (2000)