COMBINING GESTURAL AND AUDIO APPROACHES TO THE CLASSIFICATION OF VIOLIN BOW STROKES

William Wilson, Islah Ali-MacLachlan, Niccolò Granieri
Birmingham City University
{william.wilson,islah.ali-maclachlan,niccolo.granieri}@bcu.ac.uk

ABSTRACT

This paper details a brief exploration of methods by which gestural and audio based approaches may be used in the classification of violin performances. These are based upon a multimodal dataset. Onsets are derived from audio signals and used to segment synchronous gestural recordings, allowing for the classification of individual bow strokes utilising data of either type — or both. Classification accuracies for the purposes of participant identification ranged between 71.06% and 91.35% for various data type combinations. Classification accuracies for the identification of bowing technique were typically lower, ranging between 53.33% and 77.35%. The findings of this paper inform a number of recommendations for future work. These are to be considered in the development of a principally similar dataset, for the analysis of traditional fiddle playing styles.

1. INTRODUCTION

1.1 Background

The audible content of a violin performance may be considered a product of the performer's gestural execution. These performance aspects may be quantified, respectively, through the use of Music Information Retrieval (MIR) and gestural analysis techniques. Prior studies have proven the implementation of machine learning techniques to be efficacious in the automation of classification tasks based upon both audio and gestural data. This preliminary study aims to assess the viability of employing such methods, with a view towards further analyses into traditional fiddle playing styles.

1.1.1 Gestural Technologies

Gestures may be recorded and quantified through the use of gestural sensors, yielding time-series gestural data. Gestural sensors are typically Inertial Measurement Unit (IMU) or Electromyography (EMG) based.

An object within a three-dimensional space has both a location and an orientation; these may both be described in three dimensions. An object's location may be described by its translational position, relative to a set of X, Y, Z axes. The orientation of an object may be similarly described by the object's rotation around each axis. Conventionally, these may be referred to as 'Roll', 'Pitch', and 'Yaw' (Craig, 2005). Each of these metrics are termed 'Degrees of Freedom' (DoF). IMU sensors quantify movements through recording changes in each DoF over time, yielding respective acceleration and gyroscopic data.

Castellini & van der Smagt (2009) summarise EMG as "a technique by which muscle activation potentials are gathered by electrodes placed on the [...] skin". Raez et al. (2006) describe raw EMG signals as consisting of electrical wave-packets ranging between -5 and +5 mV, attributing these to the electrical field generated by muscle fibres during contraction. Citing a number of prior studies, Castellini & van der Smagt (2009) discussed the efficacy of using forearm surface EMG in combination with machine-learning algorithms for the classification of hand posture. They attributed the success of prior implementations to an existing relationship between the force applied by a muscle and the amplitude of the resultant EMG signal; in implementation, the use of multiple sensors allows for the identification of 'precise force configurations' associated with specific hand postures.

1.1.2 MIR Feature Extraction Techniques

Feature-extraction derived low-level descriptors provide a representation of a signal's timbral and rhythmic characteristics. Schedl et al. (2014) note limitations surrounding the interpretability of these descriptors, favouring instead their implementation within computational classification systems.

Ali-MacLachlan (2019) asserts the utility of Mel Frequency Cepstral Coefficients (MFCCs) as a representation of timbre, terming these a "compact feature representation used in audio signal classification". Zheng et al. (2001) define MFCCs as "the results of a cosine transform of the real logarithm of the STFT expressed on a Mel-frequency scale". A noted benefit of the Mel-scale's application in such tasks is its approximation of human-frequency perception

While acknowledging their usefulness as an indicator of timbrality, McFee et al. (2015), contend that MFCCs are flawed in their depiction of pitch, considering them to offer "poor resolution of pitches and pitch-classes". Instead, the authors suggest the use of Chroma representations in the depiction of these, purporting them to "encode harmony while suppressing variations in octave height, loudness, or timbre." Stein et al. (2009) identify a number of techniques by which Chroma representations may be calculated. In doing so the authors noted each as derived from the Pitch-Class-Profile (PCP) technique. An FFT of an input signal is first taken, after which the frequency bin magnitudes within each semitone boundary are summed. The subsequent semitone magnitudes are summed by pitch with those of different octaves, providing an instantaneous indicator of perceived pitch.

1.1.3 Neural Network Classification

Alpaydin (2020) describes a Single-Layer Perceptron as "the basic processing element" of any neural network, comprising of a single node which may receive any number of numerical inputs. To each numerical input, a weight is ascribed. Through summation of the product of each input and ascribed weight, the node produces an output value. Russell & Norvig (2020) describe the Multi-Layer Perceptron (MLP) as an expansion of the Single-Layer Perceptron comprising of multiple layers of nodes, decreasing in quantity and linked by interconnected weights. Weights are initialized randomly, and refined through training upon labelled data, through which the input data may be classified to an output. While a conventional MLP comprises of weights connecting in only one direction, (and is thus termed a feed-forward network) Russell and Norvig (2020) identifies a Recurrent Neural Network (RNN) as a variant of the MLP incorporating recurrent connections, wherein the output of an intermediate node may be fed back towards the input of itself, or other preceding nodes.

Dalmazzo et al. (2021) demonstrated a high degree of accuracy while using Convolutional Neural Networks (CNN) trained upon IMU data for the classification of eight bowing techniques: martelé, staccato, detaché, ricochet, legato, trémolo, collé, and col legno. Reported recognition rates ranged between 97.147% and 99.234% for a variety of CNN based models - the prior a conventional CNN and the latter a CNN Long Short-Term Memory Network.

Dalmazzo & Ramirez (2017) investigated the efficacy of employing forearm-surface EMG alongside IMU data for the off-line recognition of fingering gestures during violin performance. Classification using a Hidden Markov Model (HMM) yielded gestural recognition accuracies of between 89.44% and 99.23%.

Zheng et al. (2001) demonstrated the efficacy of using extracted MFCCs to train a HMM for the purposes of speech recognition.

Miotto & Orio (2008) utilized Chroma representations in their development of an automated music identification system, proposing their use as 'indexes' in an HMM-based retrieval system.

2. METHOD

A series of multi-class classification tasks were completed using the open source *Violin Gesture Dataset* published by Sarasúa et al. (2017). The dataset contains simultaneous IMU (50Hz), EMG (200Hz) and audio (48kHz) recordings for 880 performances of an excerpt from Kreutzer's Etude No. 2 in C Major, with a typical duration of around 11 seconds. Each recording is labelled by both participant and a bow-articulation condition (*martelé, staccato, detaché, spiccato, legato*) - to be subsequently termed 'Style'. While the dataset in its entirety contains recordings of 8 participants, one of these was excluded during execution of the classification task due to a corrupted EMG recording. This decision was made with the intention to maintain a consistent amount of data per participant.

Figure 1 depicts a chronology of the implemented method, which aimed to classify data associated with an isolated bow stroke by participant or style. The three data types employed are depicted at the far left of the figure, as sourced from the dataset. Subsequent processes applied to these are identified, as were applied in preparation of the data for classification — depicted at the far right of the figure.

2.1 Gestural Data Processing

Processing of the gestural data was performed upon the signals in their entirety, prior to their segmentation.

A linear de-trend function was first applied to each channel of IMU data, given the tendency of IMU sensors to exhibit drift over time - a result of accumulated error (Kok et al., 2017).

Proportional normalisation was applied to both the IMU and EMG data such that the maximum magnitude of a signal was bounded by 1, while the proportional difference in maximum magnitude between concurrent channels of data (e.g. EMG channels 1-8) was maintained.

A low-pass filter was subsequently applied to the EMG data, with a cutoff of 10Hz; in implementation providing a simple amplitude envelope (Tanaka & Ortiz, 2017).

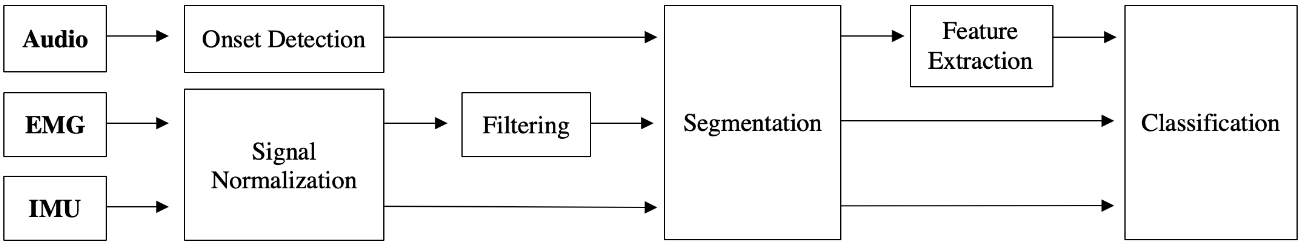

Figure 1: Flow Chart Depicting Constituent Sub-Tasks of the Implemented Method.

2.2 Data Segmentation

Note-onset positions were first identified within the audio data through use of the onset detection functionalities of the *Librosa*[1] Python library. Each onset position was returned as an index of the audio sample array. Given sample-rate discrepancies between the three data types, proportional scaling of each index was necessary in identifying temporally equivalent indices within the gestural data.

The recordings of each data type were then split, using their respective onset indices, into a series of inter-onset-intervals; these were considered to be representative of singular bow strokes.

2.3 Audio Feature Extraction

Feature extraction techniques were subsequently employed through use of *Librosa*, for the purposes of calculating low-level descriptors derived from the segmented audio data; namely 13 MFCCs, 13 Delta-MFCCs, 13 Delta-Delta-MFCCs, and 12 Chromas. The mean values for each feature were calculated, such that for each bow stroke singlar sets of MFCCs, Delta-MFCCs Delta-Delta-MFCCs and Chromas were produced.

2.4 Neural Network Classification

MLP networks were used for the completion of 12 separate multi-class classification tasks; two tasks per input data type.

The number of input and output nodes of the MLP varied between tasks, given variation in the number of input data points per data type, and the number of classes per classification task. Despite this, the fundamental architecture of the MLP remained consistent. This comprised of an input-layer and two densely-connected hidden layers, each with an equal number of nodes to the number of input data points. The output layer comprised of an equal number of nodes to the number of classes.

A stratified, 80:20 train-test split was used for the creation of train and test subsets.

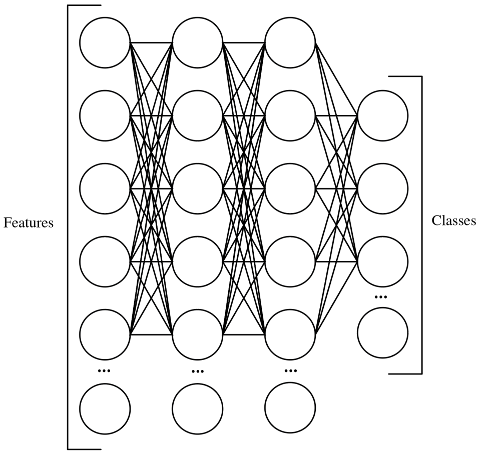

Figure 2: Multi-Layer Perceptron Network Showing Input and Output Layers with Two Hidden Layers

3. RESULTS

Data Type	Participant	Style
EMG	87.78%	70.31%
IMU	89.66%	77.35%
EMG+IMU	89.63%	74.20%
MFCC	71.06%	44.32%
MFCC+Deltas	86.68%	53.33%
MFCC+Deltas+IMU	91.35%	73.13%

Table 1: Test Classification Accuracies per Data Type.

Consistently higher classification accuracies were exhibited in completion of the participant classification task, with an average accuracy of 85.52% across all data type combinations. 'Style' classification accuracies were considerably lower per data type combination, with an average classification accuracy of 65.44%.

Lone MFCC data demonstrated comparatively low accuracies in the completion of both classification tasks, although the inclusion of additional feature-extracted low-level descriptors (Delta-MFCCs, Delta-Delta-MFCCs, Chromas) resulted in an accuracy comparable to that of the gestural data types for the purposes of participant classification. The inclusion of these did not prove similarly beneficial for the purposes of style classification; while a significant increase in accuracy was observed, the resultant classification accuracy was far below that of the gestural data types. It should be noted that this implementation of MIR feature extraction techniques preserved no temporal information contained within the original audio signals, in contrast with the gestural data types which remained time-series. Consideration of this, in the context of the aforementioned results, may suggest temporality to be a more crucial aspect in the classification of bowing technique than in participant identification.

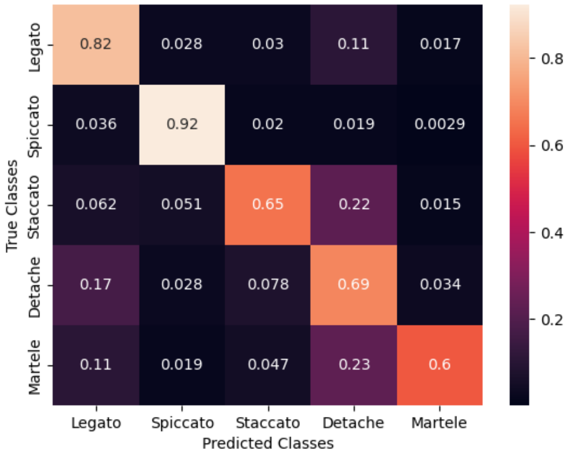

Figure 3: IMU Style Classification False Negative Rate Confusion Matrix.

Figure 3 depicts the proportion of false-negative predictions in the classification of 'Style' based upon IMU data - the highest performing data type for this task. *Martelé* was most the frequently misclassified, with 40% of test-data

[1] https://librosa.org

classified incorrectly; most commonly as *detaché* (23%) or - with around half the frequency - *legato* (11%).

In contrast, *Spiccato* test-data frequently exhibited a far lower rate of misclassification (8%) across all data types, indicating a greater dissimilarity between data of this type and that of the remaining classes.

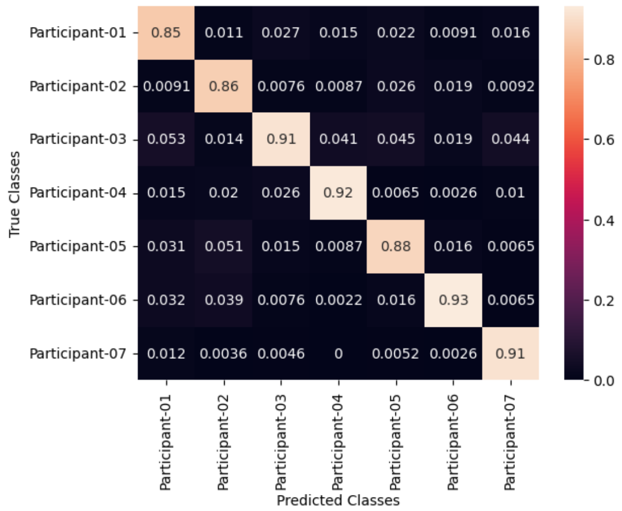

Figure 4: IMU Participant Classification True Positive Rate Confusion Matrix

Figure 4 depicts the proportion of true-positive predictions in the classification of participant based upon IMU data, denoting a comparatively consistent classification accuracy across all classes.

4. EVALUATION

Inaccuracies in the implemented onset detection algorithm initially lead to the inclusion of some inter-onset-intervals that were either too long or short to be representative of a single bow stroke. A condition whereby these were discarded was implemented with the intention of negating this. Given the onset detection algorithm performed more accurately for recordings of specific 'Style' classes (particularly *Spiccato*), this lead to the exclusion of disproportionate amounts of data between them. This, in turn, resulted in a significantly smaller amount of *Martelé* and *Staccato* data across all data types. Respectively, these classes were found to comprise of 76% and 66% fewer features than the average of the remaining three classes. While a direct causation has not been established, these were also the two most frequently misclassified across all data types.

As discussed in Section 3, a true parallel between the employed gesture- and audio-based approaches cannot be drawn, given the lack of temporal information preserved while calculating the MFCC data.

5. CONCLUSIONS AND FUTURE WORK

Gestural data may be used effectively for player identification within violin playing. Participants may be identified with reasonable accuracy through classification of their movement alone, indicating consistent quantitative distinctions in gestural execution. An established audio-based approach proved similarly effective for the purposes of participant-identification, but considerably less effective in the identification of bow-articulation.

Analyses concerning the gestural content of traditional fiddle playing are considered viable through the use of principally similar methodologies, provided a suitable dataset is compiled. Given the efficacy of the implemented gestural approaches, it is the intention of the authors to do so. This would facilitate quantitative, gestural analyses of traditional fiddle playing in unprecedented depth.

A system is in development whereby a pair of MYO-Armband gestural sensors are used to capture IMU and EMG data from each forearm synchronously. This allows for the capture of gestural data associated with both fingering and bowing techniques. An additional bow-mounted IMU sensor is used to record changes in the position and orientation of the bow over time.

The findings of this paper informed the following recommendations for further analyses based upon the proposed dataset. The implementation of an improved onset detection method is expected to minimise discrepancies in the proportion of data per class; these posed an extraneous variable as a result of the existing method employed. An alternative audio-based approach is also proposed, whereby a series of low-level descriptors are used in the classification of each bow stroke. This is expected to facilitate greater accuracy in the classification of bow-articulation, through maintenance of the audio data's temporality. The investigation of more sophisticated classification algorithms is expected to further increase classification accuracies across all tasks. Having reviewed current literature, those of specific interest include HMM and RNN models.

6. REFERENCES

Ali-MacLachlan, I. (2019). *Computational Analysis of Style in Irish Traditional Flute Playing*. PhD Thesis, Birmingham City University, Birmingham, UK.

Alpaydin, E. (2020). *Introduction to Machine Learning* (4th ed.). Massachusetts, USA: The MIT Press.

Castellini, C. & van der Smagt, P. (2009). Surface EMG in advanced hand prosthetics. *Biological Cybernetics*, *100*(1), 35–47.

Craig, J. J. (2005). *Introduction to Robotics - Mechanics and Control* (3rd ed.). New Jersey, USA: Pearson Education, Inc.

Dalmazzo, D. & Ramirez, R. (2017). Air violin: A Machine Learning Approach to Fingering Gesture Recognition. In *Proceedings of the 1st ACM International Workshop on Multimodal Interaction for Education*, (pp. 63–66)., Glasgow UK. ACM.

Dalmazzo, D., Waddell, G., & Ramírez, R. (2021). Applying Deep Learning Techniques to Estimate Patterns of Musical Gesture. *Frontiers in Psychology*, *11*, 575971.

Kok, M., Hol, J. D., & Schön, T. B. (2017). Using Inertial Sensors for Position and Orientation Estimation. *Foundations and Trends® in Signal Processing*, *11*(1-2), 1–153. arXiv: 1704.06053.

McFee, B., Raffel, C., Liang, D., Ellis, D., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and Music Signal Analysis in Python. (pp. 18–24)., Austin, Texas.

Miotto, R. & Orio, N. (2008). A Music Identification System Based on Chroma Indexing and Statistical Modelling. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, (pp.$\tilde{6}$)., Pennsylvania, USA.

Raez, M., Hussain, M., & Mohd-Yasin, F. (2006). Techniques of EMG Signal Analysis: Detection, Processing, Classification and Applications. *Biological Procedures Online*, *8*, 11–35.

Russell, S. & Norvig, P. (2020). *Artificial Intelligence - A Modern Approach.* (4th edn. ed.). Pearson series in Artificial Intelligence. New Jersey, USA: Pearson Education, Inc.

Sarasúa, , Caramiaux, B., Tanaka, A., & Ortiz, M. (2017). Datasets for the Analysis of Expressive Musical Gestures. In *Proceedings of the 4th International Conference on Movement Computing*, (pp. 1–4)., London United Kingdom. ACM.

Schedl, M., Gómez, E., & Urbano, J. (2014). Music Information Retrieval: Recent Developments and Applications. *Foundations and Trends in Information Retrieval*, *8*(2-3), 127–261.

Stein, M., Schubert, B. M., Gruhne, M., Gatzsche, G., & Mehnert, M. (2009). Evaluation and Comparison of Audio Chroma Feature Extraction Methods, 9.

Tanaka, A. & Ortiz, M. (2017). Gestural Musical Performance with Physiological Sensors, Focusing on the Electromyogram. In *The Routledge Companion to Embodied Music Interaction* (pp. 422–430). Oxfordshire, England: Routlege.

Zheng, F., Zhang, G., & Song, Z. (2001). Comparison of Different Implementations of MFCC. *Journal of Computer Science and Technology*, *16*(6), 582–589.