Automatic assessment of violin performance using dynamic time warping classification

Sergio Giraldo*, Ariadna Ortega*, Alfonso Perez*, Rafael Ramirez*, George Waddell†, Aaron Williamon†,
* Music and Machine Learning Lab
Pompeu Fabra University
Barcelona, Spain
{sergio.giraldo, alfonso.perez, rafael.ramirez}@upf.edu, ariadna.ortega01@estudiant.upf.edu
† † Centre for Performance Science
Royal College of Music
London, UK
{george.waddell, aaron.williamon}@rcm.ac.uk

***Abstract*—The automatic assessment of music performance has become an area of special interest due to the increasing amount of technology-enhanced music learning systems. However, in most of these systems the assessment of the musical performance is based on the accuracy of onsets and pitch, paying little attention to other relevant aspects of performance. In this paper we present a preliminary study to assess the quality of violin performance using machine learning techniques. We collect recording examples of selected violin exercises varying from *expert* to *amateur* performances. We process the audio signal to extract features to train models using clustering based on Dynamic Time Warping distance. The quality of new performances is evaluated based on the level of match/miss-match to each of the recorded training examples.**

***Keywords*—Machine learning, Dynamic Time Warping, Automatic assessment of music performance, Violin performance.**

I. Introduction

The qualitative assessment of music performance is an essential task in music education. Music information retrieval technologies can play an important role in music education [1]. Thus, an important amount of systems are being developed to enhance the learning process of musical instruments, which make use of audio signal processing technologies for music information retrieval [1], some of them providing automatic assessment tools. Music Plus One [2] uses a Hidden Markov Model to generate orchestral accompaniments able to follow the soloist expressive variations. Antescofo [3] provides a score following system that allows the recognition of the player position and tempo in a score. Song2see [4] is an application gaming software for music learning and practicing that make use of Music Information Retrieval tools such as pitch detection and source separation to return feedback based on a rating system. Also, others commercial systems such as Yousician[1] and Smart Music[2], are able to provide real-time feedback rating of music performance. In the contex of automatic assessment of music performance the automatic characterization of dynamics and articulation from low level audio features has been studied [5] in the context of expressive music performance. Other approaches make enphasis in the automatic assessment of tone quality in trumpet sounds using machine learning techniques [6]. Good-Sounds [7] make use of machine learning techniques to identify good and poor quality notes of trumpet, clarinet and flute, given training data consisting of low and high level audio features extracted from performed musical sounds with each instrument. In our previous work, within the TELMI project [8], [9], we proposed a system to automatically asses the quality of timbre of violin sounds, using machine learning techniques. However, several complications arise when attempting to obtain reliable models for the assessment of music performance. On one hand, in most of these technologies the models to evaluate a musical performance rely on the accuracy of pitch and onset, leaving a side important performance aspects in terms of musical interpretation, communication, and expressiveness. On the other, the subjectiveness of such high hierarchical performance attributes (e.g semantic labels for timber quality assessment) represents a significant complication in generating models that agree with the general consensus of music experts (or evaluators). Furthermore, such consensus some time is unachievable among experts.

In this paper we present a preliminary study for the automatic assessment of music performance, using Dynamic Time Warping Classification [10]. We have collected recordings from a selection of violin exercises, ranging from *amateur* to *expert* performances. Frame based audio features have been extracted from the audio, and machine learning models have been trained using clustering based on Dynamic Time Warping distance. The assessment of the quality of a performance is obtained by the level of match/miss-match between expert or non-expert recording examples.

[1] www.yousician.com
[2] www.smartmusic.com

II. Methodology

A. Data acquisition

The framework of our methodology is depicted in Figure 1. Firstly we obtained recordings of 10 violin exercises, each of them recorded by 2 expert, 2 mid level, and 2 beginner musicians. We obtained a total of six performance examples per exercise, as well as three classes per exercise (i.e. expert, mid-level, and beginner). Expert musician recordings were performed by well known professional violinists, with both academic and performance consolidated careers. Mid-level musician recordings were performed by students form a recognized music institution, who were following the performance degree program in violin. Finally beginner performance recordings were done by amateur non professional musicians, able to read musical notation.

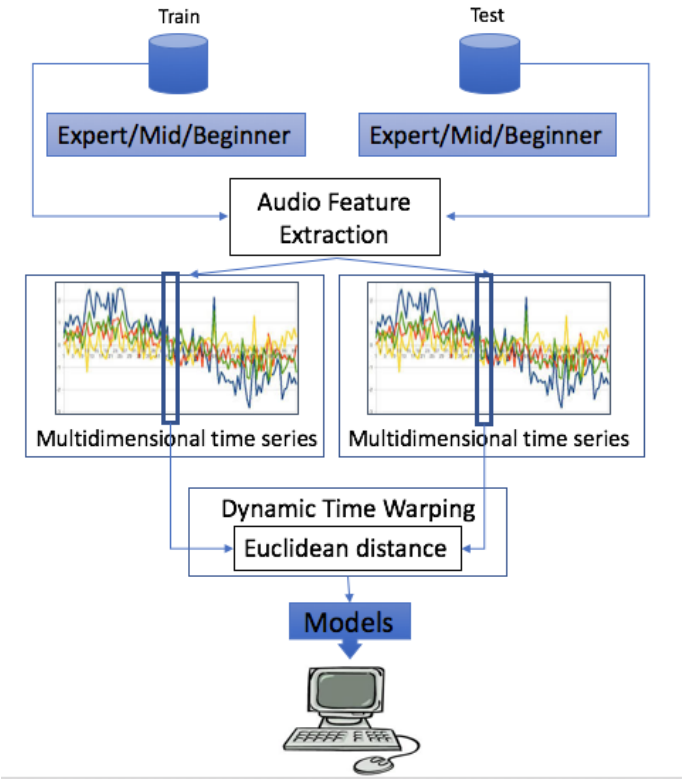

Fig. 1: Overall framework of the proposed automatic music performance assessment approach.

B. Audio feature extraction

For each of the audio pieces we extracted low-level frame-based audio descriptors (see [11] for an overview), using the Essentia library [12]. Low level audio descriptors included pitch, spectral (e.g. spectral centroid, spectral kurtosis, MFCCs, etc.), and energy descriptors (e.g. RMS). Each audio sequence was then represented by time series of the extracted audio descriptors, where each frame consists of a multidimensional point (one dimension per descriptor) in time.

C. Dynamic time warping for time series classification

We applied Dynamic Time Warping (DTW) distance to classify between the three aforementioned classes: Expert, Mid-level, and Beginner level. For the purpose of this study, one of the performances of each class/exercise was (randomly) chosen for training, whereas the other was used for testing. The classification task is done in several steps. Firstly, we apply DTW to obtain a distance measure of a test example to each of the train examples. The *euclidean distance* is used to compute the distance between each multidimensional time series point inside the DTW cost function. Thus, for each of the test examples we obtained three distance measures, one for each of the three training examples. Finally we applied a softmax function to obtain a distance as a probability number. The final score is obtained by weight-averaging each distance probability with a class-ranking, where rankings were set to 10, 5, and 0 for master, mid-level, and beginner performances, respectively.

III. Evaluation

A. Experiment set up

Each of the recorded exercises was tested on a train/test basis, using a real-time audio signal processing scenario depicted in Figure 2.

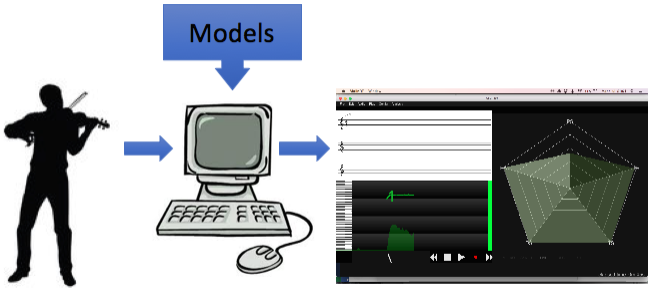

Fig. 2: Experiment set up using the real-time audio analysis tool

From our previous approach in [8] and [9], our application was able to extract in real-time frame-based descriptors (see Section II-B), when capturing the audio signal with the computer built-in microphone. For the experiments, each recorded example was reproduced on a high quality audio system. This setting was chosen as this is the expected end-user scenario. For each exercise, the system initially loaded with each of the three training examples (e.g. exercise 1: expert, mid-level, beginner training recordings), by recording the time series of each of the audio descriptors. Later, the same process was applied to each of three test examples (e.g. exercise 1: expert, mid-level, beginner test recordings). After a test example was parsed by the system, a training stage took place automatically where the system performed the DTW and output the match/miss-match probabilities for each of the three classes. This set up was repeated for all ten train/test exercises.

IV. Results

The results of the aforementioned experiment set up is summarized in Table I. In the first column presents the list of training recordings, in which for each exercise we had the three labels. In the other columns we present the match and miss-match probabilities in a confusion matrix. High probabilities were obtained for the true positive cases (e.g. expert test example and expert train example), whereas lower probabilities were obtained for the opposite examples (e.g. beginner test example and expert train example).

TABLE I: Match/miss-match probabilities.

		Test Data		
Exercise	Train Data	Expert	Mid-level	Beguiner
1	Expert	90	6	4
	Mid-level	6	90	4
	Beginner	2	6	92
2	Expert	88	8	4
	Mid-level	10	83	7
	Beginner	5	12	83
3	Expert	94	4	2
	Mid-level	5	91	4
	Beginner	6	14	80
4	Expert	93	5	2
	Mid-level	4	94	2
	Beginner	2	4	94
5	Expert	81	13	6
	Mid-level	12	80	8
	Beginner	2	6	92
6	Expert	90	7	3
	Mid-level	11	81	8
	Beginner	2	3	95
7	Expert	86	10	4
	Mid-level	9	85	6
	Beginner	4	10	86
8	Expert	86	10	4
	Mid-level	8	87	5
	Beginner	4	9	87
9	Expert	92	5	3
	Mid-level	7	88	5
	Beginner	0	1	99
10	Expert	91	6	3
	Mid-level	11	82	7
	Beginner	2	4	95

V. Conclusions and future work

In this study we have presented a framework to automatically assess music performance in violin. We have obtained expert, mid-level and beginner level recordings of 10 selected exercises for violin. Each recording was performed by two different musicians, therefore, each class/exercise performance was grouped into train and test sets. We have extracted low-level audio features from the audio recordings, which included pitch, spectral and energy frame-based descriptors. We have applied DTW to measure the distance between each test example with the corresponding three training examples. Later the classification probability of each class was used to map a score from zero to ten. Initial experiments show an average classification accuracy of 90% for the test set. For future work, we plan to extend this approach by obtaining ratings by music experts of specific aspects of performance over the data set of recordings, to later apply the automatic evaluation strategy to each of the performance aspects considered. Later, a feature selection optimization step will be performed for each of the performance aspects being assessed. Finally, we plan to perform a large validation of the automatic assessment within real master-apprentice scenarios.

Acknowledgment

This work has been partly sponsored by the Spanish TIN project TIMUL (TIN 2013-48152-C2-2-R), the European Union Horizon 2020 research and innovation programme under grant agreement No. 688269 (TELMI project), and the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502).

References

[1] C. Dittmar, E. Cano, J. Abeßer, and S. Grollmisch, "Music information retrieval meets music education," in *Dagstuhl Follow-Ups*, vol. 3. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012.

[2] C. Raphael, "Music plus one and machine learning," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 21–28.

[3] A. Cont, "Antescofo: Anticipatory synchronization and control of interactive parameters in computer music." in *International Computer Music Conference (ICMC)*, 2008, pp. 33–40.

[4] E. Cano, C. Dittmar, and S. Grollmisch, "Songs2see: learn to play by playing," in *12th International Society for Music Information Retrieval Conference (ISMIR 2011)*. Miami, 2011.

[5] E. Maestre and E. Gómez, "Automatic characterization of dynamics and articulation of expressive monophonic recordings," in *Proc. 118th Audio Eng. Society Convention*, 2005.

[6] T. Knight, F. Upham, and I. Fujinaga, "The potential for automatic assessment of trumpet tone quality." in *ISMIR*, 2011, pp. 573–578.

[7] O. Romani Picas, H. Parra Rodriguez, D. Dabiri, H. Tokuda, W. Hariya, K. Oishi, and X. Serra, "A real-time system for measuring sound goodness in instrumental sounds," in *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.

[8] S. Giraldo, R. Ramirez, G. Waddell, and A. Williamon, "A computational approach for measuring performance quality in violin tones," in *International Symposium in Performance Science (ISPS 2017)*, Reikiavik, Iceland, August 2017.

[9] ——, "A realtime feedback learning tool to visualize sound quality in violin performances," in *10th International Workshop on Machine Learning and Music (MML 2017)*, Barcelona, Spain, October 2017.

[10] Y.-S. Jeong, M. K. Jeong, and O. A. Omitaomu, "Weighted dynamic time warping for time series classification," *Pattern Recognition*, vol. 44, no. 9, pp. 2231–2240, 2011.

[11] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," 2004.

[12] D. Bogdanov, N. Wack, E. Gómez Gutiérrez, S. Gulati, P. Herrera Boyer, O. Mayor, G. Roma Trepat, J. Salamon, J. R. Zapata González, and X. Serra, "Essentia: An audio analysis library for music information retrieval," in *Britto A, Gouyon F, Dixon S, editors. 14th Conference of the International Society for Music Information Retrieval (ISMIR); 2013 Nov 4-8; Curitiba, Brazil.[place unknown]: ISMIR; 2013. p. 493-8.* International Society for Music Information Retrieval (ISMIR), 2013.