Towards Quantifying Differences in Expressive Piano Performances: Are Euclidean-like Distance Measures Enough?*

Carlos Cancino-Chacón[1,3], Silvan Peter[1], Emmanouil Karystinaios[1] and Gerhard Widmer[1,2]
[1]Institute of Computational Perception, Johannes Kepler University Linz, Austria
[2]LIT AI Lab, Linz Institute of Technology, Linz, Austria
[3]RITMO Centre for Interdisciplinary Studies in Time, Rhythm and Motion, University of Oslo, Norway

Abstract

In this work, we present some preliminary results about quantifying differences in expressive piano performances in the context of computational generative models.

1 Introduction

Computational models of expressive music performance can be used to generate an expressive (i.e., *human-like*) performance of a piece given its score (Cancino-Chacón et al., 2018). For models of piano performance, these generated performances would typically include deviations in *performance parameters* such as tempo/timing, dynamics and articulation. A central question with generative computational models that produce some sort of *artistic* output is how to evaluate the quality of this output. Unfortunately, large scale evaluation of such models through listening tests would be too time consuming in the context of common research practice (Bresin & Friberg, 2013).

Advances in machine learning have led to a renewed interest in generative data-driven models of expressive performance (Cancino-Chacón et al., 2017; Jeong et al., 2019; Maezawa et al., 2019). Evaluation of such models usually involves quantitatively comparing performances generated by these models to performances by (expert) human pianists. This comparison is done by measuring the *reconstruction error*, i.e., the *distance* between curves of performance parameters (tempo, dynamics, etc.) generated by the model and those of *reference* human performances, using a metric (distance measure)

*This paper is a slightly updated version of the extended abstract presented at the Rhythm Production and Perception Workshop 2021 (RPPW2021), Oslo, Norway.

such as the Euclidean distance or its derivatives (such as the mean squared error). It is easy to see that such an approach might present several issues:

1. Quantitative metrics of closeness do not necessarily imply perceptual nor musical (or aesthetic) closeness: not all errors (i.e., points/notes at which the generated performance is not identical to the human performance) would be perceived by listeners as equally important. For example, speeding up at the end of a phrase (instead of slowing down) might come as more unexpected/unmusical than simply playing the entire piece a little bit faster, yet both cases might result in the same numerical distance.
2. Choice of the reference human performance(s) to compare the output of the model to: pianists interpret music in different but *musically equally valid* ways.

In this study we investigate the limitations of reconstruction error-like functions to evaluate models of expressive piano performance using a validity/reliability framework.

2 Methods

Using two datasets of precisely measured performances of classical piano music (recorded on computer controlled grand pianos), we present tests that indicate that evaluating performance models using commonly used distance measures do not necessarily satisfy even moderate reliability and validity requirements. We compare multiple metrics (Euclidean, cosine, L_p norm, etc.), including metric learning (i.e., using data-driven methods to learn appropriate distance measures), as well as aggregating distances at different time-scales, and different transformations/scalings of the curves of performance parameters.

We complement this experiment with a listening experiment in which participants are asked to identify randomly generated (i.e., unmusical) performances from human performances. This experiment aims to validate the use of the validity/reliability framework to compare distance measures for measuring the similarity between human performances.

3 Expected Results

Data collection is ongoing. We expect the results to show the limitation of standard measures of similarity between performances and serve as a basis for discussing implications for quantitative evaluation of generative models of performance.

Furthermore, we aim to derive recommendations for the evaluation of computational models of expressive performance. Future work will include investigating the contribution and interaction of different performance parameters (tempo, dynamics, articulation, etc.) to the perception of similarity in performances.

Acknowledgements

This research has received support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 670035 (project "Con Espressione") and by the Research Council of Norway through its Centers of Excellence scheme, project number 262762 and the MIRAGE project, grant number 287152.

References

Bresin, R., & Friberg, A. (2013). Evaluation of Computer Systems for Expressive Music Performance. In A. Kirke & E. R. Miranda (Eds.), *Guide to computing for expressive music performance* (pp. 181–203). London, UK: Springer-Verlag.

Cancino-Chacón, C., Gadermaier, T., Widmer, G., & Grachten, M. (2017). An Evaluation of Linear and Non-linear Models of Expressive Dynamics in Classical Piano and Symphonic Music. *Machine Learning*, *106*(6), 887–909.

Cancino-Chacón, C., Grachten, M., Goebl, W., & Widmer, G. (2018). Computational Models of Expressive Music Performance: A Comprehensive and Critical Review. *Frontiers in Digital Humanities*, *5*, 25. Retrieved from `https://www.frontiersin.org/article/10.3389/fdigh.2018.00025` doi: 10.3389/fdigh.2018.00025

Jeong, D., Kwon, T., Kim, Y., Lee, K., & Nam, J. (2019). VirtuosoNet: A Hierarchical RNN-based System for Modeling Expressive Piano Performance. In *Proceedings of the 20th international society for music information retrieval conference (ismir 2019)* (pp. 908–9015). Delft, The Netherlands.

Maezawa, A., Yamamoto, K., & Fujishima, T. (2019). Rendering Music Performance with Interpretation Variations using Conditional Variational RNN. In *Proceedings of the 20th international society for music information retrieval conference (ismir 2019)* (p. 855-861). Delft, The Netherlands.