Predicting Dynamics in Violin Pieces with Features from Melodic Motifs

Fábio Jose Muneratti Ortega(✉), Alfonso Perez-Carrillo, and Rafael Ramírez

Music Technology Group, Machine Learning and Music Lab, Department of
Communication and Information Technology, Pompeu Fabra University,
Barcelona, Spain
fabiojose.muneratti@upf.edu

Abstract. We present a machine–learning model for predicting the performance dynamics in melodic motifs from classical pieces based on musically–meaningful features calculated from score–like symbolic representation. This model is designed to be capable of providing expressive directions to musicians within tools for expressive performance practice, and for that reason, in contrast with previous research, all modeling is done on a phrase level rather than note level. Results show the model is powerful but struggles with the generalization of predictions. The robustness of the chosen summarized representation of dynamics makes its application possible even in cases of low accuracy.

Keywords: Expressive music performance · Machine learning · Violin

1 Introduction

The development of computer models of music expression has been an active field of research for over 30 years with a wide range of approaches [9]. Most models that have been proposed by the research community share the trait of being designed with the goal of improving computer performance. Our motivation, on the other hand, is to make use of smart technologies to improve the tools available for learning to play music, and in particular, to help musicians improve their expressive performance skills. In our envisioned scenario [12], a computer system powered by a meaningful expressive performance model could give musicians expressive directions during performance or visual feedback regarding a recording based on information extracted from a musical score. As an effort to enable such scenario, this paper presents a machine learning model for predicting the dynamics of an ensemble based on high-level features extracted from a score–like symbolic representation of the musical piece. The proposed method focuses on modeling long-term dynamics variations, so as to allow a musician to follow the modeled dynamics suggestions during performance.

1.1 Related Work

Several computer models of expression have been successful in the generation of convincing performances, particularly of classical piano pieces, as could be witnessed in the RENCON competition [8]. Most recent are the automatic compositions in the context of project Magenta [7] but the nature of the model makes composition and performance inherently inseparable whereas our learning scenario primarily requires producing performances for already composed and well-established pieces. An approach more related to our own is seen in the YQX system [16], which predicts timing, dynamics and articulation variations in classical piano pieces, and in [6] where a system for predictions of ornamentations in jazz guitar melodies is described. In both cases, melodic lines play an important role, characterized by their Narmour Implication/Realization model classes [11]. In our case, the same type of information is presented to our machine–learning algorithm using a different representation based on pitch curve coefficients. We differ from both, however, by predicting phrase–level instead of note–level expression. Most applicable to our desired scenario are the models reported in [3], which, as in our case, are able to output predictions of expressive parameters based on score information, but a sensible difference is that these models use dynamics markings from the score as a starting point whereas ours seek to generate predictions without using any input indication of expression.

The final authenticated version is available online at https://doi.org/10.1007/978-3-030-43887-6_46.

2 Materials and Methods

2.1 Materials

An adequate dataset for the intended model design required a wide variety of melodic themes in both audio and synchronized symbolic representation. Corpi of solo piano pieces such as the MAESTRO [7] were not optimal for the problem since we were interested in mapping the relationship between melody and harmony in the modeled features, and these elements tend to be fully blended in piano parts. The MusicNet dataset [14] provides audio–to–score synchronization as well as the necessary melodic diversity and still allows a clear distinction between main melodic lines and harmonies thanks to the abundance of chamber music pieces with individual instrument parts, and was thus chosen for the task. To distinguish the main melody from harmony, violin parts were treated as melodies and all other instruments, as harmony. Only the subset of pieces which contained a violin were used, resulting in 122 pieces and a total of 874 minutes of recordings. For estimating the dynamics performed by the ensembles, the momentary loudness in windows of 0.1s according to the EBU R128 standard [4] was computed with the help of the Essentia library [1].

2.2 Methods

The designed model consists of a feed–forward neural network trained to predict the dynamics curve of a musical *motif*, that is, a short phrase of roughly one or two bars. An important aspect of the modeling is that each training instance represents a motif rather than a single note. This design decision is motivated by two beliefs: first, that musicians plan and execute their expressive movements considering a horizon of a few notes rather than momentarily focusing on each one; and second, that in our music learning scenario, performance suggestions based on model outputs can be best visualized and interpreted in that level of granularity. As a consequence of choosing to train the model on motifs, it is necessary to determine musically-relevant motif boundaries in the pieces as well as appropriate features for this representation. The motif boundary detection is done by applying the LBDM [2] algorithm to estimate boundary probabilities, and recursively dividing the piece until one of two conditions is met: either no boundary probability is two standard deviations larger that the rest, or the resulting segment has fewer than 10 notes. Table 1 summarizes the input features used for training. Piece keys and modes were estimated from pitch profiles as detailed in [13]. The output features of the model should represent the dynamics of the motif and its variation on time. We have summarized that information by approximating the performance loudness curve extracted with Essentia by a parabola, fit using the least-squares method. This is consistent with the observation by Todd [10] and other researchers [5,15] that dynamics variations tend to follow a quadratic profile. Given this approximation, the task of the neural network is optimizing the three coefficients that define the dynamics curve.

To facilitate the optimization task, some data conditioning was performed. Loudness measurements of each piece were normalized to zero mean and unit variance to eliminate differences caused by inconsistent recording conditions. Motifs with less than 4 notes and outliers (z-score above 10 in any feature) were discarded, all nominal features were converted to “one-hot” format and all numeric features were standardized. The resulting dataset had around 10.000 instances, which were divided into training and test sets containing 90% and 10% of instances, respectively.

The feed-forward network was programmed in the PyTorch[1] framework and built with two hidden layers of 25 nodes each, using ReLU as an activation function and standard mean-squared error as a loss function. The training was run for 1800 epochs in stochastic gradient descent optimization with batches of 100 instances, learning rate of 0.2 and momentum of 0.1. The learning rate was decreased by a factor of 10 every 600 epochs. All parameters were cross-validated using a subdivision of the training set prior to the final training round.

3 Results and Discussion

3.1 Results

Table 2 shows the obtained correlation coefficients for each of the dynamics coefficients predicted for all instances in the test set. Examples of the loudness curve, ground-truth quadratic approximation and predicted curve for three motifs can be seen in figure 1. Table 3 provides some perspective on the accuracy of the modeled dynamics by indicating root-mean-square errors for a deadpan prediction, for the ground-truth approximations, and for the model’s output.

[1] http://pytorch.org

Table 1. Input features of the model.

Feature	Data type	Description
Beat in Measure	$x \in [0, 4]$	The beat where the motif begins
Metric strength	$x \in \{3, 2, 1, 0\}$	How strong the start beat is e.g.: down beat = 3
Number of notes	$x \in \mathbf{N}$	Total of notes in motif
Duration	$x \in [0, \infty)$	Motif duration in beats
Location in piece	$x \in [0, 1]$	Where in the piece the motif is played
Pitch curve coefficients	$x_0, x_1, x_2 \in \mathbf{R}$	Quadratic coefficients approximating the MIDI pitches of motif notes
Pitch contour coefficients	$x_0, x_1, x_2 \in \mathbf{R}$	Quadratic coefficients approximating the variation in MIDI pitches of motif notes
Rhythm Drops	Boolean	Whether a note with higher duration follows another with shorter duration in the motif
Rhythm Rises	Boolean	Whether a note with shorter duration follows another with higher duration in the motif
Rhythm contour coefficients	$x_0, x_1, x_2 \in \mathbf{R}$	Quadratic coefficients approximating the variation in duration of motif notes
Strongest note location	$x \in [0, 1]$	Where in the motif is the note with highest metric strength.
Piece key	A - G#	Tonality estimation of motif piece
Piece mode	Major/Minor	Mode estimation of motif piece
Chord probabilities	$x_0..x_6 \in [0, 1]$	Estimated diatonic chords presence probabilities
Initial chord degree	I - VII	Most likely chord in motif start
Final chord degree	I - VII	Most likely chord in motif end
Has Dissonance	Boolean	Whether there are notes from a different tonality
Dissonance Location	$x \in [0, 1]$	Location of first occurence of dissonant note
Is solo piece	Boolean	Solo or ensemble piece

Table 2. Correlation coefficients for output features.

Output coefficient	Pearson's r (test set)	Pearson's r (training set)
x^2	0.2177	0.5222
x^1	0.2375	0.7054
x^0	0.2383	0.6818

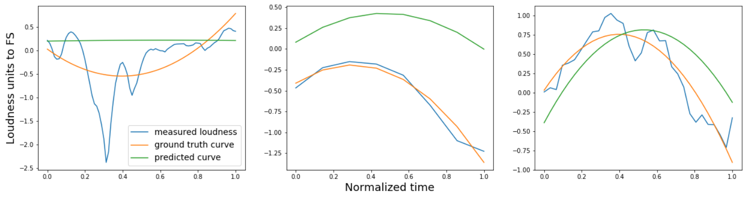

Fig. 1. Comparison of loudness values measured in performance, their ideal (ground-truth) approximation, and model output for three motifs.

3.2 Discussion

The Pearson correlation coefficients obtained for the training set show that the model is sufficiently powerful to predict the complex relationships present in this scenario, but the lower correlation values seen in the test set indicate that some overfitting occurred, and the meaningful correlations detected in the data only partially explain the observed dynamics. The deadpan-level (E_d) and ground-truth-level (E_g) errors in table 3 can be seen as lower and upper boundaries of accuracy, indicating that this modeling approach offers a potential reduction in prediction error of up to $E_d - E_g = 2.17$ dB. The 3.39 dB value obtained with our predictions implies an error reduction of $E_d - 3.39 = 0.47$ dB compared to the deadpan baseline, which corresponds to $0.47/2.17 = 21.65\%$ of the predicted potential. That is consistent with the correlation coefficient values and shows that the prediction of coefficients translates well into prediction of dynamics levels.

Table 3. RMS error in loudness levels prediction.

Prediction type	Error level
Deadpan performance	3.86 dB
Ground-truth approximation	1.69 dB
Model prediction	3.39 dB

The prediction examples highlighted in figure 1 illustrate some relevant conclusions: It can be seen that most of the short–term variation in loudness levels occurs on note boundaries due to note articulation, and in terms of perceived dynamics can be understood as noise. The quadratic approximation (labeled ground-truth) provides a cleaner and more intuitive visualization of the variation of loudness in a phrase, and in most individually–inspected cases represents it quite well. The leftmost example is an exception, as it shows a case in which the phrase boundaries chosen by the algorithm don't seem to match the performer's choice, hence the silence during the phrase and the poor results even in the proposed ground-truth approximation. In many observed cases, as shown in the middle and rightmost graphs, the predicted curve shows robustness, especially with relation to the x^2 and x^1 coefficients, since some variation in their predictions doesn't affect the character of the interpretation. It is reasonable to assume that despite the difference between ground-truth and predicted values in such cases, performances executed according to instructions from the latter could be considered just as pleasing.

Logical improvements to our proposed approach under consideration include adding information that relates to the repetition of motifs, detecting key modulations or modal harmony in pieces, augmenting the training set with multiple different divisions of motifs per piece and experimenting with different treatments of time–series data such as training with long short–term memory networks.

References

1. Bogdanov, D., Wack, N., Gómez Gutiérrez, E., Gulati, S., Herrera Boyer, P., Mayor, O., Roma Trepat, G., Salamon, J., Zapata González, J.R., Serra, X.: Essentia: An audio analysis library for music information retrieval. In: 14th Conference of the International Society for Music Information Retrieval (ISMIR). International Society for Music Information Retrieval (ISMIR) (2013)
2. Cambouropoulos, E.: The local boundary detection model (LBDM) and its application in the study of expressive timing. In: Proceedings of the International Computer Music Conference ICMC01. pp. 232–235. Havana, Cuba (2001)
3. Cancino-Chacón, C.E.: Computational Modeling of Expressive Music Performance with Linear and Non-linear Basis Function Models. Ph.D. thesis, Johannes Kepler University Linz (2018)
4. EBU TC Committee: Tech 3341: Loudness metering: 'EBU mode' metering to supplement EBU R 128 loudness normalization. Tech. rep., EBU, Geneva (2016)
5. Gabrielsson, A., Bengtsson, I., Gabrielsson, B.: Performance of musical rhythm in 3/4 and 6/8 meter. Scandinavian Journal of Psychology **24**(1), 193–213 (1983). https://doi.org/10.1111/j.1467-9450.1983.tb00491.x
6. Giraldo, S.I., Ramirez, R.: A machine learning approach to discover rules for expressive performance actions in jazz guitar music. Frontiers in Psychology **7**(DEC), 1965 (12 2016). https://doi.org/10.3389/fpsyg.2016.01965
7. Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C.Z.A., Dieleman, S., Elsen, E., Engel, J., Eck, D.: Enabling factorized piano music modeling and generation with the MAESTRO dataset. In: International Conference on Learning Representations (2019)

8. Katayose, H., Hashida, M., De Poli, G., Hirata, K.: On evaluating systems for generating expressive music performance: the Rencon experience. Journal of New Music Research **41**(4), 299–310 (2012). https://doi.org/10.1080/09298215.2012.745579
9. Kirke, A., Miranda, E.R.: An Overview of Computer Systems for Expressive Music Performance. In: Guide to Computing for Expressive Music Performance, pp. 1–47. Springer London, London (2013). https://doi.org/10.1007/978-1-4471-4123-5_1
10. McAngus Todd, N.P.: The dynamics of dynamics: A model of musical expression. The Journal of the Acoustical Society of America **91**(6), 3540–3550 (1992). https://doi.org/10.1121/1.402843
11. Narmour, E.: The analysis and cognition of melodic complexity: The implication-realization model. University of Chicago Press (1992)
12. Ramirez, R., Ortega, F.J.M., Giraldo, S.I.: Technology enhanced learning of expressive music performance. In: Proceedings of the 16th Brazilian Symposium on Computer Music (2017)
13. Temperley, D.: What's Key for Key? The Krumhansl-Schmuckler Key-Finding Algorithm Reconsidered. Music Perception: An Interdisciplinary Journal **17**(1), 65–100 (oct 1999). https://doi.org/10.2307/40285812
14. Thickstun, J., Harchaoui, Z., Foster, D.P., Kakade, S.M.: Invariances and data augmentation for supervised music transcription. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2018)
15. Tobudic, A., Widmer, G.: Relational IBL in music with a new structural similarity measure. In: Proceedings of the 13th International Conference on Inductive Logic Programming. pp. 365–382 (2003)
16. Widmer, G., Flossmann, S., Grachten, M.: YQX Plays Chopin. AI Magazine **30**(3), 35 (2009). https://doi.org/10.1609/aimag.v30i3.2249