Phrase-Level modeling of expression in violin performances

Fábio J. M. Ortega, Sergio I. Giraldo, 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. A model is proposed for predicting expressive variations in dynamics for violin performances with the purpose of facilitating expressive performance learning by students. The model uses phrases rather than single notes as units of analysis in a lazy learning approach: each phrase in a new score is matched to phrases from expressive performances by experts, adapting the experts' transformations to render an expressive performance of the new score. In preliminary tests, the model approximates the dynamics of actual performances better than an unexpressive baseline model whenever the reference dataset contains melodies similar to those being predicted.

Keywords: expressive music performance, machine learning, music information retrieval

1 Introduction

Expression in music can be understood as the variations in timing, dynamics, pitch, timbre, and other features introduced by musicians as they play. Teaching musical expression traditionally relies on the continuous feedback that a face–to–face setting can provide [1]. When practicing an instrument on one's own, the absence of expert supervision makes acquiring this skill much harder, leading to frustration and high abandonment rates among students [2,3]. If, however, the information about how to play expressively could be generalized by a model based on some large set of recordings by professional musicians, a system could be devised that would be able to provide real–time feedback to students practicing any piece, even if no sample performance of it exists.

In this paper we propose a model for predicting dynamics for violin performances which mimics the process by which a musician would choose to interpret a melody based on their memory of a similar one. Our aim is to determine whether an automatic recognition of phrasing and melodic content present in a score can be used for selecting adequate examples of performance, and, if so, whether having these examples is enough to generate a plausible rendition of a piece.

2 Background

Several models of musical expression have been proposed for various instruments and purposes [4]. We highlight the most relevant comparisons. The *DISTALL* system [5], though designed for piano, can also produce dynamics predictions based on phrase–level analysis of performances, though they define phrasing hierarchically whereas we only focus on short *motivs*. Also, since the musicological analysis of scores is reportedly manual, it does not fit our requirement for being used in an expressive tutor system. Ramirez *et al.* [6] design a model for jazz saxophone that produces performance rules based on data via genetic algorithms. Besides focusing on note–level instead of phrase–level predictions, their approach is different by being rooted in classification (e.g. *piano*, *mezzo–forte*, *fortissimo*), with numerical values for synthesizing audio resulting from an *a posteriori* approximation. These same remarks apply to the model by Giraldo and Ramirez [7] for jazz guitar. Lastly, the basis–functions approach by Grachten and Widmer [8] relies on expressive markings in scores and their interpretation, whereas our model does not require annotated scores and applies very little musicological knowledge. Furthermore, all the discussed models assume a previous training step often very time–consuming before producing performances, whereas we are interested in taking a lazy learning approach that can be used to our favor for selecting the most relevant references for each performance prediction.

3 Materials and Methods

All data used in development come from recordings which were made as part of experiments on ensemble expressive performance [9,10]. A dynamics curve was calculated from the audio extracted from the pickup of the first violin in a performance of the fourth movement of Beethoven's String Quartet no. 4, Op. 18, purposely exaggerated in its expressiveness.

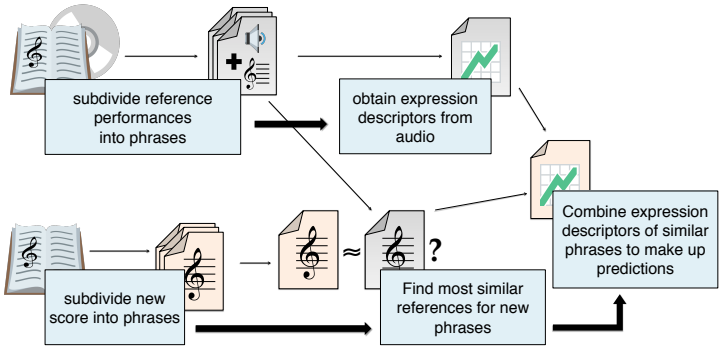

Fig. 1. Diagram of the steps taken to predict the dynamics for a new score.

The method for generating a prediction is depicted in figure 1. First, the score of the target piece is automatically segmented into phrases in a top–down approach based on the method by Cambouropoulos [11]. Then, ratings of melodic similarity are computed between each phrase from the score and every phrase in the database of references. A dynamic time warping algorithm is used for determining the degree of similarity between two phrases as proposed by Stammen and Pennycook [12]. The warping cost between phrases is interpreted as the distance between them, and the cost function takes pitch contour and note duration ratios into consideration. Predicted dynamics for each phrase may then be computed based on its closest matches.

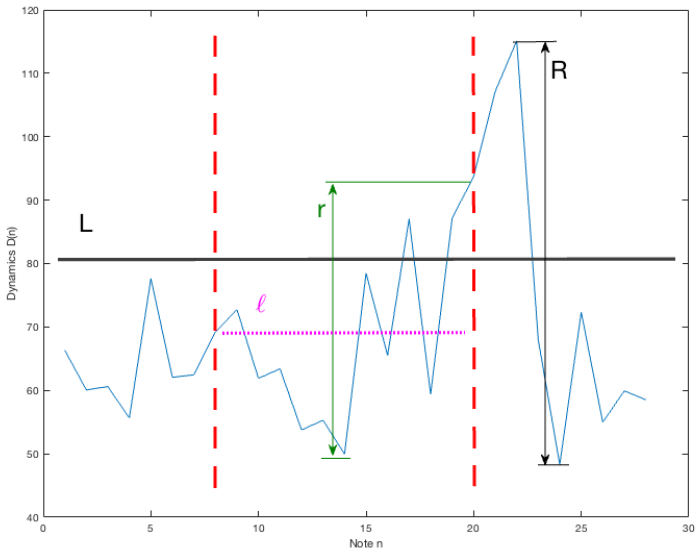

Fig. 2. Performed dynamics for a section of a piece and some key measurements.

Figure 2 represents a dynamics curve plot note by note where the dynamics at each note n is $D(n)$. Dynamics values are obtained as the logarithm of RMS values of audio samples within the (manually segmented) duration of each note, adjusted to the $0 - 127$ scale commonly applied to MIDI velocities. Between dashed lines is the section of a particular phrase in that piece. L is the mean level of the piece, whereas ℓ is the mean level of the phrase. The dynamic range is given by R for the piece and r for the phrase and we use the standard deviation of loudness values as their measure. Considering that pieces may be performed at widely different mean levels and dynamic ranges, if we intend to use phrases from multiple pieces as references for prediction it makes sense to measure their values relative to L and R and allow these to be set by the user for the predicted

rendition. Therefore, the characterization of each phrase p in our model is given by three components:

$$\alpha_p = \frac{\ell_p - L}{R} \quad \beta_p = \frac{r_p}{R} \quad \Gamma_p(n) = \frac{D(n) - \ell_p}{r_p}$$

Where α_p represents the overall salience of p in relation to the piece, β_p is the relative dynamic range of the phrase, and Γ_p represents the relative dynamics contour, that is, a function which describes how each note in a phrase contributes to its dynamics. Consequently, the dynamics at each note $k \in p$ can always be written as:

$$D(k) = L + R \cdot (\alpha_p + \beta_p \cdot \Gamma_p(k))$$

These three components are measured for each reference phrase and make up the target variables for our learning step. By predicting α, β and Γ for all phrases of a target score, the above equation gives us the output prediction for freely chosen values of L and R.

4 Results

A preliminary analysis was conducted using a leave–one–phrase–out setting on the data from the Beethoven recording. Figure 3 shows the distributions of mean absolute error for each note using k–NN ($k = 1$), k–NN ($k = 1$) predicting Γ as a quadratic polynomial, and k–NN ($k = 3$). For this last case, target variables take the mean values of the three nearest neighbors. The baseline is a mechanical, unexpressive prediction. From the plot it is visible that the quadratic polynomial was effective as an aproximation, and that k–NN ($k = 3$) is successful in reducing the instances with larger error values while maintaining mean absolute error MAE = 19.6 (15.44% of full–scale), which is lower than the baseline ($p = 1.56 \times 10^{-6}$ for one–sided t–test). Though it performs better than baseline, it should be noted that this is an advantageous case for the model, since the available reference phrases were part of the same performance.

In order to validate the hypothesis that phrases rated as melodically similar share similar dynamic profiles, we split the phrases predicted using k–NN ($k = 1$) in half, separating phrases with *closest* nearest neighbors in the training set from phrases with *farthest* ones. The mean absolute error in dynamics prediction for the phrases with closest neighbors is 8.83% of full–scale on average, whereas phrases with farthest neighbors show 15.77 %FS error on average, meaning predictions were more accurate for phrases which had instances in training data with a higher melodic similarity to them. This is an indication that the adopted measure of melodic similarity can be used as a predictor for performance dynamics, and also that given a larger number of performance examples, the model has a good margin for improving the quality of its predictions, since a wider range of melodies would be available as references.

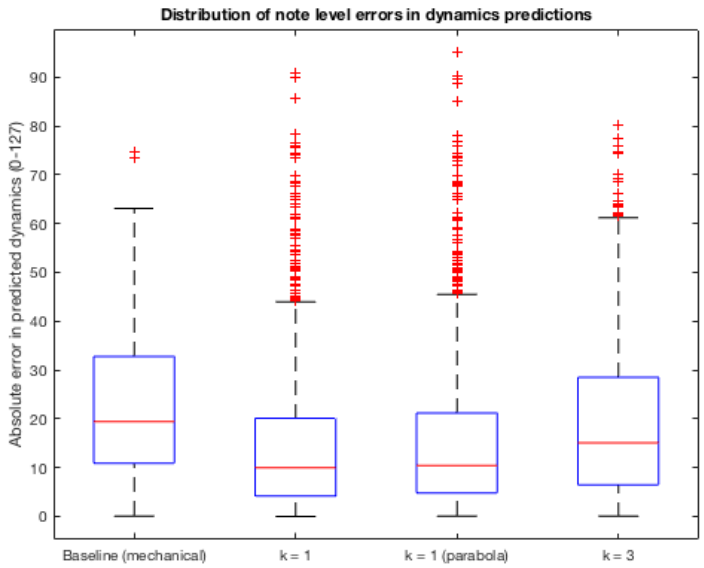

Fig. 3. Boxplot of the prediction errors in a leave–a–phrase–out approach vs. baseline.

5 Conclusions

We have devised a model for expressive dynamics prediction of solo violin performances based solely on the adaptation of performances of similar melodies. Both the proposed approach to characterizing expressive deviations in dynamics and the adopted measure of melodic similarity have been evaluated favourably, though with a limited dataset. Given the model's instance–based nature and the generality of its musicological assumptions, it should also be applicable to performances of other instruments. Furthermore, our data treatment for interpreting the contributions of each phrase to the expression of a musical piece may benefit other models as well. As previous models have shown, including other musical aspects deducible from the scores such as metrical strength and harmonic content should improve the quality of results. Also, an indirect dependency exists between generated predictions and phrasing boundaries, so including multiple phrasing interpretations could be an advantageous trait. As a logical next step, the authors are now preparing a perceptual evaluation of the predictions made by the model to verify if they sound pleasant to listeners.

Acknowledgements

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. Woody, R.H.: The effect of various instructional conditions on expressive music performance. J. Res. Music Educ. 54(1), 21–36 (2006), http://jrm.sagepub.com/cgi/doi/10.1177/002242940605400103
2. Covington, M.V.: The self-worth theory of achievement motivation: Findings and implications. Elem. Sch. J. 85, 4–20 (1984)
3. Juslin, P.N., Karlsson, J., Lindstrm, E., Friberg, A., Schoonderwaldt, E.: Play it again with feeling: computer feedback in musical communication of emotions. J. Exp. Psychol. Appl. 12, 79–95 (2006)
4. Kirke, A., Miranda, E.R.: A survey of computer systems for expressive music performance. ACM Computing Surveys 42, 1–41 (2009)
5. 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)
6. Ramirez, R., Hazan, A., Maestre, E., Serra, X.: A genetic rule-based model of expressive performance for jazz saxophone. Comput. Music J. 32, 38–50 (2008)
7. Giraldo, S.I., Ramirez, R.: A machine learning approach to discover rules for expressive performance actions in jazz guitar music. Front. Psychol. 7, 1965 (2016)
8. Grachten, M., Widmer, G.: Linear basis models for prediction and analysis of musical expression. J. New Music Res. 41, 311–322 (2012)
9. Marchini, M., Ramirez, R., Papiotis, P., Maestre, E.: The sense of ensemble: a machine learning approach to expressive performance modelling in string quartets. J. New Music Res. 43, 303–317 (2014)
10. Papiotis, P., Marchini, M., Perez-Carrillo, A., Maestre, E.: Measuring ensemble interdependence in a string quartet through analysis of multidimensional performance data. Front. Psychol. 5, 963 (2014)
11. 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 ICMC. pp. 17–22 (2001)
12. Stammen, D.R., Pennycook, B.: Real-time recognition of melodic fragments using the dynamic timewarp algorithm. In: Proceedings of the International Computer Music Conference ICMC. pp. 232–5 (1993)