Automatic Recognition of Famous Artists by Machine

Gerhard Widmer[1] and **Patrick Zanon**[2]

1 INTRODUCTION

The work presented here is part of a large investigation into the use of computational methods for studying basic principles of expressive music performance [3, 2]. One of the questions we study is whether and to what extent aspects of *individual artistic style* can be quantified. And one of the possible approaches to this question is to investigate whether machines can learn to distinguish and recognise different performers based on their style of playing.

This short paper describes our latest results along these lines. Learning algorithms are applied to the task of identifying the performer in audio recordings, by famous pianists, of several Mozart piano sonatas. It is shown that in pair-wise discrimination settings good results can be obtained, and these results partly carry over to music of a very different style. This is quite surprising, given the very limited information contained in the available measurement data.

2 DATA AND METHODOLOGY

Expressive music performance is the art of shaping a musical piece by continuously varying important parameters like tempo, dynamics, etc., particularly in classical music. The expressive nuances added by an artist are what makes a piece of music come alive, and what distinguishes great artists from each other. One of the goals of the research presented here is to use AI techniques to get a better understanding of what factors really contribute to personal artistic style, by applying machine learning methods to real performance data. In the following, we focus on the two most important expressive dimensions, at a rather coarse level: changes of *tempo* and *dynamics* (i.e., loudness).

For the experiments, commercial recordings by six renowned concert pianists (Daniel Barenboim (DB), Roland Batik (RB), Glenn Gould (GG), Maria João Pires (MP), András Schiff (AS), and Mitsuko Uchida (MU)) of piano sonatas by W.A. Mozart were collected, and a sizeable number of pieces (12 complete sonata movements from the sonatas K.279, 280, 281, 282, 330, and 332) were selected for performance measuring and analysis.

From the audio recordings, rough measurements characterising the performances were obtained. Changes of *tempo* and *general loudness* were measured at the level of the beats, by determining and marking the precise onset time of each beat, e.g., each 8th note position in a piece written in 6/8 time. From the varying time intervals between successive beats, the beat-level tempo changes can be derived. Overall loudness of the performance at these time points was extracted from the audio signal and is taken as a very crude representation of the dynamics applied by the pianists. No more detailed information (e.g., about articulation, individual voices, or timing details below the level of the beat) is available.

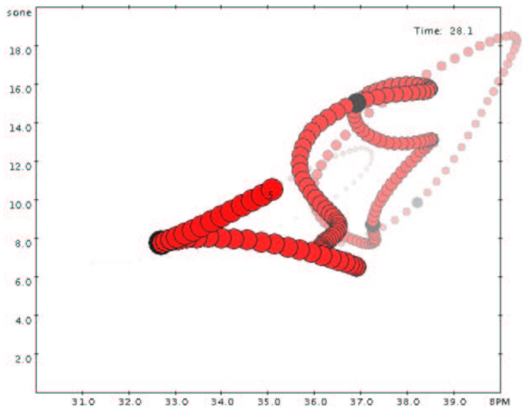

Figure 1. First four bars of Daniel Barenboim's performance of Mozart's F major sonata K.332, 2nd movement. Horizontal axis: tempo in beats per minute; vertical axis: loudness in *sone*.

These sequences of measurements can be represented as two sets of performance curves — one representing variations in beat-level tempo over time, the other beat-level loudness changes — or in an integrated two-dimensional way, as trajectories over time in a 2D tempo-loudness space [1]. A graphical animation tool called the *Performance Worm* displays such performance trajectories in synchrony with the music. A part of a performance as visualised by the Worm is shown in Figure 1.

The raw data for our experiments is thus local tempo values T_i and loudness values L_i measured at specific time points i in a performance. Each measured time point, along with its context, is used as a *training example*. In other words, an example or *instance* for the learners is a subsegment of a tempo-loudness trajectory (see Fig. 1), centered around a specific time point.

The instances are represented by a set of *features* that are calculated over a window w_i of two bars, centered on the time point of the instance. The following features are computed both for tempo and loudness: the average value within the window $\mu(w_i)$, the standard deviation $\sigma(w_i)$, and the range $R(w_i) = max(w_i) - min(w_i)$. For each of these features, mean-normalised versions are also computed. Additional features include derivatives of tempo and loudness, correlations between tempo and loudness, and a feature called *directness* that captures the curvature of a trajectory segment.

Some of these features were actually excluded from the learning experiments. Features relating to absolute loudness could trivially reveal the performer via the absolute recording level of the records. Care was taken to include only features that only relate to performance style. More on this can be found in a longer version of this paper [4]. We eventually ended up with a data set of some 23,000 instances for all the six pianists, described by 20 numeric attributes.

[1] Department of Medical Cybernetics and AI, Medical University of Vienna; Email: gerhard@ai.univie.ac.at
[2] Austrian Research Institute for Artificial Intelligence, Vienna

For the experiments, the original n-class problem was converted to $n(n-1)/2$ two-class discrimination problems, one for each possible pair of pianists. For each pair A-B, the performances by the two pianists of the selected training pieces were used for learning, and the task was then to identify the correct pianist in a new test piece, where only recordings by A and B were used for testing. Recognition accuracy was tested via *cross-validation* at the level of sonata movements: the learner was trained on all of the sonata movements except one, and the learned classifier was then tested on recordings of the remaining movement. This was repeated in a circular fashion, so that each piece served as test piece exactly once for each classifier.[3]

3 RESULTS

A variety of learning algorithms were tested on this data set. In the following, we only report on the results of *logistic regression* as implemented in the WEKA toolbox [5]. Detailed results for all other algorithms can be found in [4]. Table 1 shows the numbers and percentages of correct predictions achieved by logistic regression. For each pair of performers, the classifier was tested on 24 recordings (12 pieces, played by each of the two pianists). Thus, the maximum possible number of correct predictions is 24, and the *baseline* — the success rate corresponding to pure guessing — is 12, or 50%.

Essentially, we see that the machine manages to distinguish all pianists (with the exception of the pair Pires-Uchida (MP-MU)) at a level way above the baseline. Recognition rates of 75 or 80% or higher are quite impressive, given that we only extract very high-level performance information from the recordings. That indicates that our performance features do contain information about the style of the artists. Again, a more detailed discussion that gives more insight into what the results reveal about the pianists is given in [4].

The above results can be improved significantly if we slightly change the classification scenario. Remember that classification of a piece is done by classifying all the instances (time points) that make up the piece, and predicting the majority class. The *ratio* of votes for class A vs. class B, over all instances, is thus a measure of the relative *confidence* in a prediction. This can be exploited in what might be called *'closed-world classification'*, where the classifier is always given a *pair* of recordings and is told that one is by pianist A and the other by pianist B. In such a case, the learner first makes the prediction of which it is more confident, and then predicts the opposite class for the other recording. The results achieved by this procedure (Table 2) are dramatically better, with the classifier achieving perfect identification rates for more than half of the pianist pairs. The average number of correct predictions per pair is 22.67 or **94.4%**.

An interesting question is how general the induced classifiers are relative to *different styles* of music. We happen to also have measured recordings, by two of the pianists considered above, of pieces by the Romantic composer *Frédéric Chopin*: the *Nocturnes* op.9/2, op.15/1, op.27/1, and op.27 No.2. The two pianists are Barenboim (DB) and Pires (MP). The four pieces were segmented into sections of different musical character, giving 11 sections overall. Thus we have 22 test cases: 11 Chopin pieces played by 2 pianists.

When testing the DB-MP classifier learned from Mozart on these Chopin recordings, we still achieve a recognition rate of $15/22 =$ **68.2%**. And again, the results improve considerably when we consider a 'closed-world' classification scenario (see above): here the recognition rates rise to an astounding $18/22 =$ **81.1%**. This is quite remarkable, given the differences between the training corpus (Mozart) and the test pieces (Chopin). It shows that the machine seems to capture something significant about artists' musical style.

[3] Note that, as explained above, the actual training examples are not entire pieces, but individual time points in pieces, characterised by a set of features. To make a classifier predict the pianist for a complete piece, it was applied to all instances making up the piece, and the class (pianist) predicted most often was then chosen as the final prediction.

Table 1. Pairwise recognition results: classification accuracy in terms of correctly classified pieces.

Pair	hits (%)	Pair	hits %)	Pair	hits (%)
AS-DB	16 (66.7)	DB-GG	18 (75.0)	GG-MU	18 (75.0)
AS-GG	17 (70.8)	DB-MP	15 (62.5)	GG-RB	17 (70.8)
AS-MP	16 (66.7)	DB-MU	15 (62.5)	MP-MU	13 (54.2)
AS-MU	16 (66.7)	DB-RB	19 (79.2)	MP-RB	20 (83.3)
AS-RB	17 (70.8)	GG-MP	17 (70.8)	MU-RB	17 (70.8)

Table 2. 'Closed-world' classification. Perfect results are printed in bold.

Pair	hits	Pair	hits	Pair	hits
AS-DB	22	DB-GG	**24**	GG-MU	**24**
AS-GG	20	DB-MP	**24**	GG-RB	22
AS-MP	**24**	DB-MU	22	MP-MU	22
AS-MU	18	DB-RB	**24**	MP-RB	**24**
AS-RB	22	GG-MP	**24**	MU-RB	**24**

4 CONCLUSION

This paper has presented experimental evidence that machines may be capable of recognising famous artists on the basis of their style, to a certain extent. Given the very limited information contained in our performance measurements (only beat-level changes in tempo and total loudness), the results are actually quite surprising. We believe it would be hard for most human listeners to achieve such recognition rates under comparable conditions.

Future work will focus on an analysis of the learned models to gain more insight into relevant aspects of artists' personal style. We are also currently conducting a parallel investigation using a different representation extracted from performance trajectories — performance strings — and applying new string kernel techniques. Preliminary experiments show that further improvements are possible.

ACKNOWLEDGEMENTS

This research was supported by the Austrian *FWF* (grant no. Y99-INF), an ERASMUS scholarship to the second author, and the EU Project SIMAC (6th FP, 507142). The Austrian Research Institute for Artificial Intelligence acknowledges basic financial support by the Austrian Federal Ministries BMBWK and BMVIT.

REFERENCES

[1] J. Langner and W. Goebl, 'Visualizing expressive performance in tempo-loudness space', *Computer Music Journal*, **27**(4), 69–83, (2003).

[2] G. Widmer, 'Discovering simple rules in complex data: A meta-learning algorithm and some surprising musical discoveries', *Artificial Intelligence*, **146**(2), 129–148, (2003).

[3] G. Widmer, S. Dixon, W. Goebl, E. Pampalk, and A. Tobudic, 'In Search of the Horowitz Factor', *AI Magazine*, **24**(3), 111–130, (2003).

[4] G. Widmer and P. Zanon, *Automatic Recognition of Famous Artists by Machine*, Technical Report OeFAI-TR-2004-04, Austrian Research Institute for Artificial Intelligence, Vienna (http://www.ai.univie.ac.at/cgi-bin/tr-online?number+2004-04), 2004.

[5] I. H. Witten and E. Frank, *Data Mining*, Morgan Kaufmann, San Francisco, CA, 1999.