WHO IS WHO IN THE END? RECOGNIZING PIANISTS BY THEIR FINAL RITARDANDI

Maarten Grachten
Dept. of Computational Perception
Johannes Kepler University, Linz, Austria
maarten.grachten@jku.at

Gerhard Widmer
Austrian Research Institute for
Artificial Intelligence, Vienna, Austria

Dept. of Computational Perception
Johannes Kepler University, Linz, Austria
gerhard.widmer@jku.at

ABSTRACT

The performance of music usually involves a great deal of interpretation by the musician. In classical music, final ritardandi are emblematic for the expressive aspect of music performance. In this paper we investigate to what degree individual performance style has an effect on the form of final ritardandi. To this end we look at interonset-interval deviations from a performance norm. We define a criterion for filtering out deviations that are likely to be due to measurement error. Using a machine-learning classifier, we evaluate an automatic pairwise pianist identification task as an initial assessment of the suitability of the filtered data for characterizing the individual playing style of pianists. The results indicate that in spite of an extremely reduced data representation, pianists can often be identified with accuracy significantly above baseline.

1. INTRODUCTION AND RELATED WORK

The performance of music usually involves a great deal of interpretation by the musician. This is particularly true of piano music from the romantic period, where performances are characterized by large fluctuations of tempo and dynamics. The expressive interpretation of the music by the musician is crucial for listeners to understand emotional and structural aspects of the music (such as voice and phrase structure) [1–3]. In addition to these functional aspects of expressive music performance, there is undeniably an aspect of personal style. Skilled musicians tend to develop an individual way of performing, by means of which they give the music a unique aesthetic quality (a notable example of this is the legendary pianist Glenn Gould). Although the main focus in music performance research has been on functional aspects of expression, some studies also deal with individual performance style. Through analysis of listeners ratings on performances, Repp characterized pianists in terms of factors that were mapped to adjective pairs [4]. In [5], a principal component analysis of timing curves revealed a small set of significant components that seem to represent performance strategies that performers combine in their performances. Furthermore, a machine learning approach to performer identification has been proposed by Stamatatos and Widmer [6], where performers are characterized by a set of features relating to score-related patterns in timing, dynamics and articulation. Saunders et al. [7] represent patterns in timing and dynamics jointly as strings of characters, and use string-kernel classifiers to identify performers.

It is generally acknowledged in music performance research that, although widely used, the mechanical performance (implying constant tempo throughout a piece or musical part) is not an adequate performance norm for studying expressive timing, as it is not the way we generally believe the music should sound. As an alternative, models of expressive timing could be used, as argued in [8]. However, only few models exist that model expressive timing in general [9, 10]. Because of the complexity and heterogeneity of expressive timing, most models only describe specific phenomena, such as the timing of grace notes [11], or the final ritardando [12, 13].

This paper addresses systematic differences in the performance of final ritardandi by different pianists. In a previous study [14] on the performance of final ritardandi, a kinetic model [13] was fitted to a set of performances. Although in some cases systematic differences were found between pianists, in general the model parameters (describing the curvature and depth of the ritardando) tend to reflect primarily aspects of the piece, rather than the individual style of the pianist. Given this result, a possible approach to study performer-specific timing in ritardandi would be by subtracting the fitted model from the modeled timing data and looking performer-specific patterns in the residuals. A problem with this approach is that the kinetic model is arguably too simple, since it models tempo as a function of score time only, and is ignorant of any structural aspects of the music, which also have an effect of the tempo curve [15]. As a result of this, residuals in the data with respect to the fitted model are likely to contain patterns related to piece-specific aspects like rhythmic grouping.

In this study, in order to minimize the amount of piece-specific information present in the residuals, we compute the average performance per piece and subtract it from each performance of that piece. In addition to this, we filter the residual data based on an estimation of its significance. This estimation is obtained from an analysis of data annotation divergences for a subset of the data. The resulting data contain the deviations from the common way of playing the ritardandi that are unlikely to be due to measurement errors.

Our long-term goal is to develop a thorough and sensible way of interpreting deviations of performance data with respect to some *performance norm*, be it either a model, or as in this study, a norm derived from the data. To obtain a first impression of the potential of characterizing artists by this method of analyzing the data, we defined a pairwise pianist identification task (as in [6]). Using a data set consisting of performances of ritardandi in Chopin's Nocturnes by a number of famous pianists, we show that pianists can be identified based on regularities in the way they deviate from the performance norm.

In section 2, we describe the acquisition and content of the data set. Section 3 documents the data processing procedure. Results of the pianist classification task are presented and discussed in section 4, and conclusions and future work in section 5.

2. DATA

The data used here consists in measurements of timing data of musical performances taken from commercial CD recordings of Chopin's Nocturnes. The contents of the data set are specified in table 1. We have chosen Chopin's Nocturnes since they exemplify classical piano music from the romantic period, a genre which is characterized by the prominent role of expressive interpretation in terms of tempo and dynamics. Furthermore, the music is part of a well-known repertoire, performed by many pianists, facilitating large scale studies.

Tempo in music is usually estimated from the interonset intervals of successive events. A problematic aspect of this is that when a musical passage contains few events, the obtained tempo information is sparse, and possibly unreliable, thus not very suitable for studying tempo. Therefore, through inspection of the score, we selected those Nocturnes whose final passages have a relatively high note density, and are more or less homogeneous in terms of rhythm. In two cases (Op. 9 nr. 3 and Op. 48 nr. 1), the final passage consists of two clearly separated parts, both of which are performed individually with a ritardando. These ritardandi are treated separately (see table 1). In one case (Op. 27 nr. 1), the best-suited passage is at the end of the first part, rather than at the end (so strictly speaking, it is not a *final* ritardando).

The data were obtained in a semi-automated manner, using a software tool [16] for automatic transcription of the audio recordings. From the transcriptions generated in this way, the segments corresponding to the final ritardandi were extracted and corrected manually by the authors, using *Sonic Visualizer*, a software tool for audio annotation and analysis [17].

3. METHOD

As mentioned in section 1, the expressive timing data is expected to have a strong component that is determined by piece-specific aspects like rhythmical structure and harmony. In order to focus on pianist-specific aspects of timing, it is helpful to remove this component. In this section, we first describe how the IOI data are represented. We then propose a filter on the data based on an estimate of the measurement error of IOI values. Finally, we describe a pianist identification task as an assessment of the suitability of the filtered data for characterizing the individual playing style of pianists.

3.1 Calculation of deviations from the performance norm

The performance norm used here is the average performance per piece. That is, for a piece k, Let M be the number of pianists, and N_k be the number of measured IOI times in piece k. We use $\mathbf{v}_{k,i}$ to denote the vector of the N_k IOI values of pianist i in piece k. Correspondingly, $\mathbf{u}_{k,i}$ is the IOI vector of pianist i for piece k, centered around zero ($\bar{\mathbf{v}}_{k,i}$ being the mean of all IOI's in $\mathbf{v}_{k,i}$):

$$\mathbf{u}_{k,i} = \mathbf{v}_{k,i} - \bar{\mathbf{v}}_{k,i} \tag{1}$$

The performance norm $\mathbf{a}_k$ for piece k is defined as the average over pianists per IOI value:

$$\mathbf{a}_k(j) = \frac{1}{M} \sum_{i=1}^{M} \mathbf{u}_{k,i}(j) \tag{2}$$

where $\mathbf{a}_k(j)$ is the j-th IOI value of the average performance of piece k.

Figure 1 shows the performance norms obtained in this way. Note that most performance norms show a two stage ritardando, in which a gradual slowing down is followed by a stronger decrease in tempo, a general trend that is also observed in [12]. The plots show furthermore that in addition to a global slowing down, finer grained timing structure is present in some pieces.

3.2 Estimation of measurement error

An inherent problem of empirical data analysis is the presence of measurement errors. As described above, the timing data from which the tempo curves are generated is obtained by measurement of beat times from audio files. The data is manually corrected, but even manually the exact time of some note onsets is hard to identify, especially when the pianist plays very softly while using the sustain pedal. Therefore, it is relevant to investigate to which degree different beat time annotations of the same performance differ from each other. This gives us an idea of the size of the measurement error, and allows us to distinguish significant deviations from the performance norm from the non-significant deviations.

Pianist	Year	Op.9 nr.3 rit1	Op.9 nr.3 rit2	Op.15 nr.1	Op.15 nr.2	Op.27 nr.1	Op.27 nr.2	Op.48 nr.1 rit1	Op.48 nr.1 rit2
Argerich	1965			X					
Arrau	1978	X	X	X	X	X	X	X	X
Ashkenazy	1985	X	X	X	X	X	X	X	X
Barenboim	1981	X	X	X	X	X	X	X	X
Biret	1991	X	X	X	X	X	X	X	X
Engerer	1993	X	X	X	X	X	X	X	X
Falvai	1997	X	X	X	X	X	X	X	X
Harasiewicz	1961	X	X	X	X	X	X	X	X
Hewitt	2003	X	X	X	X	X	X	X	X
Horowitz	1957			X		X			
Kissin	1993					X	X		
Kollar	2007	X	X	X	X		X	X	X
Leonskaja	1992	X	X	X	X	X	X	X	X
Maisenberg	1995			X					
Mertanen	2001	X	X	X	X	X	X		
Mertanen	2002							X	X
Mertanen	2003							X	X
Ohlsson	1979	X	X	X	X	X	X	X	X
Perahia	1994			X					
Pires	1996	X	X	X	X	X	X	X	X
Pollini	2005	X	X	X	X	X	X	X	X
Richter	1968			X					
Rubinstein	1937	X	X	X	X	X	X	X	X
Rubinstein	1965	X	X	X	X	X	X	X	X
Tsong	1978	X	X	X	X	X	X	X	X
Vasary	1966	X	X	X	X	X		X	X
Woodward	2006	X	X	X	X	X	X	X	X
d´Ascoli	2005	X	X	X	X	X	X	X	X

Table 1. Performances used in this study. The symbol "X" denotes the presence of the corresponding combination of pianist/piece in the data set. The additions "rit1" and "rit2" refer to two distinct ritardandi within the same piece

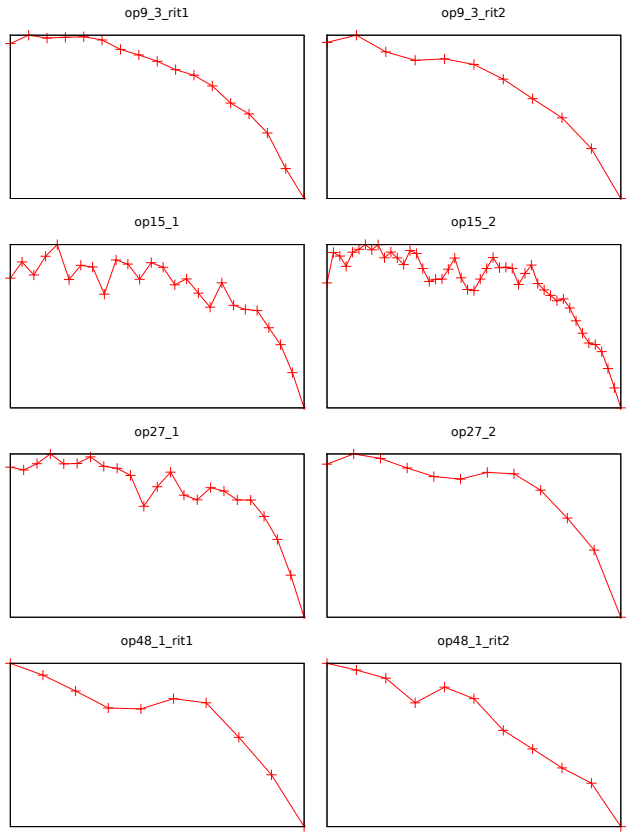

Figure 1. The average performance per ritardando. Both score time (horizontal axis) and tempo (vertical axis) are normalized

To this end, a subset of the data containing seven performances of various performers and different pieces has been annotated twice, by two different persons. [1] This set in total contains 304 time points to be measured. For each beat a pair of annotated beat times was available after annotation by both annotators, from which the absolute pairwise differences were calculated.

Figure 2 shows a scatter plot of absolute pairwise differences of measured IOI time versus the beat duration. [2] Note that beat durations have been calculated from note interonset times that were sometimes at a substantially faster pace than the beat. Hence, a beat duration of, say, 14 seconds does not imply that two measured points are actually 14 seconds apart. It can be observed from the plot that at slower tempos, there is more agreement between annotators about the onset times of notes. This is likely to be either because the slower parts tend to be played in a more articulate way, or simply because of the lower note density, which makes it easier to determine note onsets precisely.

The line in figure 2 shows the function that we use as a criterion to either accept or reject a particular IOI data point for further analysis. More specifically, the function specifies how far a data point must be away from the performance norm in order to be considered as a significant deviation. Conversely, we consider deviations of points closer to the norm too likely caused by measurement errors. The criterion is rather simple, and defines .2 seconds as an absolute minimum for deviations, with an increasing threshold for measurements at higher tempos (shorter beat durations), to accommodate for the increasing measurement differences observed in the data. The constants in the the function have been chosen manually, ensuring

[1] Because of the size of the data set, and the effort that manual correction implies, it was not feasible to annotate the complete data set multiple times

[2] by *beat* we mean score unit duration, rather than a perceived pulse

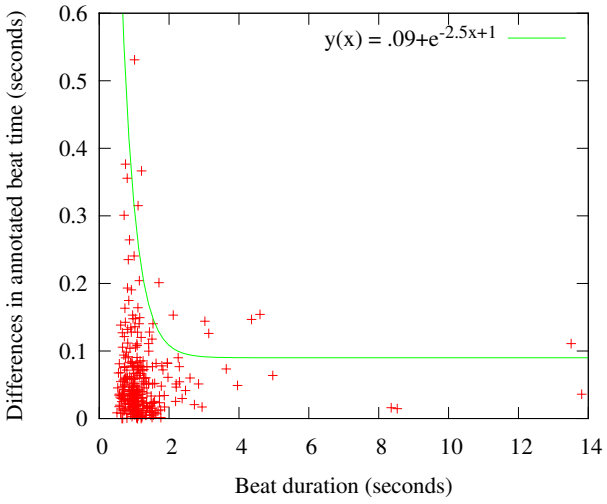

Figure 2. Scatter plot of absolute beat time annotation differences versus beat duration, between two annotators

that a substantial amount of the measurement deviations in the scatter plot ($> 95\%$) are excluded by the criterion. This approach can admittedly be improved. Ideally, taking into account the significance of deviations from the performance norm should be done by a weighting of data points that is inversely proportional to the likelihood of being due to measurement errors.

With the current criterion we filter the data by keeping only those data points that satisfy the inequality:

$$\mathbf{u}_{k,i}(j) > 0.09 + \exp\left[-2.5(\mathbf{a}_k(j) + \bar{\mathbf{v}}_{k,i}) + 1.0\right] \quad (3)$$

The set of data points after filtering is displayed for two pianists in figure 3. The left plot shows the significant deviations from the performance norm over all ritardandi performed by Falvai. The right plot shows those of Leonskaja. In order to compare the ritardandi from different pieces (with differing length and different number of measured IOI's), time has been normalized per piece. Note that a large part of Falvai's IOI deviations has been filtered out based on their size. This means that Falvai's ritardandi are are mostly in agreement with the performance norm. Interestingly, the endings of Falvai's ritardandi deviate in a very consistent way by being slightly faster than the norm until the last few notes, which tend to be delayed more than normal. Leonskaja's IOI deviations are more diverse and appear to be more piece dependent. A more in-depth investigation seems worthwhile here, but is beyond the scope of this article.

3.3 Evaluation of the data: automatic identification of pianists

In order to verify whether the residual timing data after subtracting the norm and filtering with the measurement error criterion in general carry information about the performing pianist, we have designed a small experiment. In this experiment we summarize the residual timing data by four attributes and apply a multilayer perceptron [18] (a standard machine learning algorithm, as available in the *Weka* toolbox for data mining and machine learning) to perform binary classification for all pairs of pianists in the data set. [3] The training instances (ritardandi of a particular piece performed by a particular pianist) containing varying numbers of IOI deviation values, each associated with a normalized score time value, describing where the IOI deviation occurs in the ritardando (0 denoting the beginning of the ritardando, and 1 the end). In order to use these data for automatic classification, they must be converted to data instances with a fixed number of attribute-value pairs. We choose an extremely simple approach, in which we represent a set of IOI deviation / score time pairs by the mean and standard deviation of the IOI values and the mean and standard deviations of the normalized time values. Thus, we effectively model the data by describing the size and location of the area where IOI deviation values tend to occur in the plots of figure 3.

4. RESULTS AND DISCUSSION

The pairwise pianist classification task is executed as follows: for each possible pair of pianists, the ritardandi of both pianists are pooled to form the data set for evaluating the classifier. The training set in most cases contains 16 instances, one for each of the eight pieces, for each of the two pianists. The pianists from whom less than 6 performances were contained in the data set were not included in the test. The data set was used to evaluate the multilayer perceptron using 10-fold cross-validation. This was done for all 171 combinations of 19 pianists. The results are compared to a baseline algorithm that predicts the mode of the target concept, the pianist, in the training data.

The classification results on the test data are summarized in tables 2 and 3. Table 2 shows the proportion of pairwise identification tasks where the multilayer perceptron classified above, at, and, below baseline classification, respectively. The top row presents the results for the condition where the IOI deviation data has been filtered using the measurement error criterion, as explained in subsection 3.2. The bottom row correspond to the condition where no such filtering was applied.

The measurement error filtering clearly leads to an improvement of classification accuracy. With filtering, the percentage of pianist identification tasks that are executed with an accuracy that is significantly ($\alpha = .05$) above baseline accuracy, is 32%. Although this percentage does not seem very high, it must be considered that the amount of information available to the classifier is very small. Firstly, the ritardandi are only short fragments of the complete performances. Secondly, the training sets within a 10-fold cross-validation never contain more than seven ritardandi of a single pianist. Lastly, the IOI deviation information available has been summarized very coarsely, by a mean and standard deviation of the values in the time and IOI dimension. This result implies that larger deviations from

[3] For some pianists less than six performances were available; Those pianists have not been included in the experiment.

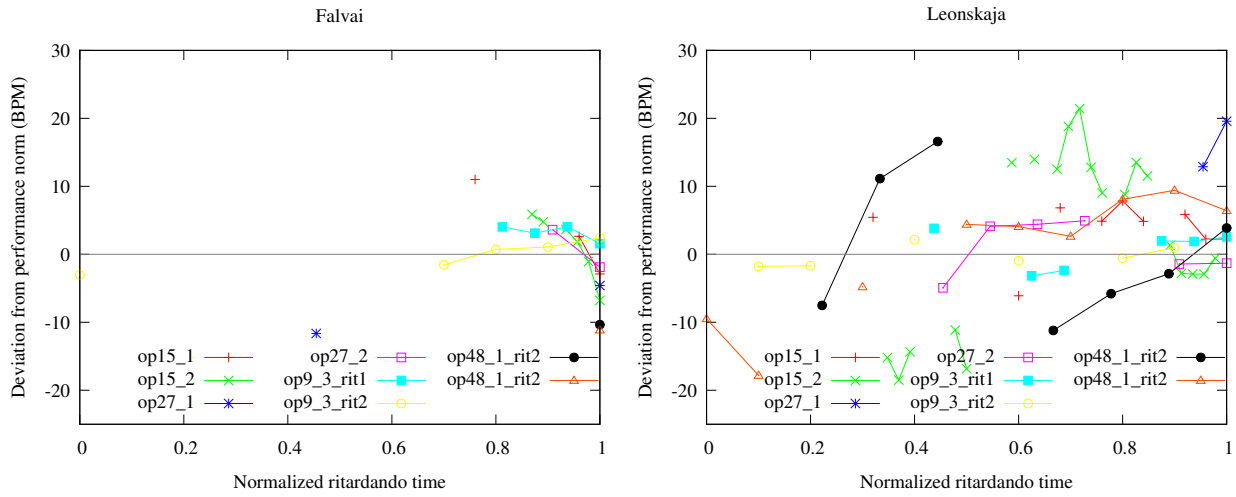

Figure 3. Deviations from the performance norm after applying the measurement error criterion; Left: Falvai; Right: Leonskaja

the performance norm by individual pianists are at least to some degree pianist specific, and not just piece specific.

We wish to emphasize that by no means we claim that the specific form of the measurement error criterion we proposed in subsection 3.2 is crucial for the success of of pianist identification. Other filtering criteria might work equally well or better. Note however that there is a trade off between avoiding the disturbing effect of measurement errors on the one hand, and a reduction of available data on the other. A more elegant approach to canceling the effect of measurement errors would be to use a weighting criterion rather than a filtering criterion.

Without filtering, accuracy is even significantly *below* the baseline in 19% of the cases. The fact that under this condition accuracy does not often surpass the baseline is not surprising, since the unfiltered data contains all available IOI deviation values, equally distributed over time. A consequence of this that mean and standard deviation of the normalized times associated to the IOI data are constant. This reduces the available information so much that it is unrealistic to expect above baseline accuracy. That the prediction accuracy is significantly below baseline is more surprising. Given that the performance norm is subtracted from the original timing data per piece, a strong interference of the piece with the pianist identification is not to be expected. A possible explanation for this result could be that there are multiple distinct performance strategies. Obviously, the average performance as a performance norm is not adequate for this situation, where multiple performance norms are present. If two pianists choose a similar strategy, their residual IOI values after subtracting the average performance may still be more similar to each other than to their own IOI values in a different piece.

Table 3 shows the average identification accuracy over all identification-tasks that involve a specific pianist. High accuracy could indicate that a pianist plays both consistently, and distinctively. By playing consistently we mean that particular IOI deviations tend to occur at the similar

Procedure	$<$ baseline	baseline	$>$ baseline
with filtering	0 (0%)	116 (68%)	55 (32%)
without filtering	33 (19%)	131 (76%)	7 (4%)

Table 2. Number of 10-fold cross-validated pairwise pianist classification tasks with results over, at, and below baseline results, respectively ($\alpha = .05$)

positions in the ritardando, as observed in the case of Falvai, in figure 3 (see also [19] for a discussion of performer consistency). Playing distinctively means that no other pianist has similar IOI deviations at similar positions. Conversely, a low identification accuracy could point to either a varied way of performing ritardandi of different pieces, or playing ritardandi in particular pieces in a way that is similar to the way (some) other pianists play them, or both.

5. CONCLUSIONS AND FUTURE WORK

Ritardandi in musical performances are good examples of the expressive interpretation of the score by the pianist. We have investigated the possibility of automatically identifying pianists by the way they perform ritardandi. More specifically, we have reported an initial experiment in which we use IOI deviations from a performance norm (the average performance) to distinguish pairs of pianists. Furthermore, we have introduced a simple filtering criterion that is intended to remove parts of the data that are likely to be due to measurement errors. Although more sophisticated methods for dealing with measurement error can certainly be developed, the filtering method improved the accuracy of pianist identification substantially.

Continued work should include the development of a more gradual way to deal with the significance of IOI deviations, rather than an all-or-nothing filtering method. Also, better models of expressive timing and tempo are needed to serve as a performance norm. In this work we have employed the average performance as a substitute norm, but it

Pianist	avg. % correct
Leonskaja	65.31
Pollini	64.83
Vasary	63.50
Ohlsson	62.28
Mertanen	62.06
Barenboim	61.69
Falvai	57.42
Engerer	54.33
Hewitt	53.50
Woodward	53.47
Biret	51.47
Pires	51.03
Tsong	50.17
Harasiewicz	49.78
Kollar	49.33
d´Ascoli	48.06
Ashkenazy	47.69
Rubinstein	45.83
Arrau	43.53

Table 3. Average identification accuracy per pianist on test data

is obvious that a norm should be independent of the data.

6. ACKNOWLEDGMENTS

We wish to thank Werner Goebl and Bernhard Niedermayer for their help in the acquisition of the timing data from the audio recordings. This work is funded by the Austrian National Research Fund (FWF) under project number P19349-N15.

7. REFERENCES

[1] E. F. Clarke. Generative principles in music. In J.A. Sloboda, editor, *Generative Processes in Music: The Psychology of Performance, Improvisation, and Composition*. Oxford University Press, 1988.

[2] P. Juslin and J. Sloboda, editors. *Music and Emotion: Theory and Research*. Oxford University Press, 2001.

[3] C. Palmer. Music performance. *Annual Review of Psychology*, 48:115–138, 1997.

[4] B. H. Repp. Patterns of expressive timing in performances of a Beethoven minuet by nineteen famous pianists. *Journal of the Acoustical Society of America*, 88:622–641, 1990.

[5] B. H. Repp. Diversity and commonality in music performance - An analysis of timing microstructure in Schumann's "Träumerei". *Journal of the Acoustical Society of America*, 92(5):2546–2568, 1992.

[6] E. Stamatatos and G. Widmer. Automatic identification of music performers with learning ensembles. *Artificial Intelligence*, 165(1):37–56, 2005.

[7] C. Saunders, D. Hardoon, J. Shawe-Taylor, and G. Widmer. Using string kernels to identify famous performers from their playing style. *Intelligent Data Analysis*, 12(4):425–450, 2008.

[8] W. L. Windsor and E. F. Clarke. Expressive timing and dynamics in real and artificial musical performances: using an algorithm as an analytical tool. *Music Perception*, 15(2):127–152, 1997.

[9] N.P. Todd. A computational model of rubato. *Contemporary Music Review*, 3 (1), 1989.

[10] A. Friberg. Generative rules for music performance: A formal description of a rule system. *Computer Music Journal*, 15 (2):56–71, 1991.

[11] R. Timmers, R.and Ashley, P. Desain, H. Honing, and L. Windsor. Timing of ornaments in the theme of Beethoven's Paisiello Variations: Empirical data and a model. *Music Perception*, 20(1):3–33, 2002.

[12] J. Sundberg and V. Verrillo. On the anatomy of the retard: A study of timing in music. *Journal of the Acoustical Society of America*, 68(3):772–779, 1980.

[13] A. Friberg and J. Sundberg. Does music performance allude to locomotion? a model of final ritardandi derived from measurements of stopping runners. *Journal of the Acoustical Society of America*, 105(3):1469–1484, 1999.

[14] M. Grachten and G. Widmer. The kinematic rubato model as a means of studying final ritards across pieces and pianists. In *Proceedings of the 6th Sound and Music Computing Conference*, 2009.

[15] H. Honing. Is there a perception-based alternative to kinematic models of tempo rubato? *Music Perception*, 23(1):79–85, 2005.

[16] B. Niedermayer. Non-negative matrix division for the automatic transcription of polyphonic music. In *Proceedings of the 9th International Conference on Music Information Retrieval*. ISMIR, 2008.

[17] C.L. Cannam, M. Sandler, and J.P. Bello. The sonic visualiser: A visualisation platform for semantic descriptors from musical signals. In *Proceedings of the 7th International Conference on Music Information Retrieval*. ISMIR, 2006.

[18] D. E. Rumelhart and J. L. McClelland, editors. *Parallel Distributed Processing*, volume 1. MIT Press, 1986.

[19] S. T. Madsen and G. Widmer. Exploring pianist performance styles with evolutionary string matching. *International Journal on Artificial Intelligence Tools*, 15(4):495–513, 2006. Special Issue on Artificial Intelligence in Music and Art.