USING EXPRESSIVE TRENDS FOR IDENTIFYING VIOLIN PERFORMERS

Miguel Molina-Solana
Comp. Science and AI Dep.
University of Granada
18071 Granada, Spain
miguelmolina@ugr.es

Josep Lluís Arcos
IIIA, AI Research Institute
CSIC, Spanish National Res. Council
08193 Bellaterra, Spain
arcos@iiia.csic.es

Emilia Gomez
Music Technology Group
Universitat Pompeu Fabra
08003 Barcelona, Spain
egomez@iua.upf.edu

ABSTRACT

This paper presents a new approach for identifying professional performers in commercial recordings. We propose a Trend-based model that, analyzing the way Narmour's Implication-Realization patterns are played, is able to characterize performers. Concretely, starting from automatically extracted descriptors provided by state-of-the-art extraction tools, the system performs a mapping to a set of qualitative behavior shapes and constructs a collection of frequency distributions for each descriptor. Experiments were conducted in a data-set of violin recordings from 23 different performers. Reported results show that our approach is able to achieve high identification rates.

1 INTRODUCTION

Expressive performance analysis and representation is currently a key challenge in the sound and music computing area. Previous research has addressed expressive music performance using machine learning techniques. For example, Juslin *et al.* [6] studied how expressivity could be computationally modeled. Ramirez *et al.* [12] have proposed an approach for identifying saxophone performers by their playing styles. Lopez de Mantaras *et al.* [9] proposed a case based reasoning approach to deal with expressiveness. Hong [7] investigated expressive timing and dynamics in recorded cello. Dovey [2] analyzed Rachmaninoff's piano performances using inductive logic programming. Work on automatic piano performer identification has been done by the group led by Gerhard Widmer. To cite some, in [14] they represent pianists' performances as strings; in [16] they study how to measure performance aspects applying machine learning techniques; and in [15], Stamatatos and Widmer propose a set of simple features for representing stylistic characteristics of piano music performers. Sapp [13] work should also be cited as an interesting proposal for representing musical performances by means of scape plots based on tempo and loudness correlation. Goebl [3] is focused on finding a 'standard performance' by exploring the consensus among different performers.

In this paper, we focus on the task of identifying violinists from their playing style using descriptors automatically extracted from commercial audio recordings by means of state-of-the-art feature extraction tools. First, since we consider recordings from quite different sources, we assume a high heterogeneity in the recording conditions. Second, as state-of-the-art audio transcription and feature extraction tools are not 100% precise, we assume a partial accuracy in the extraction of audio features.

Taking into account these characteristics of the data, our proposal therefore identifies violin performers through the following three stage process: (1) using a higher-level abstraction of the automatic transcription focusing on the melodic contour; (2) tagging melodic segments according to the E. Narmour's Implication-Realization (IR) model [10]; and (3) characterizing the way melodic patterns are played as probabilistic distributions.

The rest of the paper is organized as follows: In Section 2 we present the data collection being used. In Section 3, we describe the proposed *Trend-Based model* and the developed system, including data gathering, representation of recordings and distance measurement. In Section 4, experiments for the case study are explained and results are presented. The paper concludes with final considerations and pointing out to future work in Section 5.

2 MUSICAL DATA

We have chosen to work with Sonatas and Partitas for solo violin from J.S. Bach [8]. Sonatas and Partitas for solo Violin by J.S. Bach is a six work collection (three Sonatas and three Partitas) composed by the German musician. It is a well-known collection that almost every violinist plays during its artistic life. All of them have been recorded many times by several players. The reason of using this work collection is twofold: 1) we have the opportunity of testing our model with existing commercial recordings of the best known violin performers, and 2) we can constrain our research on monophonic music.

In our experiments, we have extended the musical collection presented in [11]. We analyzed music recordings from 23 different professional performers. Because these audio files were not recorded for our study, we have not interfered at all with players' style at the performance [5]. The scores of the analyzed pieces are not provided to the system.

3 TREND-BASED MODELING

Our approach for dealing with the identification of violin performers is based on the acquisition of *trend models* that characterize each particular performer to be identified. Specifically, a trend model characterizes, for a specific audio descriptor, the relationships a given performer is establishing among groups of neighbor musical events. We perform a qualitative analysis of the variations of the audio descriptors. Moreover, as we will describe in the next subsection, we analyze these qualitative variations with a local perspective. We will be using two trend models in this paper: energy and duration. The trend model for the energy descriptor relates, qualitatively, the energy variation for a given set of consecutive notes, while the trend model for duration indicates, also qualitatively, how note durations change for note sequences. Notice that trend models are not trying to characterize the audio descriptors with respect to an expected global behavior.

Given an input musical recording of a piece, the trend analysis is performed by aggregating the qualitative variations on their small melody segments. Thus, in advance of building trend models, input streams are broken down into segments. As most of automatic melody segmentation approaches, we will perform note grouping according to a human perception model.

Our system has been designed in a modular way with the intention of creating an easy extendable framework. We have three different types of modules in the system: 1) the audio feature extraction modules; 2) the trend analysis modules; and 3) the identification modules. Moreover, the system may work in two different modes: in a training mode or in a testing mode. Modules from (1) and (2) are used in both modes. Modules from (3) are only used in the testing mode. Figure 1 shows a diagram of the system modules. On top, audio files in *.wav* format as input.

At the training stage, the goal of the system is to characterize performers by extracting expressive features and constructing trend models. Next, at the identification stage, the system analyzes the input performance and looks for the most similar previously learnt model. The training process is composed of three main steps: 1) the extraction of audio descriptors and the division of a performance in segments; 2) the tagging of segments according to IR patterns; and 3) the calculus of probabilistic distributions for each IR pattern and descriptor (trend generation).

3.1 Feature Extraction and Segmentation

The first step consists on extracting audio features. Our research is not focused on developing new methods for extracting audio features, so that we employ existing techniques. At the moment, we consider fundamental frequency and energy, as these are the main low-level audio features related to melody. These features are then used to identify note boundaries and to generate melodic segments. The current version of our system uses the fundamental frequency estimation algorithm proposed by Arturo Camacho [1]. This module provides a vector with instantaneous fundamental frequency and energy values.

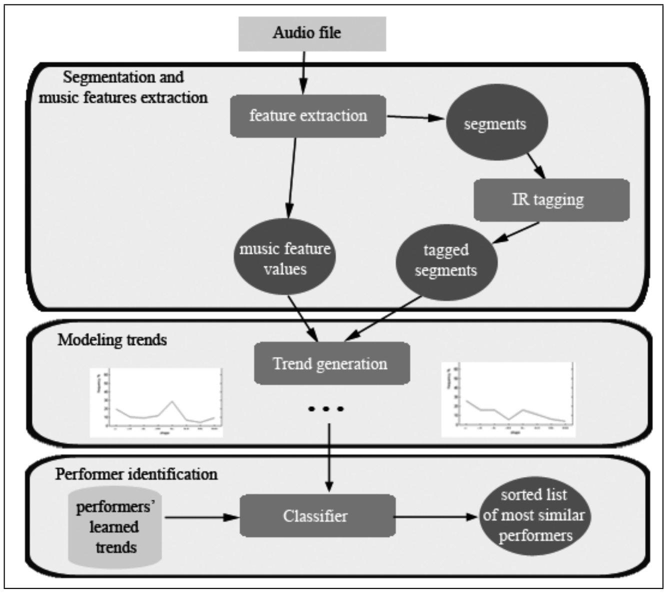

Figure 1. System's dataflow

We have developed a post-processing module for determining the possible notes played by the performers. This module first converts fundamental frequencies into quantized pitch values, and then a pitch correction procedure is applied in order to eliminate noise and sudden changes. This correction is made by assigning to each sample the value given by the mode of their neighbors around a certain window of size σ. With this process, a smooth vector of pitches is obtained. By knowing on which frames pitches are changing, a note-by-note segmentation of the whole recording is performed. For each note we collect its pitch, duration and energy.

We assume that there might be some errors in this automatic segmentation, given the hetereogenity of recording conditions. Our approach for dealing with this problem consists of using a more abstract representation that the real notes, but still close to the melody. That is, instead of focusing on the absolute notes, we are interested in modeling the melodic surface.

We use the IR model by E. Narmour [10] to perform melodic segmentation. This model tries to explicitly describe the patterns of expectations generated in the listener with respect to the continuation of the melody. The IR model describes both the continuation implied by particular melodic intervals, and whether or not this expected continuation is fulfilled by the following interval. Taking this cognitive model as the basis for the melodic segmentation, each IR pattern determine a different segment.

The IR model has been shown to be suitable for assessing melodic similarity (see MIREX'05 [4]). Since our goal is to characterize expressive trends, we analyze the way different audio descriptors change in the different IR patterns. See [10] for a detailed description of IR patterns.

3.2 Modeling Trends

A trend model is represented by a set of discrete probability distributions for a given audio descriptor (e.g. energy). Each of these probability distributions represents the way a given IR pattern is played against that certain audio descriptor.

To generate trend models for a particular performer and audio descriptor, we use the sequences of values extracted from the notes identified in each segment. From these sequences, a qualitative transformation is first performed to the sequences in the following way: each value is compared to the mean value of the fragment and is transformed into a qualitative value where + means 'the descriptor value is higher than the mean', and - means 'the descriptor value is lower than the mean'. Being s the size of the segment and n the number of different qualitative values, there are n^s possible resulting shapes. In the current approach, since we are segmenting the melodies in groups of three notes and using two qualitative values, eight (2^3) different shapes may arise. We note these possibilities as: —, –+, -+-, -++, +–, +-+, ++- and +++.

Next, a histogram per IR pattern with these eight qualitative shapes is constructed by calculating the percentage of occurrence of each shape. These histograms can be understood as discrete probability distributions. Thus, trend models capture statistical information of how a certain performer tends to play. Combining trend models from different audio descriptors, we improve each performer characterization.

Since our goal is the identification of violin performers, the collection of trend models acquired for each performer is used as the patterns to compare with when a new audio recording is presented to the system.

3.2.1 Current Trends

We have generated trend models for both duration and energy descriptors. Note durations are calculated as the number of samples between pitch changes. Qualitative deviations are calculated by comparing the real and the average duration for each note. Then, the trend model for the duration descriptor was calculated from the frequency distributions of each IR pattern. Figure 2 shows, for the duration descriptor, the frequency distributions of the eight shapes in P patterns (ascending or descending sequences of similar intervals) and for four violin performers (Ara Malikian, Arthur Grumiaux, Brian Brooks, and Christian Tetzlaff).

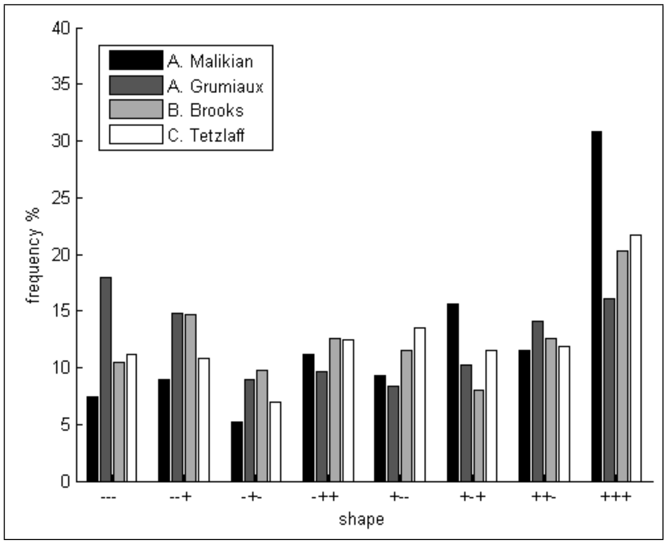

Figure 2. Frequency distribution of duration deviations for the P pattern in the Sixth movement of Partita N.1. Only four performers are shown

We can observe that the way different professional performers are playing is not equally distributed. For instance, A. Malikian has a higher propensity to extend the durations while an opposite behavior can be observed for A. Grumiaux (see his values for the two left qualitative shapes). It should be noticed that more exact ways of obtaining this measure could be used, as well as taking into account the attack and release times, as other researchers do [12].

We have also acquired trend models for the energy descriptor in an analogous way. The energy average for each fragment is calculated and, given the energy of each note, the qualitative deviations are computed. Next, from these qualitative values, the trend models are constructed by calculating the frequencies of the eight shapes for each IR pattern.

3.3 Classifying new performances

A nearest neighbor (NN) classifier is used to predict the performer of new recordings. Trend models acquired in the training stage, as described in the previous section, are used as class patterns, i.e. each trained performer is considered a different solution class. When a new recording is presented to the system, the feature extraction process is performed and its trend model is created. This trend model is compared with the previously acquired models. The classifier outputs a ranked list of performer candidates where distances determine the order, with 1 being the most likely performer relative to the results of the training phase.

3.3.1 Distance measure

The distance d_{ij} between two trend models i and j, is defined as the weighted sum of distances between the respective IR patterns:

$$d_{ij} = \sum_{n \in N} w_{ij}^{n} dist(n_i, n_j) \tag{1}$$

where N is the set of the different IR patterns considered; $dist(n_i, n_j)$ measures the distance between two probability distributions (see (3) below); and w_{ij}^{n} are the weights assigned to each IR pattern. Weights have been introduced for balancing the importance of the IR patterns with respect to the number of times they appear. Frequent patterns are considered more informative due to the fact that they come from more representative samples. Weights are defined as the mean of cardinalities of respective histograms for a given pattern n:

$$w_{ij}^{n} = (N_i^n + N_j^n)/2 \tag{2}$$

Mean value is used instead of just one of the cardinalities to assure a symmetric distance measure in which w_{ij}^{n} is equal to w_{ji}^{n}. Cardinalities could be different because recognized notes can vary from a performance to another, even though the score is supposed to be the same.

Finally, distance between two probability distributions is calculated by measuring the absolute distances between the respective patterns:

$$dist(s, r) = \sum_{k \in K} |s_k - r_k| \tag{3}$$

where s and r are two probability distributions for the same IR pattern; and K is the set of all possible values they can take (in our case $|K| = 8$).

When both audio descriptors are considered, we simply aggregate the individual corresponding distances.

4 RESULTS AND DISCUSSION

We tested our system by performing experiments using different information coming from the acquired trend models. Specifically, we evaluated each experiment setting with only duration-based trend models, only energy-based trend models, and with both trend models.

Each experiment consisted in training the system with one movement and then testing the trend models acquired presenting to the system the recordings from another movement. For experimentation, we used a collection of audio recordings from 23 different professional violinists. We performed two different types of experiments. The first experiment was focused on assessing the performance of the system by using two movements from the same piece. Specifically, we used the Second and the Sixth movements of Partita No. 1. These fragments are quite interesting for early testing because most of the notes are eighth notes, leading us to acquire a model based on many homogeneous segments. In the following, we will call **set-1** the experiment where the second movement was used for training and the sixth for testing. Analogously, we will call **set-2** the experiment where the sixth movement was used for training and the second for testing.

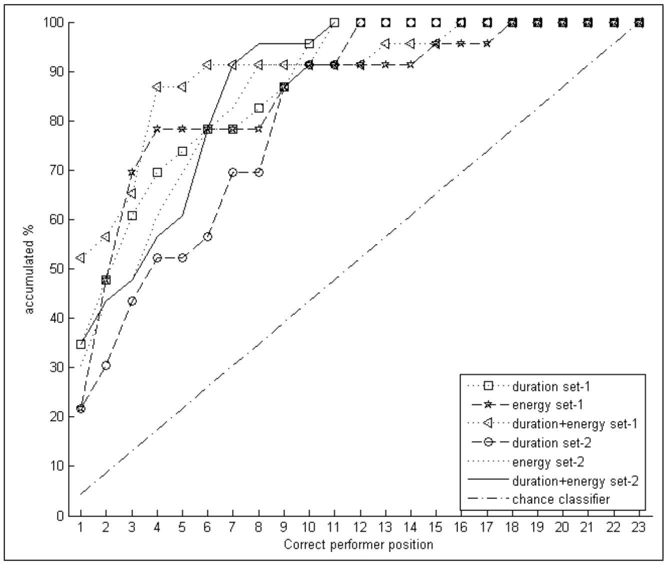

Figure 3. Accumulated success rate by position of the correct performer. **Set-1** and **set-2** are shown

The second type of experiments was focused on assessing the performance of the system by using two movements from different pieces. Specifically, we used the sixth movement of Partita No. 1 for training and the fifth movement of Partita No. 3 for testing. We will refer to this test as **set-3**.

For each input recording, the system outputs a ranked list of performers sorted from the most similar to the farthest one to the input. In experiments **set-1** and **set-2**, the correct performer was mostly identified in the first half of the list, i.e. at most in the 12th position. In **set-3**, the most difficult scenario, the 90% of identification accuracy was overcame at position 15. Regarding the different trend models, the energy model was the one that achieved a highest accuracy. This result is not surprising since the duration descriptor presents a high sensitivity with respect to the analysis precision. The combination of both trend models improves the results when focusing only on the first proposed player.

A complete view of the performance of the system is shown in Figure 3 and summarized in Table 1. Figure 3 shows the percentage of input recordings identified at each position. It provides a picture of the accuracy of the system using as a threshold the length of the proposed ranking. The results are promising, especially comparing with a random classification where the success rate is clearly outperformed. We can observe that the system achieved a 50% of success

	set-1			**set-2**			**set-3**		
	1st	3rd	10th	1st	3rd	10th	1st	3rd	10th
duration	34.8	60.9	95.7	21.7	43.5	91.3	10.5	26.3	68.4
energy	21.7	69.6	91.3	30.4	47.8	91.3	15.8	31.6	73.7
both	52.2	65.2	91.3	34.8	47.8	95.7	15.8	26.3	68.4

Table 1. Success rate (%) in all experiments taking into account three different ranking positions proposed for the correct performer: 1st, 3rd, and 10th

using the four top candidates in **set-1** and **set-2**.

Table 1 summarizes the performance of the system for the three experimental sets and the three trend models. The three columns of each experiment show, respectively, the percentage of performers identified in the first position, at least in the third position, and at least in the tenth position. We can observe that for settings **set-1** and **set-2** the correct performer is predicted, in the worst case, 20% of times as the first candidate, clearly outperforming the random classifier (whose success rate is 4.3%).

Figure 4 presents a matrix that summarizes the classifier output for **set-2** using both duration and energy trend models. The figure shows, for each input recording (row), the sorted list of predicted performers as squares. The gray scale maps to the ranking values. The black color indicates the first performer proposed and the gray degradation is used to draw all the performers predicted until the correct one. Notice that the success in the first position means a black square in the diagonal. The matrix is not supposed to be symmetric and each column can have the same color several times because a predicted performer can occur in the same position for several inputs. For instance, we can see that Garret Fischbach's performance (`gar`) for Sixth Movement is very different from the rest of performers' Second Movement performances: all values correspond to last position (i.e. the furthest). On the other hand, Christian Tetzlaff's (`chr`) and Rachel Podger's (`rac`) performances are quite similar to most of Second Movement performances since there are many squares in their columns.

Finally, Figure 5 shows in which position the correct performer is ranked for each performer in the test set. This Figure complements the former two. The results came from **set-1** using both trend models ('duration+energy set-1' curve in Figure 3). Twelve right identifications were achieved at first position (52% of success rate). The rest was correctly identified in positions 2 to 4 except three performers. Nathan Milstein was identified at position 6. Finally, Sergiu Luca and Shlomo Mintz were not clearly identified. After a detailed analysis of the distances among all performers, we observed that these two musicians are not clearly distinguished. We mean, small variations in the trend models confuse the identification process.

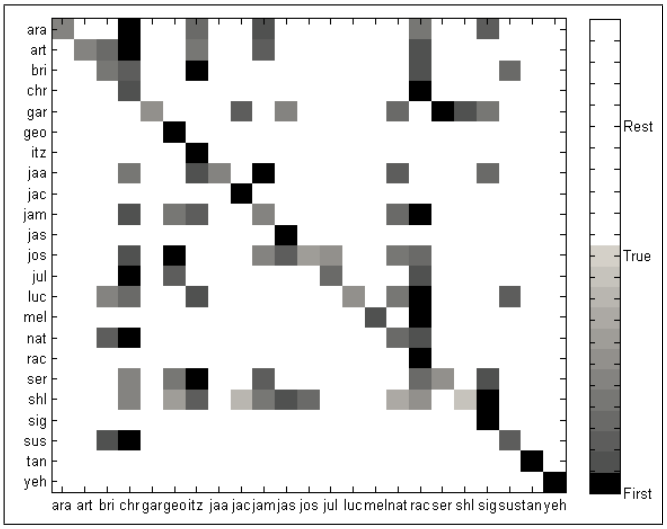

Figure 4. Classifier output in matrix form for **set-2** where both trend models were used

5 CONCLUSIONS

This work focuses on the task of identifying violinists from their playing style by building trend-based models that capture expressive tendencies. Trend models are acquired by using state-of-the-art audio feature extraction tools and automatically segmenting the obtained melodies using IR patterns. Performers were characterized by a set of probability distributions, capturing their personal style with respect to a collection of melodic patterns (IR patterns). We have shown that, without a great analysis accuracy, our proposal is quite robust.

The experiments were concentrated on identifying violinists and using note durations and energies as descriptors. We tested the system with 23 different professional performers and different recordings. Results obtained show that the proposed model is capable of learning performance patterns that are useful for distinguishing performers. The results clearly outperform a random classifier and, probably, it would be quite hard for human listeners to achieve such recognition rates. In order to assure the robustness of the system, other sets of works should be used for learning and testing.

We have presented a qualitative analysis using only two qualitative values. We want to keep our model at this qualitative level but we plan to extend the model with the use of fuzzy sets. This improvement will allow us to use the capabilities of fuzzy theory for a better assessment in the similarity measure.

Combining information from different music features have been demonstrated to improve results. We are currently working in increasing the number of descriptors. Since the predictability of a given descriptor varies depending on the

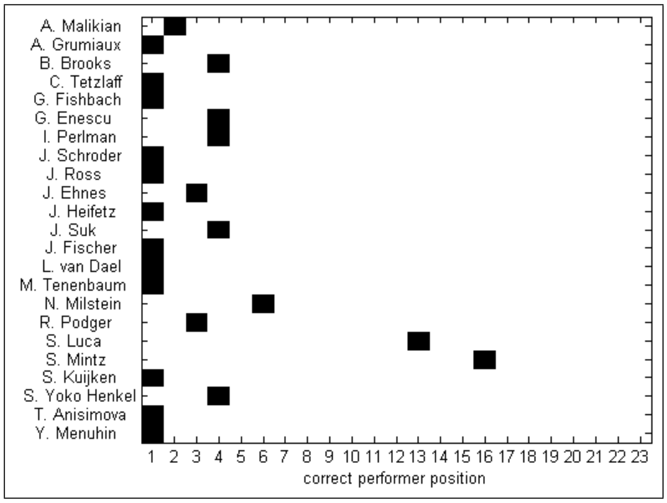

Figure 5. Correct performer position for each performance in **set-1**. Both trend models are used

performers, we are also interested in discovering relations among the descriptors. Finally, the use of hierarchical classifiers or ensemble methods is a possible way of improving the identification.

6 ACKNOWLEDGEMENTS

M. Molina-Solana is supported by the Spanish Ministry of Education and Science under the project TIN2006-15041-C04-01. J. Ll. Arcos is supported by the Spanish project MID-CBR (TIN2006-15140-C03-01), EU-FEDER funds, and by the Catalan Goverment (2005-SGR-00093). E. Gomez is supported by EU-eContent-Plus project VARIAZIONI[1] (ECP-2005-CULT-038264).

7 REFERENCES

[1] Camacho, A. *SWIPE: A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music*. PhD Dissertation, University of Florida, USA, 2007.

[2] Dovey, M.J. "Analysis of Rachmaninoff's piano performances using inductive logic programming", *Proc. Eur. Conf. Machine Learning*, pp. 279-282, 1995.

[3] Goebl, W. "Analysis of Piano Performance: Towards a Common Performance Standard?", *Proc. of the Society for Music Perception and Cognition Conference*, (SMPC99), 1999.

[4] Grachten, M., Arcos, J.L., Lopez de Mantaras, R. "Melody retrieval using the Implication/Realization Model", *Proc. of 6th Int. Conf. on Music Information Retrieval*, (ISMIR 2005), (First prize of the MIREX Symbolic Melodic Similarity Contest), 2005

[5] Juslin, P.N., Sloboda, J.A. *Musical Emotion: Theory and Research*. New York: Oxford Univ. Press, 2001.

[6] Juslin, P.N., Frieberg, A., Bresin, R. "Toward a computational model of expression in performance: The GERM model", *Musicae Scientiae*, special issue 2001-2002, pp. 63-122.

[7] Hong, J. "Investigating expressive timing and dynamics in recorded cello", *Psychology of Music*, 31(3), pp. 340-352, 2003.

[8] Lester, J. *Bach's Works for Solo Violin: Style, Structure, Performance*. First published in 1999 by Oxford University Press, Inc.

[9] Lopez de Mantaras, R., Arcos, J.L. "AI and music, form composition to expressive performance", *AI Magazine*, 23(3), pp. 43-57, 2002.

[10] Narmour, E. *The Analysis and Cognition of Melodic Complexity: The Implication Realization Model*. Chicago, IL: Univ. Chicago Press, 1990.

[11] Puiggros, M. *Comparative analysis of expressivity in recorded violin performances. Study of the Sonatas and Partitas for solo violin by J. S. Bach*. Master Thesis, Universitat Pompeu Fabra, Barcelona, 2007.

[12] Ramirez, R., Maestre, E., Pertusa, A., Gomez, E., Serra, X. "Performance-Based Interpreter Identification in Saxophone Audio Recordings", *IEEE Trans. on Circuits and Systems for Video Technology*, 17(3), pp. 356-264, 2007.

[13] Sapp, C.S. "Comparative Analysis of Multiple Musical Performances", *Proc. of 8th Int. Conf. on Music Information Retrieval*, (ISMIR 2007), Vienna, Austria, pp. 497-500, 2007.

[14] Saunders, C., Hardoon, D., Shawe-Taylor, J., Widmer, G. "Using string kernels to identify famous performers from their playing style", *15th Eur. Conf. on Machine Learning*, Pisa, Italy, 2004.

[15] Stamatatos, E., Widmer, G. "Automatic Identification of Music Performers with Learning Ensembles", *Artificial Intelligence*, 165(1), pp. 37-56, 2005.

[16] Widmer, G., Dixon, S., Goebl, E., Pampalk, W., Tobudic, A. "In search of the Horowitz factor", *AI Magazine*, 24(3), pp. 111-130, 2003.

[1] http://www.variazioniproject.org