Music Performer Verification Based on Learning Ensembles

Efstathios Stamatatos and Ergina Kavallieratou

Dept. of Audio and Musical Instrument Technology
T.E.I. of Ionian Islands
28200, Lixouri
{stamatat, ergina}@teiion.gr

Abstract. In this paper the problem of music performer verification is introduced. Given a certain performance of a musical piece and a set of candidate pianists the task is to examine whether or not a particular pianist is the actual performer. A database of 22 pianists playing pieces by F. Chopin in a computer-controlled piano is used in the presented experiments. An appropriate set of features that captures the idiosyncrasies of music performers is proposed. Well-known machine learning techniques for constructing learning ensembles are applied and remarkable results are described in verifying the actual pianist, a very difficult task even for human experts.

1 Introduction

The printed score of a musical piece provides a representation of music that captures a limited spectrum of musical nuance. This means that if the exact information represented in the printed score is accurately transformed into music by an ideal performer, the result would sound mechanical or unpleasant. The interpretation of the printed score by a skilled artist always involve continuous modification of important musical parameters, such as tempo and loudness, according to the artist's understanding of the structure of the piece. That way the artist stresses certain notes or passages deviating from the printed score. Hence, *expressive music performance* is what distinguishes one performer from another in the interpretation of a certain musical piece.

Because of its central role in our musical culture, expressive performance is a central research topic in contemporary musicology. One main direction in empirical performance research aims at the development of rules or principles of expressive performance either with the help of human experts [5] or by processing large volumes of data using machine learning techniques [12,13]. Obviously, this direction attempts to explore the similarities between skilled performers in the same musical context. On the other hand, the differences in music performance are still expressed generally with aesthetic criteria rather than quantitatively. The literature in this topic is quite limited. In [8] an exhaustive statistical analysis of temporal commonalities and differences among distinguished pianists' interpretations of a well-known piece is presented and the individuality of some famous pianists is demonstrated. A computational model that distinguishes two pianists playing the same pieces based on measures that represent the deviation of the performer from the score plus measures that indicate the properties of the piece is presented in [10]. Results of limited success in the identification of famous pianists' recordings based on their style of playing have been reported in [14].

This paper is an attempt to quantify the main parameters of expressive performance that discriminate between pianists playing the same musical pieces. Specifically, our aim is to develop a music performer verification system, that is, given a certain performance of a musical piece and a set of candidate pianists the task is to examine whether or not a particular pianist is the actual performer. To this end, machine learning techniques are used for taking advantage of different expressive performance features by combining a number of independent simple 'experts' [2]. The dimensions of expressive variation that will be taken into account are the three main expressive parameters available to a pianist: *timing* (variations in tempo), *dynamics* (variations in loudness), and *articulation* (the use of overlaps and pauses between successive notes).

The data used in this study consist of performances played and recorded on a Boesendorfer SE290 computer-monitored concert grand piano, which is able to measure every key and pedal movement of the artist with very high precision. 22 skilled performers, including professional pianists, graduate students and professors of the Vienna Music University, played two pieces by F. Chopin: the *Etude* op. 10/3 (first 21 bars) and the *Ballade* op. 38 (initial section, bars 1 to 45). The digital recordings were then transcribed into symbolic form and matched against the printed score semi-automatically. Thus, for each note in a piece we have precise information about how it was notated in the score, and how it was actually played in a performance. The parameters of interest are the exact time when a note was played (vs. when it 'should have been played' according to the score) – this relates to tempo and timing –, the dynamic level or loudness of a played note (dynamics), and the exact duration of played note, and how the note is connected to the following one (articulation). All this can be readily computed from our data.

In the following, the term *Inter-Onset Interval* (IOI) will be used to denote the time interval between the onsets of two successive notes of the same voice. We define *Off-Time Duration* (OTD) as the time interval between the offset time of one note and the onset time of the next note of the same voice. The *Dynamic Level* (DL) corresponds to the MIDI velocity of a note. The 22 pianists are referred by their code names (i.e., #01, #02, etc.).

2 Representation of Expressive Music Performance

If we define (somewhat simplistically) expressive performance as 'intended deviation from the score', then different performances differ in the way and extent the artist 'deviates' from the score, i.e., from a purely mechanical ('flat') rendition of the piece, in terms of timing, dynamics, and articulation. In order to be able to compare performances of pieces or sections of different length, we need to define features that

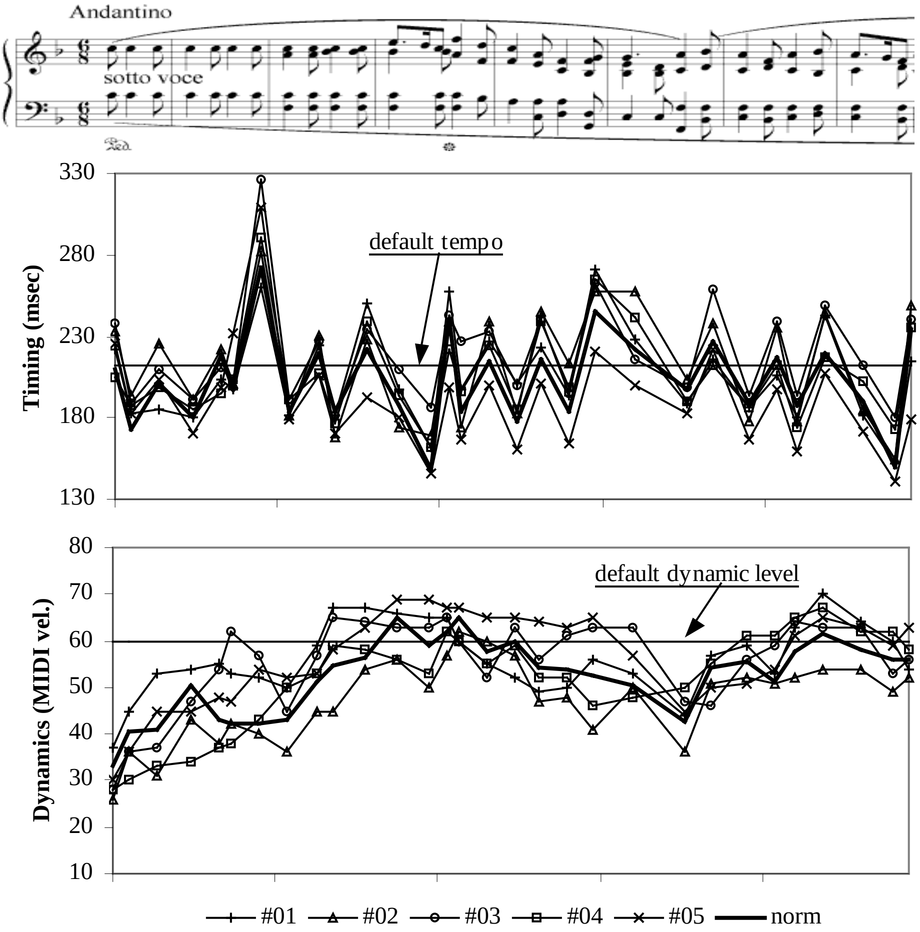

Figure 1. Timing and dynamics variations for the first 30 soprano notes of the *Ballade* (score above) as performed by pianists #01-#05. Default tempo and dynamic level, and performance norm derived by pianists #06-#10 are depicted as well.

characterize and quantify these deviations at a global level, i.e., without reference to individual notes and how these were played.

Figure 1 depicts the performances of the first 30 soprano notes of *Ballade* by the pianists #01-#05 in terms of timing (expressed as the inter-onset interval on the sixteenth-note level) and dynamics. The default tempo and dynamic level according to a pre-specified fixed interpretation of the score correspond to straight lines. As can be seen, the music performers tend to deviate from the default interpretation in a similar way in certain notes or passages. In the timing dimension, the last note of the first bar is considerably lengthened (last note of the introductory part) while in the dynamics dimension the first two bars are played with increasing intensity (introductory part) and the 2nd soprano note of the 5th bar is played rather softly (a phrase boundary).

Although the deviation of the real performances from the score can capture some general stylistic properties of the performer, it seems likely that it would heavily depend on the structure of the piece (i.e., similar form of deviations for all the performers, presenting peaks and dips in the same notes or passages).

For discriminating successfully between different performers, we need a reference point able to focus on the *differences* between them rather than on *common expressive performance principles* shared by the majority of the performers. This role can be played by the *performance norm*, i.e. the average performance of the same piece calculated using a different group of performers. Figure 1 depicts the performance norm, in terms of timing and dynamics, calculated by the performances of pianists #06-#10. As can be seen, the norm follows the basic form of the individual performances. Therefore, the deviation of a given performance from the norm is not dramatically affected by structural characteristics of the piece. Consequently, the deviations of different performers from the norm are not necessarily of similar form (peaks and dips in different notes or passages) and the differences between them are more likely to be highlighted. Norm-based features have been compared to score-based features and proved to be more reliable and stable especially in intra-piece conditions, i.e., training and test cases taken from the same musical piece [11].

Another valuable source of information comes from the exploitation of the so-called melody lead phenomenon, that is, notes that should be played simultaneously according to the printed score (chords) are usually slightly spread out over time. A voice that is to be emphasized precedes the other voices and is played louder. Studies of this phenomenon [7] showed that melody lead increases with expressiveness and skill level. Therefore, deviations between the notes of the same chord in terms of timing and dynamics can provide useful features that capture an aspect of the stylistic characteristics of the music performer.

We propose the following global features for representing a music performance, given the printed score and a performance norm derived from a given set of different performers:

<u>Score deviation features:</u>	
$D(\mathrm{IOI}_s, \mathrm{IOI}_m)$	timing
$D(\mathrm{IOI}_s, \mathrm{OTD}_m)$	articulation
$D(\mathrm{DL}_s, \mathrm{DL}_m)$	dynamics
<u>Norm deviation features:</u>	
$D(\mathrm{IOI}_n, \mathrm{IOI}_m)$	timing
$D(\mathrm{OTD}_n, \mathrm{OTD}_m)$	articulation
$D(\mathrm{DL}_n, \mathrm{DL}_m)$	dynamics
<u>Melody lead features:</u>	
$D(\mathrm{ON}_{xy}, \mathrm{ON}_{zy})$	timing
$D(\mathrm{DL}xy, \mathrm{DL}zy)$	dynamics

where $D(\boldsymbol{x}, \boldsymbol{y})$ (a scalar) denotes the deviation of a vector of numeric values $\boldsymbol{x}$ from a reference vector $\boldsymbol{y}$, IOI_s and DL_s are the nominal inter-onset interval and dynamic-level, respectively, according to the printed score, IOI_n, OTD_n, and DL_n are the inter-onset interval, the off-time duration, and the dynamic-level, respectively, of the performance norm, IOI_{m}, OTD_{m}, and DL_{m} are the inter-onset interval, the off-time duration, and the dynamic-level, respectively, of the actual performance, and ON_{xy}, and DL_{xy} are the on-time and the dynamic-level, respectively, of a note of the *x*-th voice within the chord *y*.

3 The Learning Model

The presented problem is characterized by the extremely limited size of training data as well as the instability of some of the proposed features (i.e., score deviation measures). These characteristics suggest the use of an ensemble of classifiers rather than a unique classifier. Research in machine learning [1] has thoroughly studied the construction of meta-classifiers, or learning ensembles. In this study, we take advantage of such techniques, constructing an ensemble of classifiers derived from two basic strategies:

- *Subsampling the input features.* This technique is usually applied when multiple redundant features are available. In our case, the input features cannot be used concurrently due to the limited size of the training set (i.e., only a few training examples per class are available) and the consequent danger of overfitting the training set.
- *Subsampling the training set.* This technique is usually applied when unstable learning algorithms are used for constructing the base classifiers. In our case, a subset of the input features (i.e., the score deviation measures) is unstable – their values can change drastically given a slight change in the selected training segments.

Given the scarcity of training data and the multitude of possible features, we propose the use of a relatively large number of rather simple individual base classifiers or ‘experts’, in the terminology of [2]. Each expert is trained using a different set of features and/or parts of the training data. The features and sections of the training performances used for the individual experts are listed in table 1. C_{11} is based on the deviation of the performer from the norm. C_{21}, C_{22}, C_{23}, and C_{24} are based on the deviation of the performer from the score and are trained using slightly changed training sets (because the norm features are known to be unstable relative to changes in the data). The training set (see next section) was divided into four disjoint subsets and then four different overlapping training sets were constructed by dropping one of these four subsets (i.e., cross-validated committees). Finally, C_{31}, C_{32}, C_{33}, C_{34}, and C_{35} are based on melody lead features. The last column in table 1 shows the accuracy of each individual expert on the training data (estimated via leave-one-out cross-validation). As can be seen, the classifier based on norm deviation features is by far the most accurate.

Table 1. Description of the proposed base classifiers. The third column indicates the number of training examples (and their length in soprano notes) per class. The last column refers to their accuracy on the training data.

Code	Input features	Training examples	Accuracy (%)
C_{11}	$D(\mathrm{IOI_n}, \mathrm{IOI_m})$, $D(\mathrm{OTD_n}, \mathrm{OTD_m})$, $D(\mathrm{DL_n}, \mathrm{DL_m})$	4x40	82.5
C_{21}	$D(\mathrm{IOI_s}, \mathrm{IOI_m})$, $D(\mathrm{IOI_s}, \mathrm{OTD_m})$, $D(\mathrm{DL_s}, \mathrm{DL_m})$	12x10	50.8
C_{22}	$D(\mathrm{IOI_s}, \mathrm{IOI_m})$, $D(\mathrm{IOI_s}, \mathrm{OTD_m})$, $D(\mathrm{DL_s}, \mathrm{DL_m})$	12x10	44.8
C_{23}	$D(\mathrm{IOI_s}, \mathrm{IOI_m})$, $D(\mathrm{IOI_s}, \mathrm{OTD_m})$, $D(\mathrm{DL_s}, \mathrm{DL_m})$	12x10	46.7
C_{24}	$D(\mathrm{IOI_s}, \mathrm{IOI_m})$, $D(\mathrm{IOI_s}, \mathrm{OTD_m})$, $D(\mathrm{DL_s}, \mathrm{DL_m})$	12x10	48.3
C_{31}	$D(\mathrm{ON_{1m}}, \mathrm{ON_{2m}})$, $D(\mathrm{ON_{1m}}, \mathrm{ON_{3m}})$, $D(\mathrm{ON_{1m}}, \mathrm{ON_{4m}})$	4x40	57.5
C_{32}	$D(\mathrm{DL_{1m}}, \mathrm{DL_{2m}})$, $D(\mathrm{DL_{1m}}, \mathrm{DL_{3m}})$, $D(\mathrm{DL_{1m}}, \mathrm{DL_{4m}})$	4x40	42.5
C_{33}	$D(\mathrm{ON_{1m}}, \mathrm{ON_{2m}})$, $D(\mathrm{DL_{1m}}, \mathrm{DL_{2m}})$	4x40	25.0
C_{34}	$D(\mathrm{ON_{1m}}, \mathrm{ON_{3m}})$, $D(\mathrm{DL_{1m}}, \mathrm{DL_{3m}})$	4x40	35.0
C_{35}	$D(\mathrm{ON_{1m}}, \mathrm{ON_{4m}})$, $D(\mathrm{DL_{1m}}, \mathrm{DL_{4m}})$	4x40	47.5

The classification method used for constructing the base classifiers is *discriminant analysis*, a standard technique of multivariate statistics. The mathematical objective of this method is to weight and linearly combine the input variables in such a way so that the classes are as statistically distinct as possible [3]. A set of linear functions (equal to the input variables and ordered according to their importance) is extracted on the basis of maximizing between-class variance while minimizing within-class variance using a training set. Then, class membership of unseen cases can be predicted according to the *Mahalonobis distance* from the classes' *centroids* (the points that represent the means of all the training examples of each class). The Mahalanobis distance *d* of a vector ***x*** from a mean vector ***m*** is as follows:

$$d^2 = (x - m)' C_x^{-1} (x - m)$$

where C_x is the covariance matrix of ***x***. This classification method also supports the calculation of *posterior probabilities* (the probability that an unseen case belongs to a particular group) which are proportional to the Mahalanobis distance from the classes centroids. In a recent study [6], discriminant analysis is compared with many classification methods (coming from statistics, decision trees, and neural networks). The results reveal that discriminant analysis is one of the best compromises taking into account the classification accuracy and the training time cost. This old and easy-to-implement statistical algorithm performs better than many modern versions of statistical algorithms in a variety of problems.

The combination of the resulting simple classifiers or experts is realized via a weighted majority scheme. The prediction of each individual classifier is weighted according to its accuracy on the training set. Both the first and the second choice of a classifier are taken into account. Specifically, the weight w_{ij} of the classifier C_{ij} is as follows:

$$w_{ij} = \frac{a_{ij}}{\sum_{xy} a_{xy}}$$

where a_{ij} is the accuracy of the classifier C_{ij} on the training set (see table 3). $a_{ij}/2$ is used to compute the weight for the second choice of a classifier. The classes can be ordered according to the votes they collect. Specifically, if $c_{ij}(x)$ is the prediction of the classifier C_{ij} for the case x and P is the set of possible classes (i.e., pianists) then the score for a class p is calculated as follows:

$$s_p(x) = \sum_{ij} w_{ij} \left\| c_{ij}(x) = p \right\| \quad p \in P$$

where $\|a=b\|$ is 1 if a is equal to b and 0 otherwise. The greater the score the more probable the pianist as the actual performer. Since both the first and second choices of each base classifier are taken into account, the highest possible score is 0.66 (first choice of all the classifiers) and the lowest is 0 (no first nor second choice of any classifier).

4 Music Performer Verification

In the following experiments, pianists #01-#12 will be used as the set of reference pianists to compute the ‘norm performance’, that is the average performance. The task will be to learn to distinguish pianists #13-#22. Chopin's *Ballade* op. 38 will be used as the training material, and the *Etude* op.10/3 as the test piece. Specifically, the training piece was divided into four non-overlapping segments, each including 40 soprano notes providing four training examples per class for the norm-based and the melody lead classifiers. As concerns the score-based classifiers, the training piece was divided into 16 non-overlapping segments, each including 10 soprano notes. These segments were grouped into four overlapping sets of training examples, leaving out four different segments each time (see table 1).

The task of music performer verification can be viewed as a two-class classification problem. Given a certain performance of the test piece (*Etude*) and a particular pianist (of the set #13-#22) the output of the proposed system will be either 1, i.e., the pianist in question is the actual performer, or 0, i.e., the pianist in question is not the actual performer. The implementation of a music performer verification system requires:

- The definition of a response function for a given pianist. For a given performance, this function should provide an indication of the degree at which the pianist is the actual performer. In this study, the output of the ensemble of classifiers, defined in the previous section is used as response function.
- The definition of a threshold value for this function. For a given performance, any pianist with score lower than the threshold is rejected.

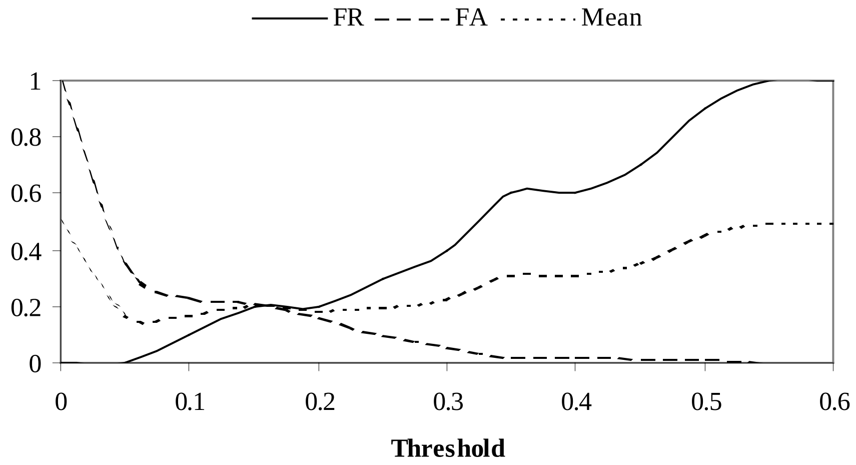

Figure 2. FR, FA, and Mean error of the ensemble model for different threshold values.

Additionally, for measuring the accuracy of a music performer verification method as regards a certain pianist, False Rejection (FR) and False Acceptance (FA) can be used. These measures have been defined in and applied to areas of similar characteristics, such as speaker verification [4] and author verification [9] and are defined as follows:

FR = rejected performances of the pianist / total performances of the pianist
FA = accepted performances of other pianists / total performances of other pianists

For the appropriate selection of the threshold value, the mean error, i.e., (FR+FA)/2, is used. Figure 2 depicts the variation of the average FR, FA, and the mean error values for the performances of the test piece by pianists #13-#22 using threshold values ranging from 0 to 0.6. Since these pianists were taken into account for calculating the discriminant functions and consequently the score function, this evaluation is considered to be a closed-set one. As can be seen, low values of threshold correspond to minimal FR while high values of threshold correspond to minimal FA. The minimal mean error corresponds to the threshold value 0.1 corresponding to FR and FA values of 0.1 and 0.23, respectively.

The results of the method based on the ensemble of classifiers can be compared to the results of the individual base classifiers. In that case, each base classifier is used alone and the response function is the Mahalanobis distance from the centroids of each class. Table 2 shows the FR and FA values for each individual base classifier for a threshold value that minimizes the mean error. As can be seen, the model coming from the learning ensemble is much better as concerns both FR and FA.

Table 2. Average FA and FR values of the base classifiers and the ensemble model. In each model, a threshold value that minimizes mean error is used.

Classifier	FR	FA
Ensemble	0.10	0.23
C_{11}	0.30	0.31
C_{21}	0.40	0.34
C_{22}	0.60	0.40
C_{23}	0.40	0.33
C_{24}	0.50	0.37
C_{31}	0.40	0.31
C_{32}	0.30	0.33
C_{33}	0.40	0.38
C_{34}	0.50	0.36
C_{35}	0.50	0.32

5 Conclusion

We have proposed a computational approach to the problem of distinguishing music performers playing the same pieces focusing on the music performer verification task. A set of features that capture some aspects of the individual style of each performer is presented. Due to the limited available data and certain characteristics of the discriminating features, we proposed a classification model that takes advantage of machine learning techniques for constructing meta-classifiers.

The results show that the proposed learning model performs much better than any of the constituent base classifiers and provides another supporting case for the utility of ensemble learning methods, specifically, the combination of a large number of independent simple ‘experts’. Moreover, it is demonstrated that the differences between music performers can be objectively quantified. While human experts use mostly aesthetic criteria for distinguishing different performers, it is shown that the individuality of each performer can be captured using machine-interpretable features.

The proposed system copes with a difficult musical task, displaying a remarkable level of accuracy. Imagine you first hear 10 different pianists performing one particular piece (and that is all you know about the pianists), and then you have to verify the hypothesis that a particular pianist is (or is not) the actual performer of a certain performance of another (and quite different) piece[1]. The comparison with human experts performing the same task is not straightforward. This is because it is very difficult to define what the similar conditions would be. How many times would the human-expert be allowed to listen to each of the training/test recordings? What would be the level of expertise of the listener? What would be the human-expert’s prior knowledge of the piece? Would such a procedure be meaningful?

[1] The interested reader can attempt to follow this procedure. The digital recordings used in this study can be accessed at: http://www.ai.univie.ac.at/~wernerg/mp3.htm

The reliability of our current results is still severely compromised by the very small set of available data. Substantial effort is required in order to collect and precisely measure a larger and more diverse set of performances by several pianists (on a computer-controlled piano). Studying famous pianists with this approach would require us to be able to precisely measure timing, dynamics, and articulation from sound recordings, which unfortunately still is an unsolved signal-processing problem.

Acknowledgement

This work was supported in part by the EU project HPRN-CT-2000-00115 (MOSART). Many thanks to Gerhard Widmer and Werner Goebl for preparing and providing the music performance data used in this study.

References

1. Bauer, E., Kohavi, R.: An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Machine Learning 39:1/2 (1999) 105-139
2. Blum, A.: Empirical Support for Winnow and Weighted-Majority Based Algorithms: Results on a Calendar Scheduling Domain. Machine Learning, 26:1 (1997) 5-23
3. Eisenbeis, R., Avery, R.: Discriminant Analysis and Classification Procedures: Theory and Applications, Lexington, Mass.: D.C. Health and Co. (1972)
4. Fakotakis, N., Tsopanoglou, A., Kokkinakis, G.: A Text-independent Speaker Recognition System Based on Vowel Spotting. Speech Communication 12 (1993) 57-68
5. Friberg, A.: Generative Rules for Music Performance: A Formal Description of a Rule System. Computer Music Journal, 15:2 (1991) 56-71
6. Lim, T., Loh, W., Shih, Y.: A Comparison of Prediction Accuracy, Complexity and Training Time of Thirty-Three Old and New Classification Accuracy. Machine Learning 40:3 (2000) 203-228
7. Palmer, C.: On the Assignment of Structure in Music Performance. Music Perception 14 (1996) 23-56
8. Repp, B.: Diversity and Commonality in Music Performance: An Analysis of Timing Microstructure in Schumann's 'Träumerei'. Journal of the Acoustical Society of America, 92:5 (1992) 2546-2568
9. Stamatatos E., Fakotakis, N., Kokkinakis, G.: Automatic Text Categorization in Terms of Genre and Author. Computational Linguistics 26:4 (2000) 471-495
10. Stamatatos, E.: A Computational Model for Discriminating Music Performers. Proc. of the MOSART Workshop on Current Research Directions in Computer Music (2001) 65-69
11. Stamatatos, E.: Quantifying the Differences Between Music Performers: Score vs. Norm. Proc. of the International Computer Music Conference (2002) 376-382
12. Widmer, G.: Using AI and Machine Learning to Study Expressive Music Performance: Project Survey and First Report. AI Communications 14 (2001) 149-162
13. Widmer, G.: Discovering Simple Rules in Complex Data: A Meta-learning Algorithm and Some Surprising Musical Discoveries. Artificial Intelligence 146:2 (2003) 129-148
14. Zanon, P., Widmer, G.: Recognition of Famous Pianists Using Machine Learning Algorithms: First Experimental Results. Proc. of the 14th Colloquium of Musical Informatics (2003)