Music Performer Recognition Using an Ensemble of Simple Classifiers

Efstathios Stamatatos[1] and **Gerhard Widmer**[2]

Abstract. This paper addresses the problem of identifying the most likely music performer, given a set of performances of the same piece by a number of skilled candidate pianists. We propose a set of features for representing the stylistic characteristics of a music performer. A database of piano performances of 22 pianists playing two pieces by F. Chopin is used in the presented experiments. Due to the limitations of the training set size and the characteristics of the input features we propose an ensemble of simple classifiers derived by both subsampling the training set and subsampling the input features. Preliminary experiments show that the resulting ensemble is able to efficiently cope with this difficult musical task, displaying a level of accuracy unlikely to be matched by human listeners (under similar conditions).

1 INTRODUCTION

The representation of music as given in the printed score is not able to capture every musical nuance. Hence, a piece played exactly as notated in the printed score would sound mechanical. *Expressive music performance* is the interpretation of a piece of music according to the artist's understanding of the structure (or 'meaning') of the piece. Every skilled performer continuously modifies important parameters, such as tempo and loudness, in order to stress certain notes or 'shape' certain passages. Expressive performance is what makes music come alive and what distinguishes one performer from another (and what makes some performers famous).

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 [6] or by processing large volumes of data using machine learning techniques [11]. Obviously, this direction attempts to explore the similarities between skilled performers in the same musical context. On the other hand, the differences between performers have not been studied thoroughly. Repp [10] presented an exhaustive statistical analysis of temporal commonalities and differences among distinguished pianists' interpretations of a well-known piece and demonstrated the individuality of some famous pianists. However, the differences in music performance are still expressed generally with aesthetic criteria rather than quantitatively.

[1] Austrian Research Institute for Artificial Intelligence, Schottengasse 3, A-1010 Vienna, Austria, email: stathis@ai.univie.ac.at

[2] Department of Medical Cybernetics and Artificial Intelligence, University of Vienna, and Austrian Research Institute for Artificial Intelligence, Schottengasse 3, A-1010 Vienna, Austria, email: gerhard@ai.univie.ac.at

In this paper, we use AI (specifically: machine learning) techniques in an attempt to express the individuality of music performers (pianists) in machine-interpretable terms by quantifying the main parameters of expressive performance. In order to avoid any subjective evaluation of our approach, we apply it to a well-defined problem: the automatic identification of music performers, given a set of piano performances of the same piece of music by a number of skilled candidate pianists. From this perspective, our task can be viewed as a typical classification problem, where the classes are the candidate pianists. A set of features that represent the stylistic properties of a performer is proposed, introducing the 'norm performance' as a reference point, while ideas taken from machine learning research are applied to the construction of the classifier. 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).

First experimental results show that it is indeed possible for a machine to distinguish music performers (pianists) on the basis of their performance style. From the point of view of machine learning, this constitutes another supporting case for the utility of ensemble learning methods (specifically, the combination of a large number of independent simple 'experts' [2]). The contribution of this work to musicology is the identification (via machine learning methodology) of a set of global characteristics of performance style that seem to be relevant to distinguishing different artists. On the other hand, it must be stressed that the current results are still very preliminary and limited because of the limited empirical data available for this investigation. Obtaining precise measurements, in terms of timing deviations, dynamics, and articulation, of performances of highly skilled artists is a difficult task. We are currently investing a large amount of effort into developing new methods for extracting expressive details from given recordings and hope to be able to report on much more extensive experiments in the near future.

2 DATA AND TERMINOLOGY

The data used in this study consists 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

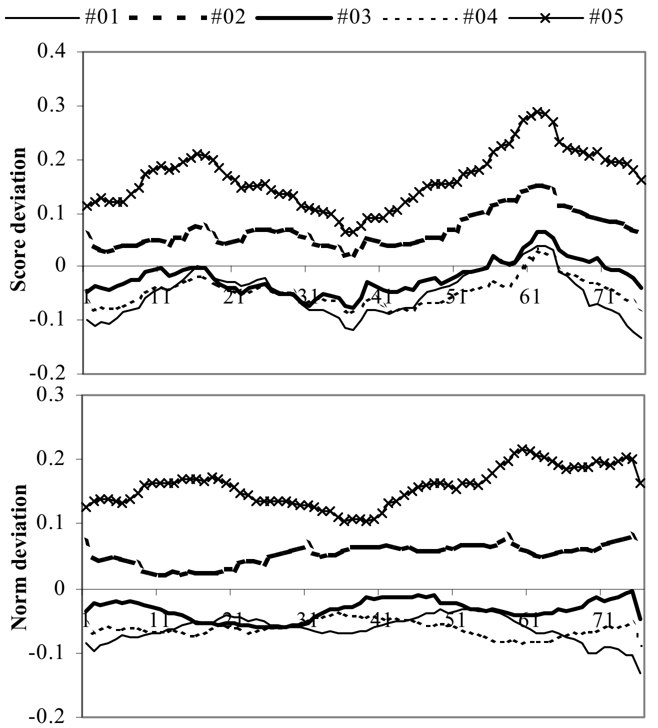

Figure 1. Smoothed timing deviation of the pianists #01-#05 from the printed score (above) and the norm of the pianists #06-#10 (below) for the soprano notes of Chopin's *Etude* op. 10/3.

transcribed into symbolic form and matched against the printed score [3]. 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 22 pianists are referred by their code names (i.e., #01, #02, etc.).

3 FEATURES FOR CHARACTERIZING PERFORMANCE STYLE

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

Figure 1 (top) shows the timing deviation of five pianists (#01-#05) from the printed score of Chopin's *Etude* op. 10/3 (measured as the difference between performed IOIs and the IOIs that would result from a mechanical performance of the piece at a pre-specified fixed tempo). It is obvious that all the pianists tend to deviate from the score in a similar way. That is not surprising. It is well known that to a certain extent, expressive variation is correlated with the structure of the piece of music (e.g., phrase structure, harmonic structure, etc.); indeed, expressive performance is a means for the performer to communicate structural information to the listener. The peaks and dips of the resulting performance curves tend to correlate (more or less strongly) with phrase boundaries and phrase centers. Thus, if we decide to rely on very global summarizations of a pianist's tempo deviations etc. and not to encode detailed aspects of the music played (such as its phrase structure, harmonic structure, etc.), these global features will strongly depend on and vary with the training set. Sampling the training set from slightly different segments of the same piece may affect the output of the classifier substantially.

This problem can be avoided by the use of what we call 'norm deviation features'. In addition to the comparison of the performance of a certain pianist with the printed score, we propose the average performance of a different set of performers as a reference point. Figure 1 (bottom) shows the timing deviation of pianists #01-#05 from the average performance (i.e., norm) of the pianists #06-#10 for the same piece as above. As can be seen, the timing deviations of the first set of pianists from the norm of the second set are more stable across the piece. This is a strong indication that the norm deviation features should not be affected by slight changes to the training set. Given a set of reference performances, the norm deviation can be easily calculated for timing, dynamics, and articulation.

Another valuable source of information comes from the exploitation of the so-called *melody lead* phenomenon [7]. Notes that should be played simultaneously according to the printed score (i.e., 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 [9] 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.

Specifically, then, 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:

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

where $D(\mathbf{x}, \mathbf{y})$ (a scalar) denotes the deviation of a vector of numeric values $\mathbf{x}$ from a reference vector $\mathbf{y}$, 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.

For measuring the deviation in each of the above features, different types of distance could be applied. We decided to choose the appropriate type of distance for each feature category according to its statistical significance in the training set. In the following experiments, Chopin's *Ballade* op. 38 will be used as the training material, and the *Etude* op.10/3 as the test piece. Pianists #01-#12 will be used as the set of reference pianists to compute the 'norm performance', and the task will be to learn to distinguish pianists #13-#22. For determining the best type of distance measure for each type of feature, the training piece (the *Ballade*) was divided into four non-overlapping segments, each including 40 soprano notes. For each segment of the performance of the piece by the pianists #13-#22, the values of the proposed features for the following different types of distance were calculated:

Simple: $D_s(x,y) = \left(\sum_{i=1}^{n}(x_i - y_i)\right) \Big/ n$

Relative: $D_r(x,y) = \left(\sum_{i=1}^{n}\frac{(x_i - y_i)}{x_i}\right) \Big/ n$

Simple absolute: $D_{sa}(x,y) = \left(\sum_{i=1}^{n}|x_i - y_i|\right) \Big/ n$

Relative absolute: $D_{ra}(x,y) = \left(\sum_{i=1}^{n}\frac{|x_i - y_i|}{x_i}\right) \Big/ n$

Then, analysis of variance (aka ANOVA) was applied to these values for extracting conclusions about the statistical significance of the different types of distance and features. The most significant features proved to be the deviation from the norm in terms of timing and articulation, the timing deviation between the first and the third voice as well as between the first and the fourth voice (the bass line), and the deviation from the score in terms of timing and articulation. As regards the different types of distances, D_r gave the best results for the score deviation features. This type of distance has been used previously for comparing different performances. D_s seems to be the appropriate selection for the norm deviation features. Finally, D_{sa} fits better the melody lead features, which indicates that information on whether a voice precedes or follows the first voice in a chord is not that important as the degree to which deviates from it.

4 THE CLASSIFICATION MODEL

4.1 Problem characteristics

Since only two pieces were available (one of which should serve as independent test piece), the training examples of the music performer classifier should consist of piece segments rather than entire musical pieces.

To determine the best mode of segmentation (equal length segments or segments based on the piece's phrase structure), a simple experiment was performed. A number of simple classifiers, based on different types of features and distance definitions, were trained (via discriminant analysis – see below) using the

Table 1. Comparison of score and norm deviation measures for different types of distance and different methods of forming training examples.

	Distance	Accuracy (%) Equal-length	Phrase-based
Score	D_s	52.5	50
	D_r	60	52.5
	D_{sa}	40	30
	D_{ra}	52.5	42.5
Norm	D_s	82.5	77.5
	D_r	57.5	45
	D_{sa}	45	45
	D_{ra}	20	20

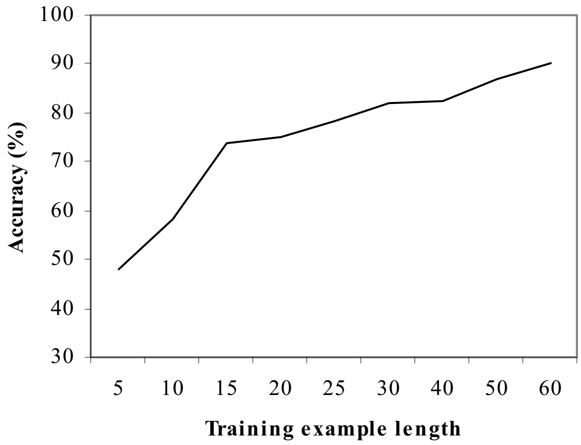

Figure 2. Classification accuracy vs. training example length (in soprano notes).

performances of the pianists #13-#22 of *Ballade* op. 38, with different methods of segmenting the piece into training examples: in one case, the piece was segmented into four parts of equal length (40 soprano notes each), in the other, it was cut into four parts according to phrase boundaries that were identified manually by a human expert. Table 1 shows the classification accuracy results (leave-one-out evaluation on the original data). As can be seen, in all the cases the classifiers based on training examples of equal length gave better or equal accuracy results in comparison with the phrase-based classifiers. The norm deviation features generally outperformed the score deviation features.

Figure 2 shows the relation of the length of the training examples (number of soprano notes) with the classification accuracy using *Ballade* op. 38 as testing ground and the norm deviation features. The longer the segments that constitute the training examples, the more accurate the classifier. This means that for constructing reliable classifiers it is necessary to have training examples as long as possible, which makes for a rather small number of examples and again means that the number of input features per example (segment) should be rather small (in order to avoid overfitting of the training data).

4.2 The proposed ensemble

All the above characteristics of the problem suggest the use of an *ensemble of classifiers* rather than a unique classifier. Recent research in machine learning [1, 4] has studied thoroughly the construction of meta-classifiers. In this study, we take advantage of these techniques, constructing an ensemble of classifiers derived

Table 2. Description of the proposed simple classifiers. The third column indicates the number of training examples (and their length in soprano notes) per class.

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

Table 3. Predictions of the individual simple classifiers on performances of the unseen test set (*Etude* op. 10/3). The first column indicates the code of the actual performer. Correct predictions are in boldface. Last row summarizes correct guesses.

Actual	C_{11}	C_{21}	C_{22}	C_{23}	C_{24}	C_{31}	C_{32}	C_{33}	C_{34}	C_{35}
#13	**#13**	**#13**	#16	**#13**	#18	**#13**	**#13**	**#13**	**#13**	**#13**
#14	**#14**	#21	**#14**	#22	#22	#21	#21	#13	#21	#15
#15	#21	#21	#14	#21	#14	**#15**	#13	**#15**	#17	#13
#16	#18	#18	**#16**	#18	#18	**#16**	**#16**	#19	**#16**	**#16**
#17	**#17**	**#17**	**#17**	**#17**	**#17**	#15	**#17**	#16	#16	#21
#18	#13	#13	#16	**#18**	**#18**	#17	#17	#22	**#18**	#14
#19	#13	**#19**	**#19**	#13	#13	#16	**#19**	**#19**	#16	**#19**
#20	#14	#21	#14	#14	#14	**#20**	**#20**	#14	#14	**#20**
#21	#14	#14	#14	#14	#14	#17	#17	#13	**#21**	#14
#22	**#22**	#17	#19	#19	**#22**	#16	#16	#15	#16	#16
Correct:	4	3	4	3	3	4	5	3	4	4

from subsampling the input features and subsampling the training data set. The former 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 latter 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 2. 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 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 *learning algorithm* used to construct the individual experts is *discriminant analysis*, a standard technique of multivariate statistics, which constructs a set of linear functions of the input variables by maximizing the between-group variance while minimizing the within-group variance [5].

The last column in table 2 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.

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 [8]. 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 2). $a_{ij}/2$ is used to compute the weight for the second choice of a classifier. The class receiving the highest votes is the final class prediction. 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 final prediction is extracted as follows:

$$\hat{c}(x) = \arg\max_{p \in P} \sum_{ij} w_{ij} \left\| c_{ij}(x) = p \right\|$$

where $\|a=b\|$ is 1 if a is equal to b and 0 otherwise.

4.3 Experimental results

The individual base classifiers as defined above were trained on the performances of the *Ballade* op.38 by pianists #13-#22; pianists #01-#11 were used to define the 'performance norm'. Both the individual base classifiers and the combined ensemble classifier were then tested on an independent test piece, the *Etude* op.10/3. Table 3 shows the classification results for the individual base classifiers. The classification accuracy of each individual classifier ranges between 30% and 50%. The errors of norm deviation and score deviation classifiers are partially correlated (i.e., common misclassifications: #16-#18, #19-#13, #20-#14, #21-#14). On the

Table 4. Predictions (first and second choice) of the ensemble of the simple classifiers on performances of the unseen test set (*Etude* op. 10/3). The first column indicates the code of the actual performer. Correct predictions are in boldface. Last row summarizes correct guesses.

Actual	1st choice	Score	2nd choice	Score
#13	**#13**	0.56	#18	0.23
#14	**#14**	0.31	#21	0.29
#15	#21	0.34	#14	0.25
#16	**#16**	0.46	#18	0.34
#17	**#17**	0.47	#15	0.16
#18	**#18**	0.30	#13	0.26
#19	**#19**	0.40	#13	0.27
#20	#14	0.42	**#20**	0.22
#21	#14	0.51	#22	0.15
#22	**#22**	0.29	#16	0.25
Correct:	7		1	

other hand, the errors of the melody lead classifiers are highly uncorrelated in comparison to the others. Note that uncorrelated errors are very crucial for constructing ensembles of classifiers [4].

Table 4 shows the classification results of the ensemble classifier. The ensemble correctly identified the pianist in 7 out of 10 cases, which gives an accuracy of 70%. The ensemble thus performs substantially better than any of the constituent classifiers. The score assigned to each prediction can be used as an indication of the classifier's certainty. Thus, the classification of the performances by pianists #14, #18, and #22 are the most difficult cases since the distance of the first choice from the second choice is less than 0.05.

Note that 70% is a high success rate in a 10-class task. Note also that this would be a very difficult task for a human: 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 identify each of the 10 pianists in a recording of another (and quite different) piece. We are planning a classification experiment with human listeners to measure the level of human performance in this type of task; we expect it to be substantially lower.

5 CONCLUSIONS

We have presented a computational approach to the problem of discriminating between music performers playing the same piece of music, and introduced a set of features that capture some aspects of the individual style of each performer. In order to cope efficiently with this problem, we proposed a classification model that takes advantage of various techniques of constructing meta-classifiers. The results show that the differences between music performers can be quantified. While human experts use mostly aesthetic criteria for recognizing different performers, it is demonstrated that the individuality of each performer can be objectively captured using machine-interpretable features.

This research is performed in the context of a large research project whose goal is to study fundamental principles of expressive music performance with AI methods. The current study can be seen as another attempt at discovering and quantifying features that are crucial to understanding and modeling this complex phenomenon.

The proposed features can be easily computed and do not make use of any piece-specific information (e.g., extracted by structural or harmonic analysis). However, the results cannot be easily interpreted in terms of the traditional music theory. Thus, the proposed features are not likely to help in the explanation of the differences between the performers. Such a task would require features associated with particular local musical contexts and piece-specific information.

The reliability of our current results is still severely compromised by the very small set of empirical data that were available. It is planned to invest substantial effort in the future into collecting and precisely measuring a larger and more diverse set of performances by a set of different pianists (on a computer-controlled piano). Studying famous concert 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.

ACKNOWLEDGEMENTS

This work was supported by the EU project HPRN-CT-2000-00115 (MOSART) and the START program of the Austrian Federal Ministry for Education, Science, and Culture (Grant no. Y99-INF). The Austrian Research Institute for Artificial Intelligence acknowledges basic financial support from the Austrian Federal Ministry for Education, Science, and Culture.

REFERENCES

[1] E. Bauer and R. Kohavi, 'An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants', *Machine Learning*, **39** (1/2), pp. 105-139, (1999).

[2] A. Blum, 'Empirical Support for Winnow and Weighted-Majority Based Algorithms: Results on a Calendar Scheduling Domain', *Machine Learning*, **26** (1), pp. 5-23, (1997).

[3] E. Cambouropoulos, 'From MIDI to Traditional Music Notation', In *Proc. of the AAAI'2000 Workshop on Artificial Intelligence and Music*, 17th National Conf. On Artificial Intelligence, pp. 19-23 (2000).

[4] T. Dietterich, 'Ensemble Methods in Machine Learning', *First Int. Workshop on Multiple Classifier Systems*, pp. 1-15, (2000).

[5] R. Eisenbeis and R. Avery, *Discriminant Analysis and Classification Procedures: Theory and Applications,* Lexington, Mass.: D.C. Health and Co., 1972

[6] A. Friberg, 'Generative Rules for Music Performance: A Formal Description of a Rule System' *Computer Music Journal*, **15** (2), pp. 56-71, (1991).

[7] W. Goebl, 'Skilled Piano Performance: Melody Lead Caused by Dynamic Differentiation', In *Proc. of the 6th Int. Conf. on Music Perception and Cognition*, (2000).

[8] D. Opitz and J. Shavlik, 'Generating Accurate and Diverse Members of a Neural-Network Ensemble', In D. Touretzky, M. Mozer, and M. Hasselmo (Eds.) *Advances in Neural Information Processing Systems*, **8**, pp. 535-541, (1996).

[9] C. Palmer, 'On the Assignment of Structure in Music Performance', *Music Perception*, **14**, pp. 23-56, (1996).

[10] B. 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), pp. 2546-2568, (1992).

[11] G. Widmer, 'Using AI and Machine Learning to Study Expressive Music Performance: Project Survey and First Report', *AI Communications*, **14**, pp. 149-162 (2001).