Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Artificial Intelligence 165 (2005) 37–56
Artificial Intelligence
www.elsevier.com/locate/artint

Automatic identification of music performers with learning ensembles

Efstathios Stamatatos [a], Gerhard Widmer [b,c,*]

[a] *Department of Information and Communication Systems Engineering, University of the Aegean, Samos, Greece*
[b] *Department of Computational Perception, Johannes Kepler University, Linz, Austria*
[c] *Austrian Research Institute for Artificial Intelligence, Vienna, Austria*

Received 29 April 2004

Available online 14 March 2005

Abstract

This article 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 very simple features for representing stylistic characteristics of a music performer, introducing 'norm-based' features that relate to a kind of 'average' performance. A database of piano performances of 22 pianists playing two pieces by Frédéric 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. Experiments show that the proposed features are able to quantify the differences between music performers. The proposed ensemble can efficiently cope with multi-class music performer recognition under inter-piece conditions, a difficult musical task, displaying a level of accuracy unlikely to be matched by human listeners (under similar conditions).

Keywords: Machine learning; Classification; Ensemble learning; Music

* Corresponding author.
E-mail addresses: stamatatos@aegean.gr (E. Stamatatos), gerhard.widmer@jku.at (G. Widmer).

0004-3702/$ – see front matter
doi:10.1016/j.artint.2005.01.007

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 and highly unmusical. Expressive music performance is the act of 'shaping' 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 formulating rules or principles of expressive performance either with the help of human experts [7] or by processing large volumes of data using machine learning techniques [25,26]. Another direction of research is based on implicit knowledge extracted from recordings of human performers using case-based reasoning techniques [13]. Obviously, these directions attempt to explore the similarities between skilled performers in similar musical contexts. On the other hand, the differences between performers have not been studied thoroughly. Repp [17] presented a 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.

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.[1] 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).

Experimental results show that it is indeed possible for a machine to distinguish music performers (pianists) on the basis of their performance style. We will show that successful learning from extremely limited training data can be achieved by maximally exploiting the given information via ensemble learning (based on subsampling both the data and the input features). From the point of view of machine learning, this constitutes another supporting

[1] In a sense, this is also related to currently ongoing efforts in the area of Music Information Retrieval (MIR), where much work is devoted to the automatic classification of music recordings according to style or artist (see, e.g., [22]).

case for the utility of ensemble learning methods, specifically, the combination of a large number of independent simple 'experts' [3]. 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 presented results are rather 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 future.

This paper is organized as follows: the next section contains a brief description of the data and terminology used in this study. Section 3 describes the proposed features for the quantification of the music performance style. Section 4 explains how preliminary experimentation was performed in order to establish the main parameters and strategy for learning. Section 5 then presents the ensemble learning experiments performed and the results achieved. Finally, Section 6 discusses the major conclusions drawn from this study and future work directions.

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 Frédéric Chopin: the Etude op.10 no.3 (first 21 bars) and the Ballade op.38 (initial section, bars 1 to 45). The digital recordings were then transcribed into symbolic form (MIDI) and matched against the printed score [4]. 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 a played note, and how the note is connected to the following one (articulation). All this can be readily computed from our data.[2]

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, and Dynamic Level (DL) as the 'loudness'

[2] Note that this is only incomplete performance information. For example, we currently do not use information about the pedalling behaviour. Also other expressive information related to the produced sound, such as the 'attack' or 'touch' of the notes (if there is such a thing—see [10]), cannot be measured in our data, because our data is derived from symbolic MIDI events.

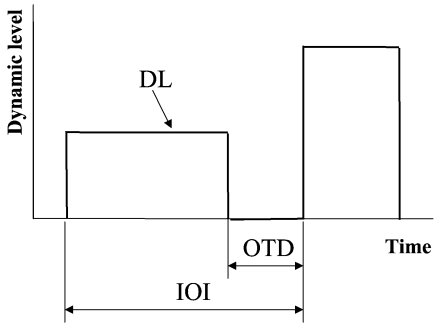

Fig. 1. The three main parameters used to characterize note-level performance details: Dynamic Level (DL), Inter-Onset Interval (IOI), and Off-Time Duration (OTD).

of an individual note (in terms of the MIDI velocity parameter).[3] The parameters are illustrated graphically in Fig. 1. The 22 pianists are referred to by their code names (#01, #02, etc.).

3. Quantifying music performance style

3.1. Score and norm

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.

Fig. 2 shows the performances of the first 30 soprano notes of the Ballade by the pianists #01–#05 in terms of timing, expressed as the real duration (played inter-onset intervals) of the melody's sixteenth notes,[4] and dynamics. The default tempo and dynamic levels of a non-expressive, purely mechanical interpretation of the score (with an arbitrarily fixed tempo and loudness level) would 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

[3] Generally, loudness is a rather complex concept. In our study, we are only interested in the relative loudness at the time of onset of a note (not, e.g., in how the acoustic tone changes over time), which is essentially what the pianist can control, and what computer-controlled pianos measure. Onset loudness is measured in MIDI in terms of the velocity with which the key was depressed—hence the name 'velocity' for the corresponding MIDI parameter.

[4] The sixteenth note is the shortest duration category appearing in the piece. IOIs longer than an eighth note were divided into the appropriate number of virtual sixteenth notes for the figure. For instance, a played quarter note is divided into 4 sixteenths of equal duration, and that duration is then plotted in Fig. 2.

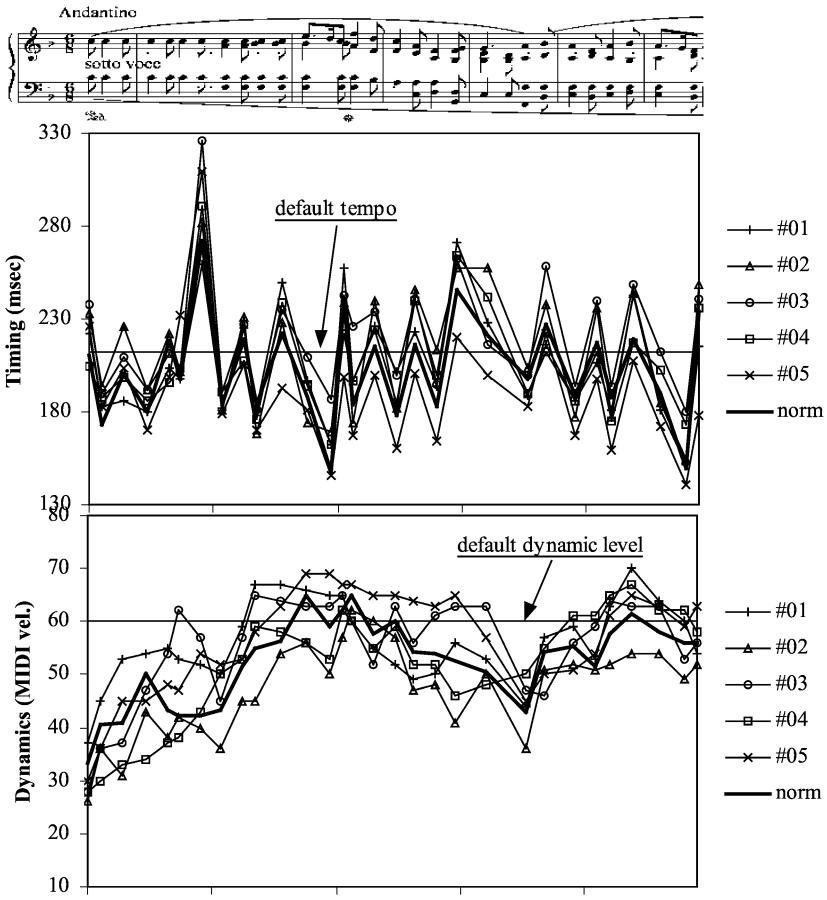

Fig. 2. Timing and dynamics variations for the first 30 soprano notes of Chopin's *Ballade* op. 38 (score above) as performed by pianists #01–#05. Default tempo and dynamic level, and performance norm derived from pianists #06–#10 are depicted as well.

(last note of the introductory part) while in the dynamics dimension the first two bars are played with increasing intensity (introductory part), and the second soprano note of the fifth bar is played rather softly (a phrase boundary).

These similarities in the performances remain when we take a more global look at curves that are smoothed over longer melodic passages. Fig. 3 (top) shows the timing deviations of five pianists (#01–#05) from the printed score of the Chopin Etude (measured as the moving average of 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

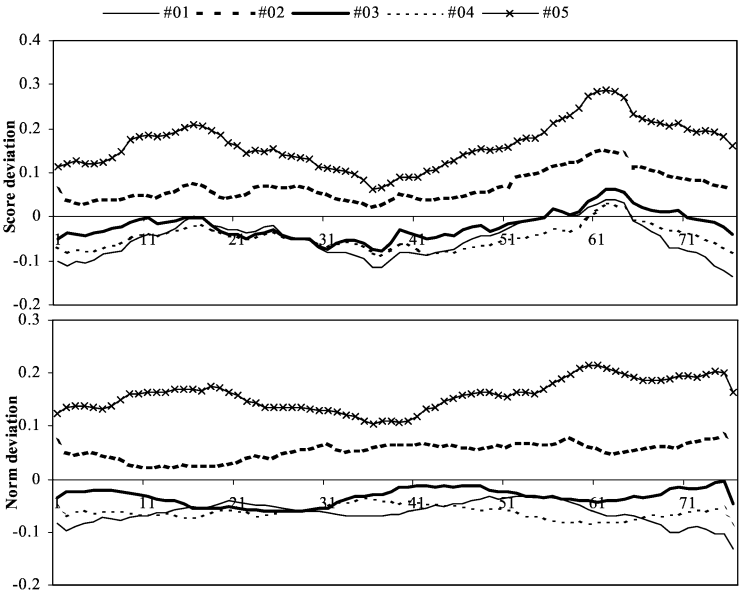

Fig. 3. Smoothed timing deviation of the pianists #01–#05 from the printed score (above) and from the norm of pianists #06–#10 (below) for the soprano notes of Chopin's *Etude* op.10, no.3.

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 (as we will in the following sections) to take entire segments of a piece as training examples and use as features global summarizations of a pianist's tempo and dynamics deviations across those segments (rather than looking at single notes and detailed aspects of the music played, such as its phrase structure, harmonic structure, etc., which would have to be produced by a labour-intensive 'manual' musical structure analysis), 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 substantially change the values of some attributes.

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 *averaged performance* of a different set of performers as a reference point. Fig. 2 shows such a *norm performance*, in terms of timing and dynamics on the single-note level, calculated from the performances of pianists #06–#10. As can be seen, the norm follows the basic form of the individual performances. Similarly, Fig. 3 (bottom) shows the timing deviation of pianists #01–#05 from the average performance (i.e., norm) of the pianists #06–#10 for the Etude, in terms of smoothed differences over multiple-note passages. 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.

3.2. Melody lead

Another valuable source of information comes from the exploitation of the so-called *melody lead* phenomenon [9]. 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 tends to precede the other voices and is usually played louder. Studies of this phenomenon [15] showed that melody lead generally 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.

3.3. The proposed features

As mentioned above, the training examples will be segments of a piece (more precisely, the melody) of a certain length, i.e., sequences of played notes. We propose the following types of global features for characterizing such performed segments, given the printed score and a performance norm derived from a given set of different performers:

– Score deviation features:[5]

$D(IOI_s, IOI_m)$ timing,
$D(IOI_s, OTD_m)$ articulation,
$D(DL_s, DL_m)$ dynamics,

– Norm deviation features

$D(IOI_n, IOI_m)$ timing,
$D(OTD_n, OTD_m)$ articulation,
$D(DL_n, DL_m)$ dynamics,

– Melody lead features:

$D(ON_{xy}, ON_{zy})$ timing,
$D(DL_{xy}, DL_{zy})$ dynamics,

where $D(\mathbf{x}, \mathbf{y})$ (a scalar) denotes the deviation of a vector of numeric values $\mathbf{y}$ from a reference vector $\mathbf{x}$, IOI_s and DL_s are the vectors of the nominal inter-onset intervals and dynamic-levels, 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 xth voice within the chord y. The same score-based features have been used in previous work for successfully discriminating two skilled performers playing the same piano pieces [18].

[5] Note that articulation deviation from the printed score is calculated based on the score IOIs because the score OTDs would always be zero.

4. Preliminary experiments

Applied machine learning is rarely confined to the simple application of an induction algorithm to a given data set. Many alternative learning algorithms are available, the data can be filtered and represented in various ways, and even the question of what should be regarded as an individual training example is often a non-trivial one. The present project is a case in point. This section gives a brief account of a variety of preliminary experiments that had to be performed in order to establish some of the basic parameters of the learning task—definition of features and training examples—and to provide an initial impression of the difficulty of the problem. This is mainly to give the reader an appreciation of the kinds of 'mundane' processing and analysis steps that are often essential to the success of empirical machine learning projects. It is the experiences gathered in these experiments that prompted us to opt for the ensemble learning approach that will be described in Section 5. The reader interested only in the final results can safely skip Sections 4.2–4.4.

4.1. Data and base-level classification algorithm

In the following experiments, Pianists #01–#12 will be used as the set of reference pianists to compute the *norm performance*. The task will be to learn to distinguish pianists #13–#22, which gives a 10-class classification problem. The 'real' experiments to be reported on in Section 5 will test learning and prediction in inter-piece conditions. There, the performances of Chopin's *Ballade* op. 38 will be used as the training material, and the performances of the *Etude* op.10/3 as the test cases. The preliminary 'parameter determination experiments' reported in the present section use only the training piece (the Ballade), so that the optimizations performed will not in any way be influenced by knowledge of the data eventually used for testing.

The *classification method* used in the following experiments 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 [6]. A set of linear functions (based on individual 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 Mahalanobis 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 $\mathbf{x}$ from a mean vector $\mathbf{m}$ is defined as follows:

$$d^2 = (\mathbf{x} - \mathbf{m})' C_{\mathbf{x}}^{-1} (\mathbf{x} - \mathbf{m}) \tag{1}$$

where $C_{\mathbf{x}}$ is the covariance matrix of $\mathbf{x}$. 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. A recent study [12] compared discriminant analysis to several more complex classification methods (from statistics, decision trees, and neural networks) and showed that discriminant analysis is absolutely competitive when considering the compromise between classification accuracy and training time cost.

4.2. Selecting the appropriate distance type

For measuring the deviation in each of the features defined in the previous section, 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 a training set. 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 features were calculated for the following different types of distance (or, more correctly, deviation measures, as some of them are not distances, mathematically speaking):

$$\text{Simple: } D_s(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i), \tag{2}$$

$$\text{Relative: } D_r(x, y) = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - y_i)}{x_i}, \tag{3}$$

$$\text{Simple absolute: } D_{sa}(x, y) = \frac{1}{n}\sum_{i=1}^{n}\big(|x_i - y_i|\big), \tag{4}$$

$$\text{Relative absolute: } D_{ra}(x, y) = \frac{1}{n}\sum_{i=1}^{n}\frac{(|x_i - y_i|)}{x_i}. \tag{5}$$

Then, analysis of variance (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 [8]. D_s seems to be the appropriate choice for the norm deviation features. D_{sa} is the best distance type for the melody lead features, which indicates that information on whether a voice precedes or follows the first voice in a chord is not as important as the degree to which the voices are separated in time and dynamics.[6]

4.3. Determining appropriate training examples

Since only two musical pieces are available in our data (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

[6] Interestingly, this observation is partly corroborated, from a different angle, by a very recent experimental listening study with human subjects [11].

Table 1
Comparison of score and norm deviation measures for different types of distance and different methods of forming training examples (accuracies computed by leave-one-out cross-validation on the training set)

		Accuracy (%)	
	Distance	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

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 on the performances by pianists #13–#22 of the 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 of different length, according to phrase boundaries that were identified manually by a human expert. Table 1 shows the classification accuracies achieved (computed via leave-one-out cross-evaluation on the training data). As can be seen, in all cases the classifiers based on training examples of equal length gave better or equal accuracy results in comparison with the phrase-based classifiers. This is a strong indication that the proposed features either cannot benefit from or are independent of traditional musicological definitions of musical structure (such as phrase structure). In addition, the norm deviation features generally outperformed the score deviation features.

Another experiment was concerned with the length (in terms of soprano notes) of the training examples. Fig. 4 shows the relation of the length of the training examples with the classification accuracy that can be obtained, again using Ballade op. 38 as testing ground and the norm deviation features. In this intra-piece condition, it turns out that the longer the segments that constitute the training examples, the more accurate the classifier. The same holds for the score deviation or the melody lead features. 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 in turn means that the number of input features per example (segment) should be rather small, in order to avoid overfitting of the training data (most learning methods—including discriminant analysis—are not able to generalize well when given too many input features in comparison to the number of training examples per class).

4.4. Performance of the simple classification model

In order to provide an initial impression of the difficulty of the problem and to reveal the most important similarities and differences between the performers, a simple classifi-

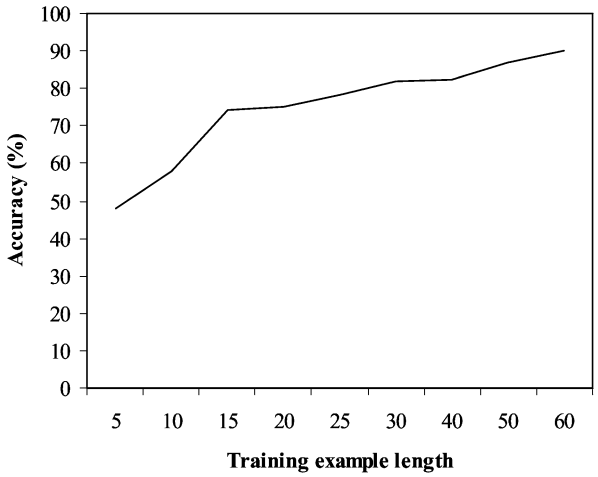

Fig. 4. Classification accuracy vs. training example length (in soprano notes).

cation model is first described. The performances of Ballade were segmented into 8 parts of equal length (20 soprano notes). These segments were then separated into two data sets, henceforth called Ballade-1 and Ballade-2, comprising the first four segments and the last four segments of each performance, respectively. Additionally, the performances of Etude were segmented into 4 parts of equal length (20 soprano notes). Thus, three data sets each one comprising four examples per class became available. This enabled us to perform both intra-piece (training and test sets taken from the same piece) and inter-piece (training and test sets taken from different pieces) experiments.

Due to the restricted number of features that should be used concurrently in a classifier (because of the danger of over-fitting), distinct classifiers for norm-based and score-based features were developed. Hence, three score-based classifiers and three norm-based classifiers were constructed based on the training sets of Ballade-1, Ballade-2, and Etude. This also enables the objective comparison between norm deviation and score deviation features. Fig. 5 depicts the class centroids in the space of the first two discriminant functions (which account for the greatest part of the total variation) derived from Ballade-1 and Etude, respectively, for both the norm-based and score-based features. Note that only the relative positions of the centroids can be compared, not the exact values of discriminant functions. As can be seen, in both cases the positions of the class centroids derived from the norm-based and the score-based features have many similarities.

However, a closer look reveals that by using the norm-based features, the centroids are distributed more widely along the first discriminant function (which by far accounts for the greatest part of the total variation). Specifically, in the case of Ballade-1, the first discriminant function values of the centroids lie between -6.8 and 3.7 for norm-based features and between -3.9 and 2.1 for score-based features. The corresponding spans in the case of Etude are between -3.8 and 5.7 for norm-based features and between -2.9 and 3.5 for score-based features. Similar observations can be made for the second discriminant function's spans. This fact means that the norm-based features are better able to produce robust

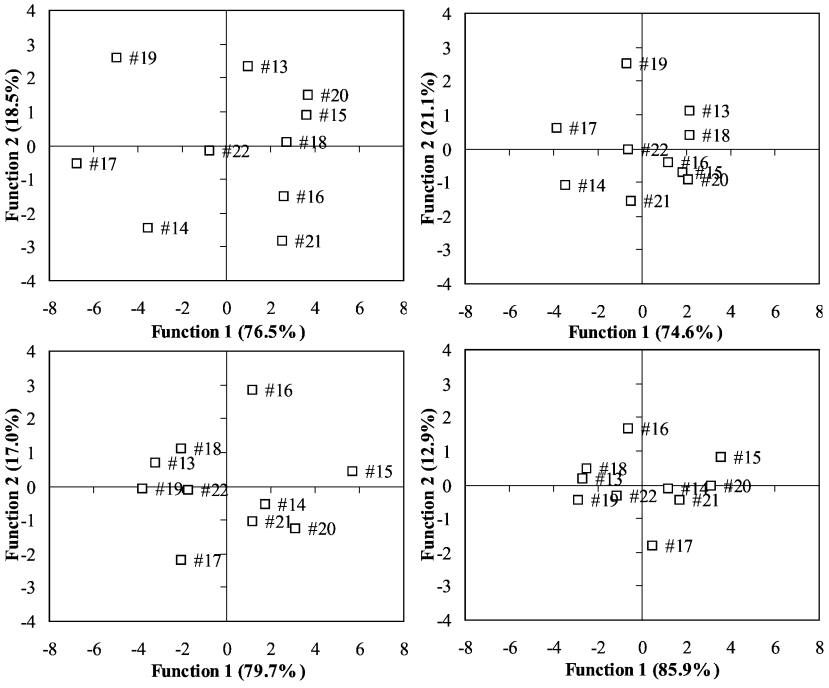

Fig. 5. The centroids of the pianists #13–#22 in the space of the first two discriminant functions for *Ballade*-1 (above) and *Etude* (below). Norm-based models are shown on the left side and score-based models on the right side. The numbers inside parentheses indicate the amount of variance explained by the corresponding function.

and reliable classifiers since the classes are more widely spread within the classification space.

The examination of the relative positions of centroids between Ballade-1 and Etude indicates that many similarities and differences between performers remain constant in inter-piece conditions. For instance, in both data sets the classification models reveal a proximity between pianists #13 and #19, #16 and #18, #14 and #22, etc. Naturally, these relations are much stronger between classification models extracted from segments of the same piece (i.e., between Ballade-1 and Ballade-2).

The results of applying the norm-based and score-based classifiers to new unseen musical parts taken either from the same piece or from a different one are given in Table 2. Each classification model derived from a training set was applied to the other two data sets. In this experiment each test set consisted of a single case per class. To illustrate this further, for instance, the classifier trained on the performances of Ballade-1 (the first half of the piece) was applied to the performances of Ballade-2 (the second half of the same piece) and the performances of Etude (a different piece), attempting to predict the most likely performer. To imitate this procedure, a human expert should first hear 10 different performances of the first half of a piece and then try to guess the performer of the second half or of another piece. The results for all possible combinations of training and test set are given in Table 2.

Table 2
Prediction results for norm-based and score-based classifiers; number of correct predictions (maximum possible is 10)

Training set		Test set		
		Ballade-1	Ballade-2	Etude
Norm	Ballade-1	–	9	4
	Ballade-2	9	–	4
	Etude	3	4	–
Score	Ballade-1	–	7	1
	Ballade-2	5	–	5
	Etude	3	4	–

As can be seen, the results in intra-piece conditions (training set: Ballade-1, test set: Ballade-2, and vice versa) of the norm-based classifiers are quasi perfect (9 out of 10 correct predictions), significantly better than the performance of the score-based classifiers.[7]

On the other hand, in inter-piece conditions, the performance of the norm-based and score-based classifiers is comparable. However, the norm-based classifiers are more robust or stable (3–4 correct predictions) in comparison to the score-based ones (correct predictions ranging from 1 to 5).

5. Ensemble learning

As shown in the previous section, the efficiency of the individual classifiers, based on only one feature source, is limited under inter-piece conditions. The characteristics of the problem suggest the use of an ensemble of classifiers rather than a unique classifier. Recent research in machine learning [2,5,21] has studied thoroughly the construction of meta-classifiers. In this study, we take advantage of these techniques, constructing an ensemble of classifiers using two basic strategies:

Subsampling the input features. This technique is usually applied when multiple redundant features are available. In our case, the input features should not 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 data set. This technique is known to produce best results when 'unstable' learning algorithms (i.e., algorithms that tend to produce strongly different models from slightly different data) are used for constructing the base classifiers. In our case, a subset of the input features (i.e., the score deviation

[7] This should not come as a surprise. Obviously, the norm-based features along with the simple distance D_s can easily capture consistent differences in tempo and loudness between pianists, and these are much more likely in intra-piece conditions: a pianist A playing the first half of a piece faster than pianist B will also tend to play the second half faster than B (although that also need not be true for each passage of the piece). Things are more complex in inter-piece conditions, and with performance parameters such as articulation and melody lead.

measures) is unstable—their values can change drastically given a slight change in the selected training segments.

5.1. The proposed ensemble

Given the scarcity of training data and the multitude of possible features, we chose to use a relatively large number of rather simple individual base classifiers or 'experts', in the terminology of [3]. 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 3. C11 is based on the deviation of the performer from the norm. C21, C22, C23, and C24 are based on the deviation of the performer from the score and are trained using slightly changed training sets (because the score 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 [16]). Finally, C31, C32, C33, C34, and C35 are based on melody lead features. They differ in whether they consider inter-chord differences in timing and/or dynamics, and between which voices of a chord. The last column in Table 3 shows the cross-validated accuracy of each individual expert on the training data (i.e., in an intra-piece condition). 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 [14]. Both the first and the second choice of a classifier are taken into account. Specifically, the weight w_{ij} of classifier C_{ij} is as follows:

$$w_{ij} = \frac{a_{ij}}{\sum_{xy} a_{xy}} \tag{6}$$

Table 3
Description of the proposed simple classifiers. The third column indicates the number of training examples (and their length in terms of soprano notes) per class

Code	Input features	Training examples	Acc. (%)
C11	$D_s(IOI_n, IOI_m)$, $D_s(OTD_n, OTD_m)$, $D_s(DL_n, DL_m)$	4×40	82.5
C21	$D_r(IOI_s, IOI_m)$, $D_r(IOI_s, OTD_m)$, $D_r(DL_s, DL_m)$	12×10	50.8
C22	$D_r(IOI_s, IOI_m)$, $D_r(IOI_s, OTD_m)$, $D_r(DL_s, DL_m)$	12×10	44.8
C23	$D_r(IOI_s, IOI_m)$, $D_r(IOI_s, OTD_m)$, $D_r(DL_s, DL_m)$	12×10	46.7
C24	$D_r(IOI_s, IOI_m)$, $D_r(IOI_s, OTD_m)$, $D_r(DL_s, DL_m)$	12×10	48.3
C31	$D_{sa}(ON_{1m}, ON_{2m})$, $D_{sa}(ON_{1m}, ON_{3m})$, $D_{sa}(ON_{1m}, ON_{4m})$	4×40	57.5
C32	$D_{sa}(DL_{1m}, DL_{2m})$, $D_{sa}(DL_{1m}, DL_{3m})$, $D_{sa}(DL_{1m}, DL_{4m})$	4×40	42.5
C33	$D_{sa}(ON_{1m}, ON_{2m})$, $D_{sa}(DL_{1m}, DL_{2m})$	4×40	25.0
C34	$D_{sa}(ON_{1m}, ON_{3m})$, $D_{sa}(DL_{1m}, DL_{3m})$	4×40	35.0
C35	$D_{sa}(ON_{1m}, ON_{4m})$, $D_{sa}(DL_{1m}, DL_{4m})$	4×40	47.5

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 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 $\hat{c}(x)$ is

$$\hat{c}(x) = \arg\max_{p \in P} \sum_{ij} w_{ij} \times eq\big(c_{ij}(x), p\big) \tag{7}$$

where $eq(a, b) = 1$ if a is equal to b and 0 otherwise.

5.2. Classification results

For testing the proposed ensemble under inter-piece conditions, the Ballade op. 38 was used as the training material, and the Etude op.10/3 as the test piece. This is because the former is the longer piece and therefore the acquisition of long training examples is possible. The training piece (the Ballade) was divided into four non-overlapping segments of 40 soprano notes each, which gives (only) four training cases per pianist, for each of the ten target pianists #13–#22. The individual base classifiers as defined above were trained on the performances of the Ballade by pianists #13–#22; pianists #01–#12 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, which was used in its entirety as one segment. Table 4 shows the classification results for the individual base classifiers. The classification accuracies of the individual classifiers range 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 other hand, the errors of the melody lead classifiers seem quite independent of the others. Note that uncorrelated errors are crucial for the success of ensembles of classifiers [5].

Table 5 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%. One

Table 4
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	C11	C21	C22	C23	C24	C31	C32	C33	C34	C35
#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

Table 5
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	

additional pianist (#20) is correctly recognised if we also admit the learner's second choice. The ensemble thus performs substantially better than any of the constituent classifiers.

Note that 70% is indeed a high success rate in a 10-way classification task with uniformly distributed classes, where the 'baseline'—the accuracy that can be achieved by intelligent guessing, i.e., by always predicting the most frequent class—is only 10%. 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.[8]

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, and the right decision was taken by the meta-classifier by a very narrow margin. Looking at Table 4, we notice that *none* of the individual base classifiers managed to correctly predict all these three pianists. Obviously, it is only by combining the expertise of the individual classifiers via a meta-classifier that this high success rate can be achieved.

More detailed insight into the stability of the ensemble learner is provided by the following investigation where, instead of just looking at prediction accuracy based on crisp predictions, we also consider the probability score (posterior probabilities) assigned to each class for a particular test case. More specifically, the *Mean Reciprocal Rank* (*MRR*) is based on the ordered list of classes (from most likely to least likely pianists) predicted by a classifier. For n test cases, the MRR of the ith classifier is defined as

$$MRR_i = \sum_{j=1}^{n} \frac{1}{R_{ij}}, \tag{8}$$

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

Table 6
Performance of the base classifiers and the ensemble on the test set. FR and FA are calculated for threshold = 0.2

Classifier	Accuracy	MRR	FR	FA
C11	0.400	0.608	0.600	0.067
C21	0.300	0.511	0.300	0.133
C22	0.400	0.555	0.400	0.122
C23	0.300	0.543	0.400	0.167
C24	0.300	0.527	0.300	0.167
C31	0.400	0.637	0.500	0.100
C32	0.500	0.618	0.500	0.111
C33	0.300	0.462	0.600	0.167
C34	0.400	0.589	0.400	0.122
C35	0.400	0.547	0.600	0.156
Ensemble	0.700	0.800	0.200	0.156

where R_{ij} is the rank of the true class of the jth case in the prediction produced by the ith classifier. For 10 classes the MRR would range from 0.1 to 1.0. The higher the MRR, the better the ranking of the true classes in the ordered list of classifier answers.

Alternatively, it is also possible to examine the types of errors committed by a classifier. Given a fixed confidence threshold, any prediction with an associated probability score above the threshold is accepted. False Rejection (FR) and False Acceptance (FA) rates then provide useful information about the classifier. FR is defined as the ratio of rejected performances by the true pianists to the total performances of the true pianists (i.e., the ratio of positive test examples missed). FA is the ratio of performances by other ('wrong') pianists accepted, to the total number of performances by other pianists. FR and FA are calculated for a given confidence threshold. There is an obvious trade-off: when the confidence threshold is too low, FR tends to 0 and FA tends to 1 (all the performances are accepted). On the other hand, when the confidence threshold is too high, FR tends to 1 and FA tends to 0 (only the performances with high score are accepted).

Table 6 shows the performance of the base classifiers and the ensemble on the test set in terms of accuracy of crisp predictions, MRR, FR, and FA (for threshold = 0.2). As can be seen, the MRR of the ensemble (consisting of 7 correct guesses, 1 second place guess, and 2 fourth place guesses) is far better than the MRR of the base classifiers (C11, C31, and C32 are the most competent ones). Moreover, the ensemble has the lowest FR value (only two of the correct choices are rejected) but many base classifiers have a better FA value (because of the trade-off between FR and FA). These results strengthen the credibility of the proposed ensemble since it is shown that the misses it produces are near-misses rather than random ones.

6. Discussion

The article has presented a computational approach to the problem of discriminating between music performers playing the same piece of music, and introduced a set of simple global features that capture some aspects of the individual style of each performer.

Due to the limited data available and to certain characteristics of the discriminating features, we proposed a classification model that takes advantage of various techniques of constructing meta-classifiers based on an ensemble of very simple classifiers, each one capturing a nuance of the style of the music performer. In particular, by subsampling the input features we manage to exploit all the different features that cannot be used concurrently (to avoid overfitting of the training set), and by subsampling the training data, we take advantage of the instability of a feature subset. Hence, it is demonstrated that a combination of AI techniques is able to deal with a very difficult problem, in effect displaying a level of accuracy unlikely to be matched by human listeners (under similar conditions).[9]

The proposed features can be easily computed and do not make use of any piece-specific information (to be extracted by structural or harmonic analysis). They are global measures extracted from multiple-note passages. An analysis of the induced classifiers and also statistical analyses [19] provide insights into the relative importance of various types of features. In particular, it turns out that features related to articulation (staccato vs. legato) and melody lead are the most informative, followed by aspects of tempo and timing and, finally, dynamics.

This result relates in interesting ways to a very recent study by a musicologist [20], which investigated, via experiments with human subjects, the subjective perception of similarity between expressive performances. The study showed quite clearly that human subjects paid most attention to global characteristics of the performances. The most important factors, according to the participants, were global tempo, rubato, and articulation, followed by loudness. In essence, the human listeners attend to the same parameters that turned out to be most informative in our experiments. The study also showed that the same parameters were important for both musicians and non-musicians. Statistical modelling of subjects' similarity ratings revealed that global measures were more often included in optimal models than local measures, and tempo features were more useful than dynamics features to explain the subjects' similarity ratings; again, this is very much in agreement with our machine learning results.

The work described here was performed in the context of a large project whose goal is to study and characterize fundamental principles of expressive music performance with AI methods [23,24]. The current study can be seen as another attempt at quantifying features that are crucial to understanding and modeling this complex phenomenon. However, from a musicological point of view, our current set of global, segment-based features, while easy to compute, do not give direct insights into the individual performance strategies of musicians. One would like to have features that explicitly describe the artist's expressive actions for every note (as, e.g., in [1]). That would require measurements at the note level, associated with particular local musical contexts and piece-specific information, as, e.g., in [26]. Especially characterizing the musical contexts is a demanding task.

[9] Actually, a comparison with human listeners is not at all straightforward. It is very difficult to define what these 'similar conditions' would be. How many times would a person be allowed to listen to each of the training/test recordings in order to reach a level similar to that of feeding the preprocessed data to a classifier? What would be the level of expertise of the listener? How could one account for the effects of previous knowledge of music and musical style (which the computer does not have)? For these reasons, we avoided to provide a direct human-machine comparison for this task.

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 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), the START program of the Austrian Federal Ministry for Education, Science, and Culture (Grant no. Y99-INF), and the project P12645-INF, sponsored by the Austrian *Fonds zur Förderung der wissenschaftlichen Forschung* (*FWF*). The Austrian Research Institute for Artificial Intelligence acknowledges basic financial support from the Austrian Federal Ministry for Education, Science, and Culture. We would like to thank Werner Goebl for providing the data used in the experiments. We are indebted to the anonymous reviewers for very helpful comments and, in particular, the suggestion to perform the MRR experiment reported in Table 6.

References

[1] J.L. Arcos, M. Grachten, R. Lopez de Mantaras, Extracting performers' behaviors to annotate cases in a CBR system for musical tempo transformations, in: Proceedings of the 5th International Conference on Case-Based Reasoning (ICCBR 2003), Trondheim, Norway, 2003, pp. 20–34.

[2] E. Bauer, R. Kohavi, An empirical comparison of voting classification algorithms: bagging, boosting, and variants, Machine Learning 39 (1/2) (1999) 105–139.

[3] A. Blum, Empirical support for winnow and weighted-majority based algorithms: results on a calendar scheduling domain, Machine Learning 26 (1) (1997) 5–23.

[4] 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, 2000, pp. 19–23.

[5] T. Dietterich, Ensemble methods in machine learning, in: First Int. Workshop on Multiple Classifier Systems, 2000, pp. 1–15.

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

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

[8] A. Friberg, Matching the rule parameters of phrase arch to performances of 'Träumerei': a preliminary study, in: Proc. of the KTH Symposium on Grammars for Music Performance, 1995, pp. 37–44.

[9] W. Goebl, Melody lead in piano performance: expressive device or artifact?, J. Acoustical Soc. Amer. 110 (1) (2000) 563–572.

[10] W. Goebl, R. Bresin, A. Galembo, Once again: the perception of piano touch and tone. Can touch audibly change piano sound independently of intensity?, in: Proceedings of the 2004 International Symposium on Music Acoustics (ISMA'04), Nara, Japan, 2004, pp. 332–335.

[11] W. Goebl, R. Parncutt, Asynchrony versus intensity as cues for melody perception in chords and real music, in: Proc. 5th ESCOM Conference, Hannover, Germany, 2003, pp. 376–380.

[12] T. Lim, W. Loh, Y. Shih, A comparison of prediction accuracy, complexity and training time of thirty-three old and new classification accuracy, Machine Learning 40 (3) (2000) 203–228.

[13] R. López de Mántaras, J.L. Arcos, AI and music: from composition to expressive performances, AI Magazine 23 (3) (2002) 43–57.
[14] D. Opitz, J. Shavlik, Generating accurate and diverse members of a neural network ensemble, in: D. Touretzky, M. Mozer, M. Hasselmo (Eds.), Adv. Neural Inform. Process. Syst. 8 (1996) 535–541.
[15] C. Palmer, On the assignment of structure in music performance, Music Perception 14 (1996) 23–56.
[16] B. Parmanto, P.W. Munro, H.R. Doyle, Improving committee diagnosis with resampling techniques, in: D. Touretzky, M. Mozer, M. Hasselmo (Eds.), Adv. Neural Inform. Process. Syst. 8 (1996) 882–888.
[17] B. Repp, Diversity and commonality in music performance: an analysis of timing microstructure in Schumann's 'Träumerei', J. Acoustical Soc. Amer. 92 (5) (1992) 2546–2568.
[18] E. Stamatatos, A computational model for discriminating music performers, in: Proceedings of the MOSART Workshop on Current Research Directions in Computer Music, 2001, pp. 65–69.
[19] E. Stamatatos, Quantifying the differences between music performers, in: Proceedings of the International Computer Music Conference (ICMC'2002), Göteborg, Sweden, 2002, pp. 376–382.
[20] R. Timmers, Predicting the subjective similarity between expressive performances of music from objective measurements of tempo and dynamics, Submitted for publication. Available as Report TR-2003-25, Austrian Research Institute for Artificial Intelligence, Vienna, 2003, http://www.ai.univie.ac.at/cgi-bin/tr-online?number+2003-25.
[21] L. Todorovski, S. Dzeroski, Combining classifiers with meta decision trees, Machine Learning 50 (3) (2003) 223–249.
[22] G. Tzanetakis, P. Cook, Musical genre classification of audio signals, IEEE Trans. Speech Audio Process. 10 (5) (2002) 293–302.
[23] G. Widmer, Using AI and machine learning to study expressive music performance: project survey and first report, AI Comm. 14 (2001) 149–162.
[24] G. Widmer, S. Dixon, E. Goebl, W. Pampalk, A. Tobudic, In search of the Horowitz factor, AI Magazine 24 (3) (2003) 111–130.
[25] G. Widmer, Machine discoveries: a few simple, robust local expression principles, J. New Music Res. 31 (1) (2002) 37–50.
[26] G. Widmer, Discovering simple rules in complex data: a meta-learning algorithm and some surprising musical discoveries, Artificial Intelligence 146 (2) (2003) 129–148.