Quantifying the Differences between Music Performers: Score vs. Norm

Efstathios Stamatatos

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

Abstract

In this study, a comparison of features for discriminating between different music performers playing the same piece is presented. Based on a series of statistical experiments on a data set of piano pieces played by 22 performers, it is shown that the deviation from the performance norm (average performance) is better able to reveal the performers' individualities in comparison to the deviation from the printed score. In the framework of automatic music performer recognition, the norm-based features prove to be very accurate in intra-piece tests (training and test set taken from the same piece) and very stable in inter-piece tests (training and test sets taken from different pieces). Moreover, it is empirically demonstrated that the average performance is at least as effective as the best of the constituent individual performances while 'extreme' performances have the lowest discriminatory potential when used as norm.

1 Introduction

Expressive music performance is a central research topic in contemporary musicology. So far, the main focus of empirical music performance research is on the exploration of *similarities* between the performers that would help the development of general rules of expressive performance. To this end, the analysis-by-synthesis methodology (Friberg, 1991) and the application of machine learning techniques to large volumes of data (Widmer, 2001) have given promising results. On the other hand, little attention has been paid to the objective detection and the quantification of *differences* between music performers.

Repp (1992) presented an exhaustive statistical analysis of temporal commonalities and differences among distinguished pianists' interpretations of Schumann's Traeumerei and demonstrated the striking individuality of Alfred Cortot and Vladimir Horowitz. However, there is still no systematic approach to automatically quantify the performers' individualities in a machine-interpretable way. In general, the detection and the interpretation of differences in music performance are defined mostly with aesthetic criteria rather than quantitatively.

A well-known notion in expressive performance research is the average performance. Many researchers have attempted to analyze the aesthetic quality of the average performance and compare it to the constituent performances in qualitatively criteria (Repp, 1997; Goebl, 1997). The results have shown that the average performance suppresses individualities but can be of high quality to the listeners.

In this paper, the average performance, calculated from a group of reference pianists, is used as a means to discriminate between another disjoint group of pianists. The average performance is considered the norm of the piece and the deviations in terms of timing, articulation, and dynamics (the three main expressive dimensions available to a pianist) from it quantify the stylistic characteristics of the performers. The proposed norm-based features are compared with corresponding features that represent deviation from the printed score and are objectively evaluated based on a series of experiments in automatic multi-class performer recognition, a very difficult musical task even for human experts. Moreover, the average performance is objectively compared with the individual constituent performances in terms of discriminatory potential.

2 Quantifying Individualities

2.1 Musical Data

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)[1]. The digital recordings were then transcribed into symbolic form and matched against the printed score (Cambouropoulos,

[1] The digital recordings can be accessed at http://www.ai.univie.ac.at/~wernerg/mp3.htm

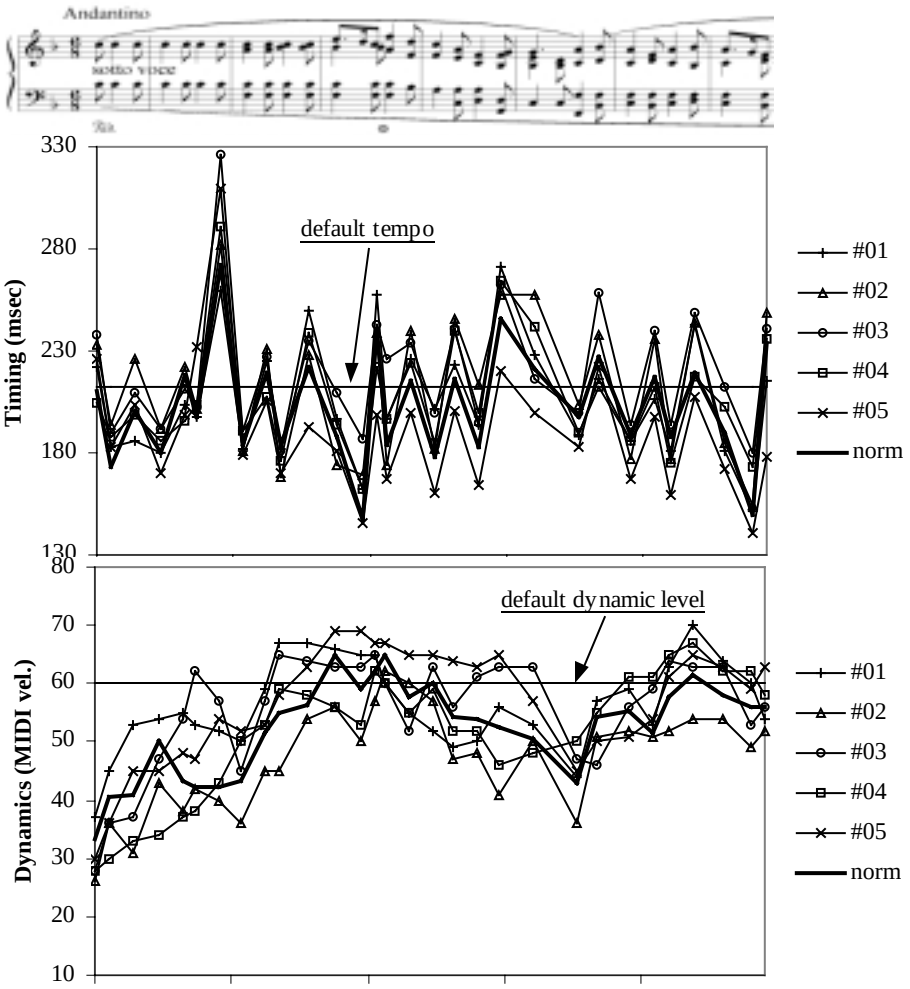

Figure 1. Timing and dynamics variations for the first 30 soprano notes of the *Ballade* (score above) as performed by pianists #01-#05. Default tempo and dynamic level, and performance norm derived by pianists #06-#10 are depicted as well.

2000). 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 sound 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. Henceforth, the 22 pianists will be referred with their code names (i.e., #01, #02,…, #22).

2.2 Feature Extraction

In order to quantify the differences between music performers, a reference point has to be defined. One obvious reference point is the printed score, which can be interpreted into a mechanical or 'flat' rendition of the piece in terms of timing, articulation, and dynamics, without any expressive nuance. Comparing real performances with the score can be viewed as comparing a waveform with a straight line. 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.

2.3 The Proposed Features

For representing the stylistic properties of the expressive performance of a melodic segment, three features are used: the average deviations from the norm in terms of timing, articulation, and dynamics. The musical context (structural or harmonic information) is not taken into account.

Given that $D(x, y)$ denotes the deviation of a vector of numerical values y from a reference vector x, the norm-based features can be expressed as $D(IOI_n, IOI_m)$, $D(OTD_n, OTD_m)$, and $D(DL_n, DL_m)$, respectively, where IOI_n, OTD_n, and DL_n, are the inter-onset interval, the off-time duration (the time between the offset time of one note and the onset time of the next note), and the dynamic level, respectively, as calculated from the performance norm, and IOI_m, OTD_m, and DL_m are the inter-onset interval, the off-time duration, and the dynamic level, respectively, as measured in a real performance. Note that only the soprano notes are taken into account for calculating these measures.

Similarly, the score-based features can be expressed as $D(IOI_s, IOI_m)$, $D(IOI_s, OTD_m)$, and $D(DL_s, DL_m)$, where IOI_s and DL_s are the default inter-onset interval and the default dynamic level, respectively, as indicated in the score. Again, only the soprano notes are taken into account. The same score-based features have been used in previous work for successfully discriminating two skilled performers playing the same piano pieces (Stamatatos, 2001).

3 Score vs. Norm

3.1 Experimental Settings

Various types of distance between two vectors of numeric values can be used for calculating the proposed measures. In preliminary experiments the statistical technique of analysis of variance (aka ANOVA) was used for measuring the statistical significance of various distance types on performer recognition tasks. According to the results of this procedure, the relative distance (D_r) best fits the score deviation features while the norm deviation features are best represented by the simple distance (D_s). For a melodic segment of n notes, these distances are defined as follows:

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

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

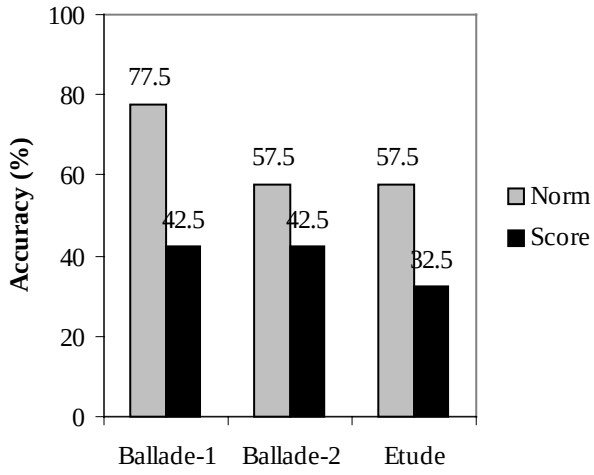

Figure 2. Classification accuracy results for norm-based and score-based classifiers within the training set (leave-one-out evaluation).

The two available pieces had to be segmented into a number of parts for providing the necessary training examples. Since the performance of a segment is represented with three features (corresponding to timing, articulation, and dynamics), at least three training examples per class (pianist) should be available for avoiding overfitting of the training set. Moreover, previous work (Stamatatos and Widmer, 2002) has shown that the longer the training examples, the more accurate the resulting classifier, and segments of equal length (measured in soprano notes) provide better training examples in comparison to phrase-based segments.

Taking all these into account, the performances of *Ballade* were segmented in 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. Pianists #01-#12 will be used as the set of reference pianists to compute the norm performance, and the

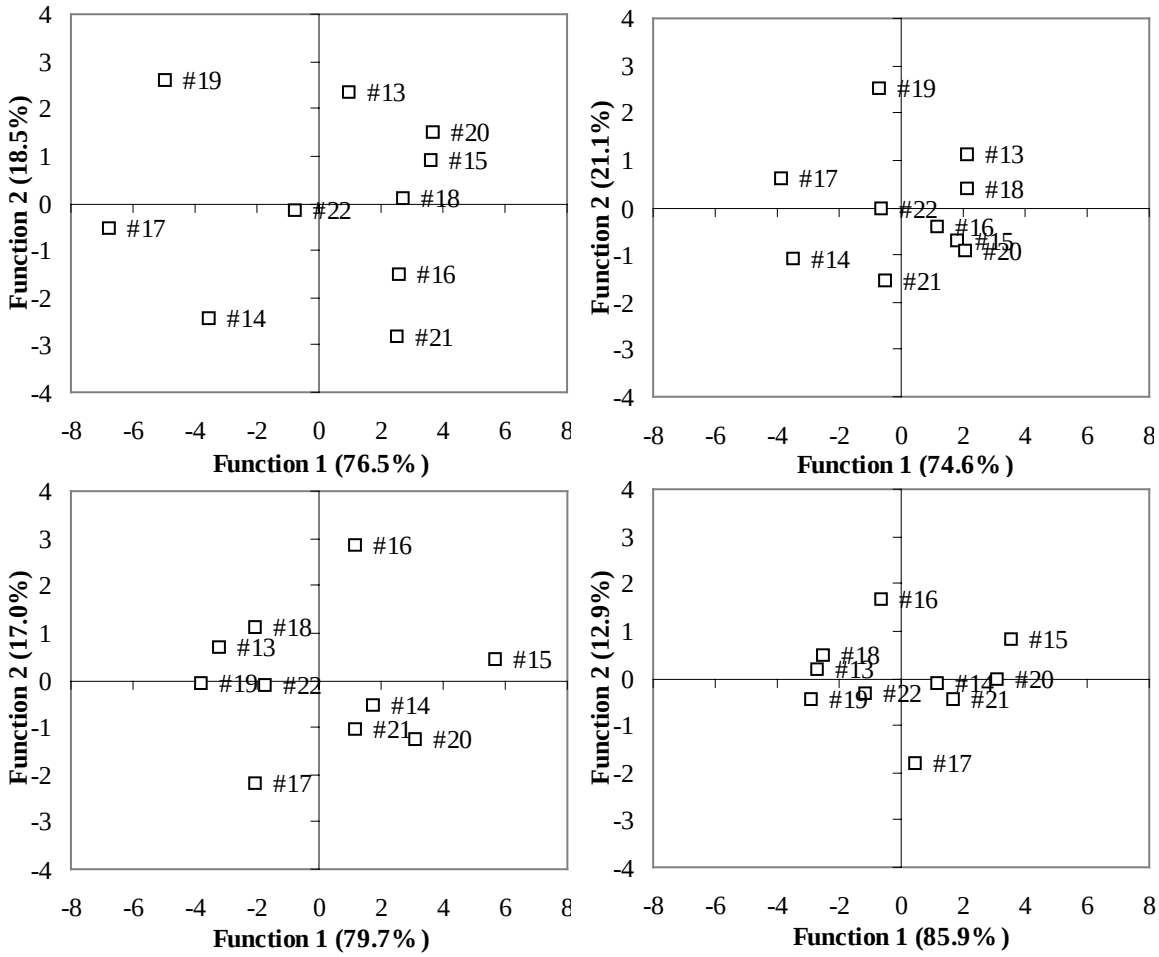

Figure 3. The centroids of the pianists #13-#22 in the space of the first two discriminant functions for *Ballade-1* (above) and *Etude* (below). Norm-based models are shown on the left side and score-based models on the right side. The numbers inside parentheses indicate the amount of variance explained by the corresponding function.

task will be to learn to distinguish pianists #13-#22 (a ten-class classification problem). The average performance is calculated on the note level.

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 (Eisenbeis and Avery, 1972). 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.

3.2 Fitting the Training Set

Three score-based classifiers and three norm-based classifiers were constructed based on the training sets of *Ballade-1*, *Ballade-2*, and *Etude*. Figure 2 shows the classification accuracy within the training set using the leave-one-out methodology (each case of the set is classified according to the classification model derived from the remaining cases). It is clear that the norm-based classifiers are much more accurate than the corresponding score-based classifiers. This is a strong indication that the norm features are better able to fit the training set and discriminate between unseen performance segments of the same piece.

Figure 3 depicts the class centroids in the space of the first two discriminant functions (which account

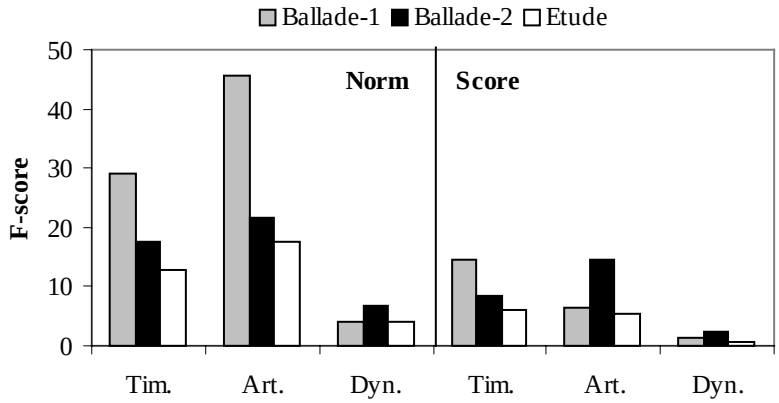

Figure 4. ANOVA F-scores for norm-based and score-based features for three data sets.

	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	

Table 1. Cross-validation results for norm-based and score-based classifiers. Number of correct predictions in a maximum of 10.

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 rather than the exact values of discriminant functions. As can be seen, in both cases the relevant 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 lay 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, though to a lower extent, can be made for the second discriminant function's spans. This fact means that the norm-based features are better able to produce robust and reliable classifiers since the classes are more widely spread within the classification space.

The examination of relative position 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*).

3.3 Cross-validation Results

The results of cross-validating the norm-based and score-based classifiers in new unseen musical parts taken either from the same piece or from a different one are given in table 1. Each classification model derived based on 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) 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 performances of the first half of a piece and then try to guess the performer of the second half or of another piece. All the possible combinations of training-test set are given in table 1.

The cross-validation 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. 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).

3.4 Stability over Different Training Sets

A very important aspect for constructing reliable classifiers is the stability of the derived classification model over slightly changed training sets. For testing the stability of norm-based and score-based features, we divided *Ballade* into four disjoint subsets and then

Norm	Acc. (%)	Tim.	Art.	Dyn.
#01	72.5	27.6	82.8	3.5
#02	70.0	26.7	18.4	11.2
#03	67.5	25.3	26.9	11.7
#04	67.5	29.0	14.2	5.2
#01-#04	72.5	33.4	51.7	9.7
#05	77.5	8.9	78.6	5.3
#06	80.0	37.4	42.1	6.6
#07	57.7	33.6	18.3	4.3
#08	77.5	14.7	15.6	4.5
#05-#08	80.0	34.0	41.7	8.7
#09	72.5	38.8	34.9	5.6
#10	67.5	14.8	13.7	5.2
#11	57.5	9.1	66.9	9.7
#12	75.0	31.8	36.1	5.0
#09-#12	82.5	28.2	78.5	7.8
#01-#12	82.5	36.4	69.5	9.3

Table 2. Classification accuracy and ANOVA F-scores for timing, articulation, and dynamics features using various norms.

four different overlapping training sets were constructed by dropping one of these four subsets (i.e., a technique known as *cross-validated committees*). Then, four different norm-based classifiers and four different score-based classifiers were constructed using one of these subsets as training set. The classification models were, then, applied to *Etude* for predicting the most likely performer. The predictions of the four norm-based classifiers were identical (4 correct predictions). On the other hand, 60% of the predictions made by the four score-based classifiers were different (2-4 correct predictions). This experiment illustrates that the score-based features highly depend on 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 is avoided with the use of norm-based features.

3.5 Significance of Features

In order to study the contribution of the performance dimensions to the classification model, we applied ANOVA to the data sets of *Ballade-1*, *Ballade-2*, and *Etude*. Figure 4 shows the resulting F-scores for norm-based and score-based features in terms of timing, articulation, and dynamics. The greater the F-score, the more important the feature for distinguishing between the classes. As can be seen, the norm-based features are by far more important than the corresponding score-based ones. Moreover, articulation seems to be the most important discriminator factor among the norm-based features whereas timing dominates (with the exception of *Ballade-2*) the score-based features. In both cases, dynamics has the least significant discriminatory potential.

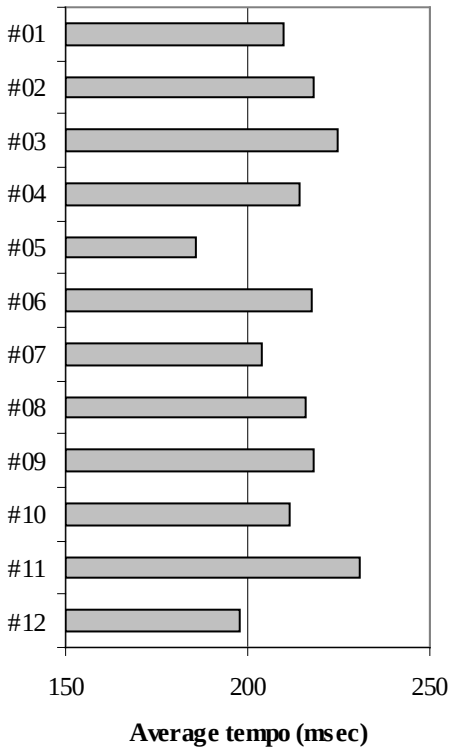

Figure 5. Average tempo (IOI on the sixteenth-note level) of performances of *Ballade* by pianists #01-#12.

4 Average vs. Constituent Performances

So far we assumed that the average performance is a better reference point than the individual constituent performances from which it is calculated. It seems plausible that the average performance suppresses individualities of the constituent performances and should be better able to represent general expressive characteristics of a musical piece. However, it has not yet been objectively demonstrated that the average performance provides better results than the constituent performances for revealing the individualities of a different group of performers on the same piece. Moreover, there should be cases where only a limited number of performances of the same piece are available. Is the calculation of the average performance (to be used as norm) necessary in such cases? Can an individual performance play this role as effectively as (or even better than) the average performance? What kind of performances may be considered effective (or ineffective) to be used as norm? The following experiment is an attempt to objectively answer to these questions.

The deviations (in terms of timing, articulation, and dynamics) of the pianists #13-#22 from each one of the pianists #01-#12 on the entire *Ballade* (*Ballade-1*+*Ballade-2*) were calculated. In other words, each one of the performances by pianists #01-

#12 (used so far to calculate the norm) was considered as norm. Additionally, three average performances were calculated from the pianist subgroups #01-#04, #05-#08, and #09-#12. The deviations of the pianists #13-#22 from these new average performances were calculated as well. Table 2 shows both classification accuracy results (based on leave-one-out evaluation) and ANOVA F-scores for each one of these new norms. It is worth noting that each subgroup norm provides equal or better classification results in comparison to its constituent performances. Moreover, the norm of the entire group (pianists #01-#12) provides equal or better results in comparison to the subgroup norms. This is a strong indication that the average performance is at least as good as the best of the constituent performances for discriminating between different performers.

As concerns the significance of the performance features (expressed with ANOVA F-scores), some individual performances may have a great discriminatory potential in one dimension but only poor abilities in an another dimension. For instance, #02 has a high score in dynamics but low score in articulation whereas #05 has a high score in articulation but a very low score in timing. In most cases, the scores of the subgroup norms may be considered as a rough approximation of the average scores of the constituent performances in each dimension. Therefore, the average performance scores are more evenly distributed among the three dimensions.

Figure 5 shows an estimation of the global tempo of the pianists #01-#12 (calculated as the average IOI on the sixteenth-note level). As can been seen, #05 has the lowest value (plays faster) and #11 the highest value (plays slower). Note that the F-scores of these two pianists for the timing dimension (table 2) are the lowest. This is an indication that 'extreme' performances are not effective norms for discriminating between different performers. However, it has to be underlined that a certain performance may be 'extreme' in one dimension and 'non-extreme' in another dimension.

5 Conclusions

This study was an attempt to objectively quantify the differences between music performers playing the same piece. Using a small set of musical data, it has been demonstrated that multi-class performer recognition can be very successful based on machine-interpretable features. It is unlikely for human-experts to achieve similar results based on aesthetic or qualitative criteria and such limited performance excerpts.

The comparison of the norm-based features with the score-based ones revealed that the average performance is a better reference point. Norm-based classifiers are more accurate in intra-piece tests and more robust (or stable) in inter-piece tests. Moreover, they are quite stable in slightly changed training sets, in contrary to score-based classifiers. However, the combination of norm-based and score-based classifiers together with classifiers derived from other feature sources (e.g., chord asynchronies) can significantly improve the classification results in inter-piece conditions (Stamatatos and Widmer, 2002).

It has also been demonstrated that the average performance is at least as effective as the best of the individual constituent performances when used as norm. The average performance, apart from suppressing individualities, distributes the discriminatory potential more evenly among the performance dimensions. Therefore, in cases where many performances of the same piece are available, it is advisable to calculate the average performance to be used as norm. On the other hand, 'extreme' performances in certain dimensions proved to be the least important ones when used as norm for quantifying the differences between performers.

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.

Acknowledgments

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

Cambouropoulos, E. 2000. "From MIDI to Traditional Music Notation." *Proceedings of the AAAI'2000 Workshop on Artificial Intelligence and Music.* 17th National Conf. On Artificial Intelligence, pp. 19-23.

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

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

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

Repp, B. 1992. "Diversity and Commonality in Music Performance: An Analysis of Timing Microstructure in Schumann's 'Traeumerei'." *Journal of the Acoustical Society of America* **92**(5):2546-2568.

Repp, B. 1997. “The Aesthetic Quality of a Quantitatively Average Music Performance: Two Preliminary Experiments.” *Music Perception* **14**(4):419-444.

Stamatatos, E. 2001. “A Computational Model for Discriminating Music Performers.” *Proceedings of the MOSART Workshop on Current Research Directions in Computer Music*, pp. 65-69.

Stamatatos, E. and G. Widmer. 2002. “Music Performer Recognition Using an Ensemble of Simple Classifiers.” *Proceedings of the 15th European Conference on Artificial Intelligence (ECAI’02)*.

Widmer, G. 2001. “Using AI and Machine Learning to Study Expressive Music Performance: Project Survey and First Report.” *AI Communications* **14**:149-162.