Time-varying estimation of parameters in rule systems for music performance

Patrick Zanon, Giovanni De Poli, Alessandro Dorigatti
CSC - DEI University of Padova - Italy
{patrick.depoli}@dei.unipd.it, excawind@inwind.it

Abstract
Most performance rule system compute deviations by using a set of weighted rules. In this paper we describe a method of parameter (or weight) estimation, so that the rule system can generate a performance that will be as similar as possible to a human interpretation of a given score. The method allows either a global and a time varying estimation on different time scales.

1 Introduction
The analysis of deviations in music performance has led to the formulation of some models that describe their structures, with the aim to be able to automatically synthesize what the player unconsciously adds to the notation of the score. Some system of rules, with different characteristics and various degrees of flexibility, have been proposed, but all of them have the aim of covering a range of “expressive” variations as wide as possible. These rules are developed mostly using the analysis by synthesis or by machine learning algorithms [5]. The most important is the KTH rule system [2]. Different rules can be weighted by the so called k parameters, allowing them to model performances more closely and adapt the rules to different situations. Moreover tuning weights could be used to model some emotional expression [1]. Weighting parameters are normally estimated by a trial and error procedure with analysis by synthesis approach. Friberg [3] attempted automatic estimation using an iterative minimization of a sort of distance between performances, by varying the values of the parameters with the intent of approaching a given performance. However the performance style or strategy can change along the piece. Thus it is interesting to have a time varying estimation of parameters. In this work we will use a suitable pre-Hilbert space to represent the scores allowing an optimal estimation of parameters on different temporal scales and following their time variations.

2 Estimation Methodology
We observed that the considered models start from the nominal performance (a literal or mechanic interpretation of the score) and introduce, in an additive way, duration and volume variations on some notes according to the rules, each of them weighted by a characteristic multiplicative coefficient \( k_j \). The idea is to estimate the parameters so that the interpretation, generated by the rule system, will be as similar as possible to the given one (e.g. a human performance), which will be called sample performance. This suggests the way to go: by representing the performances with suitable vectors and formalizing their concept of distance with a particular formulation of the Euclidean norm, it will be possible to access to the results of the theory of pre-Hilbert space\(^1\), and in particular to the theorem of the projection, that is the best approximation in the least square sense.

This can be do with some definitions: every performance of \( n \) notes was symbolized as a vector \( p \) lying in a \((3n-1)\)-dimensional space \( \mathcal{P} \), in which the elements are: \( n \) sound intensities (Sound Levels \( L_i \)), \( n \) durations (\( D_i \)), and \( n-1 \) time intervals between notes (Inter On Set Intervals \( I_i = O_{i+1} - O_i \), where \( O_i \) is the onset time of the \( i \)-th note). The vector space \( \mathcal{P} \) as it is, does not evidence the variations inserted by the performer or by the rule of the expressiveness model. Therefore, when it is necessary to refer to the deviations from the nominal performance \( p_N \) we use the symbol \( \Delta p = p - p_N \). With this notation, each rule of the model is represented by a \((3n-1)\)-dimensional vector \( r_j \), whose components are the deviations that would be applied to \( p_N \) using a unitary parameter as weight. Thus the model can generate the following performances:

\[
p_0 = p_N + \sum_{j=1}^{m} k_j r_j
\]

where \( m \) is the number of applied rules. All \( p_0 \) lay in a vector subspace \( \mathcal{R} \) of dimension \( m \) whose base is \( \mathcal{R} = [r_1 \ldots r_m] \).

The Euclidean norm should take in account characteristics of the human ear, so several psychoacoustic principles were used. On the base of these studies it has been chosen to measure the variations of intensity in dB and the variations of duration in percentage relative to the nominal duration or inter on set. At this point a criterion about how loudness variations and time variations could be considered together must be selected. Many possibilities were tested, but the best results were obtained by weighting this different kind of variations so as to balance the effects of the just noticeable variations for human ear (we used 0.5 dB on average for sound levels and 5% approximately for the durations). Thus the norm used in our space is:

\[
|| \Delta p ||_{\Delta} = \sqrt{\alpha_L \sum_{i=1}^{n} \Delta L_i^2 + \alpha_D \sum_{i=1}^{n} \Delta D_i^2 + \alpha_T \sum_{i=1}^{n-1} \Delta T_i^2 + \alpha_I \sum_{i=1}^{n-1} \Delta I_i^2}
\]

\(^1\)A pre-Hilbert space is a linear space equipped with an inner product. A complete pre-Hilbert space is a Hilbert space.
where $\alpha_L = 0.1$, $\alpha_D = 1$ and $\alpha_T = 1$ are the weights.

In order to find the performance $\tilde{p}_k$ as similar as possible to a sample one, the theorem of the best approximation is used: this means that the solution is the orthogonal projection of the vector, representing the variations of a given performance, onto $\mathcal{R}$ and that the multiplicative coefficients are the components of the projection according to the base $\mathcal{R}$. Moreover it is possible to measure how much the model of expressiveness can approach the human performance by comparing the norm of the human deviations and the norm of the synthetic ones by defining efficiency as

$$\xi \triangleq 1 - \frac{||p - \tilde{p}_h||}{||p - \tilde{p}_N||}$$  \hspace{1cm} (3)

Having defined a pre-Hilbert space, some others features, such angles among rules and among performances can be obtained. This method allows to verify if the rules or the performances are related in a geometrical way. Rules that are related are almost parallel and analysis of the angle between them allows a simplification of the model of expressiveness. In fact often rules are triggered by the same underlying principle and cooperate to introduce deviations. Care should be paid in the estimation procedure that if some rules are approximately parallel, their parameters can be estimated with very high values because one rule tends to cancel the effect of the other. The method instead try to explain deviations using the small difference between the two rules (over-fitting problem).

With this methodology it is possible to estimate the optimal coefficients that can be used to approximate a given performance according any additive model of expressiveness. We present our results with reference to the KTH rule system. Minimization of the norm (2) can be used also to estimate non additive parametric rule systems (e.g. neural networks), but in this case we cannot exploit geometric properties such angles of the performance space.

### 2.1 Tempo and amplitude scales estimation

The nominal performance is played using the nominal tempo specified in the score, but the musician can choose another value or change it according to his expressive strategy; similar considerations for the intensity can be done. On the other hand we want that the performance deviations $\Delta p$, used as input to our estimation procedure, should be really expressive deviations and not an artifact resulting by a bias caused by a wrong estimation of tempo and intensity. It would be possible to normalize tempo and loudness to the total performance duration and intensity, but the presence of short rests separating each phrases, and other phrasing effects sometime make this normalization not very reliable on the local scale. Thus we extended our estimation procedure in order to have an optimal estimation of tempo and intensity scales. To this purpose, two special vectors were added for tempo and amplitude estimation represented by two $(3n-1)$-dimensional vectors:

$$r_i \triangleq [0 \ldots 0 \ D_1 \ldots D_n \ I_1 \ldots I_{n-1}]^T$$

$$r_L \triangleq [1 \ldots 1 \ 0 \ldots 0 \ 0 \ldots 0]^T$$  \hspace{1cm} (4)

where in the rule $r_L$ the first $n$ components are set to 1 and the others to 0. The first rule $r_i$ introduces variations only for the timing components: if its relative weighting parameter $k_i = 1$, then all the durations of the synthetic performance will double. The second rule $r_L$ introduces deviations only for loudness components: if its relative weighting parameter $k_L = 1$, then intensities will increase of 1 dB. It is important to underline that $r_i$ and $r_L$, after the estimation, indicate how much the human performance is slower and louder than the nominal performance in the least square sense. For example we can use a nominal performance played at 100 bpm and at 80 dB; if the human performance we want to analyze was played at constant 60 bpm and at constant 85 dB, then the estimation of the parameters using only the rules $r_i$ and $r_L$ produce these results: $k_i = 2/3$ and $k_L = 5$.

It can be noticed that in general rules have a side effect: they introduce deviations that change the average tempo and loudness. Our method allows to compensate the side effect of the rules and to obtain a better fitting of the human performance. Thus when using all the rules different $k_i$ and $k_L$ could be obtained. This situation can be expressed in term of non-orthogonality between $r_i$ (or $r_L$) and some other rules. Notice that $r_i$ and $r_L$ are orthogonal because the non-zero components are different in the two rules.

### 2.2 Sliding Window

The methodology was extended allowing a time-varying estimation of parameters using a user-definable mobile window, called sliding window. In this way the estimation have a local meaning and so it is possible to see if parameters change during the time. For this local behaviour, attention should be payed on which rules have to be included for the estimation: many of them have a global meaning (e.g. the Phrase Arch rule) and for this reason have to be excluded from the sliding window estimation in order to avoid artifacts in the estimation. Another interesting topic is the over-fitting problem that may emerge from the $k$ parameters trend: sometimes many rules introduce, for a particular score, the same kind of deviations and for this reason only one of them were included in the estimation.

### 3 Results

The method have been employed with different musical pieces, each of them was interpreted by a pianist more times. As an example we present the results obtained with different interpretations by a professional pianist of the “Sonatina in G major” by Ludwig Van Beethoven on a Yamaha Disklavier Coda. The first 24 measures were played without repeats, for a total of 150 notes, divided in three main phrases, respectively from the 1st to the 48th note, from the 49th to the 102nd note, and from the 103rd to the 150th note. The structure of the piece is ABA. For each stimulus the pianist was asked to play many performances, and then he chose the most effective one. A first set of performances was played in a natural way: the pianist was asked to play without any expressive intention, and interpreting the score in an as similar as possible way. The values estimated (using all the KTH rules and $r_i$, $r_L$) were strictly related with the performance style and the set of coefficients was very stable (figure 1a). A second group of performances was used, in which the pianist
Figure 1: Estimation of parameters on the whole piece using (a) different natural performances and (b) different expressive intentions (Anger, Disgust, Fear, Happiness, Indifferent, Sadness)

<table>
<thead>
<tr>
<th>1-150</th>
<th>1-48</th>
<th>49-102</th>
<th>103-150</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_L$</td>
<td>-0.26</td>
<td>-0.28</td>
<td>-0.23</td>
</tr>
<tr>
<td>$k_L$</td>
<td>-3.78</td>
<td>-3.83</td>
<td>-3.90</td>
</tr>
<tr>
<td>$\xi$</td>
<td>0.54</td>
<td>0.52</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Table 1: Estimated parameters using only the $r_1$ and $r_2$ rules, over the whole piece and over the three phrases for the “Anger” performance.

was asked to play with different expressive intentions: in this case some parameters presented considerable variations (figure 1b).

A second set of results were obtained considering the estimation of parameters using only $r_1$ and $r_2$ rules. This choice allows to estimate the mean tempo and the mean loudness. From now on the discussion of the results will be carried out with reference to the “Anger” performance. Similar discussion can be done for all the other performances. The estimation were done over the whole piece and over each phrase (table 1). In general the piece was played faster ($k_L < 0$) and softer ($k_L < 0$) than the nominal performance. In particular the two instances of the phrase A were played faster and louder than the B one on average, and from the efficiency $\xi$ we can see that the A phrases were played introducing a greater number of deviations than in B that cannot be explained using only the $r_1$ and $r_2$ rules. Subsequently the analysis was carried out using a sliding window of 10 notes large (1-2 bars on average). Results of the estimation are presented in figure 2. We can see how the central phrase was played using a lower tempo than in the two other ones, as seen above: moreover for each instance of the phrase A two parabolic arcs are present. Similar considerations can be done for the profile of $k_L$ parameter: it indicates a softer interpretation of the phrase B, and in general a louder interpretation marking the boundaries between phrases; it should be noticed the absence of loudness modulation in the final phrase. Finally, as predicted by the efficiency $\xi$ previously computed, we can see that in the phrase B a lower number of deviations were introduced by the musician than in phrases A.

Finally a restricted set of rules were used using the sliding window procedure. In this case the window size was increased to 25 notes (5-4 bars on average) in order to avoid over-fitting problems. The rules used in this estimation were the following: $r_1$, $r_2$, Duration-Contrast-A0-D1, Duration-Contrast-A1-D0, Duration-Contrast-Art, Harmonic-Charge-A0-D1, Harmonic-Charge-A1-D0, High-Loud, Melodic-Charge-A0-D1, Melodic-Charge-A1-D0 and Punctuation-D0-D1 (Dur=0, Duroff=1), where A0-D1 means the duration part of the rule and A1-D0 means the amplitude part of the rule. The Punctuation-D1-D0 rule was excluded because of the high correlation with the Duration-Contrast-A0-D1 rule that produced over-fitting and over estimation of the respective $k$ pa-

412
Table 2: The average $k$ parameters an their variance for the “Anger” performance

<table>
<thead>
<tr>
<th>Rule name</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration-Contrast-A0-D1</td>
<td>-1.50</td>
<td>3.37</td>
</tr>
<tr>
<td>Duration-Contrast-A1-D0</td>
<td>0.64</td>
<td>0.13</td>
</tr>
<tr>
<td>Duration-Contrast-Art</td>
<td>0.39</td>
<td>0.06</td>
</tr>
<tr>
<td>Harmonic-Charge-A0-D1</td>
<td>-0.32</td>
<td>2.97</td>
</tr>
<tr>
<td>Harmonic-Charge-A1-D0</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>High-Loud</td>
<td>0.15</td>
<td>0.01</td>
</tr>
<tr>
<td>Melodic-Charge-A0-D1</td>
<td>0.75</td>
<td>3.49</td>
</tr>
<tr>
<td>Melodic-Charge-A1-D0</td>
<td>-0.19</td>
<td>0.08</td>
</tr>
<tr>
<td>Punctuation-D0-D1</td>
<td>0.79</td>
<td>2.94</td>
</tr>
</tbody>
</table>

4 Conclusions

A method for the time varying estimation of weighting parameters in additive rule systems for music performance was presented. It allows to characterize the changing style of the performance. The definition of an Euclidean norm allows a geometrical interpretation of the approximation problem, giving more insight on the rule system. We attempted to include perceptual issues in defining our norm. Different definitions of norm in the performance space can be experimented using the same estimation approach. It should be noticed that the hypothesis of additive rules is very interesting for an user point of view for its explanation power, but it is a oversimplification of the real performance strategies of the performer. Different factor interact in an intricate way [4]. Minimization of norms as (2) can be used also to estimate non additive parametric rule systems (e.g. neural networks), but in this case we cannot exploit geometric properties such angles of the performance space.

References