A structurally guided method for the decomposition of expression in music performance

W. Luke Windsor
School of Music and Interdisciplinary Centre for Scientific Research in Music, University of Leeds, Leeds LS2 9JT United Kingdom

Peter Desain
NICI, Radboud University, Postbus 9104, 6500 HE Nijmegen, The Netherlands

Amandine Penel
Laboratoire de Psychologie CognitiveUniversite de Provence & CNRS UMR 6146 Bat 9, Case D 3, place Victor Hugo 13331 Marseille Cedex 3 France

Michiel Borkent
NICI, Radboud University, Postbus 9104, 6500 HE Nijmegen, The Netherlands

(Received 25 March 2005; revised 12 October 2005; accepted 8 November 2005)

A method for separating, profiling, and quantifying the contributions of different structural components to expressive musical performance is described. The method is demonstrated through its application to a set of expert piano performances of a short piece from the classical period. The results show that the output of the method aids in the understanding of how the different structural components in a piece of music combine in the generation of an expressive performance. A second demonstration applies the method to performances at different tempi to illustrate its effectiveness in pinpointing the structural features responsible for small but statistically significant differences between performances. The method is compared with other approaches to the analysis and modeling of musical performance, and a number of potential applications are identified. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2146091]

PACS number(s): 43.75.Cd, 43.75.St, 43.75.Yy [DD] Pages: 1182–1193

I. INTRODUCTION

A. Expression and structure in musical performance

Much progress has been made in the development of methods for extracting and analyzing discrete and continuous expressive parameters from audio and MIDI recordings of musical performances, and such methods have been used to develop and test hypotheses regarding the cognitive and motor processes which underlie such performances. Since the research of Seashore and colleagues (Seashore, 1938), it has been understood that skilled performers manipulate expressive parameters in their performances in structured and predictable ways that are related to the structure of the music. The vast majority of researchers in this area have concluded that many aspects of expressive timing and dynamics can be predicted from an analysis of the structure of a piece of music to be performed, and that such predictions are concrete enough to be formalized in a system of rules (see, e.g., Clarke, 1988; Palmer, 1997).

B. Generative approaches to expression in performance

The idea that the expressive aspects of musical performance are created from a representation of musical structure has led a number of researchers to advance computational theories that formalize and express the mapping from score plus structure to performance in algorithmic terms. We call these computational models generative theories here. For example, Clynes (1983) predicts timing and dynamics from time signature and composer, recursively subdividing time intervals multiplicatively at each metrical level. Friberg (1991; also see Sundberg, 1988) focuses on local structure (e.g., a jump in pitch) to calculate expressive deviations from the mechanical rendition. This approach uses a wide range of rules, each instantiating a different aspect of expression, the rules' effects accumulating in ways that may be quite difficult to interpret. Todd (1985, 1992, 1995) predicts timing and dynamics from phrase structure alone, applying a single formula (a parabola) additively at each level, an approach which has a recursive elegance. Such generative theories seem to have a huge advantage over other, less precise, theories (Desain *et al.*, 1998). One of their major benefits is that they can be fitted to empirical data, yielding an estimate of their predictive power and optimal parameter values. Such comparisons have been fairly widespread in the literature (e.g., Todd, 1992; Friberg, 1995; Windsor and Clarke, 1997; Widmer and Tobudic, 2003; also see Sec. I C). In most cases an overall measure of goodness of fit, or conversely a measure of error, is used to quantify the success with which a model (and, one assumes the theory upon which it is based) can explain an individual performance or set of performances.

However, although such generative computational models have greatly helped in building and testing the theoretical concepts used in the field, and are sometimes quite satisfactory in terms of output simulation (as in Widmer and Tobudic, 2003), they are in general not very successful when fits

to real performance data are attempted. This can be caused by the fact that many models are only partial, and expressive deviations linked to ignored types of structure easily upset the fitting process. For example, a local mid-bar phrase ending that is expressed with a ritardando in a performance would easily upset the optimization of Clynes' rather subtle composer's pulse, which is linked to the metric structure alone. Moreover, as empirical findings have demonstrated, identical effects might derive from very different rules or structures. A rule which maps the score location of an event within a phrase to a local modification of tempo can produce an effect which is indistinguishable from a rule which makes a similar prediction on the basis of metrical location. Similarly, a pause at or near a phrase boundary might be the result of a rule which applies to only one event (e.g., a micropause) or might be the result of a rule which applies to more than one event (e.g., a ritardando) (see Windsor and Clarke, 1997).

Note that this is not a criticism of generative theories as such, which can and sometimes do combine many different assumptions about expression. Musical structure seems not to be made of singular and homogeneous aspects, but constitutes a bundle of interlinked properties, which are often incompatible but not independent of each other. This complexity and interdependency has to be taken into account in investigating how musical structure gives rise to the expressive signal. What this paper addresses is how to better examine and quantify these multiple contributions to expression.

C. Estimating combined and individual fit of structural parameters

One solution would be to consider many kinds of musical structure at once and fit them jointly to a performance. Not only does this solve the problem of confounding ignored factors in the fitting procedure, but it also becomes possible to assess the relative contributions of different types of musical structure for a single piece. This was proposed by Desain and Honing (1997), and the current paper is an elaboration, implementation, and test of those ideas.

Given that a single piece may be structurally ambiguous and performers may even apply different strategies in relation to the same structure (see, e.g., Clarke and Windsor, 2000), these aspects constitute the so-called "interpretation" chosen by the performer and they form a rather important aspect of the data. This solution has been adopted with some success (such as in Sundberg *et al.*, 2003; Zanon and de Poli, 2003a, b), usually with quite specific (and quite local) rules that contain elaborate domain knowledge (like generating a pause before a large melodic leap) but only few parameters per rule. Our approach is different in that we do not aim to test any such specific aspects of expression. Instead, we assume regularity (e.g., each bar is expressed by the same timing fluctuation) and an open shape with a number of parameters (piecewise linear profiles) and aims to analyze expression (in this case expressive timing) in order to reveal more global mappings between structure and expression.

The method proposed here, which is implemented in the POCO environment [a software environment for the analysis of expression, see Honing (1990, 1992)] in a module entitled DISSECT (with SECT standing for Structural Expression Component Theory), not only delivers the relative contribution of the various components to the overall expressive profile, but also yields the component profiles themselves as well, effectively decomposing expression into its structural elements (note that to run POCO requires Macintosh Common LISP; for plotting results the scriptable statistics package JMP is used). Such decomposition may help to better reveal processes underlying the relationship between structure and expression. For example, if one measures the interonset timing of a number of performances of the same piece obtained under different conditions or from multiple performers, and merely compares the data in terms of their global differences or similarities [using the kinds of statistical methods applied by Shaffer (1981) or Repp (1992)] one is left with a rather uninformative result in regard to the underlying processes. It could be that there are systematic differences between performances (1) that reflect a difference in the application of various rules (e.g., a performer not expressing the time signature by means of timing); (2) that reflect the application of the same rules with different parameter settings or weights (a performer slowing down more or less in a phrase final ritard); or (3) that reflect the operation of the same rule on a different structural interpretation (e.g., expressing a different phrase structure with the same ritards at the end of each phrase). With a technique to decompose expression and compare its elements it becomes possible to distinguish between these hypotheses in a quantitative manner. Together with a few other attempts to judge the relative contributions of different musical structures in a systematic analysis (such as Thompson and Cuddy, 1997; Penel and Drake, 1998; Chaffin and Imreh, 2002; Sundberg *et al.*, 2003; Zanon and de Poli, 2003a, b), this method is high dimensional. It can be opposed to the visualization techniques applied to performance expression as elaborated by, for example, Dixon *et al.* (2002), which aim to represent expressive variation in a single time-variant plot of a few attributes like overall tempo and loudness. Although most generative theories propose quite explicit forms or shapes that make up the expressive signal (parabolic beat intervals, micropauses, recursive metric subdivisions), DISSECT works without imposing an explicit set of *a priori* expressive rules, hence it can be seen as more data driven. It does assume that the mapping from score to performance is constrained within parameter consistency; in other words, our assumption is that if an element of musical structure maps score to performance in a particular way, this relationship will be preserved for all examples of that structure within a performance. Secondly, we have chosen to assume that tempo change is linear (although the approach is not restricted to linear mappings in principle or practice). Hence, although our method has similarities to that described by Zanon and De Poli (2003a, b), it differs in that their approach is specific to a particular rule-based model of expression, whereas our approach is more general in formulation in that it evaluates a structural analysis of a piece and a mapping between this analysis and the expression, making only few *a priori* assumptions about what form the mapping might take.

Although the focus here is on expressive timing, our

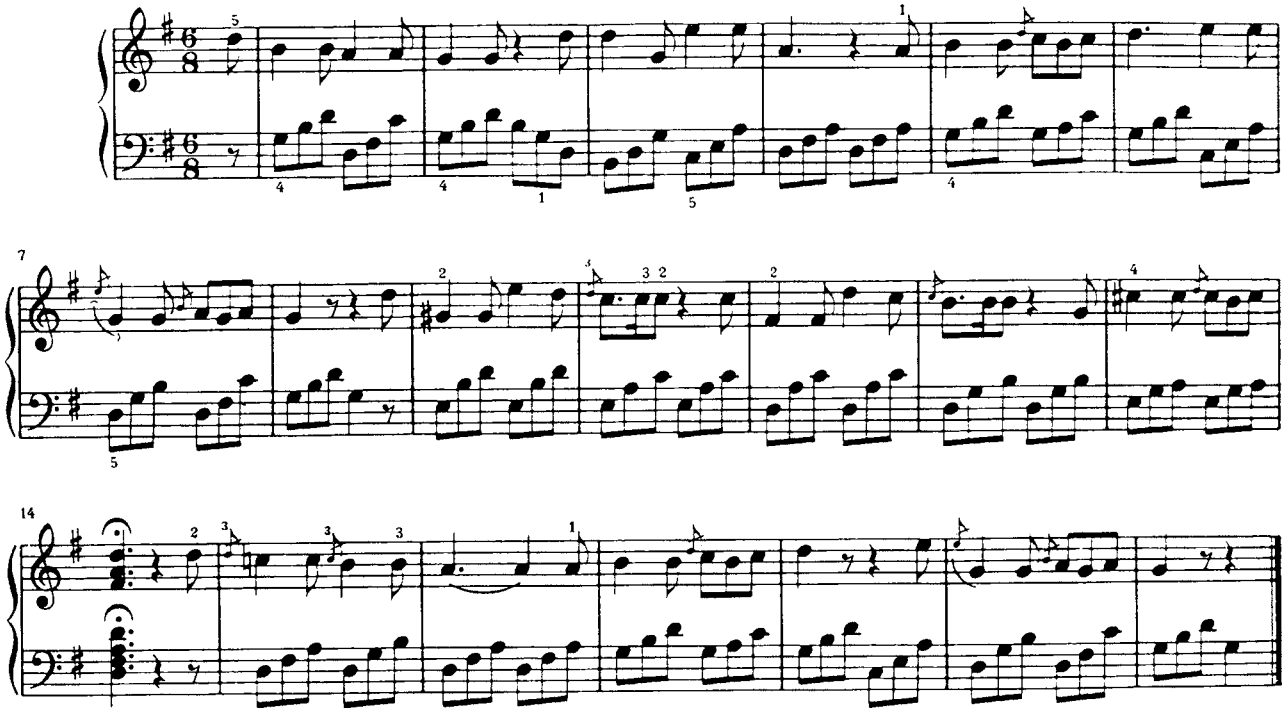


FIG. 1. Score of the Beethoven *Paisiello* theme.

method is in principle applicable to any expressive parameter, and this focus is chosen on pragmatic grounds. Moreover, although the dataset analyzed here was collected using a MIDI piano, the method can be applied to time series of measurements derived from an audio representation. The remainder of this paper demonstrates the application of DISSECT to a dataset of expert piano performance by analyzing the structural components contributing to performances at a single tempo, then showing how the deviations from proportional tempo which occur when a pianist is instructed to play at a higher or lower base tempo (see, e.g., Schmidt, 1985; Gentner, 1987; Desain and Honing, 1994; Repp, 1994; Windsor *et al.*, 2001) can be associated with differences in the interpretation of a small number or structural components.

II. THE TARGET DATASET OF PERFORMANCES

The performances modeled in this paper are derived from an earlier study which focused on grace note timing and the proportional tempo hypothesis (Windsor *et al.*, 2001) and are the same performances modeled in Penel *et al.* (1999) and Penel (2000). The piece performed has also been used in an earlier study of these issues (Desain and Honing, 1994).

A. Score

The piece used is the theme from Beethoven's six variations in G-major WoO 70 (1795) on the duet "Nel cor più non mi sento" from the opera "La Molinara" by Giovanni Paisiello (see Fig. 1).

The theme has a nominally isochronous broken-chord accompaniment in the left hand and a melody in the right, embellished by ornamental grace notes, and is notated in compound duple meter. The melodic gestures begin with an upbeat eighth note. The piece is essentially in two voices, except at the paused chord two-thirds through. Interestingly, the metrical and phrase structures of the piece are out of phase by one eighth-note unit, a common feature of music from this period. This feature alone suggests that this piece is an interesting candidate for the analysis to be carried out here, given that these two structural components might both be regarded as having a role to play in generating expression.

B. Performer and recording procedure

The performances were originally recorded for Windsor *et al.* (2001). The performer was a professional pianist and instrumental professor at the Tilburg Conservatory in the Netherlands (age 26). He was paid an appropriate hourly professional fee. The inter-onset timings of note onsets in the performances were captured using a Yamaha Disklavier MIDI grand piano and recorded via MIDI on a Macintosh PowerPC 9600/233 running a commercial sequencer package.

The performer had been given three weeks to prepare performances at nine different tempi from the score in Fig. 1. From these nine tempi, we have selected three instructed tempi for this study: "slow" [50 dotted quarter-note beats per minute (BPM)], "medium" (57 BPM), and "fast" (75 BPM). Although 50 BPM might seem rather slow for this piece, and 75 BPM rather fast, the pianist reported that they were musically acceptable. The "medium" tempo was regarded the most musically uncontroversial by the pianist.

Within the original experiment the pianist played randomized blocks of five repetitions of the theme at each of the tempi, giving a total of 45 complete performances. The pianist was allowed to practice the theme at the tempo requested

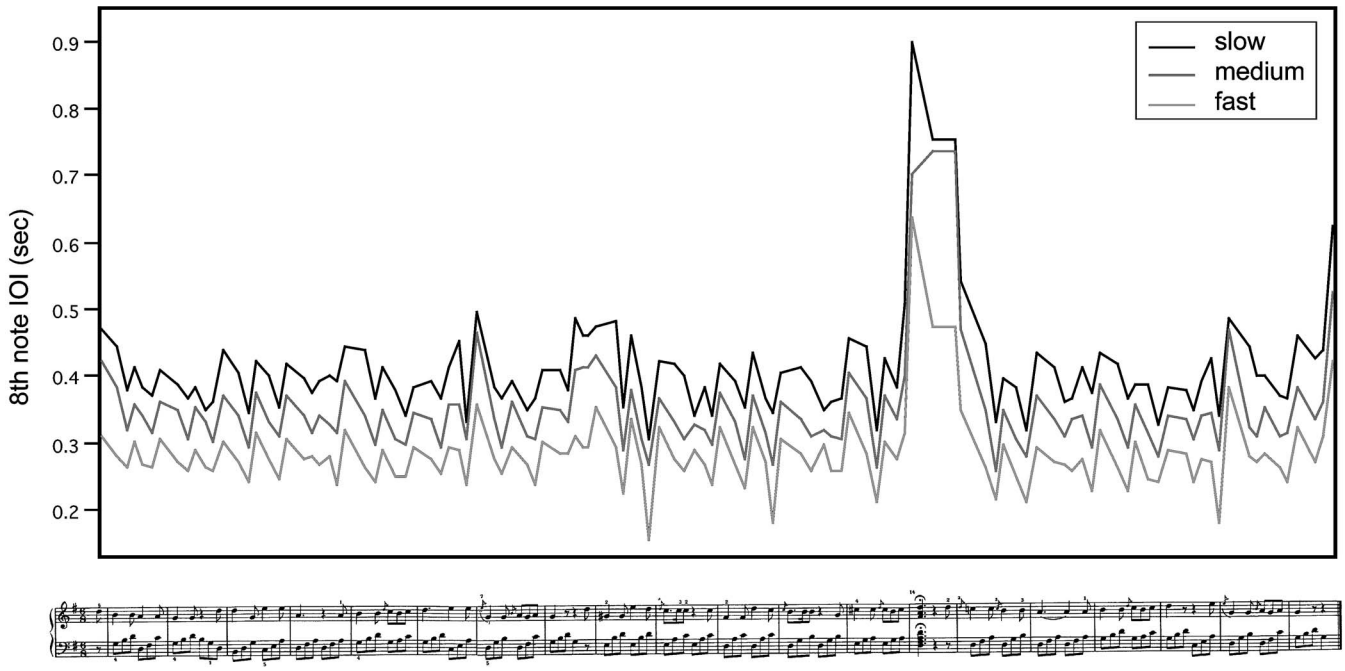


FIG. 2. Performances at each tempo, plotting score position against eighth-note IOI in seconds.

(a digital metronome was provided to remind the pianist of the tempo), and was asked to indicate whenever the next block could be recorded. Between each repetition there was a short break of about 5 s.

Using POCO (Honing, 1990) the onset times of all notes in the 15 performances were extracted, and inter-onset intervals (IOIs) were determined by onsets of melody notes (right hand) or by onsets of notes in the accompaniment (left hand) when there was no melody note. Grace note onsets were excluded from all analyses reported here [see Windsor *et al.* (2001) for an analysis of their timing].

C. Descriptive statistics for the selected performances

The 15 performances selected here were remarkably consistent within tempo condition, but show evidence of an effect of tempo on note timing. An ANOVA taking note IOI (for all onsets except those which precede rests in the score and the last onset) as the dependent variable, tempo condition and note position as factors, and repetition (five levels) as a random factor shows a significant interaction between note position and tempo condition ($F_{220,1320}=3.2153$, $p<0.0001$). Clearly, the performer did not maintain proportional timing over tempo at the note level, but was able to provide consistently timed performances within tempi. Hence, for the purposes of this paper average IOIs were calculated for each onset within each tempo, creating the three timing profiles shown in Fig. 2.

Comparison of the three profiles illustrates how well they correlate (about 0.95 between fast and medium and between medium and slow, and about 0.9 between fast and slow, $n=113$), despite having clear local differences for certain note positions and an offset due to the effect of global tempo.

III. APPLYING THE METHOD TO THE TARGET DATASET

A. Overview

The method, the statistical assumptions of which are outlined below, fits a generalized linear model to a time series of inter-onset intervals collected from a real performance. This model takes as its input a representation of the musical structures which might account for variation in expressive timing, estimates the fit, and provides prediction profiles for each element in this structure. The analysis can be thought of as a decomposition of the expressive timing into profiles associated with different kinds and levels of musical structure.

B. Assumptions and procedure

The method assumes that the expressive timing signal, expressed as beat length (inverse tempo), is a sum of several repeating (and possibly overlapping) timing profiles, each one reflecting the expression of a distinct structural unit such as a (sub)phrase or a metrical level. A subset of these units may form a hierarchical decomposition (e.g., bars and beats for a tight hierarchical structure), but this is not forced. The profiles are assumed to consist of line segments, with breakpoints specified at the first and last notes they span and, if necessary, at one or more intermediate notes (usually one extra breakpoint in the middle suffices).

Figure 3 shows a schematic depiction of a score, its structural annotation, and a profile for each structural unit. Note how each profile is determined by a set of breakpoints: the local tempo at each score time unit is estimated as a parameter in the fitting procedure to a real performance. In this sense the method is music-theoretically informed, because this structural description of the piece has to be pro-

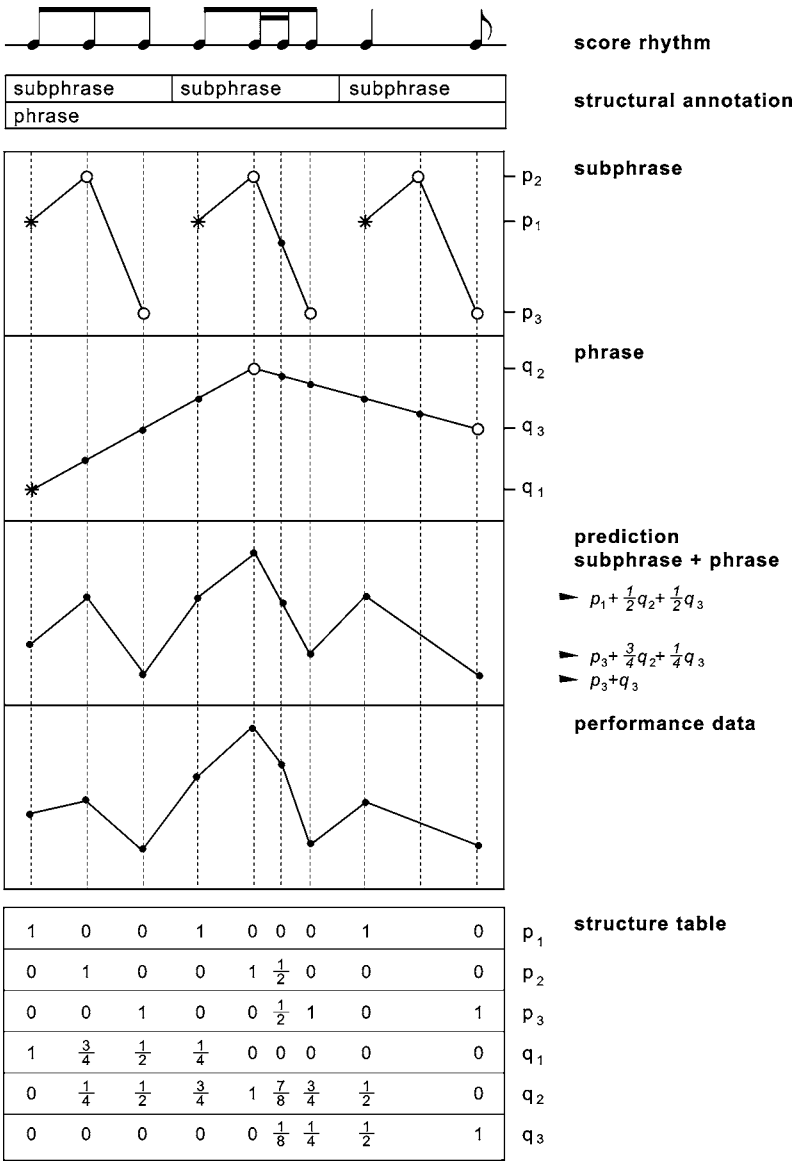


FIG. 3. A schematic depiction of a score, its structural annotation, the profiles for each structural unit, and how they combine into a prediction to be compared with performance data. The bottom panel illustrates how the structure is expressed as a matrix reflecting the linear combinations of parameters that constitute the model and is used for the regression analysis.

vided before an analysis can be done. The component profiles are defined by parameters, and the points of the overall profile are given by (weighted) sums of parameters, interpolating between them where necessary. The weights, which capture the structural description, are collected in a matrix A. There is a row in this table for each parameter, and a column for each note, i.e., each measured performance data point. The coefficients in the table specify the structural decomposition. If a note falls on a breakpoint of a profile, the corresponding coefficient in the table is 1; if it is outside the line segment starting or ending at that breakpoint, it is 0; and if it is on such a line segment, the coefficient expresses a linear combination (interpolation) of two parameter values. This table is generated from the structurally annotated MIDI score file in POCO. Now the predicted overall profile can be fit to the performance data.

If the expressive data to be predicted are expressed as vector x, with x_i being the local tempo of note i, and the parameters as vector p, the problem is to find the p_{opt} that minimizes the difference between the predicted A^*p and observed x:

$$p_{opt} = \text{argmin}_p \|A^* p - x\|.$$

Using the sum of the square errors as a measure of difference this is a linear regression problem that can be solved with simple means. The predicted overall profile is given by A^*p_{opt}. As the parameters p_i decompose into subsets, one set for each component, each component profile is calculated in a similar way, but zeroing in p_{opt} all parameters not belonging to that profile.

Since profiles repeat, we can usually create a nondegenerate matrix A and use fewer parameters than data points. However, because beginnings and ends of overlapping profiles will often coincide, the rank of A may be lower than its dimension. Clamping a few parameters to zero solves this problem.

The significance of individual parameters is not so relevant, as they form an inherent part of a profile, but the whole profiles are reanalyzed in a standard multiple regression which yields their contributions to the explained variance and their significance levels.

If optimization of free parameters leads to a good overall

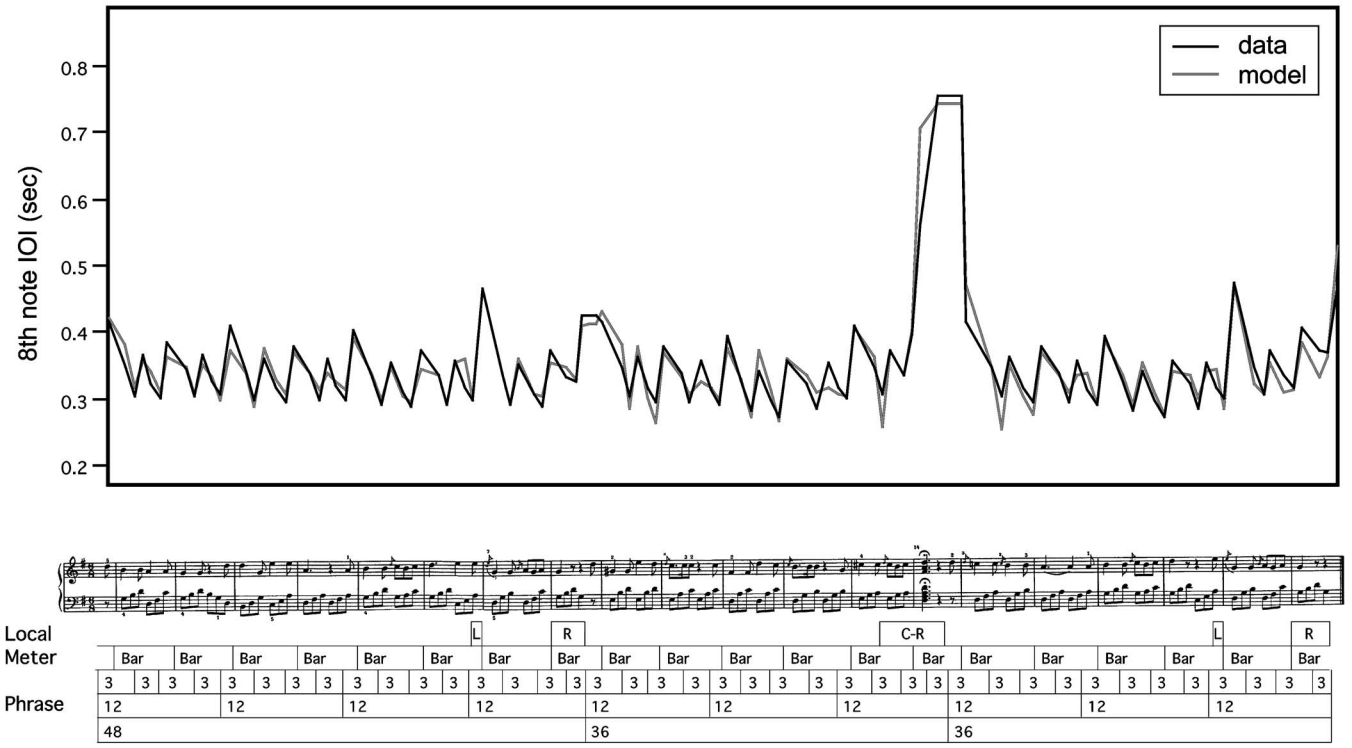


FIG. 4. Observed and predicted IOIs plotted against a score annotated with phrase, bar, and noncontiguous segments (L=leap; R=ritard; C-R= chord-ritard). Phrase segments are identified by their duration in score time measured in eighth notes. Note that the x axis is warped to align with the musical notation.

fit accounting for a large proportion of the variance, the musical structure is appropriately chosen. This means that the performance data exhibits systematic expressive features directly linked to the structural description.

C. The musical structure and constraints on the associated profiles

1. The structural representation

A set of structural units was added to the score in POCO (which adds a structural annotation capability to standard MIDI files). This structural annotation reflects the analytical intuitions of the first and third authors, breaking the piece down into a hierarchy of phrases, a metrical hierarchy, and identifying local sources for expression at phrase endings and accounting for the fermata. The analysis here is similar to that employed in Penel *et al.* (1999) and Penel (2000). The score is annotated with these structural units in Figs. 4 and 5, and further detail on the segmentation can be found in Table I.

2. Profile constraints

We have chosen here to constrain the model to a certain extent in order to reduce the number of free parameters in line with some hypotheses about patterns of expressive timing. The following constraints represent the "generative rules" we have chosen to include: (1) tempo change is linear between breakpoints (see below for our rationale for this) and (2) expressive timing is equal when structural units are repeated. In this instance each structural unit has a profile consisting of straight line segments with breakpoints at determined positions. The program allows for arbitrary complex shapes with many breakpoints, but only three point profiles (for the main phrase and metric units), two point profiles (for the final ritard), and local one point units (for local effects) were used, except for the profile for 12-phrase, which has five breakpoints to allow for expression associated with its initial upbeats and final interval. Table I specifies the extent, shape, and number of free parameters for each profile with brief descriptions. All other intermediate breakpoints are located at the midpoint of the structural unit. Other shapes or linear combinations could have been applied to the analysis of this performance, but small numbers of breakpoints and piecewise linear profiles were adopted to demonstrate the application a set of relatively simple statistical assumptions. Many generative models use nonlinear curves, but these tend to mask step tempo changes [see Windsor and Clarke (1997) for a discussion of this issue in relation to Todd (1992)].

3. Accounting for global tempo

The resulting profiles combine additively to predict the observed performance. However, the freedom in doing so is still too large: the global tempo can be explained as an offset to any profile that spans the whole piece, or be distributed between them. Hence, choices have to be made to reduce the number of parameters and make a unique solution to the optimization possible. In this instance the initial values of most profiles are clamped to zero (except for the global tempo intercept and the local pointwise parameters), and a constant intercept parameter was added to capture the global

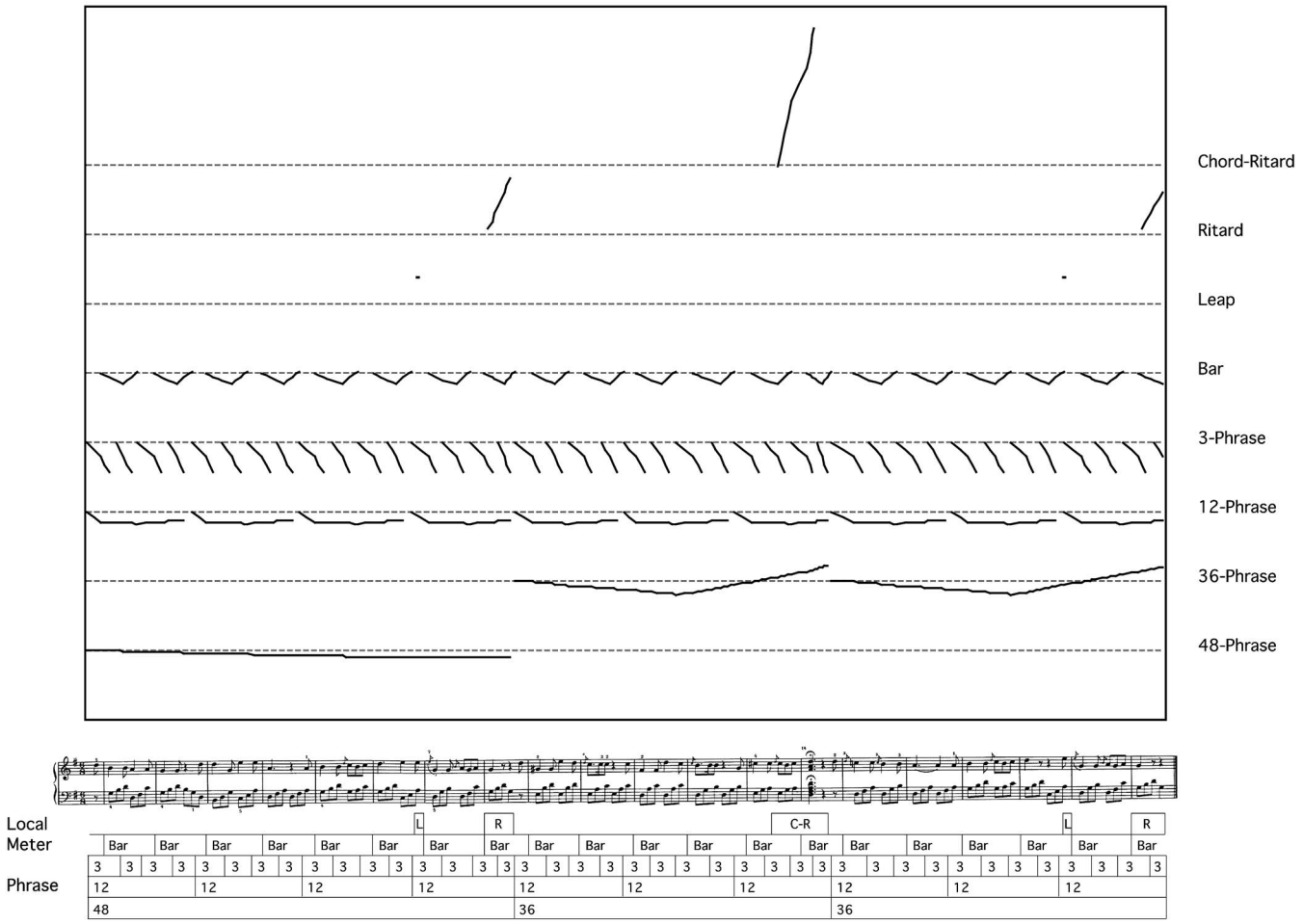


FIG. 5. The decomposition of the expressive profiles: the *x* axis shows score position, the *y* axis the parameter values (in eighth note IOIs in seconds). The names of the structural units are displayed to the right of each profile. Score annotations as in Fig. 4. Note that the *x* axis is warped to align with the musical notation, distorting the canonically regular shape of the profiles.

tempo. The 12-phrase's last breakpoint was also clamped to zero. Thus all profiles reflect timing deviations relative to the tempo specified in the intercept parameter.

D. Example analyses

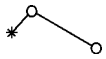
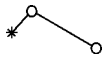
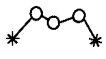
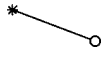
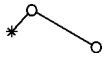
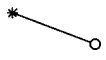
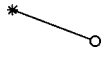
1. Structural decomposition at medium tempo

Given the structural annotation described above, the method fits the prediction to the medium tempo data quite tightly, as is shown in Fig. 4, explaining about 95% of the variance. This is a good fit considering the number of data points (122) and free parameters (13). Fitting the structure to the five individual performances rather than their average delivered only slightly worse fits, ranging between 92% and 95%, indicating that the averaging process indeed removed a small amount of unsystematic noise, possibly motor noise. Moreover, to check how far the model generalizes over repetitions, parameter values obtained in the fit to one performance were applied to all others. This full cross-validation gave again only slightly worse results as on average 92% of the variance was explained: repeated performances are so much alike that a single model generalizes over them. This also means that the successful fit cannot be attributed to overfitting the data: regularities in expressive timing are discovered by the model in a robust manner.

As well as accounting for local expressive timing, such as the pattern associated with the smallest phrase-structure units (three eighth notes long), the model captures more global aspects, such as the ritardandi at the ends of the large-scale phrases. There are some points at which the fit is less good, but a decision was made to accept a trade-off between fit and number of free parameters. It might have been possible to improve the fit at the local phrase level by associating different phrases of the same length with different parameters, but this would have significantly decreased the extent to which the model constrains the data. Similarly, it might have been possible to achieve a better fit with a somewhat more irregular phrase structure, one which allows, for example, for both upbeat and afterbeat phrasings. To test this hypothesis a fit was calculated for a near identical structure to that shown in Fig. 4. The difference here was that the continuous 3-phrase units were replaced with the following pattern: 3342 3342 3333 3333 3342 3342 3333 3333 3333 3333. The fit was near identical (to three decimal places), but requires an extra three free parameters.

However, the decomposition into expressive components that the method achieved is more interesting than the overall fit. Figure 5 shows what (and where) the various components contribute to the total profile.

TABLE I. Structural units and their duration, shape (* is a breakpoint clamped to zero, o is a breakpoint controlled by a free parameter), and number of (free) parameters.

Kind	Name	Description	Extent in 8th notes	Shape	Parameters (free)
phrase	48-Phrase	Opening phrase at same hierarchical level as 36-phrase. Allows for acceleration or deceleration towards and away from central breakpoint.	48		3 (2)
	36-Phrase	Two equal length phrases at same level as 48-phrase. Allows for acceleration or deceleration towards and away from central breakpoint.	36		3 (2)
	12-Phrase	Sub-divides the 48- and 36- phrases. Contains extra breakpoints to allow for agogic accent for anacruses and micropause for last event.	12		5 (3)
	3-Phrase	Lowest level in grouping structure. Allows for acceleration or deceleration within this span.	3		2 (1)
meter	Bar	Profile reflecting the 6/8 metre, with upbeat, and incomplete final bar. Allows for acceleration or deceleration towards and away from breakpoint.	6		3 (2)
local	Leap	Delayed note preceding a grace note to a downwards leap. Only two occurrences.	1		1 (1)
	Chord-Ritard	Slowing towards the fermata.	8		2 (1)
	Ritard	Slowing down at end of first and last long phrase.	5		2 (1)

The results of the method can now be used to detail the links between the different musical structures and the expressive timing and to learn about the interpretation of this specific piece. The most important structural unit was a large slowing down (chord-ritard), explaining three quarters of the variance. The decomposition reveals that in addition to this deceleration towards the fermata, there are less extreme ritardandi (ritard) at the ends of each major phrase, and a repeated acceleration is present over each three eighth-note unit (3-phrase). Over the beginning of the piece (48-phrase) the pianist accelerates gradually, and in each of the subsequent phrases (36-phrase) he follows a schematic acceleration-deceleration profile [familiar from work such as Todd (1992)]. A local lengthening occurs for each note preceding the downward leaps in the melody (leap), marking this distinctive feature. At an intermediate level in the phrase structure (12-phrase) there is an agogic accent on the first event (the upbeat) followed by a slight acceleration deceleration profile. Lastly, the metrical structure is marked in a highly schematic manner, with a pattern of linear acceleration/deceleration across the six beats, rather than a marking of particular beat strengths according to their hierarchical importance (such as described in Palmer and Krumhansl, 1990; Parncutt, 1994).

It has to be stressed that the method is well suited to exploratory data analysis: trying out different structural descriptions and checking out how far they help the fit to the data. In arriving at this successful structural description a number of alternatives were tried. Small increases in the goodness of fit could be achieved by adding parameters associated with additional features, but these increases were regarded as too expensive. For example, the addition of two subphrases of 24 eighth-note units duration within 48-phrase adds two free parameters but only improves the fit by less than 1%. Other structural descriptions that were tested but failed to improve the results were phrases of six and nine units. The most critical improvements in fit/parameter ratio were achieved by creating separate profiles for the ritardandi. This allows for the relatively extreme tempo change at and before the fermata. Table II shows the size of each profile, measured as relative standard deviation, the significance of their contributions, and the amount of variance that each explains. Though some effects and contributions are small, all profiles contribute significantly, and the significance of some contributions is extremely high. Note that these fits arose using 13 parameters, predicting 122 data points. Because some profiles only contribute to a time segment of the data, the amount of variance explained in the whole performance

TABLE II. Explained variance, significance of fit, size (in proportion to the sd of the observed performance), explained variance in stepwise residues, and the number of parameters of each structural component and the complete model.

Profile	r^2	$p<10^{\wedge}$	Proportional sd	Stepwise r^2	Parameters (free)
Chord ritard	0.79	−52	0.82	0.79	2 (1)
3 Phrase	0.08	−21	0.28	0.45	2 (1)
12 Phrase	0.05	−2	0.08	0.18	5 (3)
Ritard	0.01	−11	0.18	0.17	2 (1)
Bar	0.02	−4	0.09	0.15	3 (2)
36 Phrase	0.34	−5	0.14	0.17	3 (2)
Leap	0.02	−3	0.09	0.11	1 (1)
48 Phrase	0.01	−2	0.07	0.07	3 (2)
Full model	0.95	−65	0.98	Nil	21 (13)

TABLE III. The explained variance and the significance of the contribution of the structural units in each tempo.

	Fast		Medium		Slow	
Profile	r^2	$p<10^{\wedge}$	r^2	$p<10^{\wedge}$	r^2	$p<10^{\wedge}$
Chord-ritard	0.49	−17	0.79	−52	0.65	−25
3-Phrase	0.20	−13	0.08	−21	0.07	−9
12-Phrase			0.05	−2		
Ritard	0.01	−2	0.01	−11	0.02	−5
Bar			0.02	−4		
36-Phrase	0.36	−3	0.34	−5	0.43	−4
Leap			0.02	−3		
48-Phrase			0.01	−2		
Full model	0.80	−33	0.95	−65	0.84	−39

is not a very good indication of their relative importance. Otherwise one would be tempted, for example, to be satisfied with the huge contribution of the local chord-ritard, which by itself leaves the expressive timing of most of the piece undefined. In contrast, one would be tempted to underestimate the contribution of the 48-phrase to the beginning section, as the correlations are calculated over the whole piece. However, a somewhat more fair evaluation can be obtained by calculating stepwise residues and reporting the variance explained by each subsequent profile in the corresponding residue. For this, profiles are ordered by explained variance in the remaining residue. This is shown by the fifth column of Table II and demonstrates that some profiles with a small contribution to the overall model are quite good predictors after some other profiles have already been taken care of.

2. Discussion

The decomposition found supports a structural analysis that includes both global features, such as long- and short-term periodicities in metrical and phrase structure, and local features, such as the pianist's response to the fermata. Where periodic structural features are present, a model predicting that the corresponding expression will be fairly similar at each repetition succeeds in predicting this pianist's average behavior very well. Although the large and expected effect of the fermata is highlighted, almost half of the remaining variance can be explained by a repeated pattern of expressive timing at the level of the smallest subphrase (3-phrase). However, all the other profiles account for significant proportions of the variance as well, and the method helps highlight the components that make up the performance.

Two aspects here are worth commenting further on. First, the expressive timing does seem to reflect a concern with local aspects of the musical structure at the expense of more global tempo rubato over longer structural spans. This would be in line with a less "romantic" interpretation of this piece, which is, after all, from the classical period. Second, it is interesting to note that the method allows one to disambiguate between the metrical and short-duration phrase structures, which are out of phase by one eighth-note unit, but multiples of one another. Although the local phrasing accounts for much of the variance in expressive timing, the metrical structure improves the fit still further and manages to predict the expressive timing significantly on its own. Much of the expressive timing follows patterns often observed (see, e.g., Palmer, 1989) or predicted (see Todd, 1992) for classical-romantic repertoire, but our approach allows one to easily observe how such patterns are applied in a nonconsistent manner. There is evidence for both acceleration towards the middle of the phrase here, but also acceleration through such intermediate points towards a phrase boundary. This argues against any model that applies a fixed rule to similar structures across a whole performance. If performers select and combine expressive strategies in a piecemeal manner, inflexible rules cannot capture the decisions that lead to such flexibility. In other words, a model needs to cope with both the extent to which a rule is applied (its weight), but also must account for which rule to apply to any given structure. A choice between accelerating towards a goal or marking it with a gradual deceleration would be a challenging one to simulate, especially where multiple rules may be operating on the same data-points.

3. Application to the analysis of tempo and timing

As shown above, there is evidence that the performer did not maintain proportional timing across the three tempo conditions. Applying the same methods and structural analysis as above but to the data for slow and fast instructed tempi result in good fits as well ($R^2=0.80$ for the fast tempo and $R^2=0.84$ for the slow tempo). The fits and patterns for the profiles are quite similar to that for the medium tempo, although the cumulative fit is not as good (cf. $R^2=0.95$). Table III lists the amount of variance the individual profiles explain in the different tempo conditions.

To check if our choice of tempi was reasonable, the model was run on the data of all nine tempi obtained in Windsor *et al.* (2001), averaged over repetitions. The best fit ($R^2=0.95$) was indeed obtained with the medium tempo, the worst with the fastest ($R^2=0.79$) and the slowest ($R^2=0.84$). The second-fastest and second-slowest tempo, and all tempi in between, allowed the model to explain 92% of the variance or more. This may indicate that at the extreme tempi the possibility to control the performance reliably starts to break down, but that the model and the structural description hold very well for the largest part of the tempo range. The cross

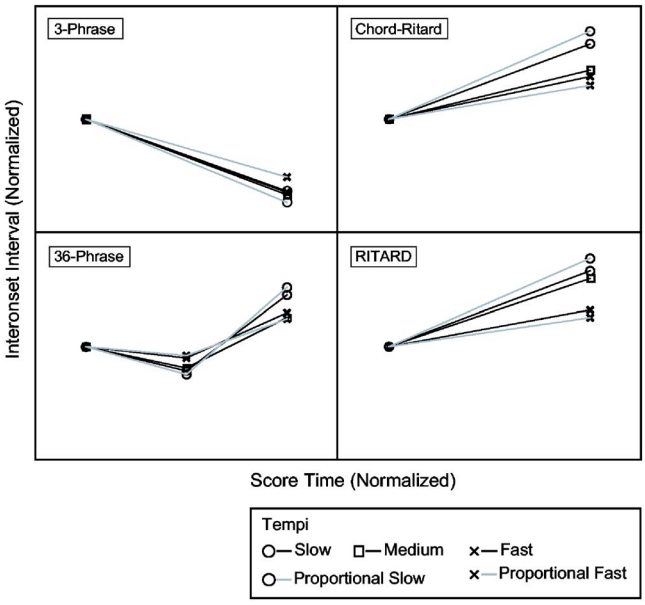


FIG. 6. Panel of the shapes component profiles in the three tempi. Here only one repetition of each profile is shown, and the magnitudes on the x and y axes are normalized for easy comparison of shapes. Reference lines have been added to show proportional tempo predictions for slow and fast tempi taking medium tempo as the baseline.

validation of the model for performances at a specific tempo, with parameters obtained from a performance at another tempo, resulted on average still in 88% explained variance. This again shows that we do not overfit: the model generalizes to a certain extent even across tempi and captures performance regularities in a robust manner. It is, however, interesting to investigate further which profiles adapt in a tempo-specific manner and which do not.

The full models at the fast and the slow tempo become simpler, as some profiles (the weakest in the medium tempo) fall below significance. An advantage of the DISSECT method is that both the relevance and the shape of the profiles can be taken into consideration as they adapt to various (tempo) conditions. The differences in expressive timing for individual profiles at each tempo are shown in Fig. 6, showing only the contributions that are significant at all tempi. In this graph the horizontal and vertical axes are normalized, losing the size of the effects and focusing only on the shape of the profiles, as they adapt to tempo.

The deviations from proportionality for each structural component can now be clearly seen in Fig. 6, as reference lines for proportional invariance are added. The profile for 3-phrase does not scale at all across tempi, while the other profiles scale with tempo, though a bit less than truly proportionally. Such analyses of the scaling behavior of individual profiles with regard to tempo could be practically applicable to the design of a "smart tempo knob" that would adapt performance timing to a set tempo, just like a human pianist would.

4. Discussion

It has been demonstrated that, although highly correlated, the performances at the different instructed tempi are not proportionally invariant. It is therefore useful to be able to show how the differences in expressive timing at the three tempi are related to the structure. Here, the differences can be attributed to subtle changes in the expression of the structural components, partly related to their size, the proportion allocated to the components in relation to others, and partly related to their shape, to the nonproportional scaling of the expression itself (3-phrase). Whether unconsciously or consciously, the pianist has reduced the contribution of the fermata, especially at the fastest tempo, and increased the relative contribution of the shortest phrase unit (3-phrase) by keeping its absolute size invariant. Without the decomposition such aspects of expressive timing are almost impossible to disentangle, and one is left with only qualitative and inductive comparisons of three almost identical tempo profiles. It is possible that the pianist's lessening of attention to the fermata and its preceding onsets and greater accentuation of low-level phrase structure might reflect a less "romantic" interpretation at the faster tempo, which, given the association of faster and less flexible tempo with more 'classical' styles of playing (see Hudson (1994), although see also Bowen (1996) for evidence that such changes are far from systematic), would not be unwarranted.

IV. GENERAL DISCUSSION

The approach to expressive timing described here may seem remarkably underspecified compared to others. Its use of free parameters allows for extremely good fits given a sensible structure, and it could be argued that this is a conceptual weakness. However, this is precisely what allows it to reveal the detailed relationships between structure and expressive timing in these performances. Other theories in this area specify explicit rules and shapes of the expressive profiles, either through experimentation (see, e.g., Sundberg, 1988) or machine-learning (as in Widmer and Tobudic, 2003). The approach developed here reflects a desire to learn these shapes in a somewhat more data-driven way, though of course the structural description is not inferred from the data, it is the given top-down assumption upon which the analysis builds. The examples reported above show the potential of this approach. This methodological concern has a matching theoretical counterpart. It is by no means a logical necessity that rule-following behavior is underpinned by symbolic rules. Even if an aspect of human behavior is, to an extent, systematic, this does not mean that it is rule governed. If expressive timing and musical structure are related, it is not the case that such a relationship need be governed by a mentally encoded set of production rules. Instead, it may be the case that performers learn somewhat similar ways of mapping structure to timing, but that these mappings are both flexible and interchangeable: a performer might choose to accelerate *or* decelerate towards the end of a phrase (see Palmer, 1989 and above) and she or he needs to decide how to combine different kinds of expression both within and between performances of the same and different pieces. These choices may be highly individual or related to stylistic or interpretative differences. Given that this is the case, models proposing a generalized set of rules mapping structure onto expression may only reveal what is least interesting in

musical performance (the way most performers play most of the time) and what is needed is a set of modeling tools that can reveal systematic patterns in performance expression in individual performances, not just those that are shared between many performances. This paper describes a methodology that focuses on this level of explanation, and yet may also discover general properties of expression which might not be captured by stricter models: not only did the parameter values for one performance generalize to an extent to repeated performances, but also across tempo conditions. Not only does this inform us about the regularities in musical performance, it also proves that we are not overfitting the data (explaining nonsystematic features of the training data with high accuracy).

Of course, the model and application presented here require further development. At present the model has only been applied to performances of a single piece by a single performer. A future aim is to show how this approach can illuminate the systematic yet individual nature of expressive timing and dynamics in performances of other pieces. Further research will show if the same approach can deal with nonlinear profile shapes (like the parabola). Another intriguing question, to be addressed in subsequent work, is if the use of a mixed additive-multiplicative model is better than the present linear version. It would separate tempo factors (which combine multiplicatively) and time shifts (which combine additively). Moreover, in principle it should be possible to generate possible structural descriptions automatically (within reasonable constraints regarding meter and phrases) and search for a compact description that explains the data well. This would make the method even more data driven and automatic.

We would argue that editing musical expression by remixing expressive profiles to create a new performance with a different expressive balance and focus would enable the same extensive and parametric control in the area of interpretation that is already commonplace for sound synthesis, filtering, and spatialization. For such practical applications an expression synthesizer has been developed as a companion to the DISSECT analysis method. The synthesizer mixes a new performance with edited expression using the profiles yielded by DISSECT and a set of weights. These weights control the extent to which a specific expressive component is present in the output. Informally, the results sound quite promising: for example, performances with exaggerated bar timing or muted final ritard timing sound to us as if they have been played by a human performer who was instructed to play in that way. A more elaborate discussion and a demonstration of the expression mixer is available at the MMM website (⟨www.nici.ru.nl/mmm⟩ under demos). Generating expression profiles with muted, exaggerated, or otherwise perturbed components provides a rich domain of stimuli that can be used to probe perceptual and motor processes (see, e.g., Clarke, 1993; Clarke and Windsor, 2000); greater and more detailed control of the parameters in such experiments would allow researchers to modify only certain aspects of expression while leaving others invariant.

V. CONCLUSIONS

This paper has presented an approach to the decomposition of expressive timing which can be thought of as a generalized model of the mappings between structure and expression that have been empirically observed since the time of Seashore (1938). We have shown how this approach can independently predict different structural components in expression, how it is sensitive to subtle changes in interpretation by a single performer, and how it provides standard and interpretable estimates of fit. Although the decomposition method only makes a few assumptions about the "rules" which map structure onto expression, it provides a sensitive and systematic method for gaining insight into what constraints operate in the domain of musical expression, a topic which, despite concerted effort and much excellent research, still seems to pose many questions. Researchers know something about what performances have in common and how they differ in gross terms (see, e.g., Repp, 1992) and are sometimes able to model some of these generalities (see, e.g., Todd, 1992). The focus here is on the subtleties of expression in a single performance and how these change under different performance conditions, and, as we have shown, such subtleties can be effectively highlighted if one systematically decomposes the expressive signal into multiple components.

ACKNOWLEDGMENTS

The authors would like to acknowledge the advice and help that they received during the course of this research from Eric Maris, Henkjan Honing, Renee Timmers, Makiko Sadakata, Diana Deutsch, and from two anonymous reviewers. This work was funded by the Netherlands Organisation for Scientific Research (NWO); the Faculty of Social Sciences, Radboud University; the Faculty of Performance, Visual Arts and Communications, University of Leeds; and a fellowship in cognitive sciences from the French Ministry of Education and Research and a PECA (Perception Et Cognition Auditive) travel fellowship.

Bowen, J. A. (**1996**). "*Tempo* Duration & Flexibility: Techniques in the Analysis of Performance," J. Music. Res. **16**(2), 111–156.

Chaffin, R., and Imreh, G. (**2002**). "Practicing perfection: Piano performance as expert memory," Psychol. Sci. **13**(4), 342–349.

Clarke, E. F. (**1988**). "Generative principles in music performance," in *Generative Processes in Music: The Psychology of Performance, Improvisation, and Composition*, edited by J. A. Sloboda (Clarendon, Oxford), pp. 1–26.

Clarke, E. F. (**1993**). "Imitating and evaluating real and transformed musical performances," Music Percept. **10**(3), 317–341.

Clarke, E. F., and Windsor, W. L. (**2000**). "Real and Simulated Expression: a Listening Study," Music Percept. **17**(3), 1–37.

Clynes, M. (**1983**). "Expressive microstructure in music, linked to living qualities," in *Studies of Music Performance*, edited by J. Sundberg (Royal Swedish Academy of Music, Stockholm), pp. 76–181.

Desain, P., and Honing, H. (**1994**). "Does expressive timing in music performance scale proportionally with tempo?" Psychol. Res. **56**, 285–292.

Desain, P., and Honing, H. (**1997**). "Structural Expression Component Theory (SECT), and a method for decomposing expression in music performance," in *Proceedings of the Society for Music Perception and Cognition Conference* (MIT, Cambridge), p. 38.

Desain, P., Honing, H., Van Thienen, H., and Windsor, L. (**1998**). "Computational Modeling of Music Cognition: Problem or Solution?" Music Percept. **16**(1), 151–166.

Dixon, S., Goebl, W., and Widmer, G. (**2002**). "Real time tracking and visualisation of musical expression," in *Proceedings of the Second International Conference on Music and Artificial Intelligence*, edited by C. Anagnostopoulou, M. Ferrand, and A. Smaill (Springer, Berlin), pp. 58–68.

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

Friberg, A. (**1995**). "Matching the rule parameters of Phrase arch to performances of 'Träumerei': a preliminary study," in *Proceedings of KTH Symposium on Grammars for Music Performance* (KTH, Stockholm).

Gentner, D. R. (**1987**). "Timing of Skilled Motor Performance: Tests of the Proportional Duration Model," Psychol. Rev. **94**(2), 255–276.

Honing, H. (**1990**). "POCO: an environment for analysing, modifying, and generating expression in music," in *Proceedings of the 1990 International Computer Music Conference* (International Computer Music, San Francisco), pp. 364–368.

Honing, H. (**1992**). "Expresso, a strong and small editor for expression," in *Proceedings of the 1992 International Computer Music Conference* (International Computer Music, San Francisco), pp. 215–218.

Hudson, R. (**1994**). *Stolen Time: The History of Tempo Rubato* (Clarendon, Oxford).

Palmer, C. (**1989**). "Mapping musical thought to musical performance," J. Exp. Psychol. Hum. Percept. Perform. **15**(12), 331–346.

Palmer, C. (**1997**). "Music Performance," Annu. Rev. Psychol. **48**, 115–138.

Palmer, C., and Krumhansl, C. L. (**1990**). "Mental representations for musical meter," J. Exp. Psychol. Hum. Percept. Perform. **16**(4), 728–741.

Parncutt, R. (**1994**). "A Perceptual Model of Pulse Salience and Metrical Accent in Musical Rhythms," Music Percept. **11**(4), 409–464.

Penel, A. (**2000**). "Variations temporelles dans l'interprétation musicale: processus perceptifs et cognitifs," unpublished doctoral dissertation, Université Paris 6. Available from the third author on request.

Penel, A., and Drake, C. (**1998**). "Sources of timing variations in music performance: A psychological segmentation model," Psychol. Res. **61**, 12–32.

Penel, A., Desain, P., Maris, E., and Windsor, W. L. (**1999**). "A decomposition model of expressive timing," in *Proceedings of the 1999 SMPC* Evanston, IL, p. 21.

Repp, B. H. (**1992**). "Diversity and commonality in music performance—an analysis of timing microstructure in Schumann's Traumerei," J. Acoust. Soc. Am. **92**, 2546–2568.

Repp, B. H. (**1994**). "Relational invariance of expressive microstructure across global tempo changes in music performance: An exploratory study," Psychol. Res. **56**, 269–284.

Schmidt, R. A. (**1985**). "The search for invariance in skilled movement behavior," Res. Q. Exerc Sport **56**(2), 188–200.

Seashore, C. E. (**1938**). *Psychology of Music* (McGraw-Hill, New York).

Shaffer, L. H. (**1981**). "Performances of Chopin, Bach and Bartok: studies in motor programming," Cognit Psychol. **13**, 326–376.

Sundberg, J. (**1988**). "Computer synthesis of music performance," in *Generative Processes in Music: The Psychology of Performance, Improvisation, and Composition*, edited by J. A. Sloboda (Clarendon, Oxford).

Sundberg, J., Friberg, A., and Bresin, R. (**2003**). "Attempts to Reproduce a Pianist's Expressive Timing with Director Musices Performance Rules," J. New Music Res. **32**(3), 317–325.

Thompson, W. F., and Cuddy, L. L. (**1997**). "Music performance and the perception of key," J. Exp. Psychol. Hum. Percept. Perform. **23**(1), 116–135.

Todd, N. P. (**1985**). "A model of expressive timing in tonal music," Music Percept. **3**, 33–58.

Todd, N. P. (**1992**). "The dynamics of dynamics: A model of musical expression," J. Acoust. Soc. Am. **91**, 3540–3550.

Todd, N. P. McA. (**1995**). "The kinematics of musical expression," J. Acoust. Soc. Am. **97**, 1940–1949.

Widmer, G., and Tobudic, A. (**2003**). "Playing Mozart by Analogy: Learning Multi-level Timing and Dynamics Strategies," J. New Music Res. **32**(3), 259–268.

Windsor, W. L., and Clarke, E. F. (**1997**). "Expressive timing and dynamics in real and artificial musical performances: using an algorithm as an analytical tool," Music Percept. **15**(2), 127–152.

Windsor, W. L., Aarts, R., Desain, P., Heijink, H., and Timmers, R. (**2001**). "The timing of grace notes in skilled musical performance at different tempi: a preliminary case study," Psychol. Music **29**, 149–169.

Zanon, P., and De Poli, G. (**2003a**). "Time-varying estimation of parameters in rule systems for music performance," J. New Music Res. **32**(3), 295–316.

Zanon, P., and De Poli, G. (**2003b**). "Estimation of parameters in rule systems for expressive rendering in musical performance," Comput. Music J. **27**(1), 29–46.