Expressive Timing and Dynamics in Real and Artificial Musical Performances: Using an Algorithm as an Analytical Tool

Author(s): W. Luke Windsor and Eric F. Clarke

Source: *Music Perception: An Interdisciplinary Journal*, Winter, 1997, Vol. 15, No. 2 (Winter, 1997), pp. 127-152

Published by: University of California Press

Stable URL: https://www.jstor.org/stable/40285746

Music Perception
Winter 1997, Vol. 15, No. 2, 127–152

Expressive Timing and Dynamics in Real and Artificial Musical Performances: Using an Algorithm as an Analytical Tool

W. LUKE WINDSOR & ERIC F. CLARKE
University of Sheffield

This paper compares timing and key-velocity data collected from a skilled performance of Schubert's G♭-major Impromptu (Opus 90) with a number of performances generated by a version of a musical expression algorithm proposed by Todd (1992). Regression analysis is used to demonstrate both the shortcomings of this model as a complete *explanation* of musical expression and how it might be more successfully used as a *tool* for analyzing data from real performances. Used in this second manner, the algorithm is shown to provide a general expressive baseline against which other aspects of expression may be highlighted. It is also suggested that such a baseline provides a method of decomposing performances into continuous and discrete forms of expression. It is concluded that using algorithmic models as heuristic tools, rather than as explanations in themselves, may better serve our increased understanding of the flexible and multiple nature of musical expression.

Introduction

In the literature on musical performance that chooses to view performance empirically or systematically, the term *musical expression* has become synonymous with those strategies used by musicians to shape their performances. Changes in instantaneous tempo or dynamic level; inflections of pitch; vibrato; and overlap or separation of successive events (articulation) can all be regarded as contributing to such musical expression. In this paper, we focus on two types of musical expression, expressive timing and dynamics, and on expert performance at the piano. There is a general background of research that is aimed specifically at expression in ex-

Address correspondence to Dr Luke Windsor, NICI, Nijmegen University, P. O. Box 9104, NL-6500 HE Nijmegen, The Netherlands. (e-mail: windsor@nici.kun.nl)

pert piano performance, and there exists a systematic model that attempts to explain expressive dynamics and timing (Todd, 1992) that will be assessed in detail in this paper. Moreover, it is relatively easy to measure the force with which piano keys are struck, which gives an index of dynamic level, and the times at which keys are depressed, as opposed to measuring the onset times and dynamic levels of acoustic events.

Overall, in this study, we sought to evaluate certain theoretical ideas developed to explain the moment-to-moment changes in dynamic level and tempo (N.B., all references to tempo in this paper are to instantaneous tempo, not to global tempo) that may be observed in expert piano performance. We seek to evaluate these ideas in two ways. First, we aim to evaluate the capacity of a particular computational model (Todd, 1992) to mimic a human performance. Second, and more importantly, we aim to use this model as a heuristic tool that might help identify and characterize particular expressive strategies in piano performance. The distinction between these two aims is important. Although it may be the case that a model of expression may fail to produce "human" results (i.e., fail a musical version of the Turing test), this does not mean that the model is of no value in explaining how musical expression is organized. Indeed, if a model in some way instantiates theoretical assumptions derived from observations of musical performance, it may allow not only a direct test of these assumptions as expressed in a set of rules, but may also point to modification of those assumptions through analysis of the model's shortcomings. Moreover, as is the case in this study, it can be accepted that a model may be incomplete, yet offer a partially accurate explanation of a phenomenon. By carefully analyzing such a model's failures, it may be possible to determine where our understanding of a complex phenomenon such as musical expression falls short.

Systematic Approaches to Musical Expression

To most musicians, the notion that dynamics and timing are used expressively would be obvious. Rubato, stress, contrast, and the like are all terms used by musicians to describe these kinds of expression. However, any suggestion that such expression is *systematic* would run the risk of ridicule. The score, whether through explicit instructions, or implicitly, through interpretation of ideological, emotional, or structural "content," might be seen as a source for a performer's expression, but this is not the same as admitting that such links are systematic. The notion that expression might be *rule-based* does not seem to match up with performers' experiences. However, considerable effort has been expended in attempting to show how musical performances share *systematic* patterns of timing and

dynamics, both between and within different performers' executions of the same pieces, and that the source of such musical expression lies within the structure of the music that is played (see, e.g., Clarke, 1988). Before an evaluation of any single example of such research, it is important to review the way in which such a view of musical expression has arisen. The evidence that has led to such a view has come from three sources: measurement of performances, models of performance, and experimental tests of such theories of expression. Since such experimental work is not a primary concern of this paper, the following discussion will focus on measurement and modeling approaches.

Direct measurement of the timing of the events in piano performance, and more recently the dynamic levels of these events, either directly as amplitude, or indirectly as the force of key depression, has resulted in a number of general observations. Since the work of Seashore (1938), the discovery that pianists intentionally and systematically (although not necessarily consciously) deviate from notated durations has been refined to suggest that such deviations from the score are systematically related to the structure of the music being played. For example, quantitative analyses of the timing of piano performances clearly identify the ways in which the metrical structure of a piece is "expressed" through systematic patterns of timing (e.g., Shaffer, Clarke, & Todd, 1985; Shaffer & Todd, 1987). Similarly, the grouping structure, or phrase structure, of a piece can be shown to correspond to systematic changes in instantaneous tempo (Gabrielsson, 1987; Repp, 1990; 1992; Shaffer & Todd, 1987) and dynamics (Gabrielsson, 1987; Todd, 1992). In this study, we focus on grouping structure (a hierarchy of phrases) to the exclusion of metrical structure (a hierarchy of beats and bars). The precise nature of these patterns of timing and dynamics and their relationship to each other, and to grouping structure, will be returned to later during a discussion of the computational model used here.

A number of different approaches have been taken to computational modeling of musical expression. For example, Sundberg and coworkers (see, e.g., Sundberg, 1988) have used a system of rules derived from, among other sources, the insights of musical performers. The rationale here is to assess the veracity of such intuitions by instantiating them as rules within an algorithm that can produce artificial performances whose acceptability can be assessed. A second approach is that of Clynes (1983): here, a complete theory of expression is instantiated in a model, rather than a set of largely independent rules. This paper is concerned with a third type of model, one that instantiates a theory of expression in algorithmic form, but one in which the theoretical basis for the algorithm is grounded in observations from human performances. Todd (1985, 1989, 1992) has proposed three such models, each of which draws on empirical data gathered from expert piano performances. The potential advantages of such a model

are not only that it provides a method of assessing the theoretical assumptions on which it is based, but also that these theoretical assumptions are based on data from actual performances. Moreover, it has potential advantages over a piece-meal analysis-by-synthesis approach (e.g., Sundberg, 1988) where interactions between different rules make it difficult to predict the consequences of changing aspects of the model.

The Todd (1992) Model of Musical Expression

The model with which we are directly concerned (Todd, 1992; also see Todd, 1985 and 1989 for descriptions of related earlier models) can be thought of as an instantiation of two aspects of what can be termed a "generative" approach to musical expression (see, e.g., Clarke, 1988). The first of these is a general characteristic of generative approaches to performance: a rule system maps structural features of the musical score onto expressive parameters in the performance. The second is its attempt to model a *particular* kind of mapping between grouping structure (i.e., of the kind proposed by Lerdahl and Jackendoff, 1983) on the one hand, and timing and dynamics on the other, in a manner suggested by empirical observations. Shaffer and Todd (1987; also Todd, 1985; Repp, 1990, 1992) have observed a tendency for performers to increase tempo toward the middle of phrases and decrease tempo toward the endpoints. Tempi seem to reach a minimum at phrase boundaries and a maximum between phrase boundaries. It has also been suggested (Todd, 1992) that dynamic levels seem to be in direct proportion to tempo: players tend to play quietest at phrase boundaries and loudest between them (Gabrielsson, 1987).

In his model, Todd (1992) proposes that a single hierarchical grouping structure should directly specify the moment-to-moment durations and dynamics of each event in such a way as to mimic the kind of behavior described earlier. At each level of the hierarchy, dynamics and tempo increase to a maximum mid phrase and decrease to a minimum at phrase end. An additive function between levels means that the fastest and loudest points in the "performance" will be in the middle of the phrase at the highest level of the hierarchy. Between maximum and minimum values, the model plots intermediate events along a curve. Todd justifies the direct link between tempo and dynamics by analogy with a physical system (such as a hammer/string interaction) where tempo change becomes acceleration, dynamics become energy (intensity), and intensity is proportional to the square of velocity (see Todd, 1992, for a fuller description of the function underlying these curves). His justification for assuming that tempo change is analogous to acceleration comes from empirical study of accelerandi and ritardandi in actual performances (Kronman & Sundberg, 1987) and the

qualitative assessment of different synthetic tempo functions by Longuet-Higgins and Lisle (1989). Todd has also claimed that this direct link between tempo and dynamics via physical motion may have a physiological basis (Todd, 1992, p. 3549). Because of the model's insistence that timing and dynamics share a single mathematical function, the interonset intervals between events and their intensity will always be positively correlated (instantaneous tempo, of course, will be negatively correlated with intensity).

The user of the algorithm as implemented here, in addition to choosing an input structure, which corresponds to a hierarchical grouping structure (*á la* Lerdahl & Jackendoff, 1983), also defines the maximum and minimum duration of the underlying pulse in the performance (the upper and lower bounds of its tempo) and the maximum and minimum dynamic levels and can choose to weight the proportion of "expression" that is assigned to each level of the phrase structure: the "level weights." Such weighting does not affect the overall changes in tempo and dynamics, a proportional decrease occurring at the other levels. For example, in a grouping structure hierarchy with only two levels, the subordinate, or lower, level of the hierarchy might be assigned a weighting of 2, the superordinate, or higher with a weighting of 1: in this case, the higher level would relate to the lower in a ratio of 1:2. The level weightings are assumed to be identical for both timing and dynamics, in order to maintain the simple relationship between expression and structure. Moreover, the distribution of expression between groups, regardless of level weightings, is such that the expression assigned to groups at the same level of the phrase hierarchy will maintain equality. The output of the model takes the form of a text stream of onset times in milliseconds and Musical Instrument Digital Interface (MIDI) key-velocity values. The model will be returned to in more detail after the rationale for the present study is outlined and data from the skilled performance with which the model output is to be compared are presented.

Rationale

This study has three specific aims. First, it aims to assess the theoretical assumptions about timing and dynamics that the model (Todd, 1992) instantiates. These assumptions can be briefly summarized as follows:

1. Performers play slowly and quietly at the beginnings and ends of phrases and increase tempo and dynamics toward the middle of a phrase.
2. Performers' expressive timing and dynamics are linked to the hierarchical nature of phrase structure such that the slowest and

quietest points in a piece will be the events that begin and end phrases at the highest level in the hierarchy.
3. These changes in expressive dynamics and tempo are generated by the same mechanism and follow the same patterns of increase and decrease.

To achieve this goal, the output of the model for a simple and relatively unambiguous phrase structure will be compared with the timing and dynamics of a human performance of the same structure.

Second, we aim to gain insight into which levels of a grouping structure might be assigned most importance by a human performer. Because the algorithm allows for adjustment of the weighting of different levels of the phrase hierarchy's contributions to expression, attempts to match the algorithmic output with a human performance afford an opportunity to ascertain which levels of a phrase hierarchy are given most expressive weight.

It is not our intention in this paper, however, merely to examine the success of the algorithm in predicting the dynamics and timing of a human performance. As a model of performance, it already instantiates theoretical principles well supported by empirical data. Hence, some degree of correspondence between the algorithm's output and the timing and dynamics of a human performance might be expected. This study's third aim is to use comparison of the algorithm with human performance as a method for highlighting those aspects of expressive dynamics and timing which the algorithm *fails* to model. The model represents only the relationship between phrase structure and expression, no other aspect of musical structure (metrical, harmonic, voice-leading, motivic) that might generate patterns of timing and dynamics is explicitly modeled, and it leaves out dynamic differentiation between voices in the musical texture and *discrete*, rather than continuous, dynamic and timing changes, such as durational or dynamic accents. Moreover, the model assumes that the links between structure and expression, being rather directly specified, are not open to conscious or unconscious suppression or exaggeration, hence such strategies may also become highlighted by the model's regularity.

It should also be added here that a more basic theoretical question can be addressed by directly comparing single human performances with those produced by an algorithm based on such a simplistic, yet empirically supported, theory of expression. If two performances, one human and one algorithmic, are generated from a relatively unambiguous representation of musical structure and are generated by the same generalized rules for mapping structure onto expressive timing and dynamics, then differences between these performances may tell us something about more idiosyncratic aspects of expressive timing and dynamics, aspects that may be based on musical structure, yet far less directly specified by a system of rules.

Moreover, it is admitted that the sources of expression may be found *outside* the score's abstract musical structure, in the ideologies of performance practice, in attempts to convey emotional or conceptual content, in the desire to challenge accepted norms of performance, or on a more mundane level, in attempts to deal with particular performance circumstances, such as a coping with a particularly reverberant performance space, or unfamiliar piano (see, e.g., Clarke, 1988). The algorithm used here cannot directly know anything about these aspects of music making, and by dint of this omission may provide evidence of their contributions to expressive dynamics and timing.

Using an algorithmic performance as a baseline in this way, to identify individual, rather than commonplace, expressive strategies, can be seen as complementary to the methods developed by Repp (1995), who uses an averaged performance as a similar baseline. Our approach is in a sense less sensitive to idiosyncrasies per se, because they might become confounded with other features that the model fails to capture. However, the approach taken here does enable one to quantify the extent to which a *single* performance differs from a baseline, without the necessity of collecting multiple performances. The model, based as it is on multiple observations, already represents a kind of average performance.

Finally, it should be noted that this study is in no way an exhaustive exploration of the model's success in capturing aspects of human musical performance. For this to occur, either a substantial number of different human performances would have to be compared with the model, or data would have to be gathered about the model's aesthetic and/or communicative evaluation by listeners. Preliminary findings from the second of these alternatives were the subject of a separate perceptual study, which demonstrated that although the algorithm may communicate alternative grouping structures for the same melody extremely well, sometimes better than skilled performers, such communicative performances are not necessarily the most aesthetically pleasing (Clarke & Windsor, 1996).

The Human Performances

The timing and dynamic data for this study were collected from a skilled professional piano soloist henceforth known as HP. HP is regarded as a specialist in the music of Beethoven and the classical and romantic repertory in general and both performs and records regularly. HP was twice asked to play the first 16 bars of Schubert's G♭-major Impromptu, opus 90 (see Appendix for the score of this extract), as if the last of these bars was the "end of the piece." He was then asked to play the piece again, but in a "restrained" manner. This latter performance was collected not as a sup-

posedly expressionless performance, but rather as a contrasting performance that might be slightly "blander" than the two unconstrained performances. HP was provided with the score a short while before the requested performances, and he reported that he knew the piece quite well, but had not performed the piece in public for some time. The only other explicit instruction provided was that dynamics and timing were the parameters we were most interested in. HP reported no difficulty with these conditions and seemed at ease with the performance conditions. The performances occurred in a practice room, played on a Yamaha Disklavier upright piano connected to a Macintosh IIfx computer via MIDI. The performances were recorded using Opcode's Vision sequencer package and stored on hard disk as standard MIDI files. After his performances had been recorded, HP was asked to comment on the phrase structure of the music played with reference to the score: the analysis offered corresponded well to that which would later form the input structure for the algorithm (shown in Figure 1). Of the two "spontaneous performances," HP suggested that the second was a better attempt.

The standard MIDI files thus obtained were converted into text files listing interonset intervals between each triplet-quaver onset and the MIDI key velocities of each of these events as a decimal fraction where 1 = maximum key velocity (127 in MIDI) using the POCO environment designed by Desain and Honing (see Honing, 1990). Where alternative onsets were available at the same score position, the highest note in the chord was taken. Regression analyses showed typical consistency across performances (see Clarke, 1982; Clynes and Walker, 1982; Shaffer, 1981; Shaffer and Todd, 1987): regressing the interonset intervals of the two spontaneous performances gave $R^2 = .848$, $df = 381$, $p < .0001$; regressing the first spontaneous performances against the restrained performance gave $R^2 = .626$, $df = 381$, $p < .0001$. Analyzing the same pairs of the performances but taking key velocities as the variables gave $R^2 = .719$, $df = 381$, $p < .0001$, and $R^2 = .736$, $df = 382$, $p < .0001$ (the corresponding correlations were all positive and significant at $p < .0001$). As might be expected, the two spontaneous performances are more highly correlated with each other than with the restrained performance in terms of timing, demonstrating that HP was able to produce consistent performances of the extract, but also to change his interpretation according to differing instructions. It should also be noted that, in agreement with Repp (1995, 1996), the dynamic regressions are slightly lower than those for timing, suggesting that timing may be more reliably preserved over multiple performances of the same piece.

Because the two spontaneous performances are so highly correlated, it was decided that this was sufficient justification for focusing the subsequent matching with the algorithm output on the second spontaneous performance, HP2. The first spontaneous performance and the restrained per-

formance allow differing criteria of acceptability for an algorithm performance to be derived. Following Todd (1992), this criterion is simply that an algorithmic performance should, when regressed against the real performance, account for as much variance as when a repeat performance from the same player is regressed against this real performance.

The Algorithmic Performances

The algorithmic performances were produced by using a LISP implementation of the Todd (1992) model of musical expression running on a Macintosh IIfx computer. The maximum and minimum dynamic levels and durations of the triplet quaver pulse of the structure were .6 and .3, 200 and 100 milliseconds. These values were chosen *before* data were collected from HP, as was the input structure shown in Figure 1. The 16 bars of the extract were hierarchically divided into groups down to the half-bar level. Each group was subdivided into two equal-sized subordinate groups. This grouping could be regarded as overly regular, but had the advantage of producing a simple pattern of expression with no overlapping phrase boundaries. Similarly, the choice of a piece with an isochronous base level of rhythmic activity gives a simple and clear representation of instantaneous tempo if one plots interonset times against score position in triplet quavers. An example of the output is shown in Figure 2. This example was obtained with neutral level weightings: henceforth this algorithmic performance will be referred to as AP (1, 1, 1, 1, 1) with each numerical value between the parentheses representing in decreasing order the weighting assigned to each

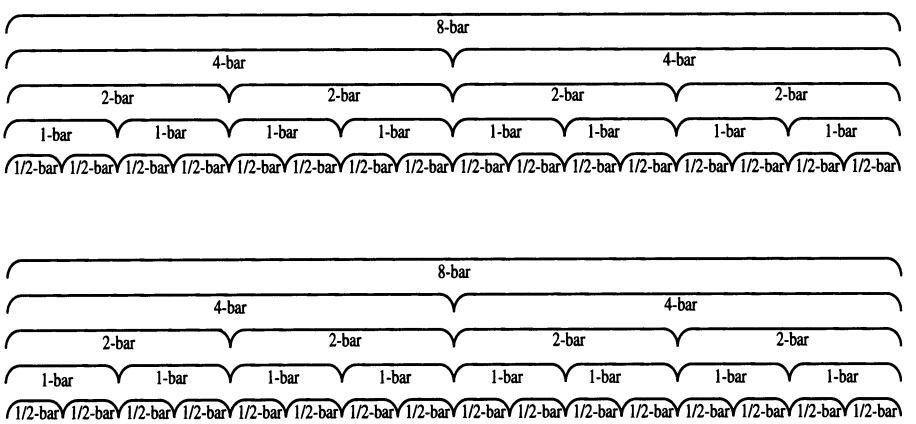

Fig. 1. A schematic illustrating the hierarchical structural representation that the model used as an input structure.

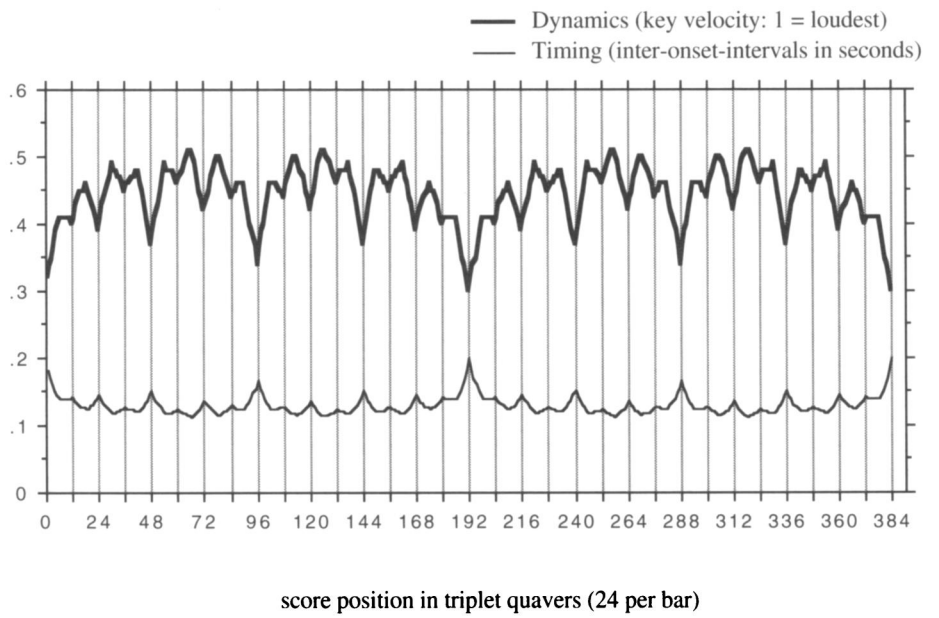

Fig. 2. A line plot showing interonset times in seconds and key-velocity values as a decimal fraction against score position in triplet quavers for the algorithm with all level weightings set to 1.

level of the structure from highest to lowest, and AP denoting an algorithmic performance. The basic form of the model output is clear: the duration of the triplet quavers increases toward phrase boundaries and decreases toward the middle of each phrase. In other words, tempo increases and reaches a maximum mid phrase. The direct links between tempo and dynamics and the links between these parameters and the input structure are plain.

Comparisons of Algorithm Output with the Human Performance

In the first instance, six algorithmic performances were generated from the same input structure and with the same range settings for dynamics and timing. Five of these algorithmic performances each assigned a double weighting to one level of the input structure, the sixth being the AP (1, 1, 1, 1, 1) performance just illustrated. In this way, it was possible to assess whether accentuating the contribution of any particular level in the phrase structure to the expressive timing and dynamics obtained a better match with the human performance HP2. The complete set of level weightings were therefore as follows: (1, 1, 1, 1, 1), (1, 1, 1, 1, 2), (1, 1, 1, 2, 1), (1, 1,

2, 1, 1), (1, 2, 1, 1, 1) and (2, 1, 1, 1, 1). Figure 3 shows the six versions of the output, illustrating the effect of adjusting the level weightings in this systematic fashion. Twelve separate regression analyses were performed, between the key velocities (six regressions) and interonset intervals (six regressions) of each algorithm version and HP2. All were positively correlated and significant at the .0001 significance level, but none approached the variance accounted for by regressing HP2 against the other human per-

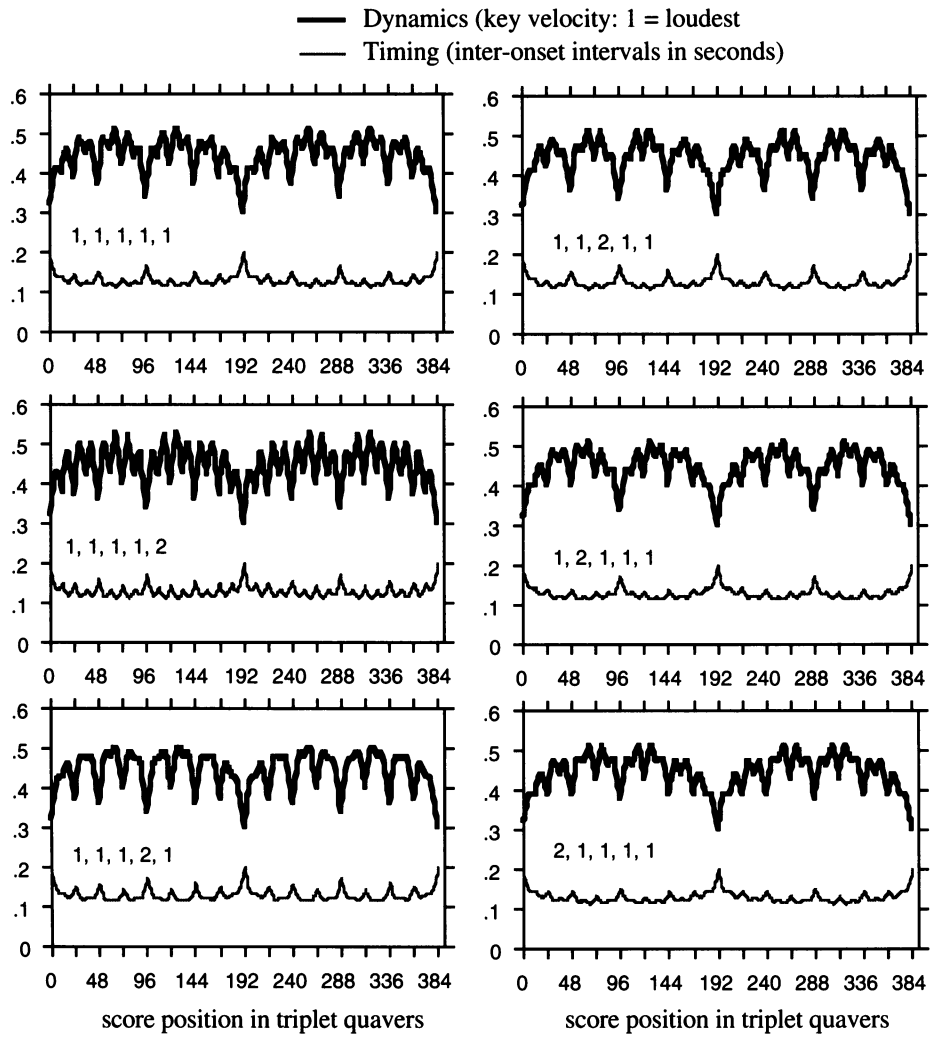

Fig. 3. A line plot showing interonset times in seconds and key-velocity values as a decimal fraction against score position in triplet quavers for the algorithm at six different level weightings. The level weightings are positioned as follows: on the left side from top to bottom (1, 1, 1, 1, 1), (1, 1, 1, 1, 2), and (1, 1, 1, 2, 1); on the right (1, 1, 2, 1, 1), (1, 2, 1, 1, 1), and (2, 1, 1, 1, 1).

TABLE 1
Summary of Regression Analyses

	Value of R^2	
Algorithmic Performance	**Timing**	**Dynamics**
AP (1, 1, 1, 1, 1)	.299	.086
AP (1, 1, 1, 1, 2)	.375	.054
AP (1, 1, 1, 2, 1)	.321	.041*
AP (1, 1, 2, 1, 1)	.235	.126
AP (1, 2, 1, 1, 1)	.258	.107
AP (2, 1, 1, 1, 1)	.217	.089

*Despite the low values of R^2, these regressions are all significant at the .0001 level: e.g. $R^2 = .041$, $df = 382$, $p = .000006701$.

formances. Table 1 shows the values of R^2 for each of these algorithmic performances regressed against HP2. Note that different algorithmic performances account for the most variance depending on whether one regresses timing or dynamics: AP (1, 1, 1, 1, 2) in the case of the timing data, and AP (1, 1, 2, 1, 1) in the case of dynamics.

This suggests that different level weights may account for the variance in the human data for timing and dynamics, a result not predicted by the Todd model, which explicitly links these parameters. However, because the amount of variance accounted for by any of these algorithmic performances is quite low (although significant), an attempt was made to produce two separate algorithmic performances that achieved a better fit with the timing and dynamics of the original performance respectively. Some intuition was necessary here, and it was decided to produce a number of algorithmic performances that weighted the middle and two higher levels of the phrase structure in the case of the "dynamics" performance and the two lowest levels of the phrase structure in the case of the "timing" performance. This decision was motivated by the general pattern of regression values shown in Table 1: accentuated levels that seemed to produce better accounts were further doubled until the variance accounted for ceased to increase. In the case of timing, for example (1, 1, 1, 2, 4) and (1, 1, 1, 4, 8), might be possible candidates because (1, 1, 1, 1, 2) and (1, 1, 1, 2, 1) produced good matches.

The highest value of R^2 obtained for timing from this process was .426 ($p = .0001$); for AP (1, 1, 1, 2, 4), the highest for dynamics was .201 ($p = .0001$): AP (4, 8, 8, 1, 1). Regressing the dynamic data from AP (1, 1, 1, 2, 4) against the dynamic data of HP2 gave $R^2 = .015$, the significance level falling to .018, and regressing the timing data from AP (4, 8, 8, 1, 1) against the timing data of HP2 gave $R^2 = .109$ ($p = .0001$). On this basis, although the variance accounted for is still not as great as that of a repeat perfor-

mance, it is clear that a greater proportion of the variance can be accounted for by manipulating the level weightings for dynamics and timing separately. Figure 4 shows the timing data from HP2 and AP (1, 1, 1, 2, 4), and

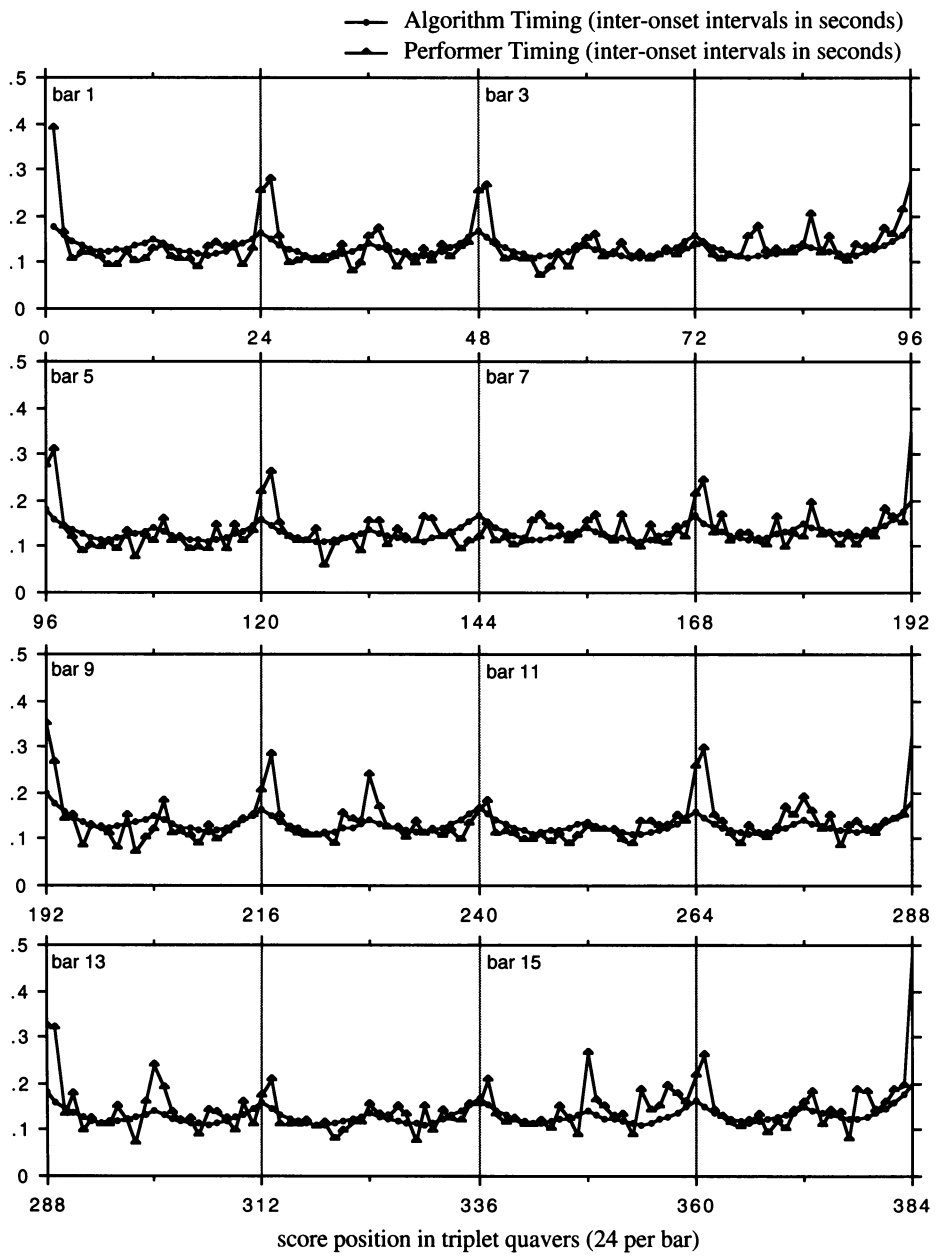

Fig. 4. A line plot showing interonset intervals in seconds for AP (1, 1, 1, 2, 4) and HP2 against score position in triplet quavers.

Figure 5 shows HP2's dynamic data and those of AP (4, 8, 8, 1, 1). Figure 6 shows what might be considered a hybrid performance, combining the timing data from AP (1, 1, 1, 2, 4) with the dynamic data from AP (4, 8, 8,

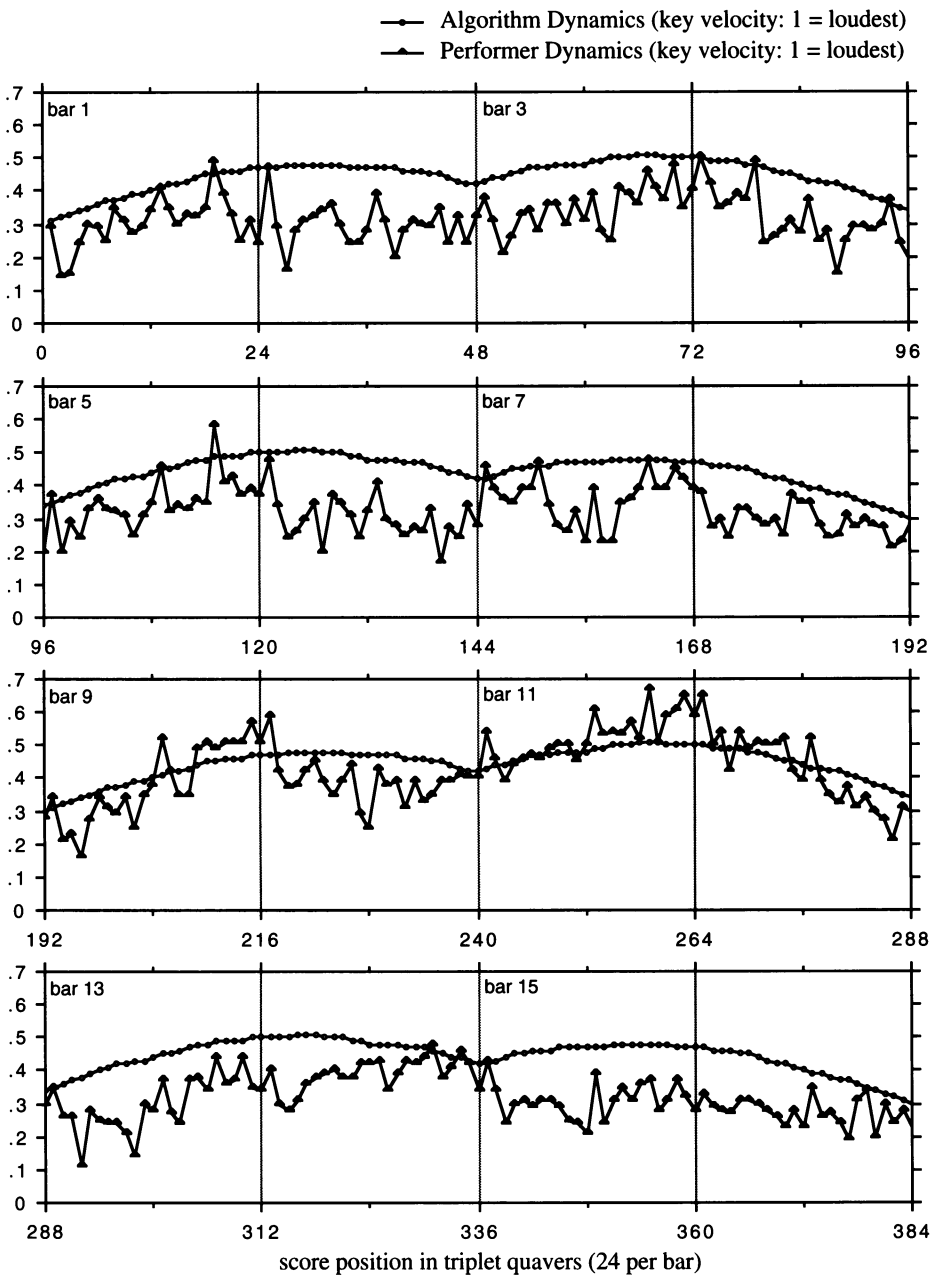

Fig. 5. A line plot showing key-velocity values as a decimal fraction for AP (4, 8, 8, 1, 1) and HP2 against score position in triplet quavers.

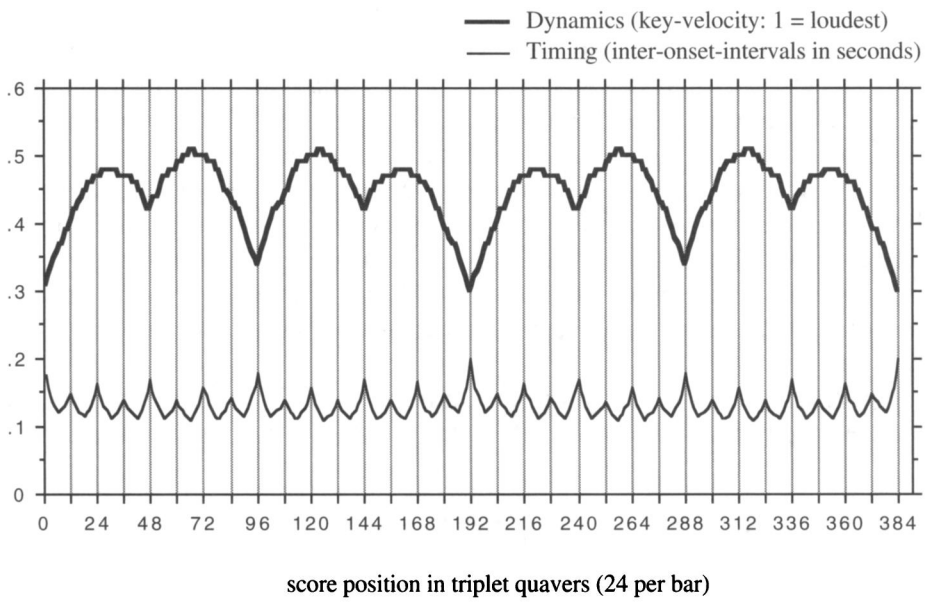

Fig. 6. A line plot showing interonset intervals in seconds for AP (1, 1, 1, 2, 4) and key-velocity values as a decimal fraction for AP (4, 8, 8, 1, 1) against score position in triplet quavers. These two profiles together constitute the "hybrid" performance.

1, 1). This hybrid performance assigns more emphasis to lower levels of structure for timing and higher levels for dynamics. Moreover, the high correlation between this hybrid with the human performance suggests that the human performer may be similarly stressing the lower levels of structure with timing, the higher with dynamics.

Visual comparison of the timing profiles in Figure 4 suggests that although the algorithm successfully predicts the more global structure of tempo variation at bar and half-bar levels, it differs from the human performance in two main respects. First, whereas the algorithm changes tempo according to a smooth curve, the human performer tends to concentrate tempo variations into large deviations at or around phrase boundaries. This must lead one to question whether the performer is (1) simply delaying the onsets of initial or terminal events in a phrase or (2) accentuating the initial notes of each phrase, which are always melody notes. This will be returned to later, in the general discussion. Second, the timing profile produced by the algorithm is far too regular in its structure to be mistaken for the human performance. This is not necessarily because the human performer is unsystematic, because it is impossible to determine whether this is the case or whether a number of interacting yet systematic processes are at work here. Again, such questions will be returned to in more detail later.

Looking at the dynamic data in Figure 5 reveals much larger discrepancies. The overall changes in the dynamic levels of the human performance

do seem to be well modeled, but local detail is simply not captured at all, partially due to the particular level weighting used, which emphasize large-scale structure. Again it is possible that the dynamics are being used not simply as a continuously varying expression of phrase structure but also as local intensity-accented markers for individual events, or smaller groups.

An obvious way to analyze these discrepancies between the model and the human performer is to plot the residuals from the regression analyses between the optimally matched human and algorithmic performances for both dynamics and interonset intervals. These residuals can be thought of as representing the variance in the human performance not accounted for by the algorithm.

Figure 7 shows residuals plotted against score position for the regression between the timing data from AP (1, 1, 1, 2, 4) (the timing weightings for

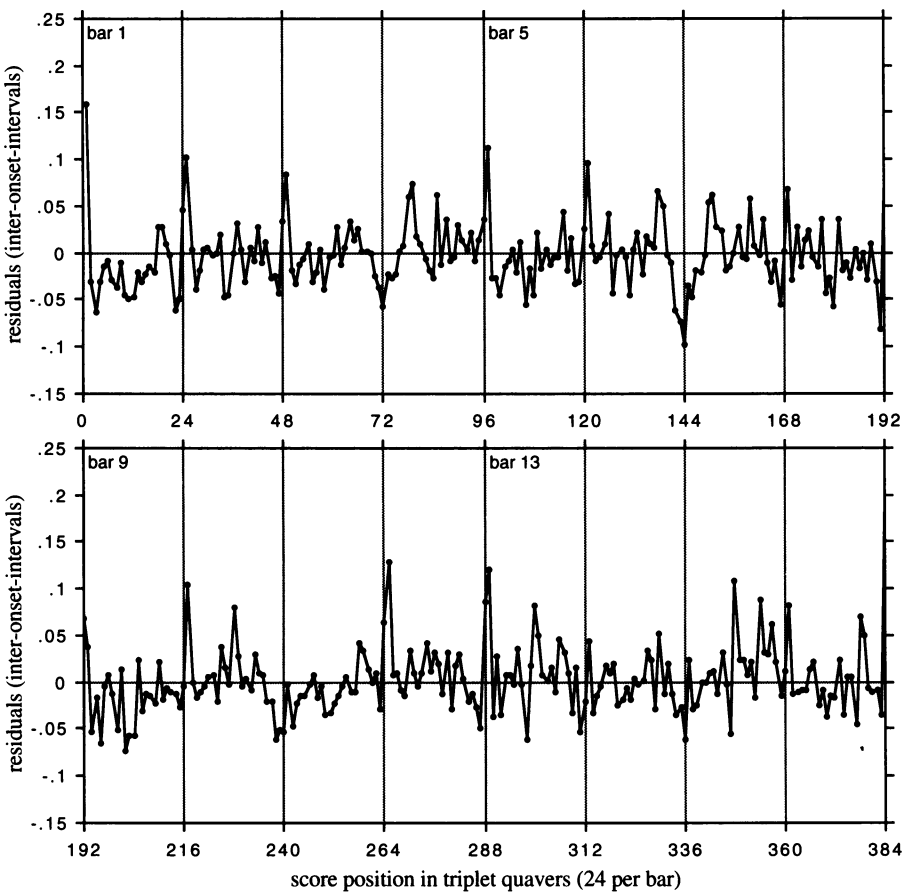

Fig. 7. A line plot showing the residuals obtained when the interonset intervals in seconds for AP (1, 1, 1, 2, 4) are regressed against those for HP2, against score position in triplet quavers.

the hybrid performance) and HP2. No large scale periodicity seems apparent except for a tendency for some melody notes (especially those that occur at the beginning of each bar and half bar) to be played longer than the model would predict. This could be accounted for by interpreting such lengthening as accentuation of the initial notes in phrases, accentuation of the melodic line as opposed to the rippling accompaniment, or a more abrupt tempo function.

Figure 8 shows residuals plotted against score position for the regression between the dynamic data from AP (4, 8, 8, 1, 1) (the dynamic weighting for the hybrid) and HP2. Here, if there is an analogous accentuation of the melody notes, it is concealed among a much greater degree of local deviation. The residuals for dynamics do suggest that the melodic climax of the

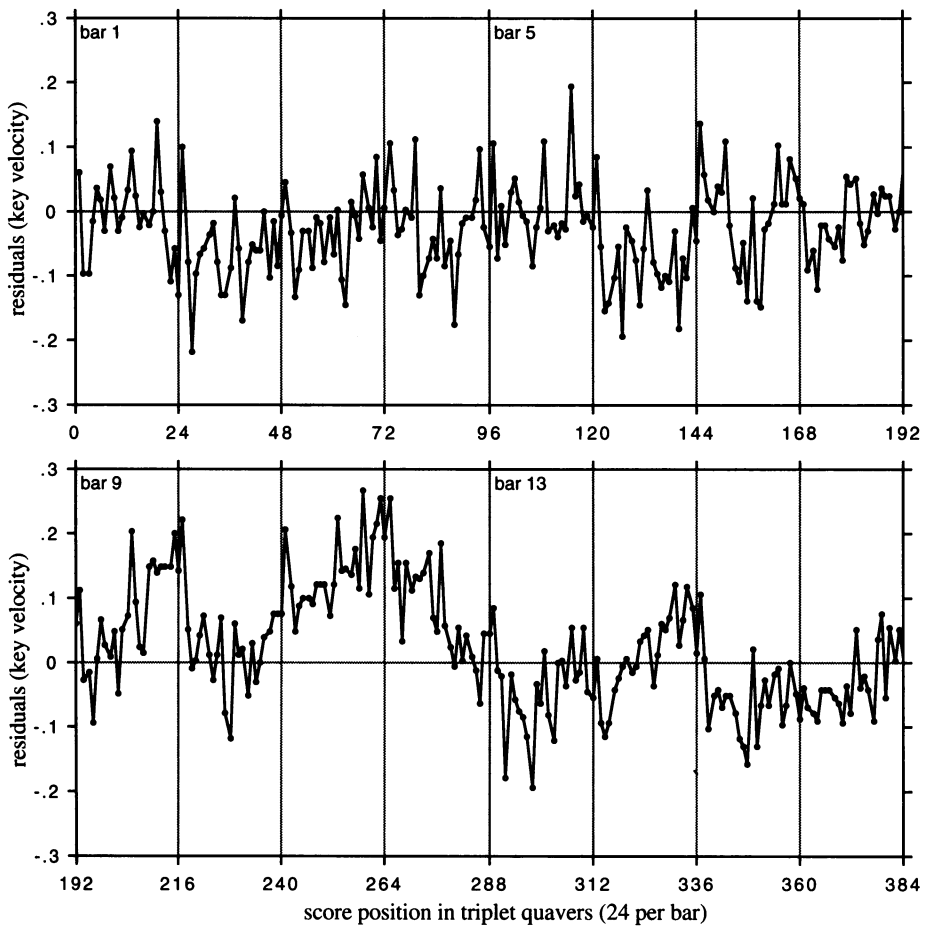

Fig. 8. A line plot showing the residuals obtained when the key-velocity values as a decimal fraction for AP (4, 8, 8, 1, 1) are regressed against those for HP2, against score position in triplet quavers.

extract (between bars 9 and 12, events 217–312), may be being given a more pronounced dynamic "curve" than the equivalent preceding four bars, something the model does not predict because of its equal distribution of dynamic variation between groups at the same hierarchic level. Indeed this is an effect that is explicitly notated by Schubert in the score (see the dynamic markings in the score for bars 9–12).

Figure 9 plots both sets of residuals on the same chart. The model predicts that interonset intervals and dynamics should be negatively correlated (that short events should be loud). Figure 9 demonstrates that the

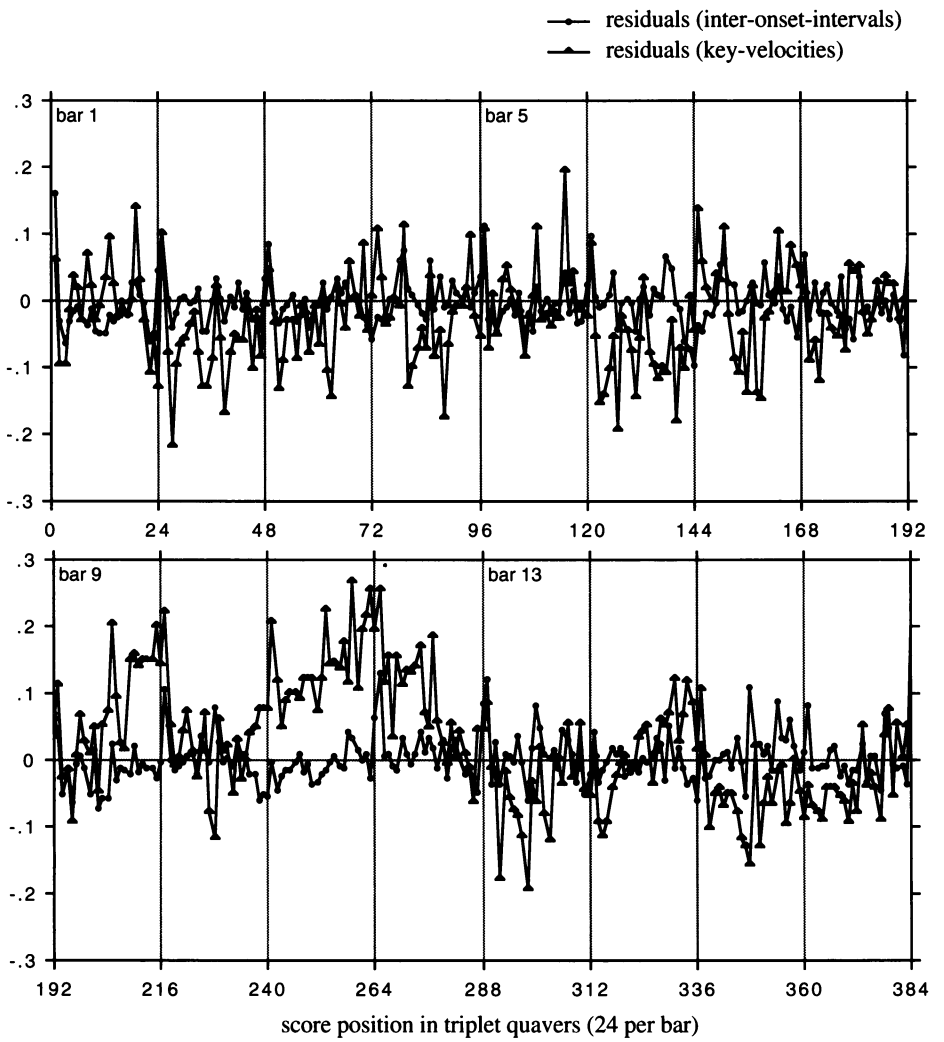

Fig. 9. A line plot showing the residuals obtained when the key-velocity values as a decimal fraction for AP (4, 8, 8, 1, 1) are regressed against those for HP2, and those obtained when the interonset intervals in seconds for AP (1, 1, 1, 2, 4) are regressed against those for HP2, against score position in triplet quavers.

residual profiles are similar and *positively* correlated (R^2 = .03577, p = .0002). If the performer is using timing and dynamic variation both as continuously varying parameters in the manner suggested by the model, but also as methods of accentuating events through intensity or length, then one might expect some events to be *both* longer and louder than others, in contradiction to the model, and as highlighted by the residual analysis. Figure 10 illustrates the residuals for the first two bars and the first event of the third, events 1–49.

Observe, for example, the high residual values of events 1, 25, and 49 for both the dynamic and timing data. These events can be interpreted as being accented because of their metrical position (first beat in bar), their being part of the melody, or their initiation of phrase units. Moreover, it is not just these particular events that are played louder and longer than the model predicts: note, for example, the way in which the surrounding events are played shorter and quieter than the model would predict, and how they increase in duration and key velocity as their score distance from the accented events decreases. Note also that the residuals of the dynamic data switch from being generally positive in bar 1 to being generally negative in bar 2, suggesting that the model overestimates the increase in dynamics that might occur in bar 2 relative to bar 1 because of its relative proximity to the center of two larger phrase units. Note also that the dynamic residu-

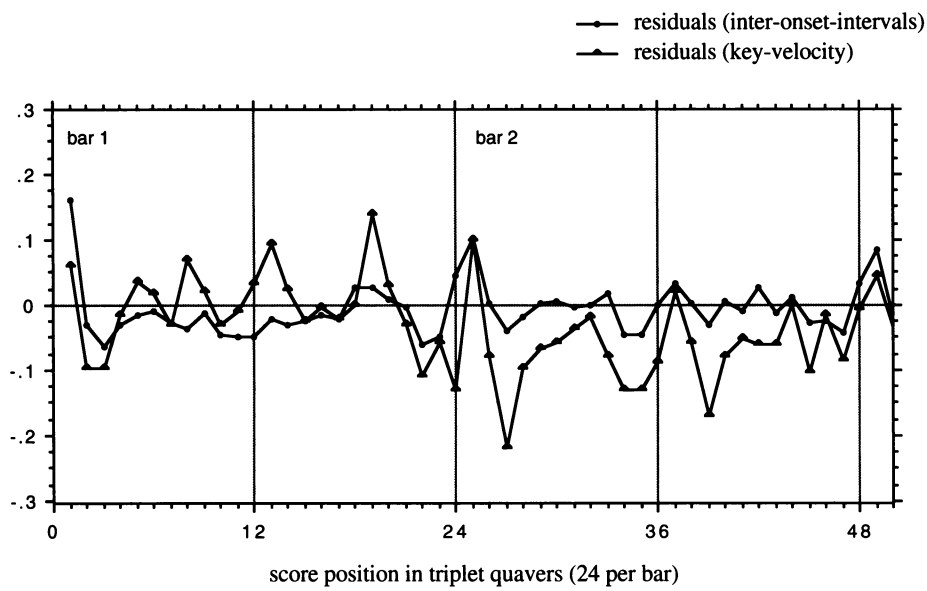

Fig. 10. A line plot showing the residuals obtained when the key-velocity values as a decimal fraction for AP (4, 8, 8, 1, 1) are regressed against those for HP2, and those obtained when the interonset times in seconds for AP (1, 1, 1, 2, 4) are regressed against those for HP2, against score position in triplet quavers for the first 50 onsets.

als show evidence of accentuation where the timing residuals do not, for example at event 20, another melody note. Hence, although the residuals suggest that timing and dynamics may have an additional, positively correlated factor not modeled by the Todd algorithm, this suggests that a performer may accent with lengthening or increased dynamic individually, or in combination—not that timing and dynamics are always positively correlated at this level of analysis. Nor do these residuals, or the analysis as a whole, suggest that the model is incorrect: rather they suggest that the model is largely accurate in respect of one kind of expressive strategy, based on continuous modulation of tempo and dynamics, but that it fails to model others: that of the accentuation of individual events by dynamic stress or lengthening, and the use of overall dynamic levels to emphasize particular sections in the music, such as the climax in bars 9–11.

General Discussion

The first theoretical issue raised by this study is that of using single or multiple rules to model performance. What are the consequences of the single rule approach taken here for the study of expression in general, and in particular, is the notion that dynamics and timing are generated by the same rule supported by the data gathered and generated here? In answer to the first question, it is argued that attempting to stretch one rule to its limits, rather than attempting to assess the impact of multiple rules (e.g., Sundberg, 1988), has advantages. Because the model used here is so single-minded in its avoidance of anything but phrase structure and continuous expression, it highlights other aspects of expression and allows their detailed analysis. Whereas the comparison of a human performance with an isochronous "score" may tell us little about the different strategies of expression used by a musician and may conceal them if one strategy accounts for most of the variance, the use of a model such as this allows one to distinguish between different aspects of expression. For example, through analysis of residual data, continuous variation (such as rubato or patterns of crescendo/decrescendo) may be stripped away in a systematic manner to allow the analysis of discrete variation or different kinds of continuous variation. The failure of the model to account for the human data as successfully as might be expected from Todd's own study (Todd, 1992) requires some comment here. Although the regression analyses reported are highly significant, the low values of R^2 do not suggest the same high degree of fit as achieved in this earlier matching study. One might, therefore, question the validity of using the model to "strip away" expressive deviations. However, it is not surprising that our data are less well accounted for: the piece used here, although similar to the Chopin prelude used in Todd's own

study in that it contains continuous isochronous patterns, is far less rhythmically undifferentiated, and hence might be expected to elicit a less uniform approach to interpretation, calling for a wider variety of expressive strategies. Indeed, it would be extremely surprising if the model accounted for all performances of all music equally well, considering the potential for stylistic and individual variability. However, this does not invalidate the use of the model as an analytical tool: despite the low values of R^2, the model is clearly capturing some underlying aspects of tempo and dynamic variation.

The links between timing and dynamics proposed in the model have also been tested here. Although this is only a single case study, it is clear even from this that timing and dynamics are not always linked by a single function. It is possible that timing and dynamics are linked by a number of different functions and are sometimes not linked at all: this last possibility is supported by the relative success of the hybrid performance and the analysis of the residuals produced by regressing real against algorithmic performances. Timing and dynamics may be (1) positively correlated where both are used to accent discrete events, (2) negatively correlated as suggested by the model, or (3) *either* timing *or* dynamics can be used to signal an accent or a phrase shape, but not both.

Just as the data analyzed here raise questions about the number of rules needed to model a human performance, they raise the possibility that human performances may derive from multiple representations of the musical score. If we assume that the human performer is responding to higher-level phrases with dynamic modulation and lower level phrases with timing modulation, this might suggest one mental representation and two separate processes controlling the rate of modulation, or two representations, one capturing high-level structure and the other low-level structure. Clearly, if timing and dynamics are generated by a single physiological system (Todd, 1992), then the current results can hardly be considered corroborative, although it is possible that more than one system of the same kind might be responsible. It is, of course, possible that performers avoid a direct link between timing and dynamics precisely because of its physiological links with the perception of self-motion, if such links exist, since emphasizing them might produce too unified a sensation of motion. Deviating from such simple patterns may in fact be one way in which performers signal structural features or express their individuality. Moreover, what kind of structure is responsible for the local accentuation of events highlighted in the analysis of residuals? It would be inaccurate to claim that an addition of a rule to handle metrical structure alone would necessarily account for the residual expression, as additional rules might also have to account for voicing, the possibility that phrase boundaries are marked by an accent, or any other kind of local detail chosen for emphasis by the performer. It

might also have to account for a decision by the performer not to use a particular kind of accent, or to fail to accent an event that might normally seem to require emphasis.

A number of more specific aspects of expression have also been highlighted here:

1. Length accentuation and ritardandi: The analyses have shown that it may be possible to distinguish between lengthening a note for local emphasis and lengthening a note as part of a tempo function.
2. Final lengthening and tempo "curves": The data support a further distinction between lengthening the final interonset interval of a phrase in order to delay the initial onset of the following phrase, thus forming a "micropause" between phrases (Clynes, 1983) and more continuous tempo shaping across a number of interonset intervals (see Clarke, 1988).
3. Dynamic accents and crescendi: The analysis of the dynamic data suggests a similar distinction between continuous and local use of dynamics: a louder onset can occur in a performance because it is in the middle of a phrase (at the apex of a crescendo) or at its beginning (as a loudness accent). Moreover, one hypothetical explanation for the possibility that higher levels of the phrase structure were emphasized by continuous dynamic change and lower levels by timing is that at lower levels crescendi and decrescendi are confusable with local accents, whereas tempo changes are not, although why this should be so remains unclear.
4. Equality across groups: Analysis of the residual data suggests that, unlike the model, the human performer might choose to play particular groups louder than others, regardless of the fact that these occur at the same level in the phrase hierarchy and in the same position relative to the highest level in the phrase structure. In the present case, this can be attributed to an explicitly notated aspect of the score, but it raises the possibility that performers may play sections louder overall for emphasis, perhaps responding to a melodic idiosyncrasy or some aspect of motivic (hence associative rather than hierarchical) structure.
5. Systematic versus unsystematic aspects of expression: Clearly the algorithm output *appears* more systematic than the data from the human performances in this study. Part of this is undoubtedly due to the simple structure of the algorithm, which cannot (and does not) claim to be a complete model of expression. However, this raises a difficult question: how complex should our explanations become before we admit that certain expressive

aspects of a human performance may be unsystematic or too idiosyncratic to model with general rules?

Before concluding it should be noted that the algorithm can be used to generate sounding algorithmic performances via MIDI. The assessment of such performances is the subject of another paper (see p. 13 of this article; Clarke & Windsor, 1996), so no attempt is made here to systematically and rigorously assess the sound of either the human performer or the algorithm. However, because a number of such algorithmic performances have been produced, some qualitative impressions would not be out of place. First, in order to create an artificial performance that sounds in any way convincing, sustain pedal has to be added in order to elide the triplet semiquavers. Second, adding a constant to the key velocities of the melody notes to bring them forward in the musical texture (see Repp, 1996, for empirical support for this move) seemed appropriate. Both of these steps can be considered legitimate because the model explicitly avoids such issues; these steps ensure, rather, that the successes of the model are not obscured by other variables. The main impression of the algorithm output, edited in this way, is that all the algorithmic performances that directly link timing and dynamics as envisaged by the model in its original form sound distinctly odd, almost unsettling, whereas dislocating them (as in the hybrid performance) provides a much more acceptable result. To test this hypothesis, nine postgraduate music students at the University of Sheffield were played the hybrid, and the two algorithmic performances on which it was based (AP [1, 1, 1, 2, 4] and [4, 8, 8, 1, 1]) in a random order and asked to indicate which performance they preferred (they were not told in advance how the performances had been arrived at). The results were striking: all nine subjects preferred the hybrid ($\chi^2 = 18$, $p < .0001$).

Conclusions

In this paper, we have demonstrated the way in which a model of performance may act as a tool in the analysis of performance data. Although the model fails to account for every aspect of a human performance, and could possibly be revised in the light of the data collected here, these failures are seen as positive because they highlight different aspects of musical expression. The model provides a baseline that is derived from a strong theoretical position, against which other expressive strategies can be assessed. In its clear and unambiguous modeling of continuous expressive strategies, the model allows one to factor out noncontinuous strategies in a manner not possible when a performance is analyzed in relation to an isochronous score.

Clearly, however, this study presents only a first attempt at this kind of analysis, as an example, rather than a conclusive piece of research. It uses only a small sample of possible model outputs, it deals with data from only one performer, and music of one particular style. All of these limitations must be addressed before too many generalizations are drawn. Within such limitations, however, the strength of the approach and the kinds of insights available have been shown.

One obvious limitation of this research, and indeed of all research of this kind, is the limited interest paid to information outside the score, whether seen as musical knowledge or information available from the performer's surroundings. These limitations are accepted, and it is hoped that research such as this may highlight those aspects of expression that demand explanations that cannot be derived from a purely score-based, generative approach to expression (see Clarke, 1988). For example, in this study, we have repeatedly returned to the possibility that different expressive strategies might be chosen to emphasize the same structural features (e.g., accenting an event on a phrase boundary rather than using continuous change in duration, or choosing to use timing and dynamics to emphasize different levels of phrase structure). The flexibility of expressive strategies and the interchange between them is not well accounted for by existing models of expression, and this study, like that of Drake and Palmer (1993), emphasizes the complexity and richness of musical expression as a mode of human behavior. This, point, however, should not obscure the value of discovering such richness and complexity by systematic means. Hence we see no conflict between approaches that start from a reductionist standpoint (such as Todd, 1992) and those that build in some notion of flexibility or diversity from the outset (such as Desain & Honing, 1991). It is surely heartening that the empirical study of musical expression provides a domain that constantly resists simple and systematic explanations and hence extends our understanding of human behavior beyond limits imposed by less challenging and complex tasks.[1]

References

Clarke, E. F. (1982). Timing in the performance of Erik Satie's 'Vexations.' *Acta Psychologica*, *50*, 1–19.

1. This study was funded under the Small Grants to the Social Sciences Scheme of the Nuffield Foundation. Thanks are due to our skilled pianist, who was more than accommodating and provided us with such excellent performances. Acknowledgments are also due to Neil Todd, without whom this work would have been impossible, and to Bruno Repp, Carolyn Drake, Greg Sandell, Peter Desain, Henkjan Honing, Alan Wing, Piet Vos, and Dirk Povel for their comments on earlier versions of this paper.

Clarke, E. F. (1988). Generative principles in music performance. In J. A. Sloboda (Ed.), *Generative processes in music: The psychology of performance, improvisation, and composition.*. Oxford: Clarendon Press.

Clarke, E. F., & Windsor, W. L. (1996). Timing, dynamics and structure in human and algorithmic performances. In *Proceedings of the 4th International Conference on Music Perception and Cognition.* Montreal: Society for Music Perception and Cognition.

Clynes, M. (1983). Expressive microstructure in music, linked to living qualities. In J. Sundberg (Ed.), *Studies of music performance* (pp. 76–181). Stockholm: Royal Swedish Academy of Music.

Clynes, M., & Walker, J. (1982). Neurobiologic functions of rhythm, time and pulse in music. In M. Clynes (Ed.), *Music, mind and brain.* New York: Plenum Press.

Desain, P., & Honing, H. (1991). Towards a calculus for expressive timing in music. *Computers in Music Research, 3*, 43–120.

Drake, C., & Palmer, C. (1993). Accent structures in music performance. *Music Perception, 10* (3), 343–378.

Gabrielsson, A. (1987). Once again: the theme from Mozart's Piano Sonata in A major: A comparison of five performances. In A. Gabrielsson (Ed.), *Action and perception in rhythm and music.* Stockholm: Royal Swedish Academy of Music.

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

Kronman U., & Sundberg, A. (1987). Is the musical ritard an allusion to physical motion? In A. Gabrielsson (Ed.), *Action and perception in rhythm and music.* Stockholm: Royal Swedish Academy of Music.

Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music.* Cambridge: MIT Press.

Longuet-Higgins, H. C., & Lisle, E. R. (1989). Modelling music cognition. *Contemporary Music Review, 3*, 15–27.

Repp, B. H. (1990). Patterns of expressive timing in performances of a Beethoven minuet by 19 famous pianists. *Journal of the Acoustical Society of America, 88*, 622–641.

Repp, B. H. (1992). Diversity and commonality in music performance: an analysis of timing microstructure in Schumann's Traumerei. *Journal of the Acoustical Society of America, 92* (5), 2546–2568.

Repp, B. H. (1995). Expressive timing in Schumann's Traumerei: an analysis of performances by graduate student pianists. *Journal of the Acoustical Society of America, 98* (5), 2413–2427.

Repp, B. H. (1996). The dynamics of expressive piano performance: Schumann's Traumerei revisited. *Journal of the Acoustical Society of America, 100* (1), 641–650.

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

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

Shaffer, L. H., & Todd, N. P. (1987). The interpretive component in musical performance. In A. Gabrielsson (Ed.), *Action and perception in rhythm and music.* Stockholm: Royal Swedish Academy of Music.

Shaffer, L. H., Clarke, E. F., & Todd, N. P. (1985). Metre and rhythm in piano playing. *Cognition, 20*, 61–77.

Sundberg, J. (1988). Computer synthesis of music performance. In J. A. Sloboda (Ed.), *Generative processes in music: The psychology of performance, improvisation, and composition.* Oxford: Clarendon Press.

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

Todd, N. P. (1989). A computational model of tempo. *Contemporary Music Review, 3*, 69–88.

Todd, N. P. (1992). The dynamics of dynamics: A model of musical expression. *Journal of the Acoustical Society of America, 91*, 3540–3550.

Appendix